Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 6 (2015) 64 – 78

Transportation Research Procedia

www.elsevier.com/locate/procedia

4th International Symposium of Transport Simulation-ISTS'14, 1-4 June 2014, Corsica, France

Investigating the mobile phone data to estimate the origin destination flow and analysis; case study: Paris region

Anahid Nabavi Larijani [a*], Ana-Maria Olteanu-Raimond [a†], Julien Perret [a], Mathieu Brédif [a], Cezary Ziemlicki[b]

a Object Design and Generalisation of Topographic Information Laboratory (COGIT), National Institute of Geographic and Forestry Information (IGN), 2-4 Avenue Pasteur, 94165 Saint-Mandé, France
b Orange Labs, department of sociology and Aconomics of Network and services, 38 avenue du Général Leclerc, Issy-les-Moulineaux, France

Abstract

This paper is an output of a French national project called iSpace&Time aiming to provide a 4 dimensional platform of an urban dynamics. In order to express the urban traffic, we took an advantage of the mobile phone data to investigate the behavior of the origin destination flow within the Paris and its suburb aiming to explore the different mode of the transportation. Indeed the spatiotemporal heterogeneities of mobile phone data make the task of mode of transportation separation very challenging, sometimes even impossible. Thus, by exploring the OD matrix in order to revealing any probable continues trends or any dominant trace of the flow stating a specific mode of transportation, the commuter trains happened to be somehow detectable. Then an individual-based step-by-step approach is proposed to estimate mode of transportation from mobile phone data. Analyzing the individual trajectory, the decision is given to a segment level with respect to different measures. An early promising outcome consists of detection of the segments in which people would take the metro.

Selection and/or peer-review under responsibility of the Organizing Committee of ISTS'14

Keywords: mobile phone data; origin-destination matrix, transportation mode, smart data processing;

[*] Corresponding name: Anahid Nabavi Larijani. Tel.: +33-658-597185
[*] *E-mail address*: anahidnl@kth.se
[†] Corresponding name: Ana-Maria Olteanu-Raimond. Tel.: +33-143-988000
[†] *E-mail address:* ana-maria.raimond@ign.fr

Selection and/or peer-review under responsibility of the Organizing Committee of ISTS'14
doi:10.1016/j.trpro.2015.03.006

1. Introduction

This work was carried out on behalf of the iSpace&Time French ANR project whose primary goal is to propose an online 4D urban geo-portal platform allowing to integrate different sources of data in order to develop a tool to visualize urban area in 4 dimensions. This platform will empower the decision makers in land-use projects and urban planning studies or equipped the engineers to employ and update reference data by crowded sourcing processes. To achieve this goal, one of the steps is to generate mobile objects (e.g. cars, pedestrians, etc.) by merging data coming from different sources (e.g. mobile phone data, sensors data, surveys). Basically generated mobile objects are the input of such urban model allowing 4D simulations in an urban network.

Mobile devices are becoming ever more common hardware choice due to the sensors they provide, their cost-effectiveness, and their high presence amongst users. The popularity of mobile phones made it relatively easy to obtain the contextual data they provide. As a matter of fact it provides a unique potential for contextually aware systems by providing them with a large user base that already carries the necessary hardware around with them regularly. Therefore, collection of these data can be extremely valuable in transportation science. They can be used for investigation of the mode choice behavior, origin-destination matrix estimation, transportation demand modeling in a more realistic manner.

Although, the distinction of transportation modes plays a key role to enrich the content of urban displacements, very few studies investigated the transportation modes from the traces of mobile phones in urban areas. This mode identification generally applies for origins-destinations estimation on a large scale rather than the small ones, for example between two cities.

Our goal is to separate the transportation mode in urban areas by using mobile phone data. The estimation of transportation mode happened to be a very difficult task due to the characteristics of mobile phone data such as location accuracy, miscommunication of the individual in which the tracking is done only if an event communications or update of position takes place. Thus the frequency of records is not constant, the location depends on antennas locations and no semantic information exists.

In this context, our idea is to propose an experimental method that gradually leads towards the derivation of the transportation mode. First, we are tempted to classify the transit modes not only taking the topology of the study area into account, but also considering the essence of urban transportation from point to point. In this purpose we first generate and then explore the OD flows in order to quantify the probabilities of distinction of those modes that supposed to likely following their fixed routings. Consequently, a comprehensive study has been done over one day anonymous mobile phone data in urban area received by a French telecommunication operator called Orange. The test area covers the whole Paris Region containing Paris and its suburbs.

The paper is structured as follow: right after in the next section the related research is presented. Section 3 briefly describes the characteristics of the mobile phone data which has been conducted in this study. Section 4 represents the OD flow construction methodology. Section 5 studied the urban dynamics related to the flow in favor of early mode split trends. Section 6 proposed ideas for a step-by-step approach of transportation mode distinction and finally the conclusion and the proposition for further studies are provided.

2. Relevant study background

Several research proposing methods to estimate transportation mode for data coming from sensors exist in the literature. Such transportation data can be used in many applications such as traffic analysis, determination of the CO_2 footprint, public transportation policy and urban planning projects. Most approaches applied on GPS data which use the speed (average, median, maximum and last percentile) as one of the most discriminating criterion (Schüssler and Axhausen, 2009; Stopher et al., 2008). Other criteria such as the travel distance in relation to the urban topography or transportation network (Biljecki et al., 2013), speed acceleration (Zheng et al., 2008) or management (Zheng et al., 2008) also has been used to discriminate transportation modes by their relatively close speeds (bus, car, tram). The lack of locations due to the loss of signal is exploited to determine the displacements taking place in metro (Biljecki et al., 2013). Many research studies have shown that the integration of different spatial data improves the quality of the mode identification. In this context, the survey data has been used to calculate the indicators who subsequently deployed to classify the segments (Gonzales et al., 2010) or to determine the transition points like the spots that mark the mode switch (Biljecki et al., 2013). Public transportation network is used to identify the path near a metro station, tram, and train (Liao et al, 2007; Biljecki et al., 2013). The location of buses in real time could be a source of information leading to distinguish between bus trips and car or tram trips (Stenneth et al., 2010).

Referring to the recent studies, the number of identified transportation modes is from three (Liao et al., 2007; Gonzales et al., 2010) to ten modes (Biljecki et al., 2013).

Among the existed proposed methods the most practical ones are those which use supervised learning approaches (Zheng et al., 2008; Biljecki et al., 2013). The supervised learning based approaches usually follow two steps: the first step (learning) is to build mobility patterns from historical data or the paths already classified by transportation mode; the second stage (inference) is to infer transportation modes using different methods of decision making. These methods require a knowledge base and therefore a manual paths classification step. In general, the knowledge base uses survey on a limited number of people.

Unsupervised learning methods are also used in few researches (Patterson et al., 2003; Liao et al., 2007; Gonzales et al., 2010). In addition to transportation mode detection, Patterson et al. (2003) and Liao et al. (2007) introduce the path following process and its goal by integrating the historical paths and parking spots for cars users. Gonzales et al. (2010) proposed a non-supervised classification based on a neural network to classify paths from an integrated GPS in smart phones. They focused on three modes of transportation: car, bus and walking. The advantage of this method is that it does not require a knowledge base though it is quite sensitive to the data quality.

Reddy et al. (2010) tested several methods of classification (k-NN, Bayesian network, Hidden Markov Models (HMM), decision tree) to extrapolate the transportation mode by traces of single user connected with five GPS sensors simultaneously. They showed, first that the best classification was obtained by combining the decision trees based method with HMM and secondly, that the sensor located in the pocket is much less sensitive in the detection process. Nevertheless, the approach proposed by Reddy et al. (2008) is rather complicated to reproduce since it requires very accurate recordings and several sensor equipments.

The most of the approaches were tested on GPS data, and a relatively small sample of users from one user (Patterson et al., 2003; Liao et al., 2007) to hundred users (Stenneth et al., 2010; Biljecki et al., 2013). Although the good results obtained using GPS data, the transportation mode detection and the use of these results for different goals still remains limited by the sample size of people accepting to wear a GPS sensor for the duration of the investigation. Mobile phone data have a significant advantage that makes a large data sample of users available on a large spatial and temporal scale. However, due to the characteristics of mobile phone data very few studies that estimate the mode transportation from mobile phone data exist in the literature.

In general, the identification of modes of transportation from mobile phone data is made for OD Flows on a large scale, for example between two cities (Doyle et al., 2011). Caceres et al. (2008) proposed a method to estimate the mode of transportation from OD flows using criteria such as travel speed, travel time and traffic information. Using CDR data (cell detailed records), Wang et al. (2013) proposed a method which determines the percentage of travelers using driving and public transportation modes for a given origin and destination. It is based on travel time criterion. Basically they focused on trips longer than 3 km, thus walking trips are not taking into account since the assumption is that a traveler would hardly walk for more than 3 km.

3. Mobile data characteristics and network structure

Each mobile phone operator collects the communication data and stores them for a given period of mobile phones customers' activities most commonly for billing or for technical measurements purposes. This type of collection is called "passive collection", since recordings are made automatically. They are three mainly types of mobile phone data collected using the passive collection: Call Detail Records (CDR) data, Probes data and Wi-Fi data. In this paper, only Probes data are described, since just this type is applied at the preliminary study. For more information about mobile phone data, see Smoreda et al. (2012).

Probes data are issued from mobile network probes, named Mobile Switching Center (MSC data). They are generally anonymous meaning that each mobile phone SIM identifier is replaced with an identifier consisting in a unique integer. In addition they contain both cell-localized communication events (i.e. calls and SMS) and itinerancy events: handover (HO) and Location Area Update (LAU). For instance, in Paris Region a location area consists of 150 base stations in average. The location of mobile phone users is limited to the base station location which is composed by a minimum of three antennas, each antenna having a spatial coverage. The records contains the following attributes: i) the anonymous SIM card identifier; ii) the antenna identifier; iii) the base station location (BTS); iv) the record type (call in/out, SMS sent/received) and v) the time of communication activity (timestamp).

Fig. 1 shows an example of mobile phone localization according to different events that occur: call, handover, LAU update and SMS.

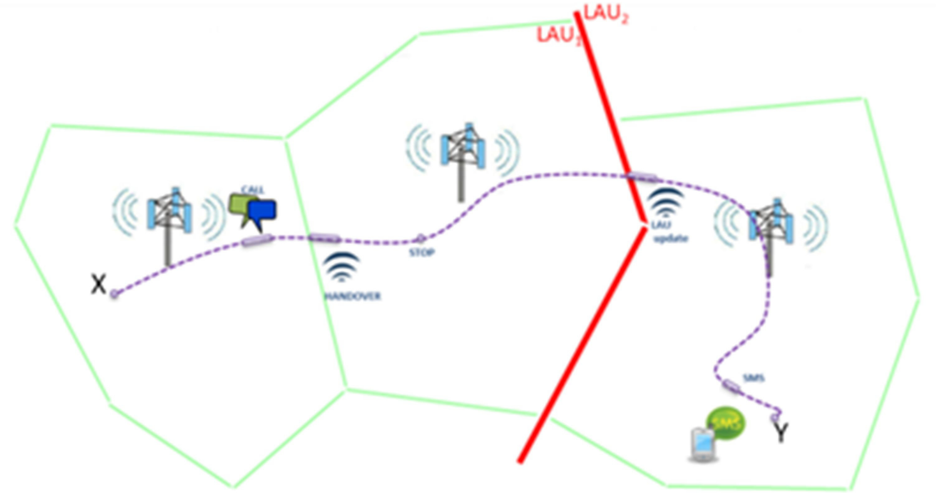

Fig. 1. An example of mobile phone network localization data types for one user travelling from a point X to a point Y

In this study the focus is in one weekday mobile phone data of the Paris Region area (12.012 km² - 4.638 km^2) (Fig. 2). Data are generated from Orange network probes. They are anonymous (attributed by a secure, random network temporal identification) and contain both cell localized communication events (calls and SMS) and update events (handover and location area update). The dataset covers 1.4 million of French mobile phone users and more than 57 million records. The location of mobile phone users is limited to the base station location, so coarser and heterogeneous grain spatial accuracy. The study areas contains 3127 BTS where 313 antennas are located at less than 50m distances one from each other and 2481 antennas stand at less than 100m from each other. The accuracy of data varies from 30 m in urban areas to 30 km in rural areas which is rather desirable in urban neighborhood than in rural areas indeed.

Notice that approximately 50% of records are due to the LAU update events, which means that data can capture the user mobility quite well. The main advantages of mobile phone data are: the big mass of located data for a long period of time and for rather a large spatial extends (i.e. country level). Moreover, records happened to represent all operator' clients, not only a sample of users. The disadvantage is due to the heterogeneity of records (limited to a communication event).

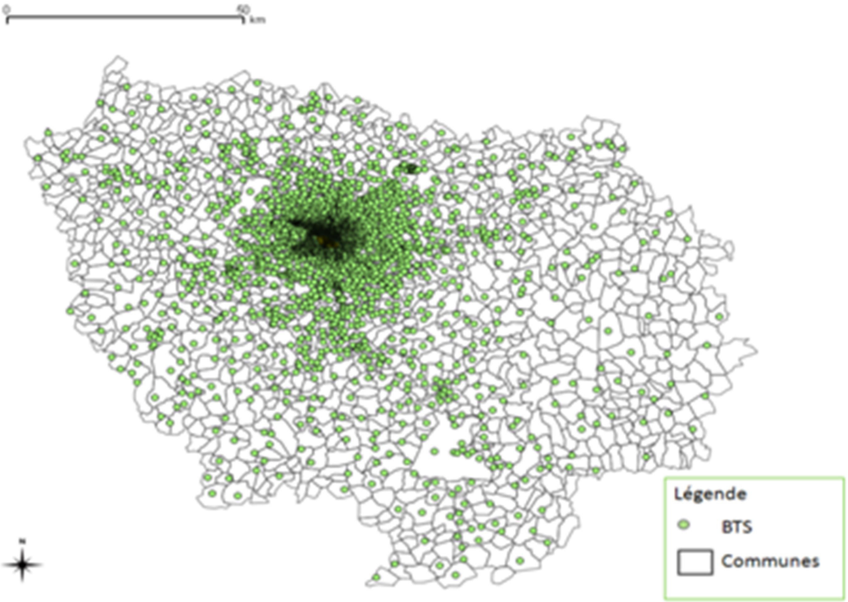

Fig. 2. Paris region study area: the municipalities and the Base Station Locations

4. Origin-Destination estimation methodology

A variety of studies, having a specific goal to infer OD flow matrix using sensor data such as mobile phone and GPS, were carried out in recent years. They basically stand on temporal variation of association rules (Friaz-Martinez et al., 2012), identifying moving and staying points (Byeong-Seok et al., 2005) or by computing stops and trips in individual trajectories and then extracting flows for each trips (Calabrese et al., 2011).

Flow can also be defined by its start spot (the first(s) point(s) of the path) in the origin region and its end spot (the last point(s) of the path) in the destination region (Caceres et al., 2007, Giannotti et al., 2011, Bahoken and Olteanu-Raimond, 2013). A time gap window is thus necessary to be defined in the essence of the fine temporal analysis.

In this project, the method proposed by of Bahoken and Olteanu-Raimond, (2007) is applied since it takes into account the spatiotemporal heterogeneities characterizing mobile phone data. The method defines a flow by taking into account the first two points and the last two points of a trajectory in a given temporal window. Let's point out that the trajectory could be a daily trajectory or a subset of the daily trajectory of a user. The origin and destination correspond to the more detailed spatial level allowed by mobile phone data which is in BTS level. Thus, first the Voronoi cells are computed using BTS positions, then a temporal filtering is applied, and finally the OD flows at Voronoi scale are computed on the trajectories belonging to the temporal window.

5. Data analysis

Now we are going to present the behavior of Paris region flows in order to have a better comprehension of data dispersion. This would guide us one step ahead to the final goal since the idea is to look for dominant trend of the flow likely representing a certain mode of transportation. Hence, we start by represent the methods applied to prepare the data for its projection on the study zone. Later on there are examples of figures and the discussion about the behavior of the flow.

As mentioned in the former section the method proposed in Bahoken and Olteanu-Raimond (2013) has been applied to estimate flows at Voronoi scale. Each Voronoi would contain the entering flow as well as the exiting flows if it is the case. To construct the OD matrix, it was sufficient to track the detected flow between zone of origin (Voronoi with exiting flow) and zone of destination (Voronoi with entering flow). Putting all the flow engaged zones against each other a matrix of 3040 x 3040 is driven. In order to represent a better illustration of such OD matrix, the following procedure is provided to project the OD matrix on the topography of the study area. Thanks to the Quantum GIS 1.8 which has been deployed to visualize the flow distribution and enables further investigation of the data.

In summary by applying two temporal filters, two OD matrixes are obtained: one from morning traffic between 6 am and 10 am the second one from 2 pm to 9 pm represented the afternoon flow. Statistically speaking, first the raw data has been investigated. Then by conducting some filtering method, the zones in the least interest has been eliminated. The second adjustment considered the new assigned flow through the area of the study. Moreover the flow has been normalized by the size of zones leading to diminish the effect of such various area sizes and the induced biases. The following discussion would go deeper in the outcomes of the processes.

5.1. OD FLOW STATISTICS

Table 1 presents the preliminary statistics such as the summation of the flow and the range for both morning and afternoon data. Variance and standard deviation has been calculated as well as minimum and maximum of the flow. The imperfection of the afternoon flow is quite dominant comparing with the amount of the morning flow.

Table 1. Basic data statistics

	Total flow	Max	Min	Mean	Mode	Variance	SD
Morning	1 256 272	885	1	2.7	1	117.2	10.8
Afternoon	11 973	14	1	1.2	1	0.29	0.51

According to basic statistics, a number of 1 256 272 flows in the morning and 11 973 flows for the afternoon has been spotted. There is no flow detected in the current records between two points in the area with rather a long distance (e.g. 80km, 140km) which was as expected. The expansion of the study region is sufficiently high. As a consequence, there are many OD pairs, in which, it would be rather meaningless to pass through in such long distances or even not feasible in some cases. They shall be called passive OD pairs against the active OD pairs for a network of this size. Nevertheless one might say that there are still active ODs among them with the likely trips which are not presented here because of the lack of the data.

The variance of the morning flow is rather high caused by the high maximum recorded flow as 885. Furthermore, it turns out that the most repetitive flow is 1 for both morning and afternoon active OD pairs. Table 2 captures the statistics about detected flow and their share. According to the data 25% of the morning flow is equal to 1, while this is true for more than 80% of the afternoon flow. This statistics are presented in the two last column of the Table 2.

Table 2. Share of the total recorded flow

	Total number of OD-pairs detected by flow records	Flow =1	otherwise
Morning	467 069	316 688 (25%)	939 584 (75%)
Afternoon	10 656	9 797 (82%)	2 176 (18%)

Summing up the flows of detected active OD pairs by each zone as origin for all exiting flow, and as destination for all entering flow, enable us to build up the matrix with accumulated flow. The share of low flow for the afternoon dropped down as much as less than the half (Table 3). This means that for just a bit less than half of the origins and the destinations, there is no flow recorded in the afternoon period at this specific date. Instead, the morning records appeared to be rather rich with almost 100% of the zones that has been provided flow more than 1.

Table 3. Share of the accumulated flow equal to 1

	Origin	Destination
Morning flow	0.4%	0.2%
Afternoon flow	44%	45%

The flow equal to one for an OD does not make a significant treat to reveal a certain conclusion. Besides, these values could contrary have a biased effect on the trend exploration. Therefore it has been decided to filter out the low flow zones out of interest at this stage (for instance zones with accumulated flow lower than 10). The rest remains to be treated for the more visual analysis on the maps. The second adjustment considers the flow normalized by the size of zones. In another word we have applied this adjustment tempting to control the effect of the various area and zone size with the new assigned flow.

The statistics of the filtered data is shown in the Table 4. The result shows that the filtering would cut off the afternoon flow as much as almost the half. But in the morning period, there is not much of the difference since 99% of the flow is higher than 1. As represented in the table, the average of the flow raised up considerably since the new sample is out of the low flow. Hence the focus will be on the filtered flow from now on.

Table 4. Filtered accumulated flow statistics

Accumulated flow > 9	Morning:		Afternoon:	
	Origin	Destination	Origin	Destination
Number of active OD pairs	2 613	2 827	383	416
Sum of the flow	1 254 791	1 255 464	6 032	6 276
Flow share	99%	99%	50%	52%
Flow average	480.2	444.1	15.75	15.1

The next part has been provided aiming to the more visual presentation of the analysis and further interpretation.

5.2. *MORNING FLOW BEHAVIOR*

After the preliminary statistics trial and data adjustment now it is time to analyze the data in various perspectives to signify the specific trend or dominant travel behavior. According to data, 1 256 272 flows have been received from morning hours representing the number of trips made during this hours.

First the flow distribution with equal interval in basis of generated trips (exiting flow) by each zone of the region has been created. It contains all the flows from the morning hours between 6 am and 10 am. At the same manner the distribution of flow based on attracted trips (entering flow) has been provided. It turns out that there is rather no specific concentration of the flow around any group of zones. This means that the data is rather dispersed in the whole region without an interesting source of clue to conclude a consistent manner of demand. This pattern is true not only for the entering flow but also for the exiting flow.

In addition no correlation has been revealed between origin-destination pairs. There are some identical zones with relative higher flow shown by darker color. One might say such repetitive appearances are the cause of the higher share of inter-zonal flow in these zones. Nevertheless the classification is varying, caused by the nature of the data. Therefore there are some lost repetitions in the entering flow diagram in consequence of a wider range of classes.

Secondly we tried to be concentrated in the higher values behavior. Therefore the flow lower than 10 has been filtered out of the morning data.

Among tree pre-examined classifications as Equal intervals, Natural breaks and the Quantile, ultimately the latter (quantile classification) has been chosen because of its better representation of this concept.

In the quantile classification method, each class contains the same number of features. As shown in the Fig. 3, the forth class dominated by the dark purple, could be easily found all around the region. Clearly the high flow zones are not limited only to be around the center. There are some continuous behaviors detected by the highest range in the same class. Exploring the Google map, this is most likely caused by the commuter trains connecting the city center to the suburb, besides the airports such as Orly and Charles-de-Gaulle in the South and Northeast respectively. It is worth to mention that there are indeed some linked attractions in suburban area, as an example the Disneyland in the East and Chateau de Versailles in the west, which also happened to be followed by the data.

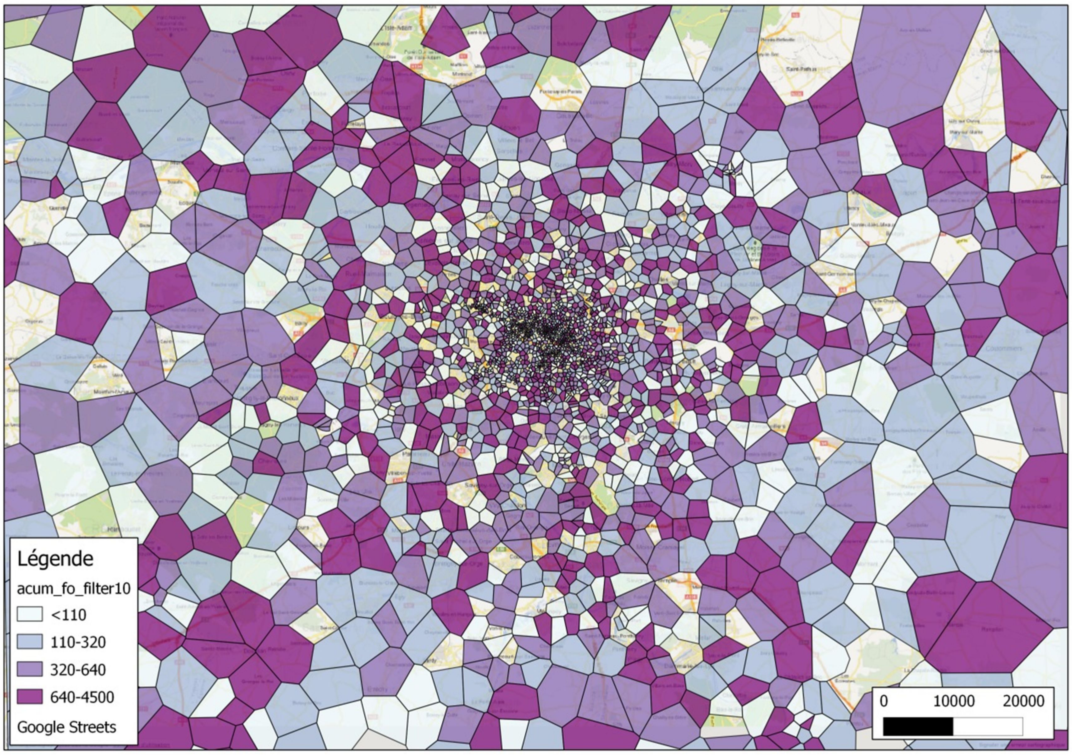

Fig. 3. Filtered Origin flow distribution by quantile classification

It has been noticed that the size of zones according to their area are quite vary. As described before, the construction of the OD zones are based on the location of the BTS. Hence, in order to decrease the inevitable bias in data, since for example one Voronoi can covers more than one municipality in rural area the flow has been normalized by the Voronoi coverage area. As a result Fig. 4 is generated for the normalized morning flows by the same classification as quantile.

Although the map shows considerable changes, it is evident that some continuous trends, which were tracked out of the complete data in Fig. 3, still remain in the Fig. 4 as well. Many high flows in the border of the region, appeared specifically in the larger zones, has been diminished in the normalized graph as expected.

Moreover moving from border to the center, the flow has an increasing trend. This could have many reasons such as socio-economical aspect besides urbanization issues. Needless to say that there are more working activities, schools and touristic attractions zooming in to the Paris area over all the territory which leads the flow to be more concentrated in the center than surroundings. It is also more convincible that the flow within the suburb is lower than the flow between suburb and the Paris.

Following this conclusion, the hypothesis of using the public transport instead of private vehicles is getting stronger. The darker separated zones in the urban area would be an evident of the intensity of the inter-zonal trips. This would increase the probability of walking, taking bike or bus.

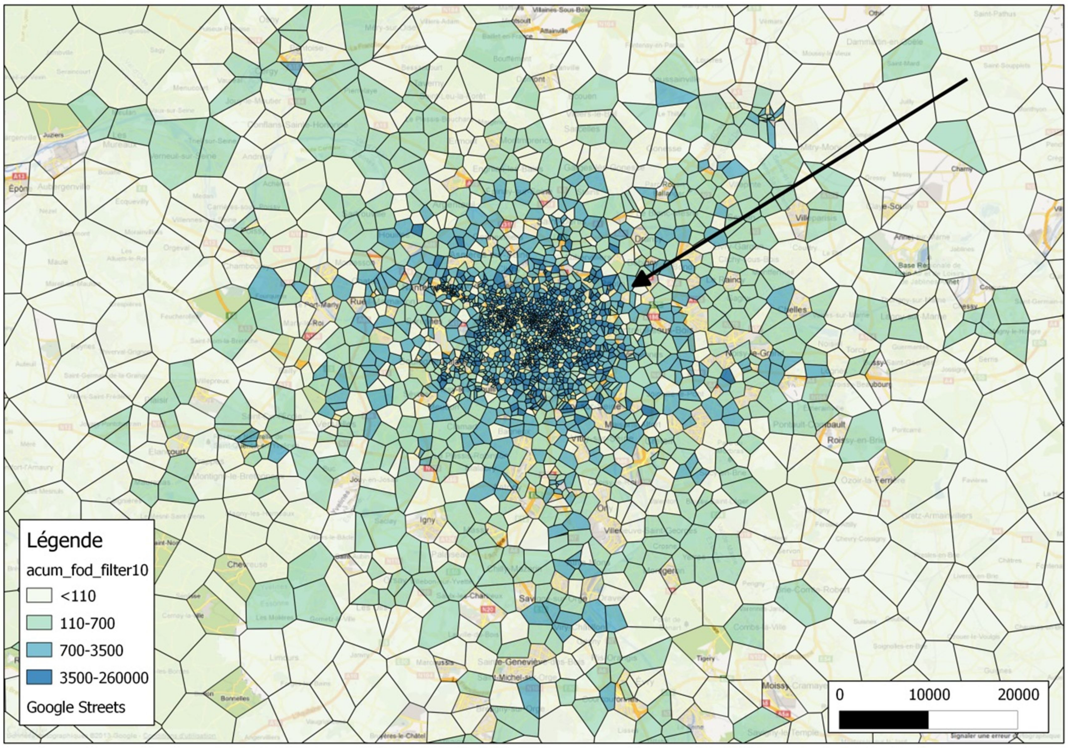

Fig. 4. Normalized filtered Origin flow distribution by quantile classification

In order to avoid the repetition of the figures we just literally confirm that the map illustrates a reasonable cover of the high flow in the whole region expansion for the destination flow as well. Therefore the discussion over the origin flow is repetitive for the destination flow.

It can be concluded that there is no considerable difference between origin flow and destination flow in terms of utilization. In other words, comparing the origin flow and the destination flow demonstrates that the zone trip attraction plays rather a same role as the zone trip generation. None of the zones expose a huge flow as generation of trips while being inactive in absorbing trips. This result plays a barrier role to deeper investigation of the mode detection at this stage.

5.3. AFTERNOON FLOW BEHAVIOR

As mentioned earlier 11 973 number of flow has been spotted for the afternoon time period. The very same procedures have been applied to treat and explore the afternoon flow data. As a result both origin distribution and destination distribution appeared more or less the same, through the study area. Since the share of registered data was much lower than morning period it has been decided not to capture more details. It appeared that further investigation of the data in this way could not reveal any more interesting clues.

6. Transportation mode detection

Following the discussion of the previous section, there are some continuous behaviors detected by the highest range of flow. Projecting the Google map, this is caused by the transit routes, more specifically commuter trains, besides the airports and some attractions in suburban area at the east and west. We realized that trend of the trips taking by commuter trains became dominant since they introduce rather longer trips than the other modes inclining to pass shorter distances.

In this context and knowing the inevitable challenge of the task, we are going to propose an individual based approach for transportation mode estimation. Thus, considering a spatiotemporal trajectory reconstructed from one day mobile phone records earlier, our approach aims to classify segments of trips by mode of transport. A segment is an estimated movement (a direct line) between two consecutive records. The idea is to distinguish between motorizes (e.g. cars, public transportation), non-motorized modes (e.g. walking mode) and uncategorized mode. This last category allows last identifies cases where the identification process is not capable of making a decision. So given the adversity of the problem due to spatiotemporal heterogeneity of mobile phone data, our idea knees down to gradually classify the segments of the trajectory. Step-by-step exploration discussion:

With respect to both public transportation and telecommunication networks in Paris region, the public transportation mode could be segmented in trains, metro and other (tramway and bus). In the first step, the metro segments are possible to be detected. Therefore, we try to reveal the segments in which the mode of transportation is the metro. Considering the Telecom network, the underground in Paris Region is characterized by a specific Local Area Code (LAC). Whenever a person gets in/out from the metro, a LAU event occurs and a new location is recorded in the database. Thus, the criteria of the metro flow detection stand on this information.

Secondly, it turns out that the commuter trains are rather traceable. Train segments could be detected using grouped handover update, topographic railways and the analysis of the daily trajectory as a whole. The same idea could be also used for tramways and buses but the difficulty occurs in urban area where same routes are shared by tramways and buses in many part of the network, their routes are highly closed to each other, their speeds are relatively closed as well. This is the drawbacks where transportation network data consisting of real time location of buses, tram lines, and bus stops spatial data is not available.

Thirdly, we look to classify car segments on none yet classified segments. Analyzing the OD flow and topographic roads network, we realised that this segmentation is only possible on long continuous distances, when the mean speed is relatively high and when consecutives segments are near to the high speed roads such a highway, internal and externals rings. For example, if a point is located close to a highway, there is more chance that it belongs to a motorized vehicle than a pedestrian.

Ultimately, the remaining segments of a daily trajectory are labeled as uncategorized. In the following sub section we explore and present the detection of the metro segments.

6.1. METRO FLOW DETECTION APPROACH

In this section, the metro detection is described and some results are presented. As explained earlier, metro segments assessment is based on metro LAC entity which is explicitly defined in Paris region. Two attributes from mobile phone data are especially used for this purpose: LAC (the identifier of the current LAC to which the point belongs) and Former-LAC (the identifier of the LAC where the point was located just a moment before). The method composes different steps stating here (Fig. 5):

- *Metro entry point detection:* for each individual trajectory and each point of the trajectory, we look for an entry point which consists to verify if the current LAC of the point is the metro LAC and the Former-LAC is different from the metro LAC.
- *Metro exit point detection:* for each point of the trajectory, registered after the entry point, we look for an exit point which consists to verify if the Former-LAC is the metro LAC and the current LAC of the point is different from the metro LAC.
- *Quality tests:* some indicators are defined in order to verify the quality of the entry and exit points detection. Thus, first we look for entry points having any exit points in a given time interval. This case occurs since our data not contain 3G users. They are included in our database where they enter in the metro, since the metro have only 2G antennas and are not captured when they exit from the metro, the 3G mobile phone is connected to a 3G antennas by default. It is worth to mention that the entry points without exit can be used in the case we are interested to study the characteristics of areas generating moves and flows. Furthermore, the consecutives entry and exit points having a time interval longer than a threshold or in a time interval less than 2 minutes are also filtered out.
- *OD flow construction:* after filtering, OD flows are defined for each trajectory treating origin and destination as the entry point and the exit point respectively. Hence, each OD flow is defined by linking two consecutives entry point and exit point with respect to the temporal sequence criteria of the entry and exit points (i.e. the exit point need to be occurred in time after the entry point). Needless to say that more than two OD flows could be detected in daily trajectory output capturing the daily home-work trips.
- *Segments labeling:* all segments between the origin (entry point) and the destination (exit point) are classified as metro.

The metro detection approach is illustrated in Fig. 5, where it can be noticed detection of the entry (A) and exit (E) points, as well as the definition of OD flows (i.e. origin = point A, destination = point E) and finally the labeling of the segments between the Origin (A) and the destination (B) as a metro segment.

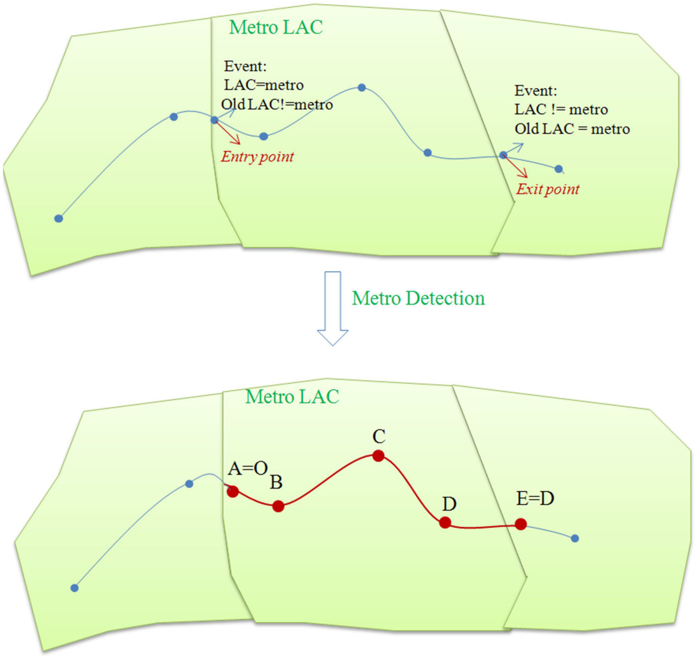

Fig. 5. Metro detection approach illustration

6.2. *METRO FLOW DISTRIBUTION*

The metro detection approach was tested on one weekday mobile phone data. From 1.4 million of French mobile phone users, 34% (485 931) of users has at least one entry in the metro area and 32% (457 871) have a minimum one exit from the metro. Generally, the 2% difference is due to the 3G users who connect to the 3G antenna where they get out from the metro. As we described in the former section, once the input and exit points are spotted, some quality tests has been conducted. As a result, the input points without exit are eliminated (29%). In addition, except the train drivers and staff who spend time in the metro area, we consider that a 3h travel by metro is way too much. Thus, trips by metro having the duration more than 3 hours, representing 11% of trips, are also filtered out. Finally, if the duration of trips is than 2 minutes and the *exit point* is in the neighborhood of the *entry point,* then the trip is removed. This case occurs mostly when people take the metro corridors to across the street or a major intersection. Once the quality tests step carried out, 181 000 OD flows in Paris region are computed at Voronoi scale.

It has been realized that 296 Voronoi areas generate flows and 1 413 Voronoi areas are the attractions areas while receiving flows. The difference is due to the fact that when a user gets out from a metro station, the mobile phone connects to the nearest antenna, belongs to a LAC different from the metro LAC which overestimates the number of areas receiving flows. To overcome this inconveniency, one idea could be to characterize each Voronoi area representing a metro station by a set of Voronoi area representing the first neighborhood, and then to aggregate the flows by metro station and his neighborhood.

Fig. 6 shows the spatial distribution of flow generated areas. The metro stations entries are illustrated by proportional circles (i.e. the size of the circles are proportional with the number of generated flows). The most active metro stations generating the flows are Chatelet, La Defense and Nation, followed by the stations that are also train stations: St-Lazare, Gare de Lyon. Let notice the importance of line 1 (La Defense-Chateau de Vincennes) who is quite a busy line among all.

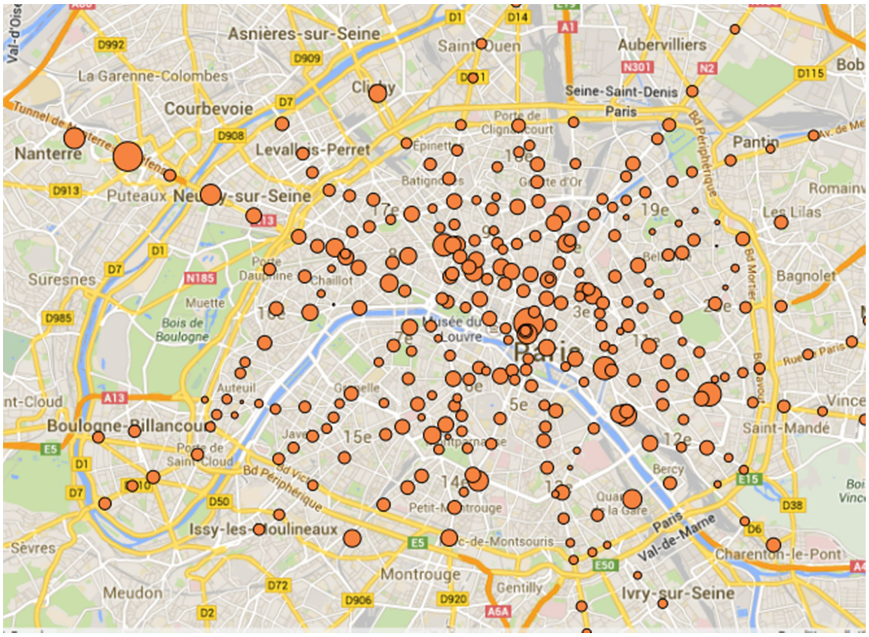

Fig. 6. Distribution of origin flows by metro

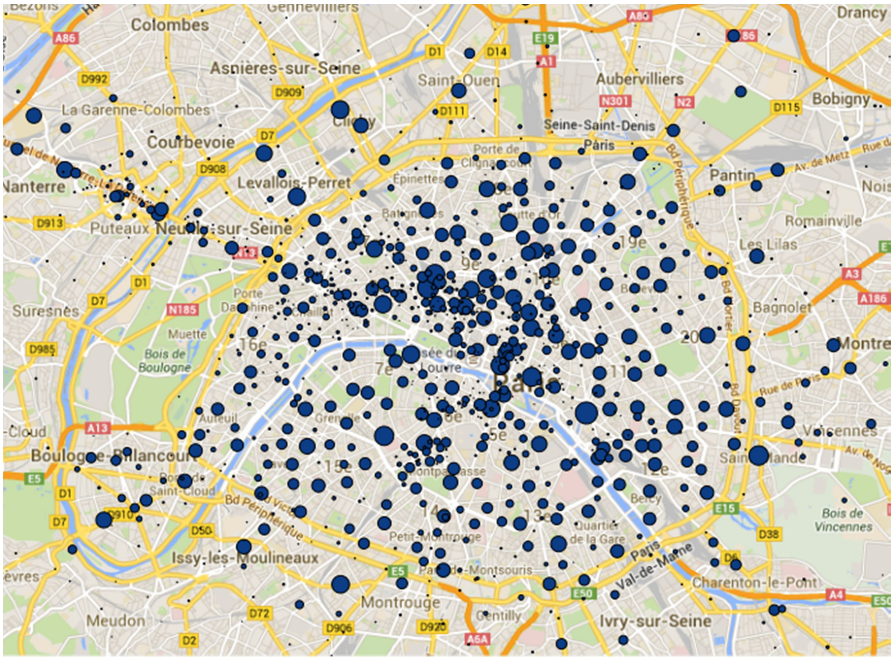

Fig. 7. Distribution of destination flows by metro

Spatial distribution of attractions areas is illustrated by proportional circles, i.e. the size of the circles is proportional with the number of arrival flows (Fig. 7). We notice here, a more homogeneous attractions area. The neighborhoods of stations such as Bastille, St-Lazare or Saint-Mandé are clearly highlighted.

One of the explicit advantages of the mobile phone data is that it provides the opportunity of dynamic analysis such as in daily basis, per hour, by every 5 minutes, etc. Our future work has been foreseen within this perspective.

7. Conclusion and proposition for future works

At this point we are pleased to discover valuable findings while facing some inevitable obstacles. Moreover, we studied the flows in Paris Region in order to guide us one step ahead to the final goal since the idea was to look for dominant trends of the flow likely revealing a certain mode of transportation. Nevertheless this paper represents the early steps of the transportation mode detection research which is still in progress. Basically the application test was focused more on metro detection. As a conclusion, here the summarized barriers are presented as well as the final results of the study.

Over two time period of the data, morning and afternoon, the focus remains on the morning flow because of the imperfection of the afternoon flow compare to the morning one. The data showed that for about half of the origins and the destinations, there was no flow recorded in the afternoon period.

It hasn’t been any detected flow between two zones with rather a long distance. Either the share of long distance trips was not significant or the trips for such trajectories are not in an interest in reality.

In general perspective there is rather no specific concentration of the flow around any group of zones. This means that the data is rather dispersed in the whole region without an interesting source of clue to conclude a consistent manner of demand. Nevertheless moving from border to the center, the flow has an increasing trend. Lack of the public transport in suburb, urbanization issues and socio-economic aspects, more working activities, schools and tourist attractions could cause this phenomenon.

There are some continuous behaviors detected by the highest range of flow. Investigating the Google map, this is likely caused by the transit routes, mostly commuter trains, linking the airports from the suburb to the city center, besides some touristic attractions in suburban area.

As we explicitly disclosed earlier, the metro users could be traced rather easy thanks to the telecom network structure in Paris. Note that some quality tests are needed to be done in order to detect erroneous entry and exit points. Our metro detection approach was tested on one day mobile phone data and encouraging results have been obtained.

The approach presented here is deeply under development. The future works go for more advanced assessment strategies in its perspective. First of all it is planned to apply the metro detection approach on mobile phone data for a longer period, more than just a day. Secondly, a step validation will be necessary by comparing the separated metro flows with reference data from RATP group, the public transport operator in Paris. The validation can be made only for origins of metro flows, since the provided data by RATP group is rather general containing only the number of entries per metro station aggregated by year. Thirdly, once the validation is done, it will be interesting to conduct the urban dynamics studies and the travel demand history with respect to the OD flows of metro users and work on disaggregation of flows using statistical data.

Another perspective is to apply this approach to detect the other modes of transportation such as trains, cars, etc. Pedestrian are the most challenging flows and hardly detectable using mobile phone data. Clearly, the records are depended on the antenna locations and pedestrian walks are quite short by its nature, therefore the probability to walk around the same antenna coverage is rather high. Nevertheless, for some users and certain segments it might be possible to manage. For example, it should be assumed that before entering in the metro and just after the user gets out from the metro, most probably he/she would be walking. Hence the segments having walking mean of transportation can be inserted by adding such hypothetical locations while the question remains as at what level this separation of pedestrian flow is significantly important.

Acknowledgements

This work is a part of the French National Research Agency (ANR) project called ISpace&Time whose primary goal is to create a 4 dimensional platform of the city dynamics including not only the 3D motion of the streets and buildings in urban area but also the dynamic movements of the vehicles and the pedestrians passing through the city. The project is funded by the ANR under the responsibility of National Institute of Geographic and Forestry Information (IGN).

References

Bahoken, F. and Olteanu-Raimond AM., 2013, 'Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths: The effect of spatiotemporal filtering on flow measurement', In proceedings of 23rd International Cartography Conference, 2013.

Biljecki, F., Ledoux, H. and Van Oosterom, Peter, 2013, 'Transportation mode-based segmentation and classification of movement trajectories', International Journal of Geographical Information Science (IJGIS), vol. 27, n°2, pp. 385-407.

Byeong-Seok, Y. and Kyungsoo C., 2005, 'Origin-destination estimation using cellular phone as information', Journal of the Eastern Asia Society for Transportation Studies, vol. 6, 2005, pp. 2574–2588.

Calabrese F., Di Lorenzo G., Liu L. and Ratti C., 2011, 'Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area', IEEE Pervasive Computing, vol.10, n°4, 2011, pp. 36-44.

Caceres N., Wideberg J. and Benitez F., 2007, 'Deriving origin-destination data from a mobile phone network, IET Intelligent Transport Systems, vol.1, n° 1, 2007, pp. 15-26.

Giannotti F., Nanni M., Pedreschi D., Pinelli F., Renso C., Rinzivillo S. and Trasarti R., 2011, 'Unveiling the complexity of human mobility by querying and mining massive trajectory data', International Journal on Very Large Data Bases, vol. 20, n° 5, pp. 695-719.

Gonzalez P., Weinstein, J., Barbeau, S., Labrador, M., Winters P., Georggi, N. and Perez R., 2008, 'Automating Mode Detection Using Neural Networks and Assisted GPS Data Collected Using GPS-enabled Mobile Phones', In Proceedings of 15th World Congress on Intelligent Transportation Systems, New York, 2008.

Liao L., Patterson D.J., Fox D., Kautz H., 2007, 'Learning and inferring transportation routines', Artificial Intelligence (AIJ), 171(5-6): 311-331, 2007.

Patterson D.J., Liao L., Fox D. and H. Kautz, 2003, 'Inferring High-Level Behavior from Low-Level Sensors', ACM UbiComp (ACM International Conference on Ubiquitous Computing) 2003.

Reddy S., Mun M., Burke J., Estrin D., Hansen M. and Srivastava M., 2010, 'Using Mobile Phones to Determine Transportation Modes', ACM Transactions on Sensor Networks, Vol. 6, No. 2, Article 13, 2010.

Schüssler, N. and Axhausen, K.W., 2009, 'Processing raw data from global positioning systems without additional information', Transportation Research Record: Journal of the Transportation Research Board, 2105 (4), pp. 28–36.

Smoreda Z, Olteanu-Raimond AM and Couronné T., 2013, 'Spatiotemporal data from mobile phones for personal mobility assessment', In Zmud J, Lee-Gosselin M, Carrasco JA, Munizaga MA (eds), Transport Survey Methods: Best Practice for Decision Making, Emerald Group Publishing, London, 2013.

Stenneth, L., Wolfson, O., Yu P.S and Bo Xu, B., 2011, 'Transformation mode detection using mobile phone', In proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp.54-63.

Stopher, P., Clifford E., Zhang J. and FitzGerald C., 2008, 'Deducing mode and purpose from GPS data', Working paper of Institute of Transport and Logistics, University of Sydney, ITLS-WP-08-06, 2008.

Zheng Y., Liu L., Wang L. and Xie X., 2008, 'Learning Transportation Mode from Raw GPS Data for Geographic applications on the Web', Proceedings of the 17th International World Wide Web Conference (WWW 2008), pp. 247-254, Beijing, China.

Wang, M.H, Schrock, S., Vander Broek, N. and Mulinazzi, T., 2013, 'Estimating Dynamic Origin-Destination Data and Travel Demand Using Cell Phone Network Data', International Journal of Intelligent Transportation Systems Research, Vol 11, N°2, 2013, pp. 76-86.