Transportation Research Record

Assessing The Potential Of Cellular Signaling Data To Generate Dynamic Travel Patterns: A Comparative Study With Travel Survey Data

--Manuscript Draft--

Full Title:	Assessing The Potential Of Cellular Signaling Data To Generate Dynamic Travel Patterns: A Comparative Study With Travel Survey Data
Abstract:	This paper proposes a framework to extract origin-destination flows and dynamic travel demand patterns from 2G and 3G cellular signaling data in the Rhône-Alpes region, France. This study describes how these passively-collected records can be processed and analyzed in order to generate low-cost valuable inputs for transport models which currently rely on expensive and infrequent travel surveys. First, we present a method for data preprocessing and filtering based on cell phone activity metrics of about 2 million mobile phone users. Stationary activities have been detected to form trip sequences of users for whom the home location is identified. Then, trips have been aggregated by time of day to estimate hourly travel flows within the region. To better characterize these flows, we propose a spatial clustering process based on temporal demand profile of each zone and combine inferred travel patterns with land use data that help to reveal meaningful and significant dynamic mobility profiles. Comparative analyses have been performed with travel survey data showing that the resulting dynamic travel demand patterns are consistent with those obtained from survey data with high correlation coefficients of about 0.9.
Manuscript Classifications:	Data and Information Technology; Urban Transp Data and Info Systems ABJ30; Data Analysis; Travel Survey Methods ABJ40; Cell Phone Data; Origin-Destination; Travel Surveys
Manuscript Number:	
Article Type:	Presentation and Publication
Order of Authors:	Mariem Fekih
	Tom Bellemans, Professor
	Angelo Furno, Ph.D.
	Loïc Bonnetain
	Patrick Bonnel, Professor
	Zbigniew Smoreda, Ph.D.
	Stéphane Galland, Professor

ASSESSING THE POTENTIAL OF CELLULAR SIGNALING DATA TO GENERATE DYNAMIC TRAVEL PATTERNS: A COMPARATIVE STUDY WITH TRAVEL SURVEY DATA

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

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

Loïc Bonnetain
IFSTTAR, ENTPE, LICIT_UMR-T9401, Université de Lyon
25, avenue François Mitterrand, Case 24, Cité des mobilités. F- 69675 Bron Cedex, France
Email: loic.bonnetain@ifsttar.fr

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

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

Stéphane Galland
CIAD, Univ. Bourgogne Franche-Comté, UTBM
F-90010 Belfort, France
Email: stephane.galland@utbm.fr

Word Count: 7496 words + 0 tables x 250 words (each) = 7496 words

Submission Date: August 1st, 2019

ABSTRACT

This paper proposes a framework to extract origin-destination flows and dynamic travel demand patterns from 2G and 3G cellular signaling data in the Rhône-Alpes region, France. This study describes how these passively-collected records can be processed and analyzed in order to generate low-cost valuable inputs for transport models which currently rely on expensive and infrequent travel surveys.

First, we present a method for data preprocessing and filtering based on cell phone activity metrics of about 2 million mobile phone users. Stationary activities have been detected to form trip sequences of users for whom the home location is identified. Then, trips have been aggregated by time of day to estimate hourly travel flows within the region. To better characterize these flows, we propose a spatial clustering process based on temporal demand profile of each zone and combine inferred travel patterns with land use data that help to reveal meaningful and significant dynamic mobility profiles. Comparative analyses have been performed with travel survey data showing that the resulting dynamic travel demand patterns are consistent with those obtained from survey data with high correlation coefficients of about 0.9.

Keywords: Cellular signaling data, Dynamic travel demand, Mobility patterns, Travel survey, Profile Clustering

INTRODUCTION

Spatiotemporal information about people movements are extremely valuable for human mobility analysis (*1*) and transportation development purposes (*2*). Emerging forms of data generated by communication pervasive systems such as cellular networks are offering new opportunities to track individual-level movements and enhance our understanding of travel behavior patterns (*3*, *4*) in different social environments. Indeed, mobile phone records are characterized by a low collection cost since they are produced automatically and passively by telecom operators. More interestingly, the existing network mechanism provides continuously temporal and spatial information about individuals' whereabouts. Therefore, massive cellular network data provide a promising source for acquiring information about travel demand, exploring the various factors that might impact community travel flows and supporting long-term policy decisions on large-scale mobility.

The traditional human mobility research relies on household travel surveys that typically record one day of travel diaries per household. Yet, there are notable limitations associated with the classical travel survey process (*5*, *6*). Collected survey data can be useful to capture cross-sectional snapshots of daily journeys and to represent static mobility behavior. However; they do not allow considering fine-grained temporal analysis of the hourly, daily or weekly variability of individual trip flows.

Understanding the dynamics of human mobility patterns is a core notion in transportation studies (*7*) related to e.g. traffic congestion management and transport infrastructure planning. Nevertheless, travel surveys typically involve high-cost process that restrict their frequency and prevent to follow the mobility dynamics. The use of communication data has the potential to completely change the current techniques to estimate behavioral transport models. The ubiquity of the data as well as the relatively cheap deployment makes it possible to conduct studies on mobility trends with various time resolution scales. Hence, dynamic origin-destination flows can be generated using methods for assigning trips into target time windows. A number of studies have been conducted to extract dynamic trip metrics using different forms of mobile phone data such as Call Detail Records (CDRs) and cellular signaling data. Most of these studies have explored CDRs and developed techniques to figure out temporal distribution of user trips. However, it has been proven that these methods perform rather poorly due to very low spatio-temporal resolution of CDRs. Moreover, few research works have validated the results against external mobility data sources. Yet, the validation process allows to identify possible biases and to have a clearer idea about the potential of cellular data.

In previous work (*8*), we have developed a full workflow to transform cell phone network logs into individual trip flows and showed the potential of the method to generate static origin-destination flow matrices. The focus of this paper is to explore cellular signaling data from 2G and 3G networks to extract dynamic travel patterns of involved mobile users within large scale area. Although the potential of these data is promising due to the involved large amounts of individual spatiotemporal traces compared to CDR data, there is still a remarkable lack of studies based on them. The outline of this research is to test whether these massive signaling data could act as reliable data source to capture real-world temporal mobility behaviors. Therefore, we enrich the proposed workflow in (*8*) by adding the temporal component for dynamic origin-destination (OD) flow estimation. We apply a clustering process to capture the different existing temporal demand patterns and combine them with the related land use attributes to highlight the impact of the latter on trip generation. We introduce techniques to cope the spatio-temporal biases in the signaling data-based demand estimation. We present a case study conducted within the Rhône-Alpes region, France, for which we were able to analyze signaling data provided by Orange, the major French mobile operator, and to conduct comparative analysis with the data obtained from the latest travel survey performed in the same region. This paper is structured as follows. The state-of-the-art is presented in the "Related work" section. In the "Methodology" section, the data used in our analyses is presented followed by an overview of the process to extract individual trips and their associated temporal profiles. In section "Case study", the results are summarized and evaluated with respect to survey data. Finally, the "Conclusions" section concludes the paper and identifies several suggestions for future research.

RELATED WORK

The potential of mobile phone data to study human mobility and understand the underlying dynamics that govern movement flows as well as travel demand patterns has been proven in several works (*4*, *9*). This field of study has recently attracted a large and diverse body of researchers from urban planning, social science and even computer science. Thus, this emerging sensing data source has inspired a new generation of data-driven approaches to study the population dynamics. Significant attempts have been made to study trip distribution differences over weekdays and weekends (*10*) , to generate O-D flows by purpose and time of day (*11*) and to reconstruct the travel mode and flows in each link of the transportation network to perform traffic assignment (*12*, *13*). Furthermore, there have been several limited-scale researches aimed at identifying temporal movements urban areas. In 2010, Ahas et al. (*14*) analyzed the diurnal rhythms of the city life and its spatial differences in Tallinn, Estonia and showed that the majority of users had a similar temporal rhythm. Kang et al. (*15*) proposed to study how mobility patterns inside cities are affected by the compactness and the size of the area. Obtained results indicate that the distribution of intra-urban travel follows the exponential law and that individuals living in large cities need to travel farther on a daily basis. More recently, in Trasarti et al. (*16*), CDR data have been used to extract interconnections between different city areas that emerge from correlated temporal variations of population local densities. In the same perspective, study on the dynamic urban activity patterns and interaction between areas has been performed in Dakar, Senegal (*17*). The authors highlighted high interactions between areas with similar land use characteristics.

Moreover, mobile phone data have been explored to generate origin-destination flows and estimate relevant temporal mobility metrics within different urban areas (*7*, *18*, *19*). From an activity-based modeling perspective, Widhalm et al. (*19*) have extracted activity behavioral patterns based on trip departure time, activity types and frequencies combined with spatial typologies and land use data. Using CDR data, they applied the method in the cities of Vienna and Boston showing similarities between conurbations. The resulting trip chains and activity patterns match well with data from surveys. Following a trip-based approach, Gundlegard et al. (*7*) proposed a process for dynamic travel demand estimation using two CDR datasets collected in Ivory Coast and Senegal. They computed relevant mobility metrics such as route and link travel flow and travels times. However, the derived estimations were not evaluated due to the lack of validation data. Also, the travel demand scaling for the full populations of the two studied areas is not discussed.

Overall, there is a large variety of existing works about methods to extract dynamic mobility metrics and travel patterns from mobile positioning data. Most of these studies focused on CDR data which are temporally and spatially sparse due to their activity-dependent nature (records generated only with call, SMS or data connections) (*20*). Few works have discussed the impact of such data on the proposed approach outcomes. Indeed, cell phone data have key attributes that are different from travel surveys and which should be carefully interpreted during the processing step. Furthermore, the fundamental question on the representativeness and biases of the analyzed data is rarely discussed. While the existing state-of-art researches employ several types of mobile phone data with different sample sizes and characteristics, they still did not provide satisfactory rules to properly deal with these passive travel data contents as well as to expand and evaluate the results against external sources to fully check the relevance of estimations for travel demand prediction and decision making purposes.

This paper is in line with our previous works (*8*, *21*) and advances the state-of-the-art on the potential of network-based signaling data to extract travel flow demand and dynamic patterns. To that aim, our method explores a mobile-network-signaling dataset, collected in 2017 from both 2G and 3G networks in the large-scale territory of Rhône-Alpes region, France. We illustrate the peculiarities of the data and present techniques to infer expanded residents' trip flows as well as their hourly temporal distribution over the day. A clustering procedure of the extracted temporal profiles is proposed to reveal meaningful patterns. For better understanding, we map the travel patterns with land use features. And for evaluation purposes, we perform comparative analyses of the different aspects with travel survey data. The following section describes in detail the used dataset and the adopted methodology.

METHODOLOGY

This study explores the developed workflow in (*8*) for data processing and trip extraction. A temporal component to estimate trip start time and a clustering step to construct dynamic travel patterns observed within the region were added. We present the overall extended framework steps in the following.

Data Source and Characteristics

In this paper, we present analyses of datasets issued from GSM and UMTS Orange mobile networks (France). The coverage concept of both networks is the same: each antenna covers a cell area, which belongs to a larger Location Area (LA). The explored dataset includes 2G and 3G signaling records from June 2017 of over 2 million mobile phone users and covers the entire Rhône-Alpes region in France. For legal privacy restrictions, data analyses are only allowed within a study period of maximum 24 hours. Hence, we have analyzed the 24-hour period data collected from 1st June 3:00 am to 2nd June 2017 3:00 am. Figure 1 presents the distribution of 2G/3G cell towers and the considered sector zoning in the region. The largest metropolitan area in the territory is the city of Lyon (zoom in Figure 1), which concentrates nearly 25% of the inhabitants of the region. The signaling data include all the events that are generated by mobile devices or by the network itself (*22*). 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 Update (LAU); 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.

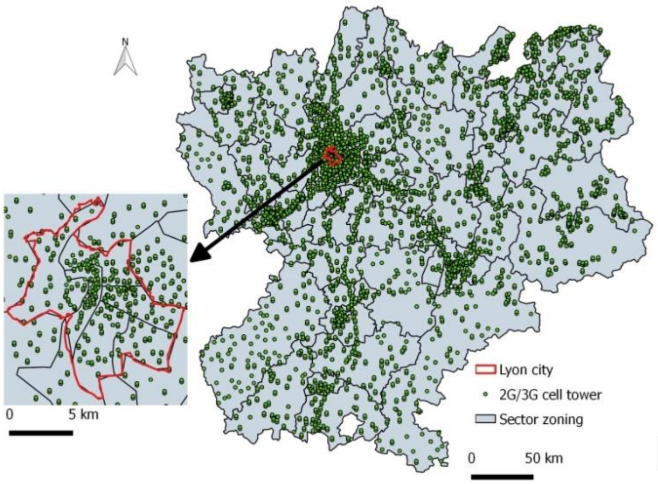


FIGURE 1. Cell tower (2G and 3G) distribution and administrative sector zoning in the Rhône-Alpes region with zoom on Lyon city

Data Preprocessing and Filtering

Due to the 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). An additional network signaling–based problem consists of rapid mobile phone cell fluctuation due to load-balancing (*23*). Hence, the location points generated from this phenomenon are considered as noise since they do not reflect the effective user's movement. Thus, before using the signaling data, the information that best corresponds to the user's tracks has to be properly selected.

It is proposed to leverage cell phone activity indicators to further filter the observed users in the dataset. 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 for each terminal over the observed day. 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).
- *Maximum Inter-event Time (MIT):*
  During night-time, devices are typically less active than the rest of the day. Thus, we introduce this MIT metric rather than the usually computed average inter-event time indicator. 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 analysis shows 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. About 5% of the devices have all observed traces in only one-hour time-slot (H=0). While 99% of devices have an entropy value less than 0.9 (more uniform behavior).

Our filtering approach requires the definition of thresholds associated to the indicators' values and consists of selection rules as follows:

- *Maximum Inter-event Time (MIT)* ≤ 180 minutes: according to the network system if a mobile phone remains inactive for 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 : all devices that have an extremely uniform distribution of observations during the 24-hour period are filtered out based on their entropy value (close to 1) since they are not handled by individuals and they do not reflect regular human mobility patterns;
- *Number of observations (NO)* ≥ 4: The trip construction method (section Trip extraction and scaling) requires at least 4 observations per individual to identify a displacement.

Home Location Detection

This work focuses on the detection of travel flows 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. Figure 2-a depicts the distribution of LAU events and shows that during the selected hours the users are stationary;
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.

By applying this method, 1.27 million resident users are identified in our initial mobile phone dataset. This corresponds to about 25% of the total region population. After the filtering process, a large sample of about 985 thousand 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 resident users and their associated event distributions are reported in Figures 2-b and 2-c. We notice that during all day, the maximum fraction of observed residents is about 63%. Additionally, during early morning (up to 7am) and from late afternoon, the hourly number of observed residents, and hence the number of observed events decreases substantially.

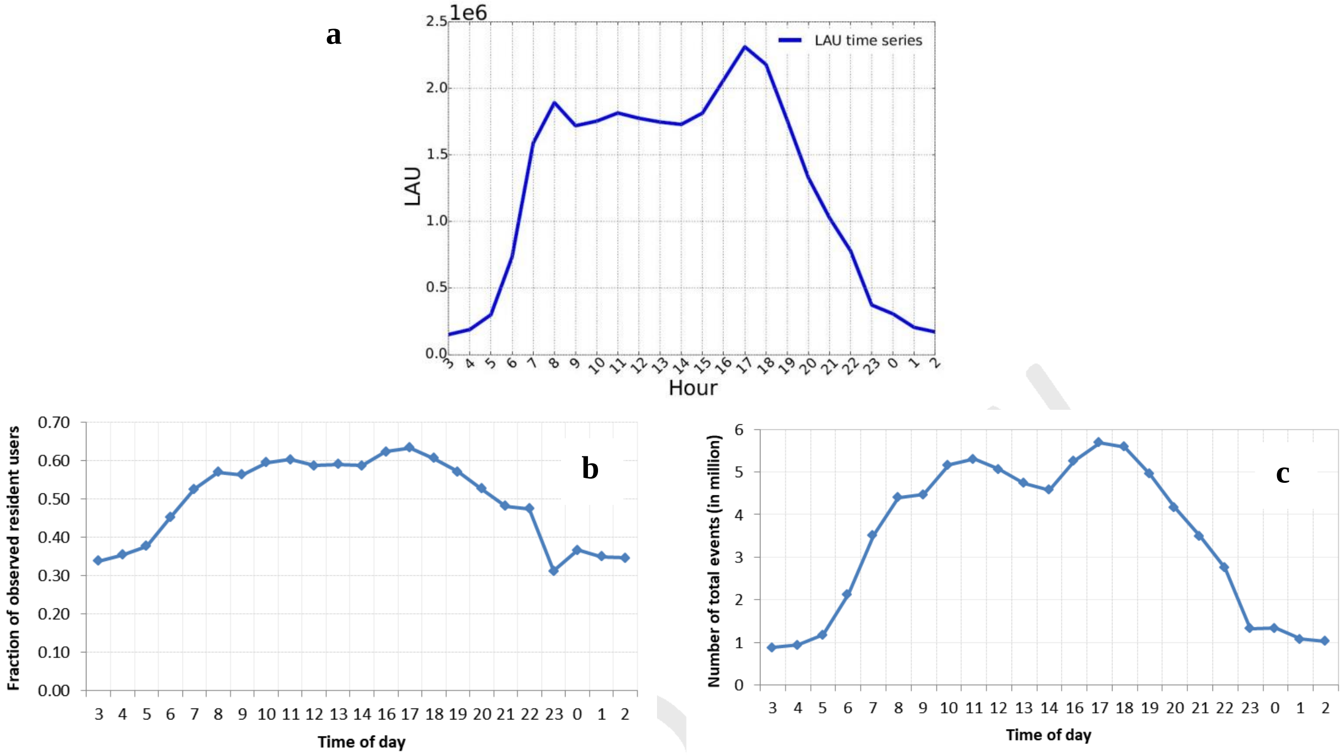


FIGURE 2 : Hourly distribution of (a) Location area update events (b) observed residents and (c) number of generated events

Trip Extraction and Scaling

After identifying and filtering the resident users who are potentially appropriate to study the dynamic travel patterns, trips can be extracted. A trip has been defined by CERTU (the French agency for transport network and urban planning) as follows (*24*): 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 a stationary activity in both the origin and the destination locations. Since the scope of this paper is to generate travel flows and to be able to validate their estimation at the same level for which survey data are available, the trip extraction method is presented in the following at the sector level (Figure 1).

To detect trips, stationary activities need to be identified first. Thus, consecutive observations of a user in a sector zone within a minimum stationary time threshold are considered. However, the size of the zones (average area of a sector is 582 km^2) and the fact that the user is traveling 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 is discussed in section results and validation).

In order to study the dynamic trip-making patterns, we need to associate for each trip a time window over the 24 hours. A very basic approach is proposed in this paper. For each user, it is assumed that a trip is made between every two consecutive stationary activities (*i, i+1*) happening within the 24 hour-period [3am-3am]. The trip occurs at a specific time spanned by the interval between the timestamp of last event *e* detected in activity *i* and the timestamp of the first event *e* detected in activity *i+1* , noted as [e_i , e_{i+1}]. Hence the start time is considered to be in [e_i, e_{i+1}]. Since the mentioned timestamps are cell phone observation-based rather than effective arrival and departure time, the start time of the trip is estimated using a probability distribution based on fuzzy logic theory. More specifically, we make the assumption that the probability of the user to be at the location of activity *i* is linearly distributed in the time interval [e_i, e_{i+1}]. And accordingly, we make the same assumption about the probability to be at the location of activity *i+1* in the time interval [e_i, e_{i+1}]. Since no other aspects seem to impact this assumption, it follows that the start time is uniformly distributed in the time window [e_i, e_{i+1}], and thus it is estimated to be in the middle of that interval.

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

- Extract all the observed location points and associate to each location a sector, e.g., with the help of a GIS tool.

  Cell tower → Sector
- Sort the sequence of extracted locations by timestamp , denoted by: *Si={si(1), si(2),...,si(n)}*, where *si(k) = (t(k), l(k))* for *k = 1,…,n*, and *t(k)* and *l(k)* are the time and location of the kth observation,
- User's stationary activities are identified from the sequence of extracted locations at sector level based on a minimum stationary time $threshold_{min}$,
- Estimate the start time for each trip based on the associated stationary activities.

Trips are then evaluated as paths between user‘s activity locations. Each *trip (U, O, D, T)* is characterized by user id *U*, origin location *O*, destination *D* and a start time T.

After applying trip extraction process, identified trips need to be properly scaled in order to be representative of the mobility of the full population. Using the resident estimations obtained from home detection process, an expansion factor can be calculated for each filtered user as the ratio of the census population and the number of residents estimated in his home sector. Hence, this expansion factor is applied to all trips of that user. 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.

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

Spatial Clustering

By aggregating all trips generated by each zone, we can derive the temporal demand profile emitted from each area. Then, our goal is to perform unsupervised clustering in order to group areas according to their temporal profile. To solve this problem, we apply a hierarchical agglomerative clustering technique on standardized hourly temporal profiles. The correlation coefficient has been used as similarity measure between profiles. We compute two indicators, i.e., the *silhouette* (*25*) and *davies_bouldin* scores (*26*), to decide the number of clusters to select. Once the clustering achieved, each area is characterized by its own temporal profile, the cluster number and the average temporal profile to which it belongs to. The library used in this step to perform clustering is scipy (python).

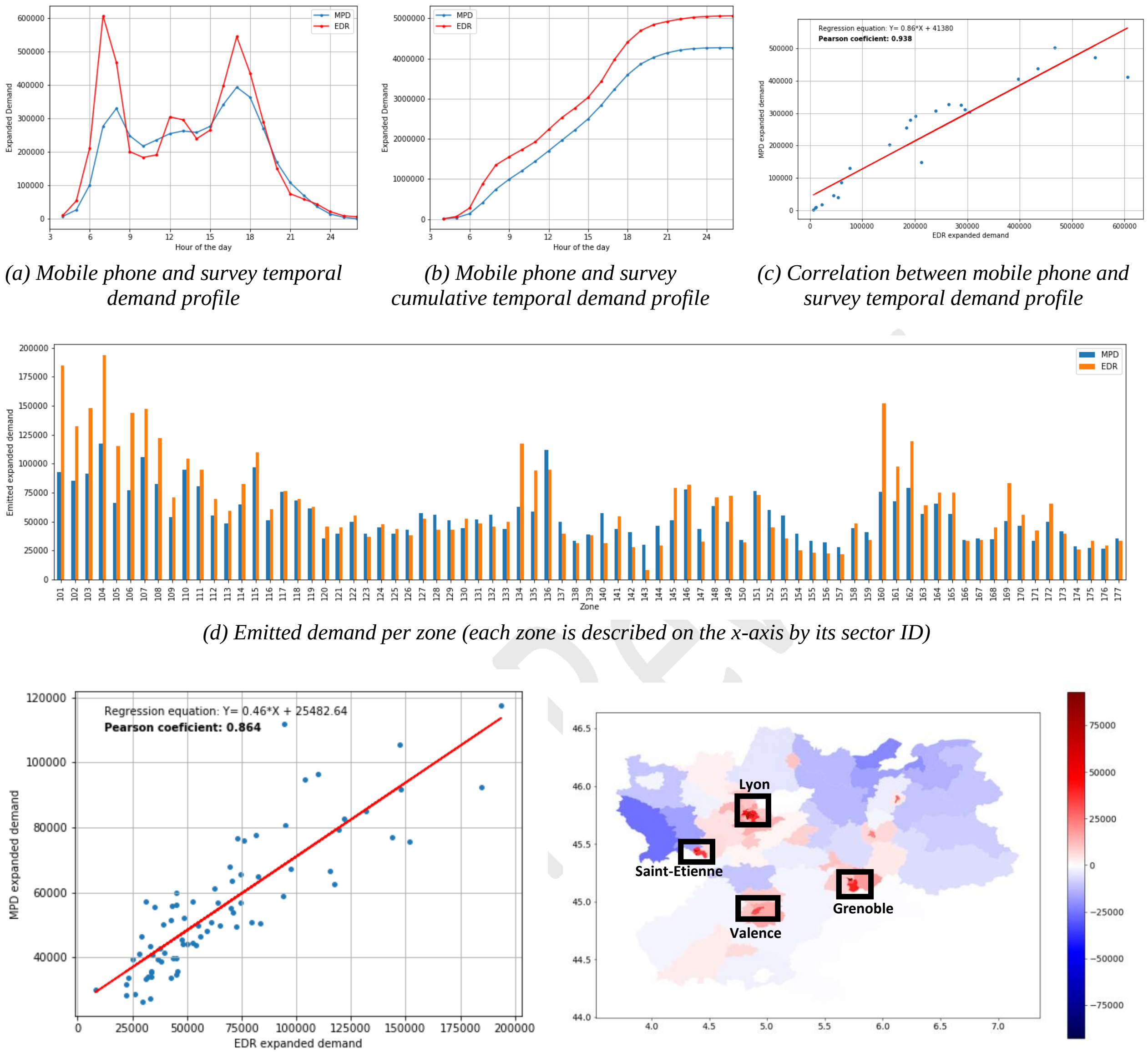
CASE STUDY

As stated in Section "Trip extraction and scaling", the assumption of the minimum activity stationary time is necessary for trip detection. In our previous work (*8*), we have tested different stationary time thresholds to show the sensitivity of the trip estimation. Relevant results were obtained with threshold of 30minutes when validating the associated static O-D matrix. Therefore, in this study, we retain an activity time threshold of 30min to detect trips.

In order to evaluate the proposed methodology, the aggregated travel demands extracted from signaling data were compared with those identified from the travel survey data. The household travel survey (called EDR 2015) was conducted in the Rhône-Alpes region between 2012 and 2015. 37,450 individuals, aged over 11 years, have been surveyed, and 143,000 trips have been identified. Given the aforementioned assumption of the minimum stationary time, we do extract information from the EDR data and apply the same time threshold to avoid considering false trips when dealing with the comparison. It is worth to remind that the following analyses are based on one typical working day (Thursday) of signaling data. Due to the bias caused by the sampling process, survey data will not be considered as ground truth but rather as a comparable reference with controlled error for our analyses. The analyses are presented at sector level by removing the intra-zone trips since the focus here is on inter-zone flows.

Figure 3-a shows the temporal demand profile for signaling and survey data. We can observe that the signaling-based demand profile has less sharp morning and afternoon peaks compared with the survey. Total demand from signaling data is lower than the one from the survey, as shown in Figure 3-b. This could be explained by the fact that we were able to detect the home sector for a group of people, referred as "static people" in the following, for whom it was not possible to detect any trip. The proportion of static people detected in the mobile phone dataset amounts to 46%. This proportion appears therefore to be overestimated probably due to reduced mobile phone observations for some non-static users (Figure 2-c): such overestimation of static behaviors inevitably leads to an underestimation of the travel demand especially during morning period.

Despite this underestimation, the hourly global demands estimated from both data sources are highly correlated (Pearson coefficient equal to 0.94) as shown in Figure 3-c. This confirms the fact that signaling data can grab, rather correctly, the well-known typical demand profile for a working day.



(a) Mobile phone and survey temporal demand profile

(b) Mobile phone and survey cumulative temporal demand profile

(c) Correlation between mobile phone and survey temporal demand profile

(d) Emitted demand per zone (each zone is described on the x-axis by its sector ID)

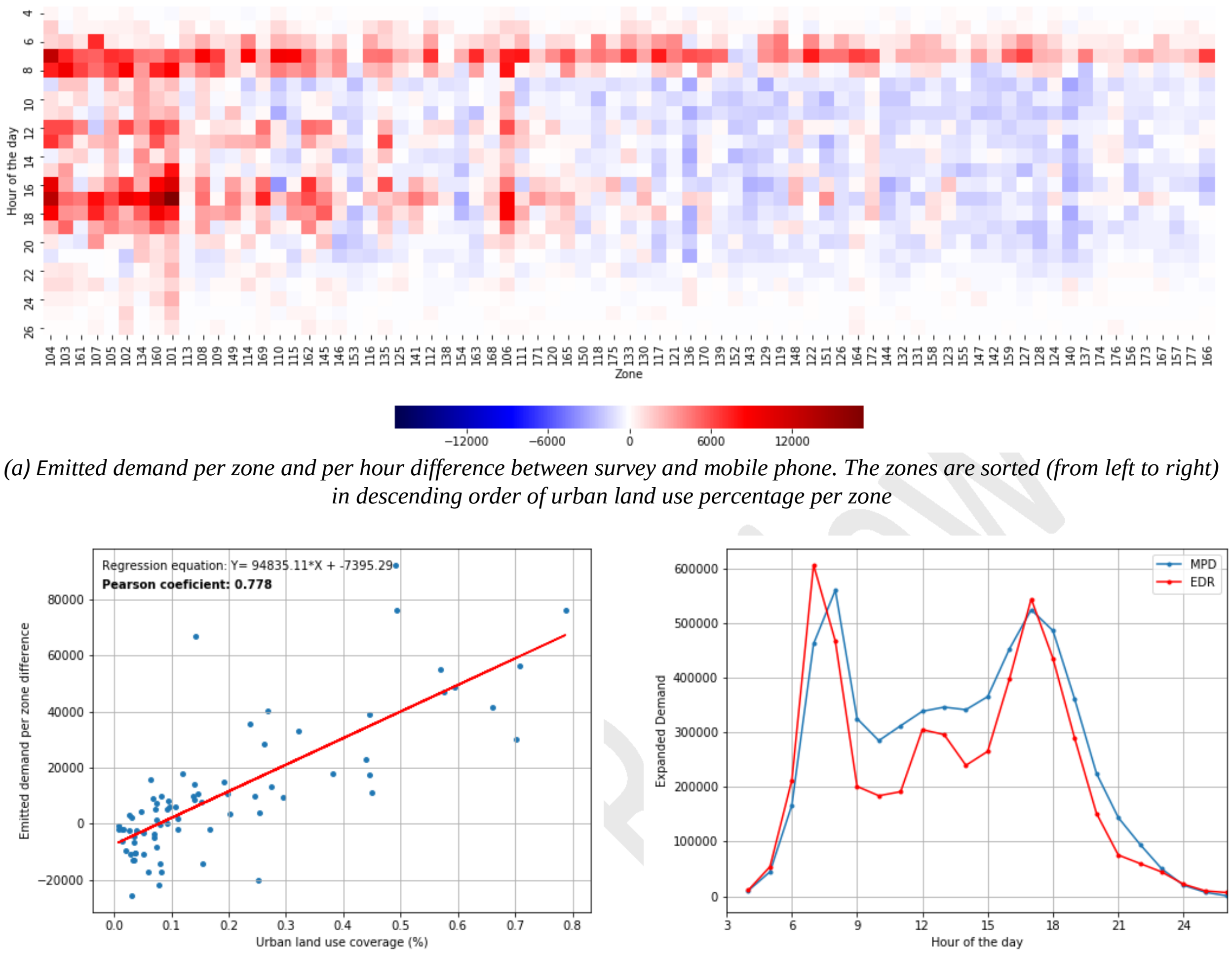
(e) Correlation between mobile phone and survey emitted demand per zone

(f) Heatmap of the difference between emitted demand per zone from mobile phone and survey data

FIGURE 3 . Comparison of total travel demand estimated from signaling and EDR survey data (a-c) and spatial distribution of signaling data-based travel demand (d-f).

After analyzing the demand estimation difference between signaling and survey data at temporal level, we have studied the difference between the demands emitted from each geographical zone. Figure 3-d shows the number of trips emitted from each sector (as the trip origin). For 45 sectors, the number of emitted trips is higher compared to survey (median relative difference of +0.20%). It is instead lower for 32 areas (median relative difference of -0.21%). These differences can be better interpreted by relying on the map shown in Figure 3-f, which is a spatial representation of the absolute difference between the demands generated by each zone from mobile phone and survey data, respectively. The emitted demand estimated with mobile phone data tends to be higher in rural areas and lower in urban dense areas. In rural areas we can reasonably assume that signaling records provide more consistent estimations since it is expected that they capture well long distance trips from a larger individual sample than in the survey in such area. Instead, in urban areas, it seems that the proposed trip extraction method is unable to capture very short distance trips, which occurs with higher frequency in urban area than rural ones. Indeed, it is not obvious to differentiate noise from short distance trips on mobile phone data. Despite such limitation, the total number of trips emitted by each zone based on cellular and survey data remains highly correlated (Pearson coefficient equal to 0.86) as shown in Figure 3-e.

Figure 4-a, representing the emitted demand per zone and per hour, confirms the spatial and temporal bias previously observed. In the figure, zones are sorted from left to right by decreasing density of urban land use (as retrieved from the 2012 CORINE land cover data (*27*)). On the one hand, the morning peak is higher in the survey compared to mobile phone regardless of the zone. On the other hand, in highly dense urban zones (on the left of the figure 4-a), the demand is higher in survey compared to mobile phone regardless of the hour of the day. After identifying this “systematic” bias, we propose a heuristic-based method to correct signaling data both on temporal and spatial dimensions. For the temporal correction, we have noticed that the observed underestimation of the demand appears also in the temporal profile of LAU (events passively generated by the device; i.e., not by explicit user’s communications, when changing the LA zone), presented in Figure 2-a. We can reasonably assume that this underestimation is due to people who have their mobile phone turned off more frequently during morning than during the afternoon. To address this bias, we have applied a uniform correction factor on mobile phone trips for the whole morning period (5-8am). This factor has been calculated as the ratio of the afternoon and morning peaks in the LAU profile by allowing the morning peak to be slightly higher than the afternoon peak. Finally, this temporal correction factor has been chosen equal to 1.3. We shall note here that the temporal correction is a de-biasing procedure, independent from the survey, thus being easily reproducible even in their absence. Concerning the spatial correction, Figure 4-b shows that the emitted demand difference between survey and mobile phone is abnormally highly correlated to the urban land use percentage. In other words, the difference is much higher for the denser urban areas compared to rural areas. This can be interpreted as an underestimation of the urban area travel demand in case of the signaling data. In order to address this bias, we have applied a spatial correction factor estimated per zone and calculated using the regression equation shown in Figure 4-b. The expression of this factor is the following: $1 + \frac{94835 * U_{la}(x)}{D_{MP}(x)}$ where $U_{la}(x)$ and $D_{MP}(x)$ represent respectively, the urban land use percentage and the mobile phone emitted demand associated to each zone x. By applying this correction factor, we are able to decorrelate this difference with respect to the urban land use percentage. After applying both spatial and temporal correction to signaling trips we have recomputed the demand profile (Figure 4-c) showing more correlated behavior against the survey.



(a) Emitted demand per zone and per hour difference between survey and mobile phone. The zones are sorted (from left to right) in descending order of urban land use percentage per zone

(b) Correlation between emitted demand difference (Survey – Mobile) and urban land use percentage per zone

(c) Mobile phone (corrected) and survey temporal demand profile

FIGURE 4 . Spatio-temporal distribution of the difference between mobile phone and survey and temporal demand distribution after correction

After de-biasing signaling trips via the proposed correction factors, we have performed spatial zone clustering based on temporal demand profile of each zone, as described in “Methodology” section. To determine the number of clusters for our analysis, we consider the *davies_bouldin* and *silhouette_index* scores, graphically depicted in Figures 5-a, b. In order to perform a finer-grained analysis with a slightly higher number of clusters, a number of 9 clusters has been selected (both indexes give better scores at 9).

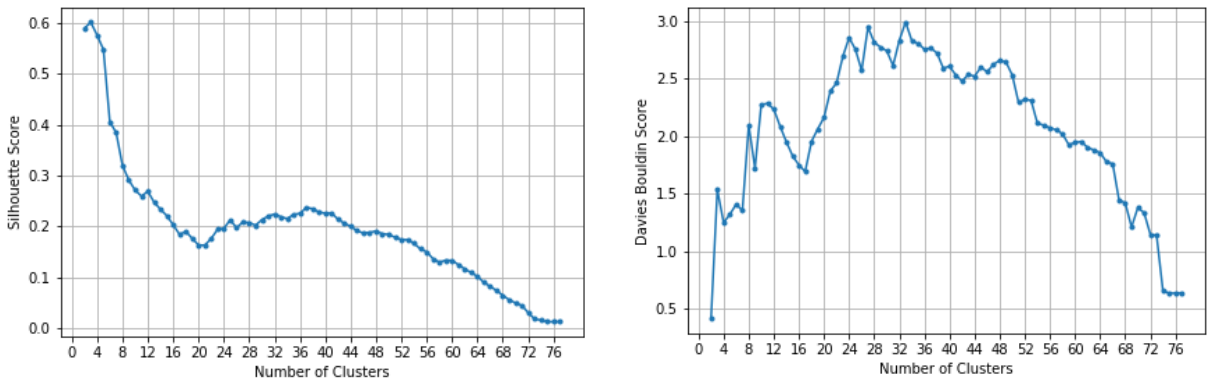
Among these clusters, there are 3 main clusters, each including at least 18 zones, and 6 minor clusters which include at most 2 zones (zones in each cluster represent trip origins). The map representing all these clusters is shown in Figure 5-i. The average temporal profile, noted as “ATP” in the following, is represented in Figures 5-c, d, e for main clusters 2, 4 and 5 respectively and Figure 6-j for the minor ones.

Based on these demand profiles, the following interpretation of the major clusters can be done:

- Cluster 2 represents rural areas. The ATP emitted by the related areas is composed of two peaks with a morning peak much higher than the afternoon peak. Given that these rural areas are mostly residential and rather unattractive in terms of business or leisure activities, we can safely state that a large amount of people leave this cluster to reach working places at the morning peak and come back at the late-afternoon peak

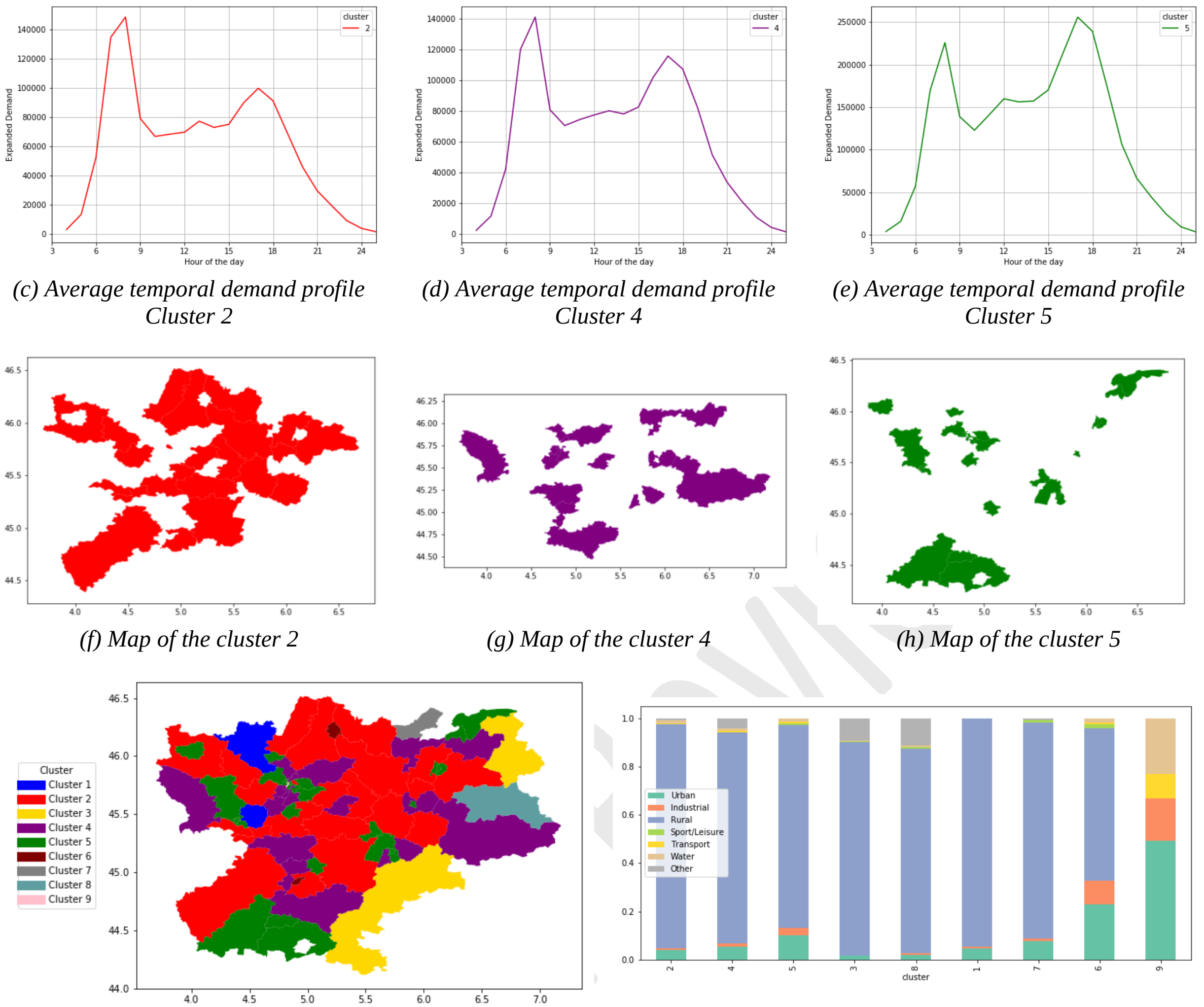
- Cluster 5 represents urban areas. The ATP emitted by these areas is composed of two peaks with an afternoon peak much higher than the morning one (rather than a symmetrical profile as the one observed for cluster 4). These zones are both residential (high population density) and attractive in terms of jobs and leisure. At the morning peak, these areas generate a high number of home-work commuting trips within the cluster and to other areas, but, at the same time, attract a significant amount of demand from the surrounding areas that is supposed to leave back the cluster (thus generating trips) later on, at the afternoon/evening peak. This appears evident from the temporal emitted-demand profile reported in Figure 5-e, highly asymmetric with a higher peak during late afternoon.
- Cluster 4 can be described as a mixture of cluster 2 and 5 in the sense that these areas are mostly neither rural nor highly dense urban zones. In this case, the ATP is more balanced (Figure 5-d) with an afternoon peak slightly higher than in cluster 2.
- The remaining clusters (1,3,6,7,8,9) have rather peculiar profiles compared to the major clusters previously discussed. The temporal demand profile of the clusters 6 and 9 (the latter being a single-sector cluster including the whole city center of Lyon) depicts a particular shape of highly urban areas (Figure 6-j) with a very sharp afternoon peak, which is consistent with the observations reported for cluster 5, mostly composed of urban areas as well.

The nature of the cluster areas (rural, urban and mix) inferred with signaling data has been validated using land use data by relying on the European CORINE land use data. Figure 5-j shows the distribution of the percentage land use coverage per cluster. The land use is divided in 7 categories: urban, industrial, rural, sport/leisure, transport, water and other. We can observe that, for all the major clusters, rural areas cover the largest part of the cluster (even urban areas have a large rural land use proportion) due to the rather large spatial extension of the analyzed sectors. However, the proportion of urban and industrial areas for clusters 6 and 9 is significantly higher compared to the one of cluster 5, which is in turn significantly higher if compared to the one of cluster 4. The land use analysis appears to clearly corroborate the previously reported interpretations of the clusters and of the associated reconstructed demand.



(a) Silhouette score

(b) Davies Bouldin score



(c) Average temporal demand profile Cluster 2

(d) Average temporal demand profile Cluster 4

(e) Average temporal demand profile Cluster 5

(f) Map of the cluster 2

(g) Map of the cluster 4

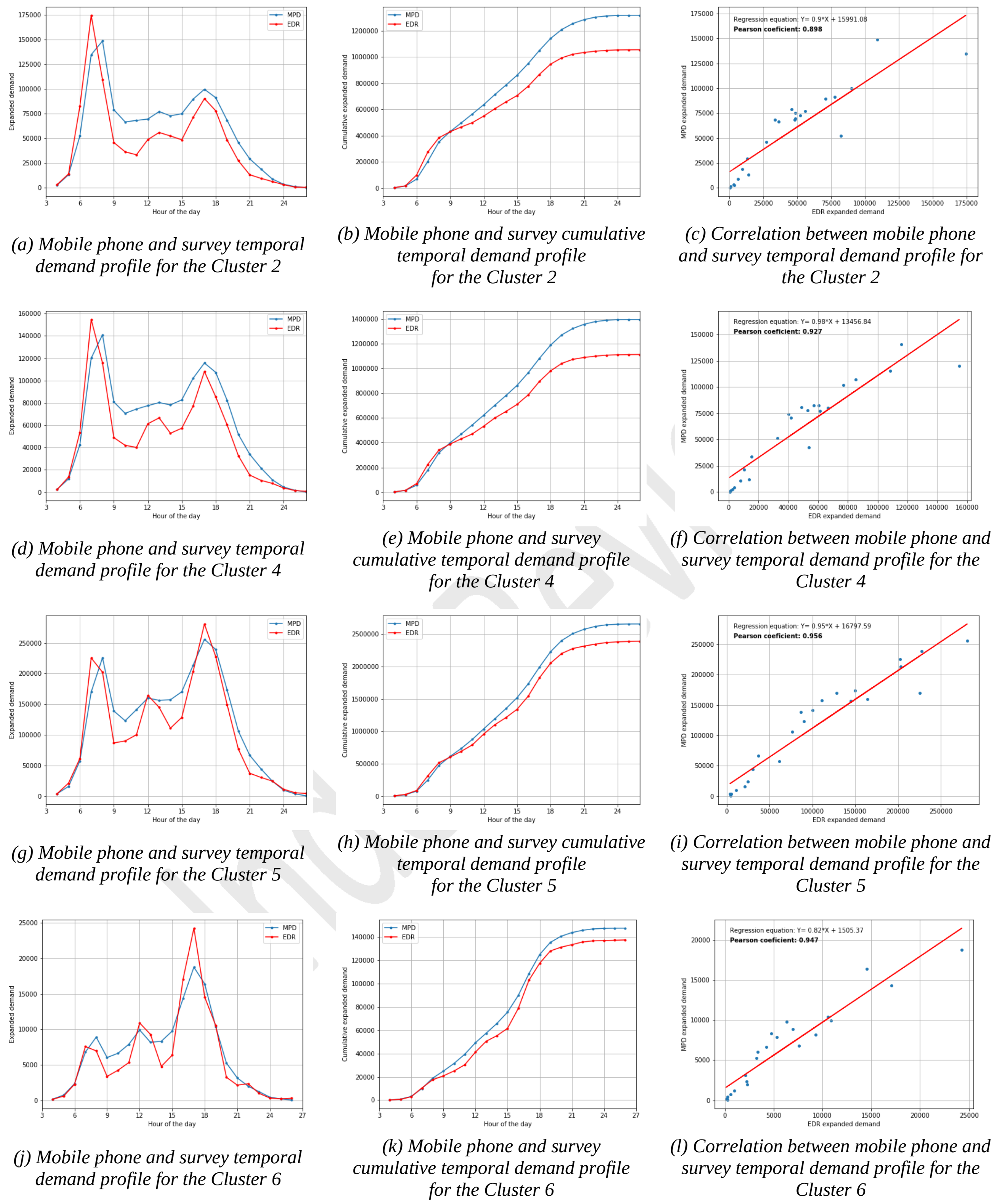
(h) Map of the cluster 5

(i) Map of all clusters

(j) Land use coverage per cluster

FIGURE 5 . Temporal demand and spatial distribution of main clusters based on signaling data

Finally, we have compared the ATPs estimated from signaling and survey data for the three main clusters (Figures 6-a,d,g) and one minor cluster (Figure 6-j). The overall patterns agree well with Pearson coefficients between 0.89 and 0.96. For high urban areas (cluster 5 and 6), signaling-based estimations are slightly lower than those estimated from survey at afternoon peak, but they preserve properly the specificity of the distribution shape. For mixed (cluster 4) and rural (cluster 2) areas, signaling-based estimations match well with those from survey with slight difference at morning peak. This confirms that signaling data can act as a good sensor and resolve the sampling rate problem of surveys in large mixed and rural areas, if properly de-biased. Also, we notice that for all clusters signaling data give higher flows surrounding the midday period, rather hardly observable via surveys. Hence, the resulting observations show that signaling data can capture unknown and more reasonable flow patterns specifically for low density and large-scale areas where accurate travel data are often not available.



(a) Mobile phone and survey temporal demand profile for the Cluster 2

(b) Mobile phone and survey cumulative temporal demand profile for the Cluster 2

(c) Correlation between mobile phone and survey temporal demand profile for the Cluster 2

(d) Mobile phone and survey temporal demand profile for the Cluster 4

(e) Mobile phone and survey cumulative temporal demand profile for the Cluster 4

(f) Correlation between mobile phone and survey temporal demand profile for the Cluster 4

(g) Mobile phone and survey temporal demand profile for the Cluster 5

(h) Mobile phone and survey cumulative temporal demand profile for the Cluster 5

(i) Correlation between mobile phone and survey temporal demand profile for the Cluster 5

(j) Mobile phone and survey temporal demand profile for the Cluster 6

(k) Mobile phone and survey cumulative temporal demand profile for the Cluster 6

(l) Correlation between mobile phone and survey temporal demand profile for the Cluster 6

FIGURE 6 . Comparison of travel demands of main clusters from signaling and survey data

CONCLUSIONS

This paper introduces a complete framework to process cellular network signaling data, infer origin-destination flows and estimate relevant dynamic travel patterns. The framework first highlights the preprocessing and filtering steps applied to signaling data in order to keep useful information for mobility extraction, an aspect typically not very well reported in the literature. Secondly, by analyzing signaling data of 2 million mobile phone users, we prove that it is feasible to use such data in order to robustly extract residents' trips and estimate their hourly distribution through the studied region, on condition that spatio-temporal biases of signaling data are properly detected and removed. Correction factors have been proposed to cope with these biases. By clustering the trip flows based on their temporal profiles and matching them with official land use data, we also unveil interesting and relevant heterogeneity in dynamic travel demand patterns related to trip production zones.

The evaluation analyses performed on both the temporal and the spatial dimensions show that the resulting travel demand profiles strongly agree with those obtained from the travel survey data with correlation coefficients higher than 0.9. Moreover, we were able to identify significant correlations between mobile phone-based dynamic patterns and area profiles. Very dense urban zones are characterized by a high afternoon peak, while low density areas depict a high morning peak where cell network signaling data exhibit more reasonable patterns and higher trip flows than survey data. This confirms that these massive data could complement conventional travel surveys as a valuable cost-effective data source especially for territories where accurate mobility data are not available or hardly collectable.

Potential improvements to the presented method consist of adding strategies to counter the underrepresentation of flows caused by cell phone observations biases. Therefore, enhancing the resident detection procedure seems to be fundamental in order to capture a more complete spectrum of trips. Specifically, adding a requirement on the minimum number of observed activities per resident can further improve the estimation of actual residents and, hence, the proposed scaling method by re-adjusting the expansion factors. Furthermore, we aim to refine the trip start time estimation process by leveraging user records that occur between successive activities. Thus, the start time estimation error could be reduced. In this study, results were obtained by exploring cellular signaling data collected during 24-hour period and cover a large territory of about 44,000km2 including different socio-demographic and economic zone profiles. Existing works usually focus on cities or spatially limited areas. Hence, this incites to go further on signaling data-based dynamic pattern extraction by analyzing the weekly, monthly or seasonally patterns which are highly required for strategic planning and travel demand modeling but very hard and expensive to investigate on them with the traditional travel surveys.

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. This work has been also supported by the French ANR project, PROMENADE, grant number ANR-18-CE22-0008. The authors alone are responsible for the paper contents.

AUTHOR CONTRIBUTIONS

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

REFERENCES

1. Barbosa, H., M. Barthelemy, G. Ghoshal, C. R. James, M. Lenormand, T. Louail, R. Menezes, J. J. Ramasco, F. Simini, and M. Tomasini. Human Mobility: Models and Applications. *Physics Reports*, 2018.
2. Arentze, T., and H. Timmermans. Data Needs, Data Collection, and Data Quality Requirements of Activity-Based Transport Demand Models. In *TRB Transportation Research Circular*, No. II-J, 2000, pp. 1–30.
3. Chen, C., J. Ma, Y. Susilo, Y. Liu, and M. Wang. The Promises of Big Data and Small Data for Travel Behavior (Aka Human Mobility) Analysis. *Transportation Research Part C: Emerging Technologies*, Vol. 68, 2016, pp. 285–299.
4. Wang, Z., S. Y. He, and Y. Leung. Applying Mobile Phone Data to Travel Behaviour Research: A Literature Review. *Travel Behaviour and Society*, 2017.
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.
6. 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.
7. Gundlegård, D., C. Rydergren, N. Breyer, and B. Rajna. Travel Demand Estimation and Network Assignment Based on Cellular Network Data. *Computer Communications*, Vol. 95, 2016, pp. 29–42. https://doi.org/10.1016/j.comcom.2016.04.015.
8. Fekih, M., T. Bellemans, Z. Smoreda, P. Bonnel, A. Furno, and S. Galland. Suitability of Cellular Network Signaling Data for Origin-Destination Matrix Construction: A Case Study of Lyon Region (France). Presented at the 89th Annual Meeting of the Transportation Research Board, Washington, D.C., 2019.
9. Blondel, V. D., A. Decuyper, and G. Krings. A Survey of Results on Mobile Phone Datasets Analysis. *EPJ Data Science*, Vol. 4, No. 10, 2015, p. 55.
10. 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.
11. Alexander, L., 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: Emerging Technologies*, Vol. 58, 2015, pp. 240–250.
12. 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.
13. Tettamanti, T., H. Demeter, and I. Varga. Route Choice Estimation Based on Cellular Signaling Data. *Acta Polytechnica Hungarica*, Vol. 9, No. 4, 2012, pp. 207-220.
14. Ahas, R., A. Aasa, S. Silm, and M. Tiru. Daily Rhythms of Suburban Commuters' Movements in the Tallinn Metropolitan Area: Case Study with Mobile Positioning Data. *Transportation Research Part C: Emerging Technologies*, Vol. 18, No. 1, 2010, pp. 45–54.
15. Kang, C., X. Ma, D. Tong, and Y. Liu. Intra-Urban Human Mobility Patterns: An Urban Morphology Perspective. *Physica A: Statistical Mechanics and its Applications*, Vol. 391, No. 4, 2012, pp. 1702–1717.
16. Trasarti, R., A.-M. Olteanu-Raimond, M. Nanni, T. Couronné, B. Furletti, F. Giannotti, Z. Smoreda, and C. Ziemlicki. Discovering Urban and Country Dynamics from Mobile Phone

Data with Spatial Correlation Patterns. *Telecommunications Policy,* Vol. 39, No. 3–4, 2015, pp. 347–362.

17. Manley, E., and A. Dennett. New Forms of Data for Understanding Urban Activity in Developing Countries. *Applied Spatial Analysis and Policy*, Vol. 12, No. 1, 2018, pp. 45–70.
18. Graells-Garrido, E., and D. Saez-Trumper. A Day of Your Days: Estimating Individual Daily Journeys Using Mobile Data to Understand Urban Flow. 2016.
19. Widhalm, P., Y. Yang, M. Ulm, S. Athavale, and M. C. González. Discovering Urban Activity Patterns in Cell Phone Data. *Transportation*, Vol. 42, No. 4, 2015, pp. 597–623.
20. Zhao, Z., S.-L. Shaw, Y. Xu, F. Lu, J. Chen, and L. Yin. Understanding the Bias of Call Detail Records in Human Mobility Research. *International Journal of Geographical Information Science*, Vol. 30, No. 9, 2016, pp. 1738–1762.
21. Bonnel, P., M. Fekih, and Z. Smoreda. Origin-Destination Estimation Using Mobile Network Probe Data. *Transportation Research Procedia*, Vol. 32, 2018, pp. 69–81.
22. 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, pp. 745–768.
23. 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.
24. CERTU. *L'enquête Ménages Déplacements Standard CERTU, Éditions Du CERTU*. 2008, p. 203.
25. Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, Vol. 20, 1987, pp. 53–65.
26. Davies, D. L., and D. W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. PAMI-1, No. 2, 1979, pp. 224–227.
27. CORINE Land Cover Data. https://land.copernicus.eu/pan-european/corine-land-cover.