Suitability Of Cellular Network Signaling Data For Origin-Destination Matrix Construction: A Case Study Of Lyon Region (France)

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

SUITABILITY OF CELLULAR NETWORK SIGNALING DATA FOR ORIGIN-DESTINATION MATRIX CONSTRUCTION: A CASE STUDY OF LYON REGION (FRANCE)

Mariem Fekih, Corresponding Author
Transportation Research Institute (IMOB), Hasselt University
Agoralaan, BE-3590 Diepenbeek, Belgium

SENSE, Orange Labs
44 avenue de la République, CS 50010, FR-92326 Chatillon Cedex, France
Email: mariem.fekih@uhasselt.be

Tom Bellemans
Transportation Research Institute (IMOB), Hasselt University
Agoralaan, BE-3590 Diepenbeek, Belgium
Email: tom.bellemans@uhasselt.be

Zbigniew Smoreda
SENSE, Orange Labs
44 avenue de la République, CS 50010, FR-92326 Chatillon Cedex, France
Email: zbigniew.smoreda@orange.com

Patrick Bonnel
LAET, ENTPE, Université de Lyon, CNRS
Rue Maurice Audin, 69518 Vaulx-en-Velin Cedex, France
Email: patrick.bonnel@entpe.fr

Angelo Furno
IFSTTAR, ENTPE, LICIT_UMR-T9401, Université de Lyon
25, avenue François Mitterrand, Case 24, Cité des mobilités. F- 69675 Bron Cedex, France
Email: angelo.furno@ifsttar.fr

Stéphane Galland
LE2I, Université de Bourgogne Franche-Comté, UTBM
13, rue Ernest Thierry-Mieg, 90010 Belfort Cedex, France
Email: stephane.galland@utbm.fr

Word count: 7,248 words text + 1 table x 250 words (each) = 7,498 words

Submission Date: July 31, 2018
Revised version submission date: November 10, 2018

ABSTRACT

Spatiotemporal data, and more specifically origin-destination matrices, are critical inputs to mobility studies for transportation planning and urban management purposes.

In this paper, we propose a methodology to infer origin-destination (O-D) matrices based on passively-collected cellular signaling data of millions of anonymized mobile phone users in the Rhône-Alpes region, France. This dataset, which consists of records time-stamped with users' unique identifier and tower locations, is used to first analyze 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 analyzed 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 signaling data-based O-D matrix carries similar estimations as those that can be obtained via travel surveys.

Keywords: Passive cellular signaling data, travel survey, home detection, trip extraction, origin-destination matrices

1. INTRODUCTION

Spatiotemporal data are extremely valuable to study human mobility for transportation and urban planning purposes *(1, 2)*. 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 behavioral transport models (e.g. travel demand and route or 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 (*3*), by declining response rates (*4*) and by unreported trips (*5*), 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 equipments. 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 (*6*). In particular, data collected from cell phones have become one of the most important data sources to study travel behavior (*7*). Their proper attributes, such as large coverage of geographic area and population, and high detailed location information have attracted researchers to analyze them to support transportation studies. A number of studies 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 signaling 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 signaling data could act as cheap and reliable alternative data source to capture individual trips. Therefore, we propose a full workflow to transform cell phone network logs into *O-D* flow matrices supported by a validation step using travel survey data. We also present a case study conducted within the Rhône-Alpes region, France, for which we were able to analyze very recent mobile phone signaling data provided by Orange, the largest French mobile operator, and compare them with the data obtained from the latest travel survey performed in the same region.

This paper is structured as follows. Section 2 describes 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 is discussed. While in Section 5, our results are summarized and validated with respect to travel survey data. Finally, Section 6 concludes the paper and identifies several suggestions for future research directions.

2. RELATED WORK

Due to the wide adoption of mobile devices (mobile phones, smartphones and tablets), the usefulness of mobile phone data has been proven for the study of human mobility for transportation research. González et al. (*8*) have proposed one of the first studies of large-scale mobility using a sample of over 100,000 mobile phone users. This study demonstrated that the distribution of users' trips is well approximated by a truncated power-law distribution. In recent years, mobile phone data have been explored for mobility pattern extraction (*9–11*), traffic flow estimation (*12*, *13*), population estimation (*14*, *15*) and route choice modeling (*13*). Furthermore, there have been several limited-scale researches aimed at analyzing the potential of mobile phone data for origin-destination estimation. In 2002, a small sample from one morning has been used to study traffic O-D matrices on specific roads (*16*). Later in 2007, Caceres et al. (*17*) calculated an O-D matrix for a road between the cities of Huelva and Seville in Spain with four possible O-D pairs. Both of these studies are based on small samples of CDR in very limited areas. More recently, 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 (*10*) and in Dhaka, Bangladesh (*18*). To validate the estimations, probe vehicle GPS data and limited traffic counts are used showing high correlations with CDR-based O-D matrices for both study areas. Calabrese et al. (*12*) and Mellegard et al. (*19*) used also CDR data for the same purpose but no detailed comparison for all the matrix cells with external data sources has been performed. Alexander et al. (*20*) have conducted analysis on triangulated CDR data (with estimated (x, y) coordinates rather than cell tower location) 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. Although CDR data have supported interesting findings for O-D extraction, their limited temporal granularity can introduce biases since they are event-driven traces (location only available when user makes call, SMS or data connections) (*12*). For this reason, cellular network signaling data (see Section 3.1) are more likely to be suitable to such O-D estimation since they capture all network-based events providing higher spatiotemporal granularity. Few existing works applied these data to extract vehicular patterns (*21*) and to estimate the route choice (*13*). In a recent study conducted in the Paris region (*22*), authors used signaling 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 with high traffic.

The aim of this research is to advance the state-of-the-art on the potential of network-based signaling data for origin-destination flow matrix extraction. To that aim, our method explores a mobile-network-signaling dataset, collected in 2017 from both 2G and 3G networks in the Lyon region. The following section describes in details the used datasets and the study area.

3. DATA SETS AND STUDY AREA DESCRIPTION

3.1. Cellular Signaling 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 network, which provides 2G services, and the UMTS network for 3G ones. Both GSM and UMTS networks have different infrastructures, but they still work with the same coverage concept with an antenna that covers a cell, which belongs to a larger Location Area (LA). Typically, tens or even hundreds of cells share a single LA. In many studies on cellular networks, the theoretical coverage area of the cells is represented by means of Voronoi polygons.

In this paper, we present analyses of datasets issued from Orange mobile network signaling probes. The explored dataset includes 2G and 3G signaling records from June 2017 of over 2 million mobile phone users. For legal privacy restrictions, data from only one day is used. Concerning the spatial dimension, this dataset covers the entire Rhône-Alpes region in France, and thus can be used to estimate origin-destination flows within this territory. Figure (1-a) 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.

The signaling data include all the events that are generated by mobile devices or by the network itself (*23*). Such dataset contains several types of events: *i)* communication events (i.e., calls and SMS); *ii)* itinerancy events: handover (i.e., cell changes during a communication) and Location Area (LA) update; *iii)* attachment/detachment events; *iv)* data/internet connections. The mentioned event types are the main characteristics of network-driven data comparing to event-driven data (e.g. CDR), which explains their higher temporal granularity. Each record in our data includes: the *anonymized user ID*, the *event type*, the *cell tower coordinates* to which the terminal is connected and the *assigned timestamp*.

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 has been collected by phone interviews using a representative sample of the region population. The 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 (Figure 1-b shows the 77 EDR-sectors and their aggregation in 14 macro-zones). Each EDR-sector involves at least 450 surveyed individuals.

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*, *begin* 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 (*24*) and elsewhere in the world.

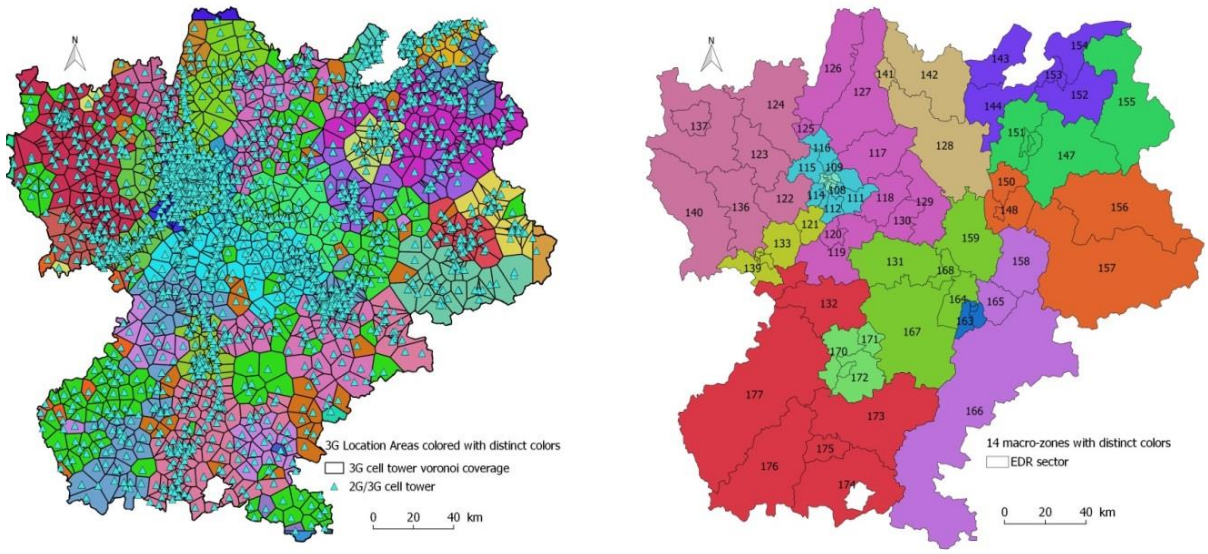

FIGURE 1 (a) Cell tower distribution and cellular network coverage (b) aggregation of EDR sectors into 14-zone zoning system in the Rhône-Alpes region

4. METHODOLOGY

In previous work (*25*), we introduced a first simple approach to generate the O-D matrices from only 3G signaling data. 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 comprehensive workflow is presented in order to transform signaling data into comprehensible O-D flow matrices supported by a validation step (Figure 2). It consists of *i)* analyzing the cell phone activity indicators to better characterize and understand the dataset; *ii)* identifying users' home locations; *iii)* filtering the detected residents based on their activity indicators to only retain users whose device traces are important enough to study their displacements; *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.

Due to user's privacy protection concern, data from at most 24-hour observation period can be used. We have analyzed the data of June 1st, 2017. It is a working day (Thursday), which is similar to the average of the working days available in the dataset. In order to be comparable to EDR, cellular network-based data is collected from 1st June 3:00am to 2nd June 2017 3:00am.

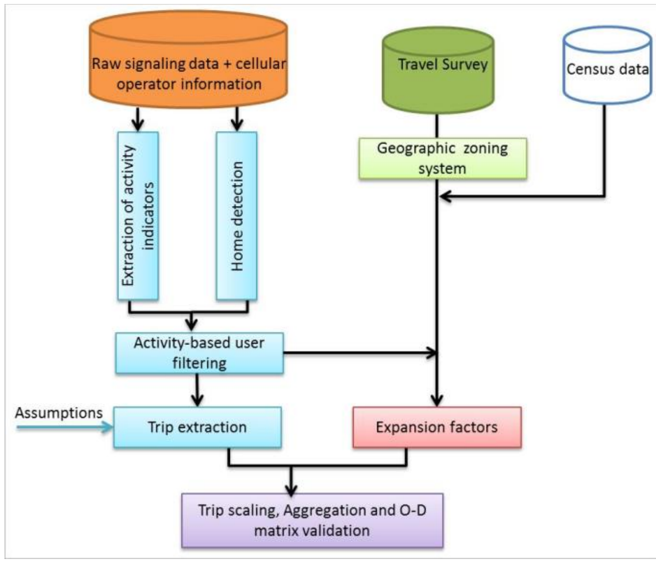

FIGURE 2 Workflow of Origin-Destination matrix construction from signaling mobile phone data

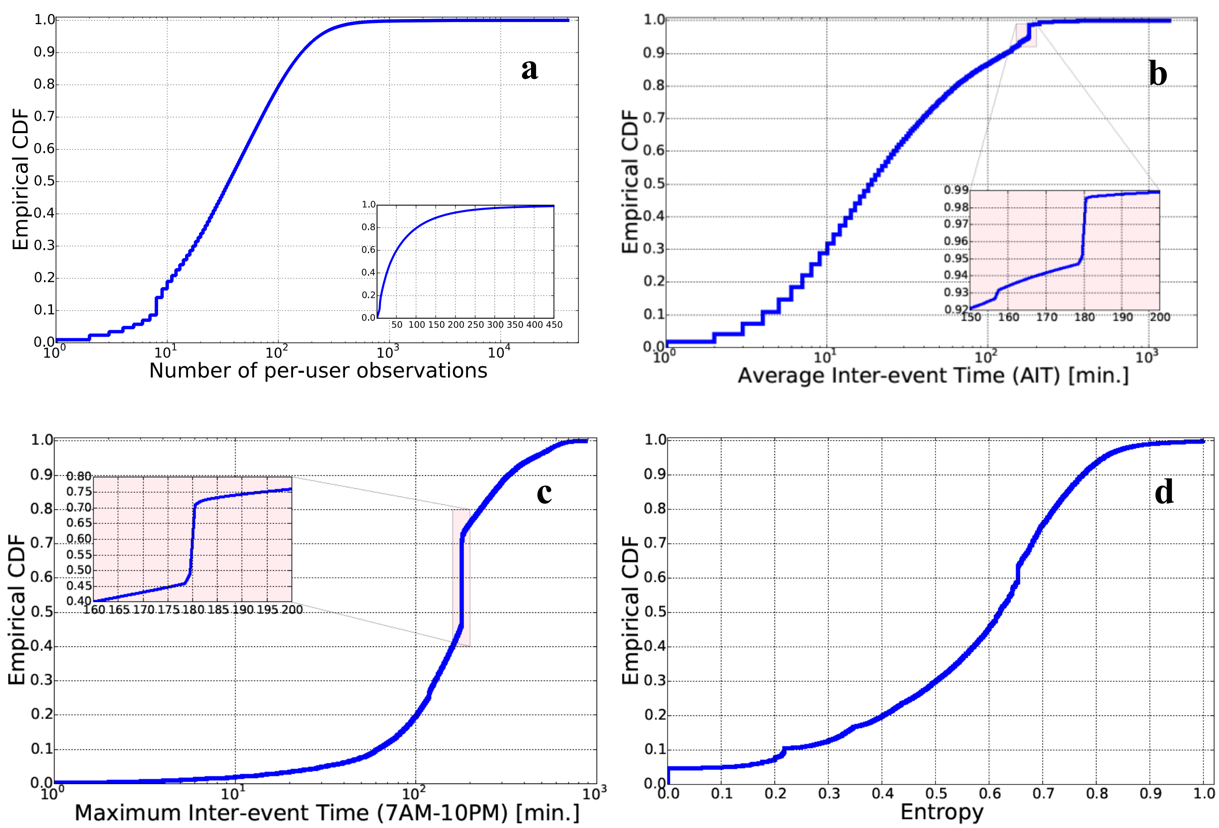

FIGURE 3 Cumulative distribution function per user of (a) number of observation (b) Average Inter-event Time (c) Maximum Inter-event Time [7am-10pm] and (d) Entropy

4.1. Cell Phone Activity Indicators

Given the nature of monitored cellular networks, the fact that digital footprints collected in real-world networks represent only expected signaling behaviors cannot be assumed. Indeed, due to increasing pervasiveness and wide adoption of embedded connected devices, telecom networks do not only capture human mobile phone communications, but also transactions from machines that use the same technology (i.e., Internet of Things). Thus, before using the signaling data, the information that best corresponds to the user's tracks has to be properly selected (*26*).

Some basic mobile phone activity indicators are introduced below 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 allows measuring the number of records (logs) for each terminal. Figure 3-a shows that records frequency on the dataset widely varies among observed devices. Around 99% of users have less than 450 events and 0.97% have only 1 record. A small part of devices (1%) seems to be extremely active with a very high number of observations (more than 1,000), which is not imputable to human behaviors, 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. Figure 3-b shows that AIT values range from 0 minute (few seconds) to 1372 minutes (22 hours), and the average value is about 40 minutes (while in CDR data the average time would typically be longer than four hours (*27*)). Most of the users (99%) are characterized by an AIT smaller than 200 minutes. The CDF shows a peak on 180 minutes that corresponds to idle mobile phones, which typically generate periodical Location Area Update (LAU) events every 3 hours.

- *Maximum Inter-event Time (MIT):*
  This indicator is different from the previous one. 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 studies, we propose to examine the maximum inter-event time during an interval of time that excludes deep night and early morning (7:00am-10:00pm). The MIT distribution is more skewed to the right (Figure 3-c), showing that 70% of users present a MIT lower than 180 minutes. This indicates that about 30% of the observed users in the dataset 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 signaling events per user over the 24 hours. It gives more precise information about the temporal distribution. The 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. Figure 3-d 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 less than 0.9 (more uniform behavior).

4.2. Home Detection

Cellular network-based data do not contain any socio-demographic information that characterizes the users due to privacy concerns. However, a large plethora of works have investigated basic mining solutions to detect users' home location (*15*, *27*) in combination with home activity identification (*14*, *28*).

In our study, the focus is on the detection of the users' home locations by considering only those that reside in the region of interest and expanding 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 keep only device events that could be generated in a stationary state such as Call, SMS, Attachment, Detachment, Data, and periodical events (e.g. periodic LAU);
2. Filter user traces to select only those occurring at night time from 3:00am to 7:00am and from 10:00pm to 3:00am;
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.

The presented method differs from the existing home detection methods by the inclusion of the event filtering step that basically considers more events that can be found in CDR data. Hence, a new approach is introduced to adapt existing algorithms to the signaling data. By applying this method, 1.27 million resident users are identified. This corresponds to 62% of all mobile phone users that are observed in the dataset, and about 25% of the total region population.

4.3. User Filtering based on Cell Phone Activity Indicators

As mentioned before, processing mobile signaling data without proper filtering can lead to estimation errors. Therefore it is proposed to leverage the cell phone activity indicators introduced in Section 4.1 to further filter the retrieved set of resident users. Our filtering approach requires the definition of thresholds associated to the indicators' values, and consists in a pipeline of selection rules as follows:

- `Maximum Inter-event Time(MIT) ≤ 180 minutes`: according to the network system if a mobile phone remains inactive during 3 hours, a periodic event (periodical Location Area Update (LAU)) is generated. We consider this value to ensure the presence of the user during the day period;
- `Entropy (H) ≤ 0.9`: as stated before, newly collected signaling data involves also Machine-to-Machine communications between objects equipped with SIM cards, which should not be considered in our analyses as they are not handled by individuals and do not reflect regular human mobility patterns. Thus, all devices that have an extremely uniform distribution of observations during the 24-hour period are filtered out based on their entropy value;

- `Number of observations (NO) ≥ 4`: this threshold value is defined with regard to the definition of a trip (see subsection of **Trip Detection**).

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. It is important to mention here that the suggested filtering process aims to filter out devices that do not generate enough traces or are not suitable to study the individual travel behavior and results. We also remark that it is not appropriate to perform a very restrictive filtering to keep only (highly) active users, as done in (*20*, *29*), since that could affect the representativeness of the users' sample and could lead to estimation biases.

4.4. Trip Detection

After identifying and filtering the resident users who are potentially appropriate to study the origin-destination matrices, trips can finally be extracted. A trip has been defined by CERTU for the purposes of the EDR as follows (*24*): a "*trip is the movement of one person conducted for a certain purpose on 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, signaling 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 a level for which EDR data are enough reliable, the trip extraction method at the EDR-sector level is presented in the following. The detailed experiments and additional data processing steps needed to study signaling 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 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 and this puts 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:

<u>For each user:</u>

- 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 extracted locations by timestamp
- Extract only locations where the user spent time $t \geq threshold_{min}$: 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 analyzed 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 applying home detection, an expansion factor can be calculated for each filtered user as the ratio of the census population and the number of residents estimated by the cellular signaling data in his home sector. It follows that users with the same home sector have the same scaling factor. Therefore, an expansion factor is defined at sector level as in Equation (1), where s_i is a sector. Moreover, given that a home location has been identified for each individual, the expansion factor of the home location sector is applied to all trips of the individual.

$$F_{exp}(s_i) = \frac{Population\ of\ s_i\ (over\ 11\ years)}{Nb\ of\ home\ locations\ detected\ in\ s_i} \qquad (1)$$

Figures 4-a and 4-b illustrate the probability distribution of the expansion factors through sectors before and after the user filtering step. The 1st, 2nd and 3rd quartiles of the expansion factors after filtering are 4.32, 5.39 and 6.71 respectively.

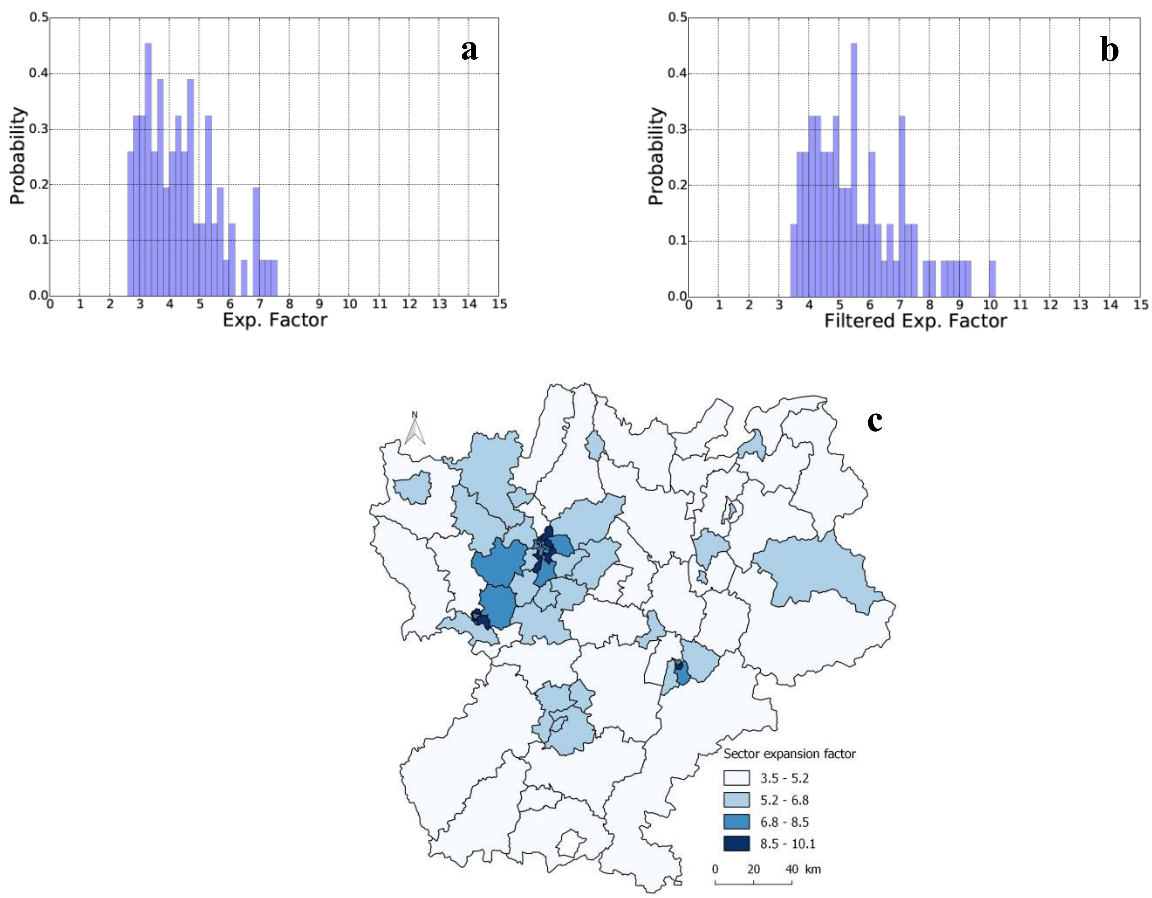

FIGURE 4 Probability distribution of sector expansion factors (a) before and (b) after filtering (c) the spatial distribution of expansion factors after user filtering

The spatial distribution of home expansion factors (Figure 4-c) 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, 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 trips can be constructed for each user on 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. Meanwhile, the sampling rate of events in mobile phone footprints should be considered; as stated in section 4, the measured average inter-event time was about 40min. 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 OD 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 would not be preferable to apply time thresholds, which are less than 30 minutes; otherwise false-positive stationary detections may occur, yielding false-positive trips.

5.1. Trip Distribution

In this section, we investigate the distribution of the number of trips on a typical weekday with respect to the users and the overall shape of OD matrix estimated from signaling data before expansion and at sector level. 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 Figure 5. 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 analyzed in more details in the next sub-section).

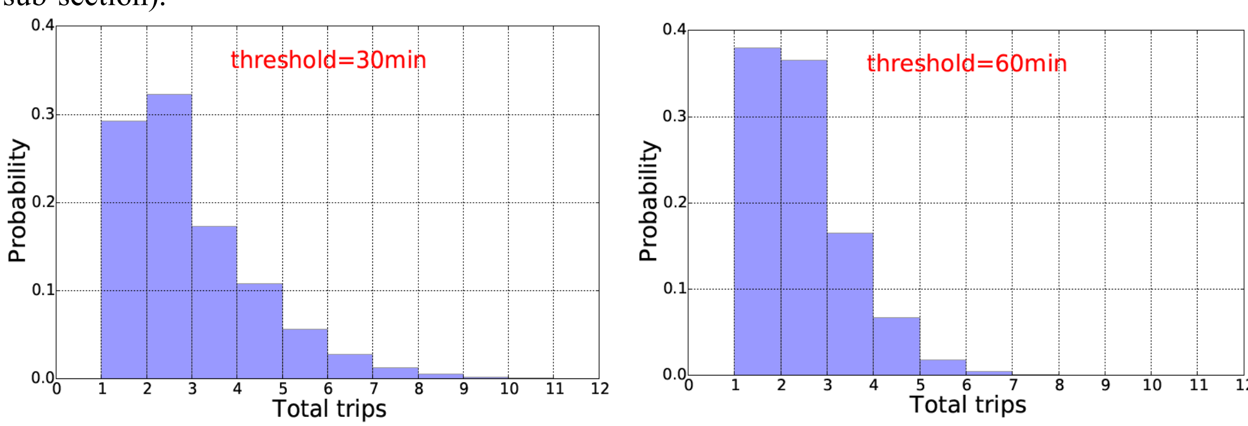

FIGURE 5 Probability distribution of total trips per user with a threshold 30min and 60min

Moreover, detailed analyses were performed to study the OD pairs inferred in the origin-destination matrix at EDR-sector level and compare them with those of the travel survey.

The data from the EDR contain all the trips made 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. We do extract therefore 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, the number of OD pairs is calculated when they are involved in both OD matrices estimated from signaling data and EDR for different thresholds (note that the total number of possible OD pairs is 5,929). We observe that in the travel survey matrix, less than the half of OD 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 OD matrices that are constructed based on travel surveys. Indeed, it is cost prohibitive to obtain sufficient observations to produce an OD 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, signaling data can cover more easily large-scale geographic areas as the collection is not dedicated but rather an operational by-product.

5.2. Origin-Destination flow matrices comparison

In this study, the aim is to test the potential of network signaling 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 OD 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 trips per matrix cell becomes too small. The confidence intervals are very wide for many O-D pairs. Therefore, the 77 EDR sectors are aggregated into 14 macro zones (Figure 1-b) 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. 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 1 summarizes the total number of trips from signaling data and the EDR.

TABLE 1 Total number of inter –zone trips from signaling 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 travelers: 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 low travel time between sectors. Therefore, in the following analyses, we retain a value of 30 minutes 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 two matrices is visually compared in Figure 6.

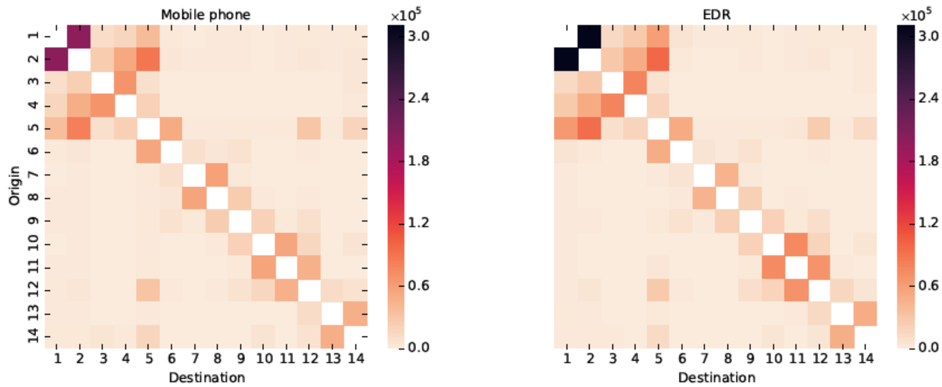

FIGURE 6 Distribution of signaling data and EDR trips over the 14 macro-zones of Rhône-Alpes region

According to the distribution of OD trips from mobile phone and EDR data with a stationary time of 30 minutes, the two 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 $\rho = 0.95$ $(p < 0.0001)$. Hence, although both 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 $R^2 = 0.96$ between macro-zone trips gives a high-level indication that the distributions of O-D flows are similar (Figure 7-a) with a regression equation $y_{ij} = 0.70 \times x_{ij} + 2{,}193$, where y_{ij} is the number of trips from signaling 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 more satisfactory than the first approach (*25*) for the same threshold (for 30 minutes, we obtained $R^2 = 0.87$).

As visually reported in the regression plot (see the two right-most points in Figure 7-a), the two O-D pairs of Lyon conurbation have very high number of trips in comparison with all other O-D. That is also shown in Figure 6 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, new regression analysis is performed without considering the O-D flows between Lyon conurbation zones (Figure 7-b).

By eliminating the outliers corresponding to the Lyon O-D pairs, the results improve. The regression provides much better parameters with a regression equation $y_{ij} = 0.85 \times x_{ij} + 877$. The slope is closer to one (0.85), and the constant (877) 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). According to R^2 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. This strong correlation is significant given that users' trips were expanded based on their home sector. Thus, this result illustrates that the applied methodology based on home detection and expansion process could serve as a proper tool to extract accurate travel patterns from cellular signaling data passively collected over a limited period of observation (e.g., 24-hour period in our case).

In addition to the previous analyses, the O-D matrices were investigated to estimate, for each O-D pair, the percentage disparity between mobile phone counts and those from EDR. As a result, some of these percentages were very high. 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). In these cases, mobile phone data generates higher flows, which illustrates that travel surveys may not reliably estimate trips due to the sample size of surveyed people and the sampling coverage. However, the under-estimation cases mainly concern very high flows corresponding to dense territories, such as 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., sectors of Lyon metropolitan area). Thus, more investigation is required on this parameter and will be matter of future work.

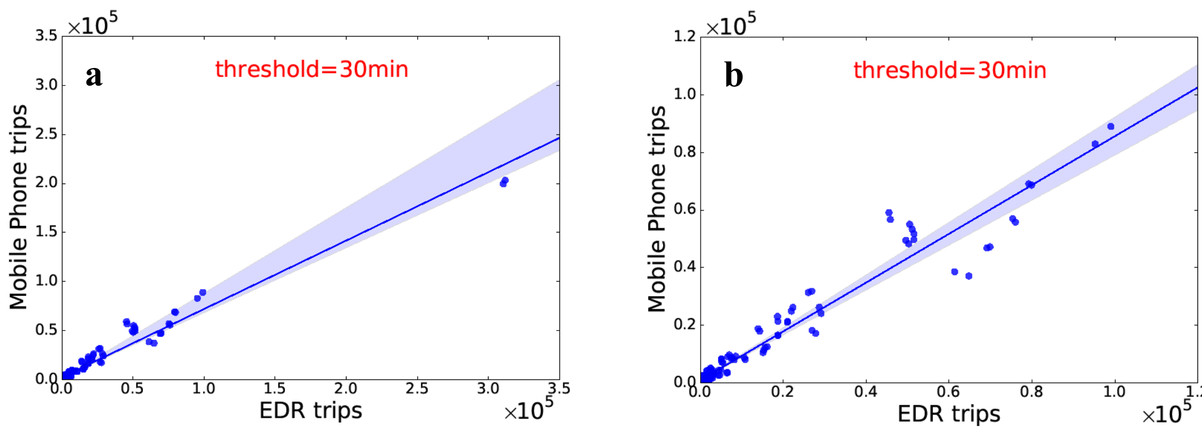

Figure 7 : Regression plot between the two matrices from signaling and EDR data with (a) all inter-sector pairs (b) with all inter-sector pairs except Lyon OD pairs

6. CONCLUSION

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 outdated and inappropriate for dynamic travel behavior studies.

In this paper, we presented a workflow of steps to generate O-D matrices from cellular network signaling data, 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 analyzing passive signaling data over a 24-hour period, we show that it is totally 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.

Detailed evaluation and validation process with the available 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 0.96 at an aggregated geographical level. That proves on one hand the efficiency of our method as only one day data are explored which reduces considerably its execution time. On the other hand, our findings show that cell network signaling data could represent a valuable, cost-effective alternative data source for origin-destination estimation. Moreover, with the increasing usage of mobile phones, cell network-based traces are expected to produce even larger-scale and higher-frequency data that will cover a growing number of people, thus allowing for estimating mobility patterns at a finer-grained spatiotemporal granularity than the one provided by travel survey estimations, which are not enough reliable due to sample bias.

Potential improvements of the proposed workflow will consist in investigating in more details the assumed hypothesis related to the stationary time threshold and the trip expansion method based on identified home locations from signaling data from a single operator. In addition to the suggested data filtering and processing steps, it will be very interesting to explore how the spatial accuracy of signaling data, which depend on cell network coverage, can compare with traditional survey accuracy according to the study area (e.g., urban or rural). Such information brings new insights for transportation practitioners to fully benefit from the new promising massive datasets at a low cost, especially with the rising transport networks complexity.

ACKNOWLEDGMENTS

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. MF' PhD is funded by Orange. The authors alone are responsible for the contents of this paper.

AUTHOR CONTRIBUTION STATEMENT

The authors confirm contribution to the paper as follows: study conception and design: MF, PB, ZS; analysis and interpretation of results: MF, TB, ZS, PB, AF, SG; MF was the lead writer of the manuscript. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. Arentze, T., and H. Timmermans. Data Needs, Data Collection, and Data Quality Requirements of Activity-Based Transport Demand Models. *TRB Transportation Research Circular*, No. II-J, 2000, pp. 1–30.
2. Giannotti, F., and D. Pedreschi, Eds. *Mobility, Data Mining and Privacy: Geographic Knowledge Discovery*. Springer, Berlin, 2008.
3. Stopher, P. R., and S. P. Greaves. Household Travel Surveys: Where Are We Going? *Transportation Research Part A: Policy and Practice*, Vol. 41, No. 5, 2007, pp. 367–381.
4. Bonnel, P. Postal, Telephone, and Face-to-Face Surveys: How Comparable Are They? In *Transport Survey Quality and Innovation* (P. Jones and P. R. Stopher, eds.), Emerald Group Publishing Limited, 2003, pp. 215–237.
5. Wolf, J., M. Oliveira, and M. Thompson. 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*, Vol. 1854, 2003, pp. 189–198. https://doi.org/10.3141/1854-21.

6. Liu, J., J. Li, W. Li, and J. Wu. Rethinking Big Data: A Review on the Data Quality and Usage Issues. *Journal of Photogrammetry and Remote Sensing*, Vol. 115, No. Supplement C, 2016, pp. 134–142.
7. Wang, Z., S. Y. He, and Y. Leung. Applying Mobile Phone Data to Travel Behaviour Research: A Literature Review. *Travel Behaviour and Society*, vol. 11, 2017, pp.141-155.
8. González, M. C., C. A. Hidalgo, and A.-L. Barabási. Understanding Individual Human Mobility Patterns. *Nature*, Vol. 453, No. 7196, 2008, pp. 779–782.
9. Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti. Understanding Individual Mobility Patterns from Urban Sensing Data: A Mobile Phone Trace Example. *Transportation research part C: emerging technologies*, Vol. 26, 2013, pp. 301–313.
10. Wang, P., T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González. Understanding Road Usage Patterns in Urban Areas. *Scientific Reports*, Vol. 2, No. 1001, 2012, pp. 1–6.
11. Asgari, F., V. Gauthier, and M. Becker. A Survey on Human Mobility and Its Applications. *arXiv preprint arXiv:1307.0814*, 2013.
12. Calabrese, F., M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome. *IEEE Transactions on Intelligent Transportation Systems*, Vol. 12, No. 1, 2011, pp. 141–151.
13. Tettamanti, T., and V. Istvan. Mobile Phone Location Area Based Traffic Flow Estimation in Urban Road Traffic. *Advances in Civil and Environment Engineering*, Vol. 1, No. 1, 2014, pp. 1–15.
14. Frias-Martinez, V., J. Virseda, A. Rubio, and E. Frias-Martinez. Towards Large Scale Technology Impact Analyses: Automatic Residential Localization from Mobile Phone-Call Data. *International Conference on Information & Communication Technologies and Development,* 2010, London, UK.
15. Ricciato, F., P. Widhalm, F. Pantisano, and M. Craglia. Beyond the "single-Operator, CDR-Only" Paradigm: An Interoperable Framework for Mobile Phone Network Data Analyses and Population Density Estimation. *Pervasive and Mobile Computing*, Vol. 35, 2016, pp. 65–82.
16. White, J., and I. Wells. Extracting Origin Destination Information from Mobile Phone Data. *Eleventh International Conference on Road Transport Information and Control*, 2002, London, UK.
17. Caceres, N., J. P. Wideberg, and F. G. Benitez. Deriving Origin–destination Data from a Mobile Phone Network. *IET Intelligent Transport Systems*, Vol. 1, No. 1, 2007, pp. 15–26.
18. Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. González. Development of Origin–destination Matrices Using Mobile Phone Call Data. *Transportation Research Part C: Emerging Technologies*, Vol. 40, 2014, pp. 63–74.
19. Mellegard, E., S. Moritz, and M. Zahoor. Origin/Destination-Estimation Using Cellular Network Data. *IEEE 11th International Conference on Data Mining Workshops*, 2011, Vancouver, BC, Canada.
20. Alexander, L. P., S. Jiang, M. Murga, and M. C. González. Origin–destination Trips by Purpose and Time of Day Inferred from Mobile Phone Data. *Transportation Research Part C*, Vol. 58, 2015, pp. 240–250.
21. Fiadino, P., D. Valerio, F. Ricciato, and K. A. Hummel. Steps towards the Extraction of Vehicular Mobility Patterns from 3G Signaling Data. In *Traffic Monitoring and Analysis* (A. Pescapè, L. Salgarelli, and X. Dimitropoulos, eds.), Springer Berlin Heidelberg, 2012, pp. 66–80.
22. Bonnel, P., E. Hombourger, A.-M. Olteanu-Raimond, and Z. Smoreda. Passive Mobile Phone

Dataset to Construct Origin-Destination Matrix: Potentials and Limitations. *Transportation Research Procedia*, Vol. 11, 2015, pp. 381–398.

23. Smoreda, Z., A.-M. Olteanu-Raimond, and T. Couronné. Spatiotemporal Data from Mobile Phones for Personal Mobility Assessment. In *Transport survey methods: best practice for decision making*, Emerald Group Publishing Limited, 2013, pp. 745–768.
24. CERTU. *L'enquête Ménages Déplacements Standard CERTU, Éditions Du CERTU*. 2008, p. 203.
25. Bonnel, P., M. Fekih, and Z. Smoreda. Origin-Destination Estimation Using Mobile Network Probe Data. *Transportation Research Procedia*, Vol. 32, 2018, pp. 69–81.
26. Wang, F., and C. Chen. On Data Processing Required to Derive Mobility Patterns from Passively-Generated Mobile Phone Data. *Transportation Research Part C: Emerging Technologies*, Vol. 87, 2018, pp. 58–74.
27. Calabrese, F., G. Di Lorenzo, L. Liu, and C. Ratti. Estimating Origin-Destination Flows Using Mobile Phone Location Data. *IEEE Pervasive Computing*, Vol. 10, No. 4, 2011, pp. 36–44.
28. Jiang, S., J. Ferreira, and M. C. González. Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Transactions on Big Data*, Vol. 3, No. 2, 2017, pp. 208–219.
29. Fiadino, P., V. Ponce-Lopez, J. Antonio, M. Torrent-Moreno, and A. D'Alconzo. Call Detail Records for Human Mobility Studies: Taking Stock of the Situation in the "Always Connected Era." *Proceedings of the Workshop on Big Data Analytics and Machine Learning for Data Communication Networks*, 2017, pp. 43-48.