Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 32 (2018) 69–81

Transportation Research Procedia

www.elsevier.com/locate/procedia

International Steering Committee for Transport Survey Conferences

Origin-Destination estimation using mobile network probe data

Patrick Bonnel [a1], Mariem Fekih [b,c], Zbigniew Smoreda [b]

[a] *Laboratoire Aménagement Economie Transports, ENTPE, Lyon, France, patrick.bonnel@entpe.fr*
[b] *SENSE, Orange Labs, Paris, France, mariem.fekih@orange.com, zbigniew.smoreda@orange.com*
[c] *IMOB, Hasselt University, Diepenbeek, Belgium*

Abstract

Mobile phone operators produce enormous amounts of data. In this paper we present applications performed with a dataset (probe data) collected by the operator Orange in 2017 in Rhône Alpes Region, France. Trips are deduced from the spatio-temporal trajectory of devices through a hypothesis of stationarity in order to define activities. Trips are then aggregated in an origin-destination matrix which is compared with traditional data (household travel survey). With some hypothesis we obtain somewhat similar origin-destination matrix, with a slope close to one when we regress the number of trips of each origin-destination from mobile phone data with household survey data.

Keywords: origin-destination matrix; mobile phone data; travel survey; passive data; Rhône-Alpes

1. Introduction

Data on spatial mobility are essential in order to build and use travel demand forecasting models, for transport planning purposes and for the appraisal of transport policies… (Arentze et al., 2000; Ortuzar, Bates, 2000). While

[1] Corresponding author. Tel.: +33 4 72 04 70 48; fax: +33 4 72 04 70 92.
E-mail address: Patrick.bonnel@entpe.fr

10.1016/j.trpro.2018.10.013

travel surveys provide extremely useful data in order to formalize and estimate behavioural choice models (for example the choice of a destination or mode of transportation), they are much less useful for constructing origin-destination (O-D) matrices due to an inadequate number of trips in many of the matrix elements. In addition, surveys are increasingly confronted by issues during the sample construction phase (Stopher, Greaves, 2007), by falling response rates (Atrostic, Burt, 1999; Ampt, 1997; Bonnel, 2003; Zmud, 2003) and by unreported trips (Wolf et al., 2003; Janzen et al., 2018), which reduce even further the quality of the resulting matrices.

Large volumes of data are produced automatically and passively such as ticketing data (Arana et al., 2014, Morency et al., 2007; Munizaga et al., 2010; Pelletier et al., 2011), bank cards… and mobile phone data which makes it possible to identify the presence of individuals in both space and time. Several researches have been performed to use the data, but little research has attempted to "validate" them by comparing them with data from other sources in order to identify possible biases and gain a clearer idea of their potential. However, the quality and accuracy of data is essential to ensure that investment or transport policy decisions are based on reliable analyses. We must therefore analyse these data in order to carefully evaluate their usability.

The aim of this paper is therefore to test the potential of mobile phone data for producing origin-destination matrices compared with other sources of available data. The analysis has been conducted within the Rhône-Alpes Region for which we were able to study the mobile phone data from the operator Orange and compare them with the data obtained from the travel survey performed in Rhône-Alpes Region.

We will begin this paper with a literature review (Section 1) before presenting the data we have used (Section 2) and the data processing methodology used to produce the origin-destination matrices (Section 3). This will allow us to compare our results with external validation data (Section 4). Finally, we will present the main lessons learned from this research, as well as some suggestions for future research directions (section 5).

2. Literature survey

Cell phone networks have existed for two decades, and mobile phones have achieved a high rate of penetration: there were 73 million active SIM (Subscriber Identity Module) cards in France in 2017, for a total population of 66 million (ARCEP, 2017). Mobile devices (mobile phones, smartphones and tablets) have become indispensable tools, bearing witness to our activities and trips. As a result of the size of the samples, and the non-intrusive way the data is collected, the exploitation of mobile phone data logs has enormous potential. Recent cases include using the data to analyse behavioural differences between men and women (Frias-Martinez et al., 2010), studying the propagation of an epidemic (Tizzoni et al., 2013), mapping activities within a city (Noulas et al., 2013; Ratti et al., 2006; Sevtsuk et Ratti, 2010; Hoteit et al., 2014; Yue et al., 2014), or improving the paging efficiency of the cellular network (Zhang, Bolot, 2007), analyse co-presence in relation with mobility profile (Picornell et al., 2015)….

The usefulness of mobile phone data has above all been proven for the study of human mobility, in spite of the fact that the localisation data associated with each log is limited to the position of the base station used, which results in a positioning uncertainty ranging from approximately a hundred metres in a dense urban zone to several kilometres in rural zones (Calabrese et al., 2013). Gonzalez et al. (2008) were amongst the first scholars to carry out a large-scale study of the mobility of users, with a sample of over 100,000 individuals. This study demonstrated that human mobility may be modelled using a random technique and that trips follow a truncated power-law distribution. They also found that individuals have a strong tendency to visit a limited number of places many times periodically and many other places just once. Cho et al. (2001) also factored in the impact of social ties, obtained from an online social network. They concluded that short journeys (less than 100 km) are in most cases periodic in nature, while long journeys are much more influenced by the individual's social network (i.e., the presence of friends). However, even if human mobility seems to comply with these laws in a generic manner, the environment has a strong influence on the parameters of the various distributions. In a series of studies, Isaacman et al. (2010, 2011) have shown that there are important differences between cities (New York and Los Angeles) and seasons (fewer trips in the winter than the summer). Temporary tourist attractions play a major role and may modify a city's normal mobility patterns (Calabrese et al., 2010). Widhalm et al. (2015) have developed activity program typologies based on duration analysis, trips and activity location and frequencies combined with spatial typologies. They applied the method in Vienna and Boston showing similarities between conurbations but also some local specificity. Xu et al. (2015) have studied spatial distribution of individual activity area from home in Shenzen. Kung et al. (2014) have tested the Zahavi (1979) hypothesis of stable daily travel time budget applying the hypothesis to home-work trips in Ivory Coast, Portugal, Saoudi Arabia and in Boston, US. Data from the mobile phone network can also be used to estimate individual trajectories. In 2009, Schlaich et al. (2010) developed an algorithm that was able to precisely identify a GSM network user's trajectory between the cities of Karlsruhe and Stuttgart in Germany. Two years later Jiang and a group of researchers (Jiang et al., 2011) went further in this area, assigning each user to the transport network in the city of Lisbon and Tettamanti, Varga (2014) in Budapest.

Mobile phone data can also be used to study mean speeds and journey times. One of the first studies to do this was led by Ygnace (2001) and carried out in the South of France on a rural motorway which became an urban motorway near Lyon. More recently, Calabrese et al. (2011, 2013), working in the Boston conurbation, used all the data collected by a telecom operator to study mean speed, mean trip length and the distribution according to the time of day. The research conducted by Bekhor et al. (2013) is without doubt the most extensive, as it concerns the analysis of the long-distance trips carried out over the entire area of Israel. It illustrates the considerable potential of mobile phone data for the analysis of long-distance trips.

The use of mobile phone data to construct origin-destination matrices in an urban region was first proposed in Italy by Bolla and Davoli (2000) and tested on a small sample in (White and Wells, 2002) with the aim of studying traffic on specific roads. In 2002, Akin and Sisiopiku (2002) selected just 500 individuals in the city of Birmingham in the United States. One of the first studies to use the whole population rather than a sample was carried out in Israel in 2007 (Bar-Gera, 2007). The research in question set out to estimate the traffic and obtain mean speed data on a 14 km road in Israel with 10 interchanges. Calabrese et al. (2011) were the first to produce O-D matrices from a detailed dataset, for the Boston region in Massachusetts. In 2002, two simultaneous research projects attempted to extract origin-destination matrices from mobile phone network data. One of these (Akin, Sisiopiku, 2002), working in the city of Birmingham (USA), developed an algorithm which calculated origins and destinations and divided the day into periods. To compute the subject's position during each time periods, they took the largest number of connections in a zone.

However, matrices obtained in the course of these studies are only representative of the individuals using the network at a given time. Representativeness is of prime importance for these data which describe the mobility of the population of a region or mobility within a region if it is envisaged to use them for planning purposes or for regulating or optimising the use of transport networks. To our knowledge, few studies have tackled this issue. Moreover, the small number of published studies frequently employs different methodologies, pursue different goals and do not always use the same types of mobile phone data.

In England, at the same time, White and Wells (2002) tested the feasibility, in the county of Kent, of creating an origin-destination matrix from billing data. They then compared the results with a survey-based origin-destination matrix. They concluded that the billing data were not accurate enough to provide a reliable origin-destination matrix. In 2007, Caceres et al. (2007) calculated an origin-destination matrix for a road between the cities of Huelva and Seville in Spain. They considered four possible origin-destination pairs based on the positioning of the motorway interchanges. The team then compared the results with those obtained from a road traffic count. The results were very satisfactory: the error did not exceed 4% on any of the possible origin-destination pairs.

More recently, Mellegard (2011) conducted a study that covered a large part of Sweden. To generate the origin-destination matrix he adapted the algorithm method described by Kang et al. (2004) to the constraints imposed by the database he used in order to obtain an origin-destination matrix. However this study made no sophisticated comparison for the entire O-D matrix, but merely compared a very small number of origin-destination pairs with the data obtained from other surveys.

In 2012, a major study was conducted in two American cities, San Fransisco and Boston, by Wang et al. (2012). This team of researchers constructed hour-by-hour origin-destination matrices in order to observe the level of saturation of the network during morning peak periods. The method only took account of journeys taking less than one hour. The results were then analysed by segmenting the population into three groups based on the amount of data collected to verify that frequency of mobile phone use did not introduce a bias. The study was based on a train/road modal split which was subsequently compared with the road traffic count data. The results were deemed to be very satisfactory.

Calabrese et al. (2013) conducted a dual analysis using data from Boston. First, they compared the number of trips per person to the data from the National Travel Survey. The results are fairly close, although the number of trips is slightly greater in the mobile phone data. The authors consider that this disparity can be explained on the one hand by the fact that the scope of the data differs in Boston from the rest of the USA and the fact that underestimates are frequent in travel surveys (Wolf et al., 2003). They then compared the estimated distances with those given by the odometer readings from the annual safety inspections of all private vehicles. The results reveal considerable differences in levels, but fairly similar structures.

Chen et al. (2014) made a contribution to data validation, but working from a sample of mobile phone data that was simulated on the basis of a household travel survey and mobile phone data. The goal was to have an “accurate” database about which everything is known (the household travel survey) and work on the simulated mobile phone database in order to identify its ability to reproduce the “accurate” data. In this way they have shown that they can reproduce the location of individuals’ home and work with a fairly high degree of accuracy, and, with less accuracy, the location of the places they visit.

Toole et al. (2015) have developed and refined algorithms performed by previous authors (Jiang et al., 2013; Zheng et Xie, 2011; Alexander et al., 2015; Colak et al., 2015; Wang et al., 2012; Iqbal et al., 2014). Furthermore, they have combined the data with other data sources like census, travel survey, traffic counts in order to impute missing attributes like transport mode or purpose and to weight mobile data. Data were then compared to household travel survey conducted in Boston and San Francisco (US), Rio de Janeiro (Brazil) and Lisbon (Portugal). Results are promising even if some differences might be important. Combining mobile phone data with other data sources, like census, is relevant in order to weight and expand the data to whole population. But we might question validation method if travel survey is both used to weight the data and impute attributes and at the same time to validate the data by comparing mobile phone and travel survey estimations.

Graells-Garrido and Saez-Trumper (2016) have compared origin-destination matrix obtained from Telefonica Chile in Santiago and from the Santiago travel survey. To identify activity stops they applied an interesting iterative end-point fit algorithm. They obtained high correlations between the two data sources in terms of trip distribution and O-D matrix.

Bonnel et al (2017) have compared Orange signaling data collected in 2009 for the Paris conurbation with the household travel survey performed on the same area in 2010. They have produced origin-destination at the level of location area which regroup from 150 to 500 antenas. They get similar matrices structures for origin-destination with high traffic, but results was less similar when traffic was lower. They hypothesis that the differences might partially be due to frontier delimitation problem which is problematic in urban area for mobile phone data.

Our aim in this research is therefore to make an additional contribution to the existing work on the representativeness of mobile phone data. Here we use the methodology similar to Bonnel et al. (2017) but we are working with more recent data, which means adding internet data (3G network), and in Rhône-Alpes Region which allows reducing frontier problem encountered in Paris Conurbation. We shall present our data in the following section.

3. Data used: mobile phone data and household travel survey data

In this section the mobile phone data and the external sources used to compare the origin-destination matrices are presented.

3.1. Orange positioning data

Mobile network data are continuously collected by telecom operators for billing and for technical measurement purposes. Among mobile networks technologies, we find the traditional GSM network which provide 2G services and the UMTS network for 3G services. The antenna in GSM is called Base Transceiver Station (BTS), however in UMTS it is denoted by Node B and is connected to internet network. Both networks have different infrastructures but they still work with the same coverage concept. Each antenna covers a cell which belongs to a Location Area (LA)[2]. In theory, cells are often represented by Voronoi polygons.

In this research work, we will present analyses by exploring datasets issued from Orange mobile network probes. These probes allow capturing user location data by monitoring every event occurring between the mobile device and the 2G or 3G network. In fact, there are two types of probes: 2G probes captures data from BTS while 3G probes captures data from Node B. Probe data represent the signaling data transiting through the cell towers. Hence every service demand request made by the users is included in this data source with its precise connectivity information. It is no surprise that the 3G probes collect more spatiotemporal logs than 2G probes as a result of extra internet services they are able to monitor.

The explored dataset includes 3G probe data from 1 to 14 June 2017 of over 1.6 million mobile phone users per day. Spatially, this dataset covers all the Rhône-Alpes region territory and thus it is used to estimate origin-destination matrices. Figure 1 presents cellular network coverage within Rhône-Alpes region and the aggregation in 3G location area.

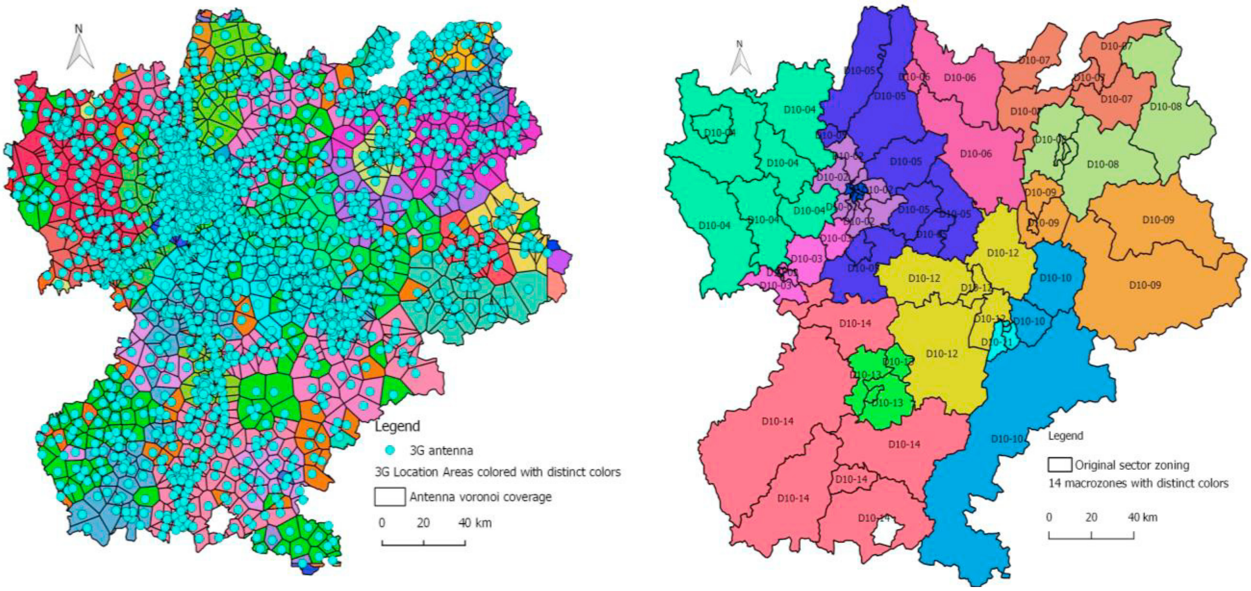

Fig. 1. Area covered by the 3G Location Areas (1a left) and aggregation of EDR sectors into 14-zone zoning system (1b right) in the Rhône-Alpes region

We have worked with the signaling data which include all the events which could be generated by mobile device or base station. Such dataset contains several types of events: communication events (calls and SMS), itinerancy events: handover and Location Area (LA) update, attach/detach events and obviously data/internet connections. Each record in the data represents a specific event within a location point with a timestamp. The Rhône-Alpes region has almost 17,000 antennas and 47 Location Areas.

To characterize the quality of dataset, we have studied the overall distribution of events generated by users. We measured the Hourly Action Rate (HAR) and we find that 50% of users generate an average of 3 records per hour or more. Moreover, to explore how uniformly the records are spread over the 24 hours, we present in Figure 2 the cumulative distribution function (CDF) and the frequency distribution of average inter-event time (AIT). We define as inter-event time the interval between a pair of consecutive records. This indicator is applied to each anonymous user with 2 records or more. AIT values range from 0 to 1400 minutes and the average value is about 28 minutes.

[2] A "Location Area" is a set of cells (antennas) that are grouped together to optimise signalling. Typically, tens or even hundreds of antennas share a single LA.

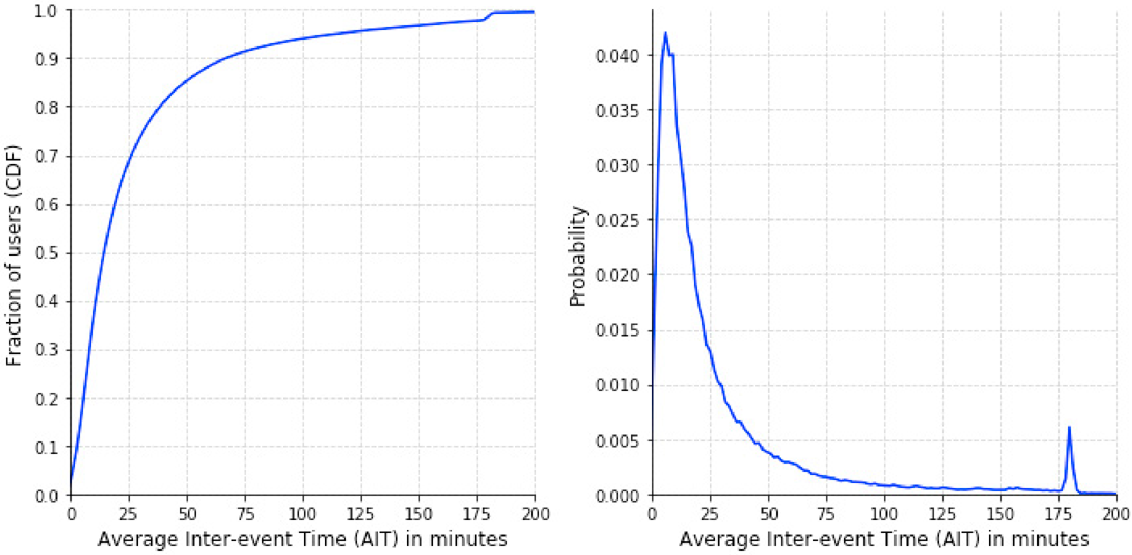

Fig. 2. Cumulative distribution function (left) and frequency distribution (right) of average inter-event time among users in the 24-hour dataset

Most of users (99%) are characterized by an AIT smaller than 200 min (Figure 2). The frequency distribution shows a peak on 180 minutes that corresponds to idle mobile phones which generate periodical Location Area Update (LAU) events every 3 hours. Mobile terminals with an AIT more than 3 hours (less than 1%) correspond to users who were not present in the Region during the whole day, or who were detached from network.

3.2. Validation data: Rhône-Alpes travel survey

Rhône-Alpes region has conducted a travel survey for the first time at this scale between 2012 and 2015 (called EDR 2015). 37,450 individuals aged over 11 years have been surveyed and 143,000 trips have been identified. Data has been collected by phone using a representative sample of the population of the Region. Sample has been constructed using geographical stratified random sampling. Geographical stratification corresponds to a zoning system of 77 zones (denoted as EDR sectors) for the whole Region. Each zone contains at least 450 surveyed individuals.

Survey contains socio-demographics of the individuals and of the households he/she belongs and all the trips made the day before of the survey. Trips are characterized with attributes (mode, beginning and finishing time in minutes, activity at the origin and at the destination, location of the origin and the destination…). Data have been collected through three waves in 2012/13; 2013/14 and 2014/15 from late autumn to early spring collecting only working day trips.

Survey methodology is similar to other travel survey conducted in urban area in France (CERTU, 2008).

4. Construction of the origin-destination matrices

A trip has been defined for the purposes of the EDR as follows by CERTU (2008): a "*trip is the movement of one person conducted for a certain purpose on infrastructure open to the public, between an origin and a destination with a departure time and an arrival time using one or more means of transport*". It is therefore necessary to specify an origin and a destination which will correspond to a purpose therefore a stationary activity in order to apply the CERTU's definition. The size of the EDR sector means that most trips between two sectors are made by motorised transport. In view of the mean speed of motorised trips in each sector as reported in the data from the EDR, we have made the assumption that if an individual is present for at least one hour[3] in a sector he/she performed a stationary activity there and therefore that the origin or the destination of a trip is located in it. In order to determine that an activity has taken place, we need at least two events. Hence, to determine a trip (activity at the origin and at the destination) we need at least 4 events.

Due to user's privacy protection concern, we have to use a 24-hour observation period. We have analysed the data of June 1st, 2017. It is a working day (a Thursday) which is similar to the average of the working days available in the 3G dataset. In order to be comparable to EDR, cellular data correspond more precisely to the period from 3am of 1st June to 3am of 2nd June 2017. The dataset contains anonymous mobile phone traces for 1.62 million Orange users. Among them only 55 062 users with 118 thousand events have strictly less than 4 events (table 1). Basically, they correspond to mobile phones which were not present in the Region during the whole day, or which were detached as mobile phone with no activity (voice, SMS or data) should have at least one Location Area Update every 3 hours = 8 events[4].

Table 1: Distribution of the number of users based on the number of events during 24 hours

Range of events	1-3	4-15	15-50	50-150	150-300	300-1000	1000+
Number of mobile phones (in thousands)	55	276	569	538	141	37	3

After this data pre-processing and applying the assumption of stationary time, we can construct an origin-destination matrix. Our first intention was to use 3G Location area zoning in order to use location area update data like for Bonnel et al. (2017) research in Paris conurbation. But as can be seen in Figure 1a, 3G location area zone might be not homogeneous. Some of the zones are split in several small zones which might be disseminated in nearly all departments of the Region. Considering this zone discontinuity, we have used EDR sector zoning system which presents the advantage for the EDR survey to produce representative data at least for population main socio-demographic characteristics.

We therefore constructed a conversion matrix to make the transition between the different zoning systems. The spatial mobility within a zone is to some extent proportional to the population of the zone and the activities conducted there. We can obtain the population of a zone, but it is not straightforward to obtain the volume of possible activities in it. We therefore used building polygons from the BDTopo database produced by IGN France (the French Mapping Agency). If we make the assumption that origin-destination pairs are uniformly distributed within the built-up zone, we can construct a conversion matrix to move between the different zoning systems. Let us take an example of a zone LA_i in the 3G LA zoning system that generates N_{LAi} trips. Let us assume that this zone straddles two EDR sectors, S_1 and S_2. Using the BDTopo database, for each LA_i, it is possible to compute the proportion of the built-up surface area that corresponds to S_1 and S_2, which are denoted respectively by $p(S_1)$ and $p(S_2)$. The N_{LAi} trips can be then distributed using the following formula:

Number of trips generated by $S_1 = p(S_1)* N_{LAi}$

Number of trips generated by $S_2 = p(S_2)* N_{LAi}$

Generalisation is straightforward for all the zoning systems for both origin and destination, which means we can construct conversion matrices in order to move from one zoning system to another and thus estimate the trip matrices obtained from the mobile data and the EDR with EDR sector zoning system.

[3] The threshold of one hour is probably too high and we will test the impact of this threshold in section 5. At this stage we have kept the same threshold than the previous research in Paris conurbation (Bonnel et al., 2017).

[4] Another problem in the data is the presence of machines equipped with SIM cards that we treat as humans. Having only one day of observation, it is not possible to detect these objects connected to the telephone network. However, this bias seems relatively small in our case: according to Orange's internal sources more than 80% of Machine-to-Machine (M2M) connections are carried over 2G networks, and their traffic on the 3G network analysed here does not exceed 0.5%.

5. Comparison between the trip matrices obtained from mobile phones with those obtained from travel surveys

The daily origin-destination matrix obtained from mobile phone data only contains the trips made by individuals who use Orange's network. However, the EDR matrices contain data for the entire population of the Rhône-Alpes Region. We therefore need to expand the mobile phone data. The penetration rate of mobiles using Orange's 3G network is not precisely known in Rhône-Alpes and we have no information about the sociodemographic characteristics of these mobile users for reasons of confidentiality and privacy. We are therefore forced to make a new assumption. As we know the (anonymised) identifier of each mobile phone, we are able to estimate the number of mobile phones which use the Orange network every day. If we assume that mobile phone users are representative of the population of Rhône-Alpes, we can determine a daily expansion factor C_{exp}:

$$C_{exp} = \frac{Population\ of\ RA\ region}{Nb\ of\ users\ using\ 3G\ network}$$

with *Population of RA region* equal to 5.2 million (population of Rhône-Alpes region aged over 11 years according to last census data of 2013) and *Nb of users using 3G network* is the number of observed telephones with at least 4 events in our database (1.56 million users). This is obviously a strong assumption and we develop here among the strongest limitations:

- We have no data that allow us to check that the travel practices of Orange 3G network users are representative of the entire population of mobile phone users. We do however know that Orange is the principal mobile phone operator in France, with more than a third of the market. We can therefore assume that the population of Orange users in Rhône-Alpes is not too atypical;
- Some people using the telecom network in the Rhône-Alpes Region do not live in the region. It is possible to identify individuals who live abroad and exclude them. Identifying where other users live is more complex with only one-day data and we have preferred to avoid this issue at this stage of the research;
- Some individuals living in the Rhône-Alpes Region may be outside the region on the day when data was collected and so have zero mobility within the region;
- Not all individuals are mobile phone users among the population aged over 11 years. It is especially the case of elderly people.

The data from the EDR contain all the trips made by residents of the Rhône-Alpes Region irrespective of the purpose and the duration of the activity on an average working day. However, we have made an assumption of a minimum of one-hour stationary time in a sector in order to identify an origin or a destination in the case of mobile phone data. We have therefore applied the same assumption to the travel survey data in order to exclude very short activities which cannot be identified as a result of our stationary time assumption.

If travel survey data are representative of the population at sector level, when we combine origin and destination, the number of observed trips for most of the origin-destination is too small. The confidence intervals are very wide for many O-D pairs. We therefore aggregated the 77 EDR sectors into a 14 macro zones (Figure 1b) in order to produce origin-destination matrices which gives a sufficient number of trips for most of the origin-destination pairs in the EDR. This makes it possible to make a comparison with the mobile phone data matrix which has also been aggregated to correspond to the 14-zone zoning system.

Table 2: Number of trips from mobile phone data and the EDR (division into 14 zones). Source of data: Orange, Rhône-Alpes Region

Threshold to determine stationary activity	60mn	50mn	40mn	30mn
EDR (in thousands)	2,211	2,260	2,344	2,448
Mobile phones (in thousands)	1,605	1,872	2,226	2,762

We have much less trips with mobile data than with EDR (table 2) with a threshold of one-hour for identifying an activity. At the size of a sector (Figure 1b shows the 14 macro zones, but also the 77 EDR-sectors), the duration of a trip to cross the sector with a motorised mode is probably lower than one hour for most trips. We have therefore tested different thresholds (table 2). Numbers of trips from the two sources are much closer for a time between 30 and 40 minutes. Of course it is not the demonstration that the right threshold is between 30 and 40 minutes, but these values are not impossible to cross most of the sectors. We will therefore keep these thresholds for the following analyses.

We have regressed the number of mobile phone trips by the number of EDR trips as the explanatory variables. All observations correspond to O-D pairs of the 14-zone zoning system. We present the results for 30 and 40 minutes thresholds:

30 minutes threshold

$\mathbf{y_{ij} = 0.79 * x_{ij} + 4{,}501}$, with R^2=0.88; Student's t for constant = 5.34 and for slope = 37.2

40 minutes threshold

$\mathbf{y_{ij} = 0.67 * x_{ij} + 3{,}552}$, with R^2=0.88; Student's t for constant = 5.17 and for slope = 37.0

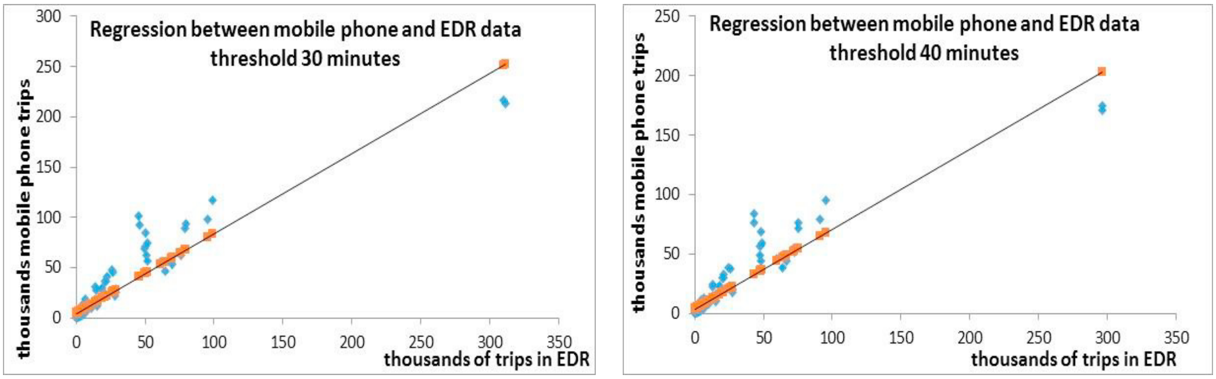

Fig. 3. Regression plot between the mobile phone data (working day) and the data from the EDR. Source of data: Orange, Rhône-Alpes Region

Regression R^2 ratio is correct. But results are not really satisfactory as we would expect to have a slope close to 1 and a constant close to 0. Analysis of the regression plot (Figure 3) shows that some points are fairly close to the regression line but other are less. We can also observe that there are two O-D with very high number of trips in comparison with other O-D. These two O-D correspond to the Lyon conurbation which is the biggest conurbation of Rhône-Alpes and the second in France with more than 1.6 million inhabitants. These two points have a strong effect on the slope. Clearly we are not able to reproduce EDR data with mobile phone data for the Lyon conurbation with a strong under-estimation for mobile phone data. We have also some O-D with quite small number of trips especially for the EDR data. The average sampling rate in Rhône-Alpes Region is 1/140. If we consider a minimum number of 30 observed trips, it represents approximately 4,200 trips when the data are expended to the whole Region. We have therefore performed new regression with only those O-D with at least 4,200 trips for the EDR in order to guarantee minimum representativeness of EDR data which are chosen as explanatory variables, and without the two O-D pairs of the Lyon conurbation:

30 minutes threshold, with at least 4,200 trips for EDR data and without Lyon conurbation O-D pairs

$\mathbf{y_{ij} = 1.01 * x_{ij} + 6{,}517}$, with R^2=0.79; Student's t for constant = 2.31 and for slope = 14.5

40 minutes threshold, with at least 4,200 trips for EDR data and without Lyon conurbation O-D pairs

$\mathbf{y_{ij} = 0.87 * x_{ij} + 4{,}912}$, with R^2=0.79; Student's t for constant = 2.13 and for slope = 14.39

Results are more satisfactory with slope closer to one as expected, especially for the first regression without Lyon conurbation O-D pairs (Figure 4). Constant should be close to 0 which is not the case even if the size of the constant is not so high regarding the total number of trips observed in the Region (table 2). But there are still some observations which are not very close to the regression line. When we analyse these O-D with high differences in number of trips between the two data sources, we notice that they mainly concern adjacent zones (O-D 5-6 and 6-5;

7-8 and 8-7; 8-9 and 9-8; 5-12 and 12-5) with similar differences for both orientations of the O-D (from O to D and from D to O). For these O-D we always observed more trips for the mobile phone data than the EDR data. This problem might be due to ping-pong effects which do not appear in short duration time and are therefore much more complicated to correct than the one we have treated (see section 4). On the reverse, mobile phone data under-estimations mainly concern O-D which are not adjacent or with very short frontier (O-D 10-11 and 11-10; 11-12 and 12-11). As can be seen on regression plot, the O-D analysis is nearly the same for the activity duration thresholds.

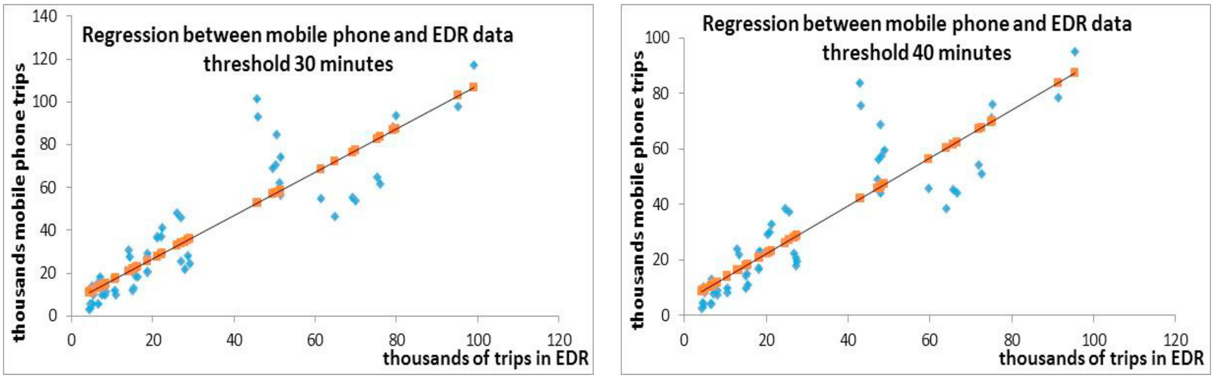

Fig. 4. Regression plot between the mobile phone data (working day) and the data from the EDR, with at least 4,200 trips for EDR data and without Lyon conurbation O-D. Source of data: Orange, Rhône-Alpes Region

6. Discussion and conclusion

As many studies have already shown, mobile phone data allow constructing origin-destination matrices. These matrices were generated from the 3G mobile phone network operated by Orange in the Rhône-Alpes Region. However, to our knowledge, and as has been very recently highlighted by Chen et al. (2014), the origin-destination matrices generated with this type of data are rarely validated at the scale of a region like Rhône-Alpes. More generally, world-wide, few studies have been undertaken to validate the travel or traffic data obtained from mobile phone data in comparison with other data sources (White, Wells, 2002; Caceres et al., 2007; Mellegard, 2011; Wang et al., 2012; Bekhor et al., 2013; Calabrese et al., 2013; Chen et al., 2014; Graells-Garrido, Saez-Trumper, 2016).

This work has been performed following a similar methodology than the one performed in the Ile-de-France Region which include Paris (Bonnel et al. 2015, 2017). In this previous research the 2009 2G network data were analysed at the level of location area and then compared with the last household travel survey (called EGT) available for the whole Ile-de-France Region performed in 2010. The time threshold for activity duration was one hour given the traffic condition in Paris conurbation and the size of LA (22 LA for the whole Ile-de-France Region). We were able to obtain a total number of trips in Ile-de-France from the mobile phone data that was similar to that given by the EGT (a difference of 9%). Above all, the linear regression we performed on the number of trips in each element in the two matrices showed that the structure of the two matrices were very similar with an R^2 value of 0.96 and a slope that was very close to 1. But these very encouraging results was accompanied by other results are less satisfactory in the case of some origin-destination pairs for which the disparities attained 70 to 80%, even if in terms of numbers, the disparities are smaller as the largest percentage disparities are for those origin-destination pairs with a fairly small number of trips.

We obtained less good results in the case of Rhône-Alpes Region even if R^2 values are not too bad (between 0.79 and 0.88 depending on O-D taken into account in the regression analysis). In case of Ile-de-France we have worked with signalling database at LA-zoning level which contains information each time a phone crosses a LA frontier whatever the phone activity. With this LA zoning we have at our disposal continuous spatiotemporal data at LA level, which is no more the case for Rhône-Alpes data where we have had to use travel survey zoning system because of the discontinuity of 3G LA zoning in Rhône-Alpes (Figure 1a). This might be one of the reasons of less satisfactory results in case of Rhone-Alpes as compared to Ile-de-France Region. We can also observe that for adjacent zones with long border we often have trips overestimation with phone data compared to EDR data and on contrary the most important under-estimations are observed for non-adjacent zones. This problem was less apparent in case of Ile-de-France due to low number of zones for the regression analysis (7 zones only due to low number of trips for some O-D).

A large number of hypotheses need to be made to construct trip matrices from mobile phone data. In order to identify possible approaches for further investigation, we shall restate these below:

- The mobile phone data related to all the trips made by individuals who were present in the Rhône-Alpes Region. However, the data from the EDR only covered Rhône-Alpes residents. It would therefore be interesting to attempt to identify where the mobile phone owners in the database live. This would make it possible to extract solely the residents of the Rhône-Alpes Region in order to improve the validity of the comparison with household travel survey data. But we have to face privacy concerns and the obligation made by the CNIL (National committee for informatics and liberty) to change every day the mobile phone identifier and therefore seriously limit the possibility to identify home location without strong hypothesis;
- We have applied a uniform assumption of minimum stationary time (between 30 to 40 mn in Rhône-Alpes and one hour in Ile-de-France) to identify stationary activity. It would certainly be possible to refine this and vary it according to the characteristics of each sector in terms of surface area and travel speeds. Moreover, the duration threshold is necessarily somewhat arbitrary, even if it was based on an analysis of the data from travel survey. The sensitivity analysis conducted on this threshold (table 2) shows that the results are highly sensitive to this assumption;
- The boundaries of the cells coverage area are identified by analysing Voronoi polygons. This means there is a high degree of uncertainty about the boundaries. Actual base station coverage limits vary depending on mobile phone traffic, weather and local topography. It would be interesting both to refine the base station boundaries and to study the impact of boundary uncertainty on the construction of origin-destination matrices. Results also seem to indicate that we have overestimation of the number of trips from phone data for adjacent zone with long common boundaries and inversely underestimation for zones which are not adjacent. This result questions both cell coverage boundaries, but also the treatment of ping-pong effects which might occur not only in short time and also with long time intervals between phone location events;
- The expansion of matrices obtained from mobile phone data is based on the very simple assumption that the mobile phone users for whom we were able to constitute at least one trip are representative of the general population. It is unlikely that we will be able to access demographic data on the users of the Orange network for obvious commercial reasons, but it is not impossible to try to collect information from other sources. Calabrese et al. (2011) and Bekhor et al. (2013) have analysed the spatial distribution of mobile phone users by comparing it to census data. Bekhor et al. (2013) have also used travel survey data which contained questions about mobile phone use. These data could be used to identify any bias affecting the samples of mobile phone data in order to adjust the data using travel data from household travel surveys;
- We undertook no analysis of the data for mobile phone owners for whom we had fewer than four events. It would nevertheless be useful to identify those who switch their mobile phone on or off during the study day in order to distinguish between them and individuals who entered or left the study zone during the day;
- As indicated in footnote 4, we cannot distinguish between telephone users and machines equipped with SIM cards. This introduces a small bias in the analyses (less than 0.5% of 3G traffic concerned), which could be filtered if equipment codes could be kept in the data;
- Finally, we have considered only one day which is a typical day. From what we know there was no special event or meteorological event on this day. But it could be interesting to reproduce the analysis on several days to analyse the variability of the results.

Mobile phone data therefore seem promising for the analysis of spatial mobility, but a considerable amount of further research is required in order to be able to fully validate their use in order to construct origin-destination matrices for transport modelling or transport planning purposes.

Acknowledgements: The authors would like to thank Orange for giving them access to the mobile phone data collected in the Rhône-Alpes region and Region Rhône-Alpes for making the EDR data available. However, the authors alone are responsible for the contents of this paper.

References

Akin D, Sisiopiku V (2002), *Estimating Origin-Destination Matrices Using Location Information from Cellular Phones*, Proc. NARSC RSAI, Puerto Rico, USA.

Alexander LP, Jiang S, Murga M, González MC (2015), Validation of origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C*, in press.

Ampt ES (1997) Response Rates - Do they matter? In: Bonnel P, Chapleau R, Lee-Gosselin M, Raux C (eds.) *Les enquêtes de déplacements urbains: mesurer le présent, simuler le futur*, Programme Rhône-Alpes Recherches en Sciences Humaines, Lyon, pp 115-125

Arana P, Cabezudo S, Peñalba M (2014) Influence of weather conditions on transit ridership: A statistical study using data from Smartcards, *Transportation Research part A*, 59 pp. 1-12.

ARCEP (2017) Observatoire des marchés des communications électroniques en France, 1er trimestre 2017, https://www.arcep.fr/index.php?id=13652.

Arentze T, Timmermans H, Hofman F, Kalfs N (2000), Data needs, data collection, and data quality requirements of activity-based transport demand models, In: *Transport surveys, raising the standard*, TRB transport circular E-C008, pp. II-J/1-30.

Atrostic BK, Burt G (1999) *Household non-response: what we have learned and a framework for the future*, Statistical Policy working paper 28, Federal Committee on Statistical methodology, Office of Management and Budget, Washington, pp 153-180.

Bar-Gera H (2007), Evaluation of a cellular Phone-Based System for Measurement of Traffic Speeds and Travel Times: A Case Study from Israel, *Transportation Research Part C*, 15 (6) pp. 380-391.

Bekhor S, Cohen Y, Solomon C (2013), Evaluating Long-Distance Travel Patterns in Israel by Tracking Cellular Phone Positions, *Journal of Advanced Transportation*, 47 (4) pp. 435-446.

Bolla R, Davoli F (2000), *Road Traffic Estimation from Location Tracking Data in the Mobile Cellular Network*, Proc. IEEE WCNC, Chicago, USA.

Bonnel P (2003) Postal, telephone and face-to-face surveys: how comparable are they? In: Stopher PR, Jones PM (eds.) *Transport Survey Quality and Innovation*, Elsevier, London, pp 215-237.

Bonnel P, Hombourger E, Smoreda Z (2013), *Quel potentiel des données de la téléphonie mobile pour la construction de matrices origines-destinations de déplacement – application à l'Ile-de-France*, Rapport de Recherche, LET, Orange Labs, 133p.

Bonnel P, Hombourger E, Olteanu-Raimond A-M, Smoreda Z (2015), Passive mobile phone dataset to construct origin-destination matrix: potentials and limitations. *Transportation Research Procedia*, vol.11, pp. 381-398. doi: 10.1016/j.trpro.2015.12.032.

Bonnel P, Hombourger E, Olteanu Raymond A-M, Smoreda Z (2017), Apports et limites des données passives de la téléphonie mobile pour la construction de matrices origine-destination, *RERU Revue d'Economie Régionale et Urbaine*, 2017-4, pp. 5-29.

Brisson P (2008), *Global system for mobile communication.* Université de Montreal.

Caceres N, Wiedeberg JP, Benitez FG (2007), Deriving Origin-Destination Data from a Mobile Phone Network, *IET Intelligent Transport System*, 1 (1) pp. 15-26.

Calabrese F, Diao M, Di Lorenzo G, Ferreira Jr J, Ratti C (2013), Understanding individual mobility patterns from urban sensing data: A mobile phone trace example, *Transportation Research Part C*, 26, pp. 301-313.

Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011) Estimating Origin-Destination Flows using Mobile Phone Location Data, *IEEE Pervasive Computing*, 10 (4) pp. 36-44.

Calabrese F, Pereira F, Di Lorenzo G, Liu L, Ratti C (2010), *The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events*, Proc. Pervasive Computing, Helsinki, Finlande.

CERTU (2008), *L'enquête ménages déplacements standard CERTU*, éditions du CERTU, Lyon, 203p.

Chen C, Bian L, Mac J (2014), From traces to trajectories: How well can we guess activity locations from mobile phone traces? *Transportation Research Part C*, 46, pp. 326–337

Cho E, Myers SA, Leskovec J (2011) *Friendship and Mobility: User Movement in Location-based Social Networks*¸ Proc. ACM SIGKDD, San Diego, USA.

Colak S, Alexander LP, Alvim BG, Mehndiretta SR, González MC (2015), Analyzing cell phone location data for urban travel: current methods, limitations, and opportunities. In *Transportation Research Board Annual Meeting*, 94th, 2015, Washington, DC, USA, 2015

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

Frias-Martinez V, Frias-Martinez E, Oliver N (2010) *A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records*, Proc. AAAI AI-D, Palo Alto, USA.

Graells-Garrido E, Saez-Trumper D (2016), A day of your days: estimating individual daily journeys using mobile data to understand urban flow, *Urb-IoT '16*, *Proceedings of the Second International Conference on IoT in Urban Space,* May 24 - 25, 2016, Tokyo, Japan, pp. 1-7. DOI: http://dx.doi.org/10.1145/2962735.2962737

Gonzalez MC, Hidalgo CA, Barabasi A-L (2008), Understanding Individual Human Mobility Patterns, *Nature*, 453 (7196) pp. 779-782.

Hoteit S, Secci S, Sobolevsky S, Ratti C, Pujolle G, (2014) Estimating human trajectories and hotspots through mobile phone data. *Computation Network,* 64, pp. 296–307.

IDATE (2008), Observatoire économique de la téléphonie mobile – faits et chiffres 2008, *Mobile et société*, 9, pp. 6-15. http://www.fftelecoms.org/sites/default/files/contenus_lies/mobile_et_societe_9.pdf

INSEE (2012), *Bases sur les flux de mobilité : documentation.* INSEE, Paris .

Iovan C, Olteanu-Raimond A-M, Couronné T, Smoreda Z (2013), Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies, In: *Geographic Information Science at the Heart of Europe*, Springer, pp. 247-265

Iqbal MS, Choudhury CF, Wang P, González MC, (2014) Development of origin–destination matrices using mobile phone call data, *Transportation Research part C*, 40, pp. 63-74.

Isaacman S, Becker R, Caceres R, Kobourov S, Rowland J, Varshavsky A (2010) *A Tale of Two Cities*, Proc. ACM HotMobile, Annapolis, USA.

Isaacman S, Becker R, Caceres R, Kobourov S, Martonosi M, Rowland J, Varshavsky A (2011), *Ranges of Human Mobility in Los Angeles and New York*, Proc. IEEE PerCom Workshops, Seattle, USA.

Janzen M, Vanhoof M, Smoreda Z, Axhausen KW (2018), Closer to the total? Long-distance travel of French mobile phone users, *Travel Behaviour and Society*, 11, pp. 31-42.

Jiang S, Fiore GA, Yang Y, Ferreira Jr J, Frazzoli E, González MC (2013), A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD *International Workshop on Urban Computing*. August 11–14, 2013, Chicago, Illinois, USA. pp. 1-9. ACM Press.

Jiang S, Vina-Arias L, Ferreira J, Zegras C, Gonzalez MC (2011), *Calling for Validation: Demonstrating the use of Mobile Phone data to Validate integrated land use Transportation models*, Proc. 7VCT, Lisbon, Portugal.

Kang JH, Welbourne W, Stewart B, Borriello G (2004), *Extracting Places from Traces of Locations*, Proc. ACM WMASH, Philadelphia, USA.

Kung KS, Greco K, Sobolevsky S, Ratti C (2014), Exploring universal patterns in human home-work commuting from mobile phone data, *PlosOne*, 9(6), e96180.

Mellegard E (2011), *Obtaining Origin/Destination-Matrices from Cellular Network Data.* Master's Thesis, Chalmers University of Technology.

Morency C, Trépanier M, Agard B (2007), Measuring transit use variability with smart-card data. *Transport Policy,* 14 (3), pp. 193–203.

Munizaga M, Palma C, Mora P (2010), Public transport OD matrix estimation from smart card payment system data. In: *12th World Conference on Transport Research*, Lisbon, Paper No. 2988.

Noulas A, Mascolo C, Frias-Martinez E (2013) *Exploiting Foursquare and Cellular Data to Infer User Activity in Urban Environments*, Proc. IEEE MDM, Milan, Italy.

Ortuzar J de D, Bates J (2000), Workshop summary, in: *Transport surveys, raising the standard*, TRB transport circular E-C008, pp. II-J/31-35.

Pelletier MP, Trépanier M, Morency C (2011), Smart card data use in public transit: A literature review, TRC*,* 19, pp. 557–568.

Phithakkitnukoon S, Smoreda Z, Olivier P (2012), Social-geography of human mobility: a study using longitudinal mobile phone data. *PLoS ONE* 7 (6), e39253.

Picornell M, Ruiz T, Lenormand M, Ramasco J, Dubernet T, Frias-Martinez E (2015), Exploring the potential of phone call to characterize the relationship between social network and travel behavior, *Transportation*, 42, pp. 647-668.

Pollini GP (1996), Trends in handover design, *IEEE Commun Mag,* 34(3), pp. 82-90.

Ratti C, Williams S, Frenchman D, Pulselli RM (2006), Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B*, 33(5), pp. 727-748

Schlaich J, Otterstatter T, Friedrich M (2010), *Generating Trajectories from Mobile Phone Data*, Proc. TRB Meeting, Washington D.C, USA.

Sevtsuk A, Ratti C (2010), Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. Journal of Urban Technology 17(1), pp. 41–60.

Smoreda Z, Olteanu-Raimond A-M, Couronné T (2013), Spatiotemporal data from mobile phones for personal mobility assessment, In: Zmud , Lee-Gosselin M, Carrasco JA, Munizaga MA (eds), *Transport Survey Methods: Best Practice for Decision Making*, Emerald, pp. 745-767.

Stopher PR, Greaves SP (2007), Household travel surveys: where are we going? *Transportation Research Part A,* 41, pp. 367–381.

Tettamanti T, Varga I (2014), Mobile phone location area based traffic flow estimation in urban road traffic, *Advances in Civil and Environmental Engineering,* 1(1), pp. 1–15.

Tizzoni M, Bajardi P, Decuyper A, King GKK, Schneider C, Blondel V, Smoreda Z, Gonzalez MC, Colizza V (2014) On the Use of Human Mobility Proxies for Modeling Epidemics, *PLOS Computational Biology*, 10(7), e1003716.

Toole JL, Colak S, Sturt B, Alexander LP, Evsukoff A, Gonzalez MC (2015), The path most traveled: Travel demand estimation using big data resources. *Transportation Research Part C*, http://dx.doi.org/10.1016/j.trc.2015.04.022

Wang P, Hunter T, Bayen A, Schechtner K, Gonzalez MC (2012) Understanding Road Patterns in Urban Areas, *Scientific Reports,* 2 (1001) pp. 1-6.

White J, Wells I (2002) *Extracting Origin Destination Information from Mobile Phone Data*, Proc. IEEE RTIC, London, UK.

Widhalm P, Yang Y, Ulm M, Athavale S, Gonzalez MC (2015), Discovering urban activity patterns in cell phone data, *Transportation*, 42, pp. 597-623.

Wolf J, Oliveira M, Thompson M (2003), Impact of underreporting on mileage and travel time estimate – results from Global Postionning System enhanced household travel survey, *Transportation research record*, 1854, pp. 189-198.

Xu Y, Shaw S-L, Zhao Z, Yin L, Fang Z, Li Q (2015), Understanding aggregate human mobility pattern using passive mobile phone location data: a home-bazsed approach, *Transportation*, 42, pp. 625-646.

Ygnace J-L (2001) *Travel Time/Speed Estimates on the French Rhone Corridor Network using Cellular Phones as Probes*, INRETS STRIP Project Technical Report.

Yue Y, Lan T, Yeh AGO, Li Q-Q (2014), Zooming into individuals to understand the collective: a review of trajectory-based travel behaviour studies. Travel Behaviour Society 1(2), pp. 69–78.

Zahavi Y (1979), *The 'UMOT' Project*, report prepared for the U.S. Department of Transportation and the Ministry of Transport of Federal Republic Of Germany, 267p.

Zhang H, Bolot J (2007) *Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks*, Proc. ACM MobiCom, Montreal, Canada.

Zheng Y, Xie X (2011), Learning travel recommendations from user-generated GPS traces. ACM *Transactions on Intelligent Systems and Technology*, 2 (1), article 2.

Zmud J (2003) Designing instruments to improve response: keeping the horse before the cart, In: Stopher PR, Jones PM (Eds) *Transport Survey Quality and Innovation*, Elsevier, Pergamon, Oxford, pp 89-1