Passive Mobile Phone Dataset to Construct Origin-destination Matrix: Potentials and Limitations

Patrick Bonnel, Etienne Hombourger, Ana-Maria Olteanu-Raimond, Zbigniew Smoreda

Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 11 (2015) 381 – 398

Transportation Research Procedia

www.elsevier.com/locate/procedia

10th International Conference on Transport Survey Methods

Passive mobile phone dataset to construct origin-destination matrix: potentials and limitations

Patrick Bonnel [a]*, Etienne Hombourger [b], Ana-Maria Olteanu-Raimond [c], Zbigniew Smoreda [d]

[a] *Laboratoire d'Economie des Transports, ENTPE, Lyon, France, patrick.bonnel@entpe.fr*
[b] *DTecITM, CEREMA, Paris, France, Etienne.hombourger@cerema.fr*
[c] *COGIT, IGN, Paris, France, Ana-Maria.Raimond@ign.fr*
[d] *SENSE, Orange Labs, Paris, France, zbigniew.smoreda@orange.com*

Abstract

Mobile phone operators produce enormous amounts of data. In this paper we present applications performed with a dataset (communication events + handover and Location Area Up-date) collected by the operator Orange from 31 March to 11 April 2009 for the whole Paris Region. Trips are deduced from the spatio-temporal trajectory of devices through a hypothesis of stationarity within a Location Area in order to define activities. Trips are then aggregated in an origin-destination matrix which is compared with traditional data (census data and household travel survey).

Peer-review under responsibility of International Steering Committee for Transport Survey Conferences ISCTSC

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

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). They must also be of good quality and, in particular, accuracy, to ensure that investment or transport policy decisions are

* Corresponding author. Tel.: +0-000-000-0000 ; fax: +0-000-000-0000 .
E-mail address: author@institute.xxx

Peer-review under responsibility of International Steering Committee for Transport Survey Conferences ISCTSC
doi:10.1016/j.trpro.2015.12.032

based on reliable analyses. While 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), which reduce even further the quality of the resulting matrices. Consequently, trip matrices are also generated from other sources, in some cases in combination with survey data, examples being roadside traffic counts, cordon or screen-line surveys and public transport surveys. The resulting data are useful for improving the quality of matrices, but do not always contain the necessary information. This applies, for example, to road traffic counts which provide information on the traffic volumes at a given point on a road, but not on trip origins and destinations. A variety of techniques have been developed for processing and combining the data from different sources. However, the reliability of the resulting matrices is uncertain and cannot always be measured statistically.

The advent of large volumes of data that are produced automatically and passively such as ticketing data (Arana et al., 2014, Morency et al., 2007; Munizagua et al., 2010; Pelletier et al., 2011), bank cards… and mobile phone data makes it possible to identify the presence of individuals in both space and time in a way which, while admittedly irregular is becoming less so. A number of techniques have been developed for converting these data into trips, 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. The aim of this paper is therefore to test the potential of these data for producing origin-destination matrices compared with other sources of available data. The analysis has been conducted within the Greater Paris Region (Ile-de-France) for which we were able to analyse the mobile phone data from the operator Orange and compare them on the one hand with the census data on commuting trips from home to place of work or study, and on the other hand with the data obtained from the "*Enquête Globale Transport*" (EGT), which is the name given to the household travel surveys conducted in the Ile-de-France Region.

We shall begin this paper with a literature survey (Section 1) before presenting the data we have used (Section 2) and (Section 3) the data processing methodology employed to produce the origin-destination matrices, which allow us to make the comparison with external data (Section 4). Finally, we shall present the principal lessons from this research, and some suggested directions for the future (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 76.8 million active SIM (Subscriber Identity Module) cards in France at the end of 2013, for a total population of 65 million (ARCEP, 2014). Mobile devices (mobile phones, smartphones and tablets) have become indispensable tools, bearing witness to our activities and trips. Mobile phone operators, who are obliged for legal or billing purposes to record information about the use of these devices, therefore find themselves with increasingly informative databases. The reason for this is that each time a mobile terminal is used to make a call, send an SMS (Short Message Service), the operator generates a call detail record (CDR) that contains the timestamp, the terminal's identifier of the base station to which the user is connected and quantitative data about the call (call duration, volume of data exchanged).

As a result of the size of the samples, which in the case of some operators can involve as many as 40-50% of a country's population, 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), or improving the paging efficiency of the cellular network (Zhang, Bolot, 2007).

But 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 (Calabrese et al., 2013) to several kilometres in rural zones. 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.

The authors of this paper 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).

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.

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.

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. The findings showed that in rural areas the data from the mobile phone network matched those obtained from road traffic surveys but there was a great difference between the data from the two sources in urban areas. 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.

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 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 three periods:

- from midnight to 8 am, when the person was theoretically at home;
- from 8 am to 4 pm when the person was theoretically at work;
- from 4 pm to midnight when the person was theoretically engaged in activities.

To compute the subject's position during these three time periods, they took the largest number of connections in a zone. Next, during each time period, they considered the largest number of connections as an origin-destination pair, and thus generated an origin-destination matrix. This study has certain limitations, as the matrix which is generated only takes account of trips which are identified as home-work, work-leisure and leisure-home. However, no verification was conducted in this paper on the basis of a comparison with data obtained from other processing methods.

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 (Call Detain Record, CDR). 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. To construct the origin-destination matrix, it was deemed that as soon as a user left the road he/she was no longer visible on the studied network. Moreover, road users had to change zones at a speed which was compatible with below the speed limit in the area. 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 the algorithmic method described by Kang et al. (2004) was applied. The method extracts from position data, and with a high degree of accuracy, the places where an individual has stayed for some time or the places an individual has passed through. (Kang et al. (2004) then applied the algorithm to GPS (Global Positioning System) data. Mellegard (2011) adapted the algorithm 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) emphasised the shortcomings of the work conducted to validate the mobility data obtained from mobile phone data, and mention that Calabrese's research is the most sophisticated in this respect. Chen's team 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.

Our aim in this research is therefore to make an additional contribution to the existing work on the representativeness of mobile phone data. Our analysis relates to the validation of origin-destination matrices obtained from mobile phones by comparing them to external data sources. We shall present our data in the following section.

3. Data used: mobile phone data, commuting 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

The mobile phone network is made up of a set of base stations each of which has a coverage zone (Figure 1). In practice, the zone area is variable with small zones in dense area and much larger zones in low density area as indicated in Figure 2. Furthermore coverage zone is not fixed as it depends on the activity of each base station, as an overloaded base station can pass on some traffic to its neighbours. It also depends on the topography and meteorological conditions. In theory the base station coverage is often represented by Voronoi polygons. The base stations are grouped together to form Location Areas (LA) for reasons to do with management of the mobile phone network as this makes it possible to identify mobile phones more rapidly in the case of a call or an SMS. The LA in which a mobile phone is located is known all the time, while at base station level its position is only known in the event of a call.

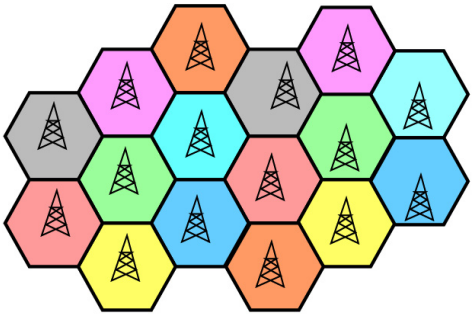


Fig. 1. The cell phone system. Source: (BRISSON, 2008)

For the purposes of this research, Orange has given us access to the data from the network of base stations in Île-de-France, which covers an area that approximates to the administrative region. Data were collected directly from the base stations between 31 March 2009 and 11 April 2009. The Ile-de-France region, which includes the conurbation of Paris, has 12 million inhabitants and covers an area of 12,000 km^2. The mobile phone data take two forms:

- Billing data (CDR-Call Detail Records), these list the base station through which the information was transmitted every time an individual receives or sends a call or an SMS;
- Signalling data, which are all the data that pass through the base stations. In addition to the billing data, these contain details of handovers (i.e. changes of base station during a call), LA updates, and logs of when the mobile phone is switched on or off. This data is collected via network quality probing systems.

e have worked with the signalling data which have the advantage of informing us in which LA the mobile phone is located on a permanent basis. However, its spatial resolution is much lower (Figures 2 and 3). The Ile-de-France region has almost 10,000 base stations, and 32 LA, each of which has between 150 and 500 base stations.

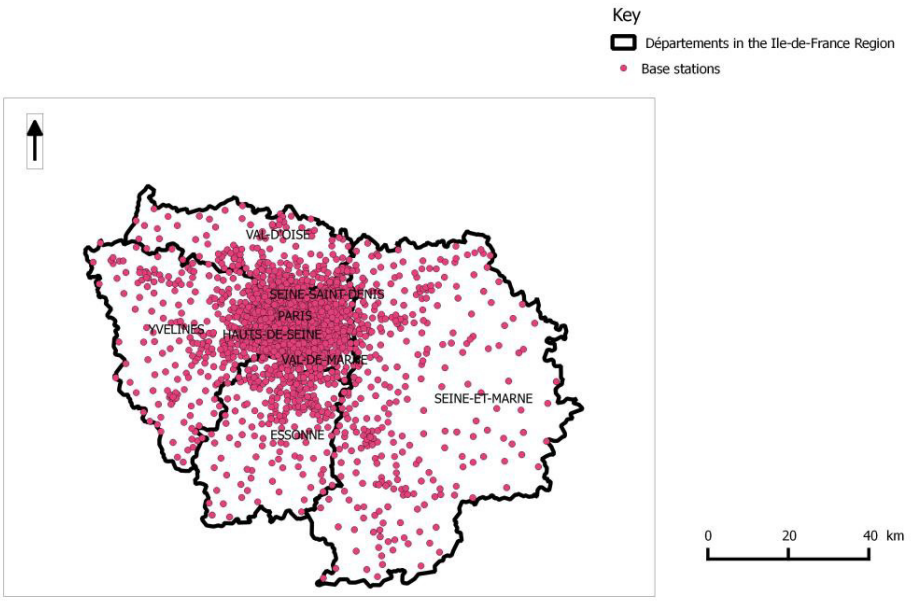


Fig. 2. Position of the base stations in the Île-de-France Region. Source: Orange

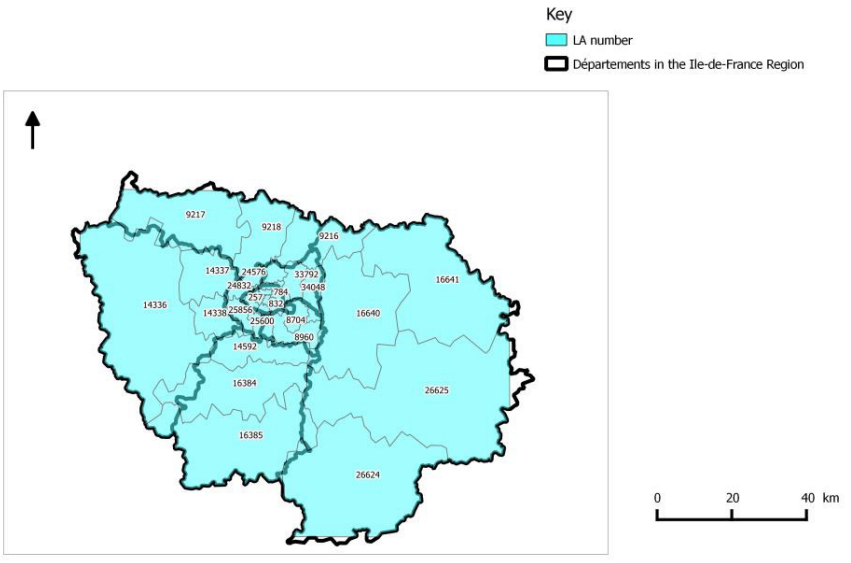


Fig. 3. Area covered by the LA in the Île-de-France region. Source: Orange

The database contains 1.5 billion logs. The availability of a unique anonymised code for each mobile phone means it is possible to find out the number of users who use the Orange network on a given day (Figure 4). According to IDATE (2009), Orange had 43.5% of the SIM card market at the end of 2008 and according to the Sofres TNS, almost 80% of the French population aged over 12 years owned a mobile phone in 2008. In view of the

total population of the Ile-de-France Region and the fact that the figures may be slightly different for the Ile-de-France Region than for the rest of France, the mean number of phones identified per day seems to be consistent with the available statistics.

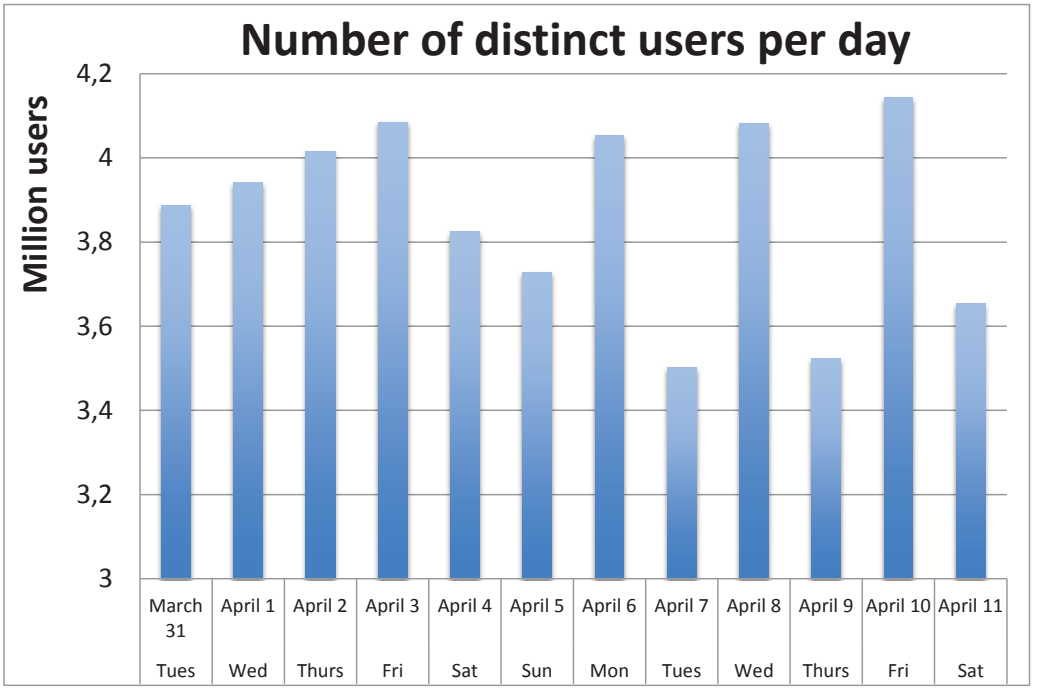


Fig. 4. Daily number of individuals using the network. Source of data: Orange

3.2. Data validation: commuting data and the Enquête Globale Transport

Commuting and travel data are the two main sources of mobility data available in France for urban areas.

The commuting data was provided by the population census conducted by the French National Institute for Statistics and Economic Studies (INSEE - *Institut National de la Statistique et des Etudes Economiques*). Since 2007, the French census is only conducted continuously each year on a part of the population. The data for each year are obtained by processing that collected during five years. The collected data include the location of the individual's home and place of work or study. These are used to produce the commuting matrix which lists the municipality in which the individual lives and that in which he/she works or studies. Based on the assumption that individuals travel every weekday, it is possible to produce a matrix that contains the flows from home to work and place of study at the municipal level. INSEE† provides an order of magnitude for the statistical precision of this flow data.

The *Enquête Globale Transport* (EGT) is "*the main source of information about the trips made by the population of Ile-de-France since 1976"* (STIF, 2010). It is a household travel survey which was last performed in 2010 when it included 18,000 households and a total of 43,000 individuals and 150,000 trips. The survey covered the population of the Île-de-France Region aged over five years. The region was divided into 109 sectors each of which contained approximately 100,000 inhabitants. Between about 400 and 500 households were surveyed in each of the 109

† "*However, in view of, in particular, the sampling, low flows (less than 200 individuals) should be considered merely as orders of magnitude"* (INSEE, 2012)

sectors, so as to construct a geographically stratified random sample. The survey collected sociodemographic data on the household and each of its members. All individuals aged five years and over were then interviewed personally in order to collect all the trips that were made the day before the survey day. The principal characteristics of each trip were collected, in particular the time it started and ended and the origin and destination zone using a grid with 100 metre squares.

The commuting data are produced at the national level and therefore make it possible to identify all the home-based trips to work or study made by individuals residing in France with at least one end in Ile-de-France. However, they have the shortcoming of only covering home-based trips that are made for the purposes of work or study to the exclusion of all other purposes. Conversely, the data from the EGT relate to all trip purposes, but only cover residents of the Ile-de-France Region. Neither of the databases therefore completely matches the mobile phone data which cover all individuals using the Orange network who are present in the Ile-de-France region, irrespective of where they live or the purpose of their trip. These differences need to be taken into account when the origin-destination matrices obtained from each database are compared.

4. Construction of the origin-destination matrices

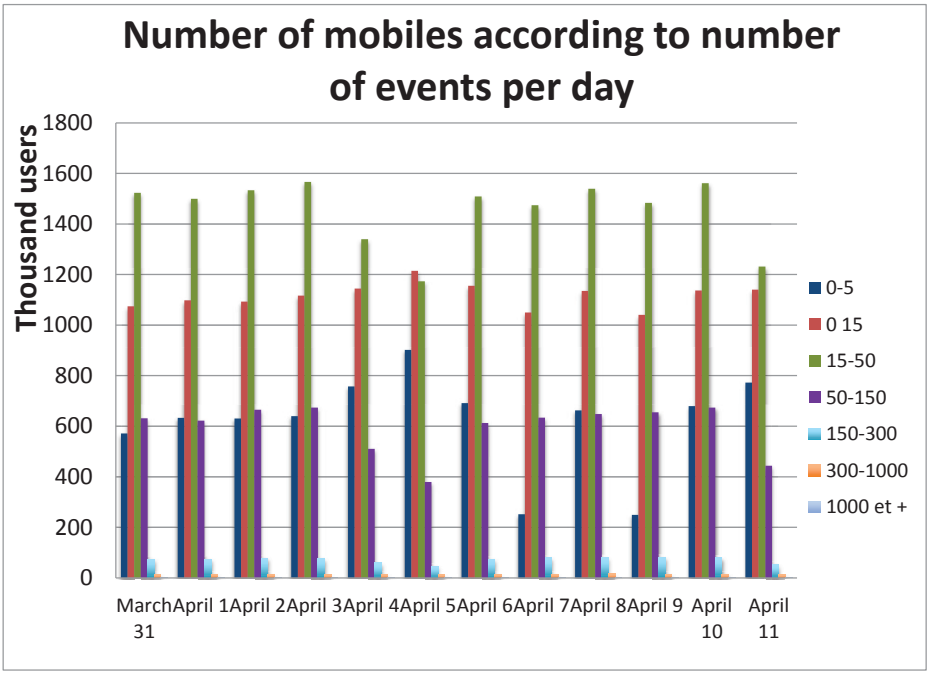


Fig. 5. Distribution of the numbers of users based on the number of events per day. Source of data: Orange

A trip has been defined for the purposes of the EGT 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. We shall deal first of all with our decision to use the signalling database rather than the billing database. The second provides data for each base station, but only when the mobile phone is communicating (sending or receiving an SMS or call in the case of the mobile phone data for 2009). Consequently the amount of information available depends to a very high degree on the amount of mobile phone activity. As the amount of mobile phone activity an individual engages in is strongly correlated to sociodemographic characteristics such as age, there is a risk of bias if this data is used to construct origin-destination matrices. Moreover, the location is only

known when communication takes place. It is therefore not possible to determine the precise location of the mobile phone throughout the day. It is consequently much more difficult to develop hypotheses in order to identify the individual's stationary activities, particularly for phones which are not frequently used, in view of the fact that there are less than 15 events per day for almost half the mobile phones (Figure 5).

However, the signalling data gives us the position of the phone on a permanent basis, but only at LA level. When a mobile phone changes LA an event is generated (a location area update or LAU) in the signalling file, but not in the billing file. The signalling file also contains an LAU every six hours if the mobile phone has remained inactive (new probing systems do this every 3 hours). This database thus allows us to track the mobile phone in a spatially continuous manner and with a maximum time step of six hours on condition that it does not move outside Ile-de-France and remains connected.

The way the trip is defined means that we have to identify an origin and a destination, and hence a stationary activity at the origin and another at the destination. The size of the LA means that most trips between two LA are made by motorised transport, except for adjacent LAs, but in this case trip duration is in general relatively short. In view of the mean speed of motorised trips in each LA as reported in the data from the EGT, we have made the assumption that if an individual is present for at least one hour in an LA 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 therefore need at least two events. To determine a trip has been made (therefore an activity has been performed at the origin and at the destination) we therefore need at least 4 events. To reduce the size of our events database mobile phones with three events or less were excluded (Figure 6).

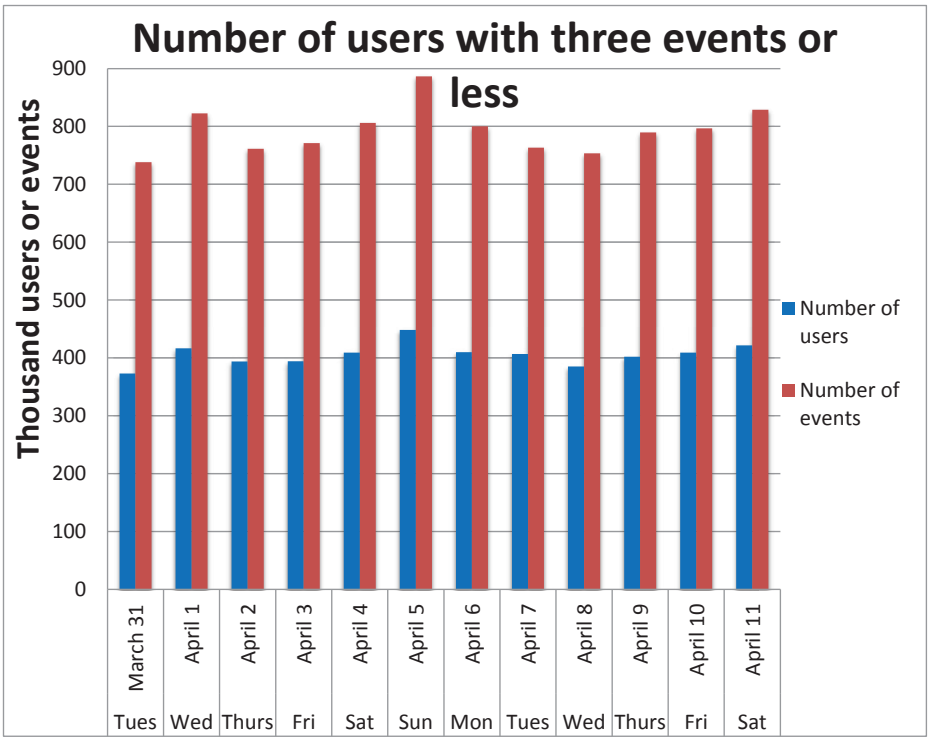


Fig. 6. Number of users with 3 events or less and their respective number of events. Source of data: Orange

‡ We shall return to this assumption in the conclusion

We have also made allowances for the ping-pong effect (Pollini, 1996, Pierdomenico et al., 2011) which occurs when a stationary mobile phone changes base station (this may be for a number of reasons, in particular base station traffic load). The method we have used is similar to that described by (Iovan et al., 2013).

After this data processing and applying the assumption of stationary time, we can construct an origin-destination matrix. It matches the zoning of Ile-de-France into LA. Obviously, this zoning is the outcome of the telecom operator's needs and has nothing in common with the zonings used in travel surveys whether these are intended to consider commuting (municipality level zoning) or the EGT (a grid with 100 metre squares). 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 C_i in the EGT that generates N_{Ci} trips. In order to simplify the calculation, let us assume that this zone straddles two location areas, LA_1 and LA_2. Using the BDTopo database, for each C_i, it is possible to compute the proportion of the built-up surface area that corresponds to LA_1 and LA_2, which are denoted respectively by $p(LA_1)$ and $p(LA_2)$. The N_{Ci} trips can be then distributed in the zone C_i using the following formula:

Number of trips generated by $LA_1 = p(LA_1) * N_{Ci}$

Number of trips generated by $LA_2 = p(LA_2) * N_{Ci}$

Generalisation is straightforward for all the zoning systems for both generation and attraction, 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 commuting data and the EGT with LA zoning.

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 other two matrices contain data for the entire French population or the entire population of the Ile-de-France Region. We therefore need to adjust the mobile phone data. The penetration rate of mobiles using Orange's network is not precisely known in the Ile-de-France Region 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 Ile-de-France, we can determine a daily expansion factor f_i thus:

$$f_i = \frac{population\ of\ Ile-de-France}{number\ of\ persons\ using\ network}$$

This is obviously a strong assumption:

- We have no data that allow us to check that the travel practices of Orange 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 about a third of the market. We can therefore hope that the population of Orange users in Ile-de-France is not too atypical;
- Some members of the population, children and older people in particular, do not own a mobile phone. These people have a much lower level of mobility than the rest of the population;
- Some of the people who use the telecom network in the Ile-de-France 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 and we have preferred to avoid this issue to begin with, as the first straightforward analysis which we attempted gave poor results. Once again, it is likely that the mobility within Ile-de-France of individuals who do not live in the region differs from that of individuals who do;

- Some of the individuals who live in the Ile-de-France Region are not in the region on the days when data was collected and so have zero mobility within the region;
- Last, we do not have a sufficient number of events to be able to identify a trip in the case of a non-marginal proportion of mobile phones (of the order of 10%, Figures 6 and 4). We therefore excluded these mobile phones from the working database. Some of these mobile phones necessarily belong to individuals who do not live in Ile-de-France and who are just entering, leaving or passing through the region, but other mobile phones might also be switched off throughout the day (because a mobile phone that is in Ile-de-France and switched on throughout the day should generate, even if it is stationary and not making calls, at least a location area update every 6 hours and therefore at least 4 events a day). This would lead to an underestimation of mobility.

These general comments aside, it would be risky to attempt to precisely estimate biases. As the effects of biases are to some extent contradictory, we have made the (strong) assumption that they compensate for each other. We therefore have matrices that can be compared in order to analyse the number of trips they contain, but also their structure in terms of origin-destination pairs.

5.1. Comparing the mobile phone data with the commuting data

Table 1 shows that the mobile phone data lead to a marked overestimation of the number of trips. This overestimation is however understandable insofar as commuting trips consist only of trips between an individual's home and their place of work or study. On average, these trips involve longer distances than trips for other purposes and therefore have a greater likelihood of resulting in a change of LA. However, analysis of the EGT data show that a high proportion of long trips are made for purposes other than commuting.

Table 1: Number of trips in the Ile-de-France Region based on the commuting data and the mobile phone data (working days only). Source of date of the data: Orange, INSEE

	Commuting matrix	**Mobile phone matrix**
Number of trips	8,926,000	13,494,000
Mean for each origin-destination pair	9,000	13,600

Looking beyond this major disparity, we attempted to analyse the structure of the two matrices in order to identify any similarities, by means of a variety of analyses (Bonnel et al., 2013). We shall present here the analysis of the correlation between the two matrices after linearisation. The aim was to identify a coefficient of proportionality between the number of trips in each cell of the two matrices (Figure 7). We obtained the following result, where y_{ij} is the number of trips given by the mobile phone matrix for the O-D pair ij and x_{ij} is the number of commuting trips for the same O-D pair:

$$\mathbf{y_{ij} = 1.36 * x_{ij} + 1\,332}, \text{ with } R^2=0.82; \text{ student t for constant} = 4.5 \text{ and slope} = 66.9$$

The constant is relatively small compared to the mean number of trips on the O-D pairs, but it is not null. The R^2 value of 0.82 is acceptable, but when the regression plot (Figure 7) is analysed, we can see there are a large number of origin-destination pairs which are at some distance from the regression line. Even if in very general terms the structure of the O-D matrix is similar, there are clearly quite major deviations from the regression line.

However, the commuting data do not cover all trip purposes, and we should therefore expect deviations. Moreover, analysis of the comparison with the data from the EGT which contains all trip purposes strikes us as being more promising.

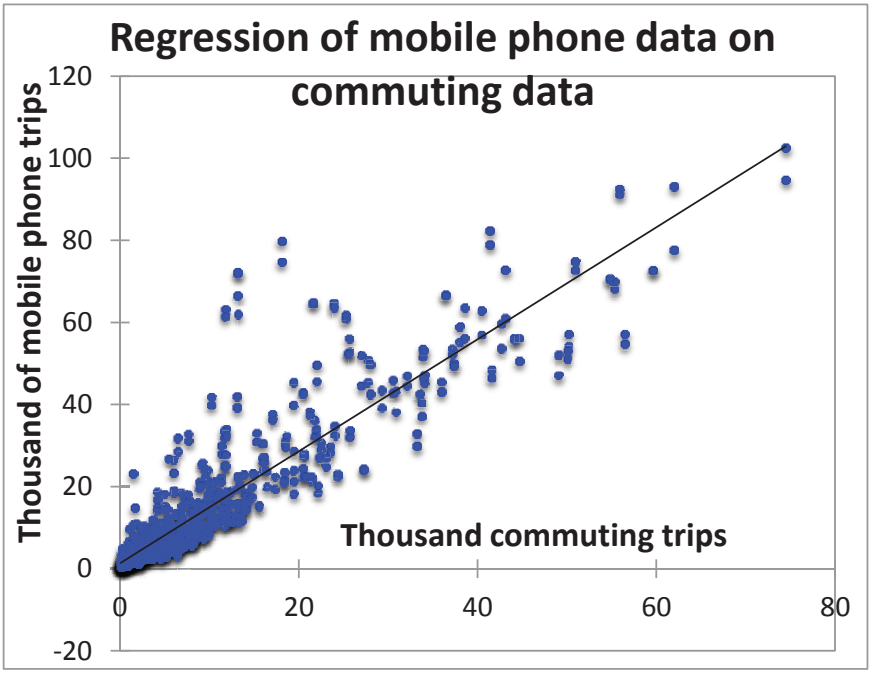


Fig. 7. Linear regression of mobile phone data (working days only) on commuting data. Source of data: Orange, INSEE

5.2. Comparison between the mobile phone data and the data from the travel survey (EGT)

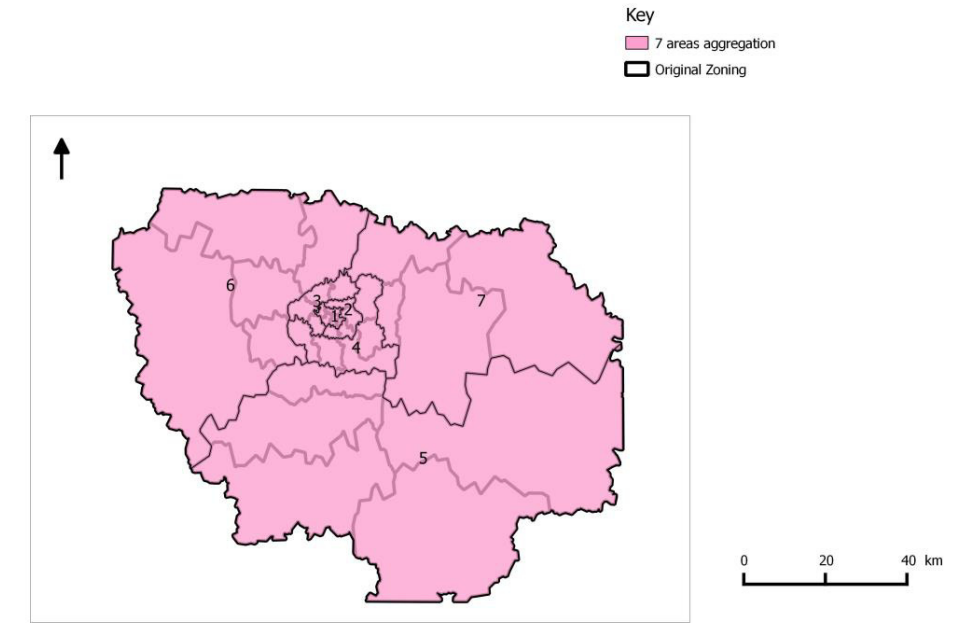
The data from the EGT contain all the trips made by residents of the Ile-de-France Region irrespective of the purpose and the duration of the activity on an average working day. However, we have made an assumption of a minimum of one hour stationary time in an LA in order to identify an origin or a destination in the case of mobile phone data. We have therefore applied the same assumption to the data from the EGT in order to exclude very short activities which cannot be identified as a result of our stationary time assumption. Last, analysis of the total survey sample of the 32-zone EGT zoning system shows that the confidence intervals are very wide for many O-D pairs. We therefore aggregated the location areas in order to produce seven-zone origin-destination matrices (Figure 8) which give a sufficient number of trips for almost all the origin-destination pairs in the EGT. This makes it possible to make a comparison with the mobile phone data matrix which has also been aggregated to correspond to give the seven-zone zoning system.

We can state that the number of trips estimated on the basis of mobile phone data is similar to the estimation based on the EGT, as the difference between the two is less than 10% (Table 2). Furthermore, the regression provides excellent results, and according to the R^2 value almost l00% of the variance is explained by the regression. The slope is close to 1, even if the constant is not null. It is nevertheless relatively small compared to the mean value for the O-D pairs.

$\mathbf{y_{ij} = 0.963 * x_{ij} + 28\ 230}$, with R^2=0.96; student t for constant = 3.46 and slope = 30.8

Table 2: Number of trips from mobile phone data (working days only) and the EGT (division into 7 zones). Source of data: Orange, STIF

	7 zones
EGT	8,739,000
Mobile phones	9,601,000



Analysis of the regression plot (Figure 9) shows that all the points are fairly close to the regression line. However, when we calculated the percentage disparity between the mobile phone data and those from the EGT for each O-D pair (Table 3), we observed that some of these percentage differences were very high. In all cases these corresponded to low flows, which explains why their values in absolute terms are weak. Most of them relate to flows in the outer suburbs or between the second suburban ring and the centre of the Paris conurbation.

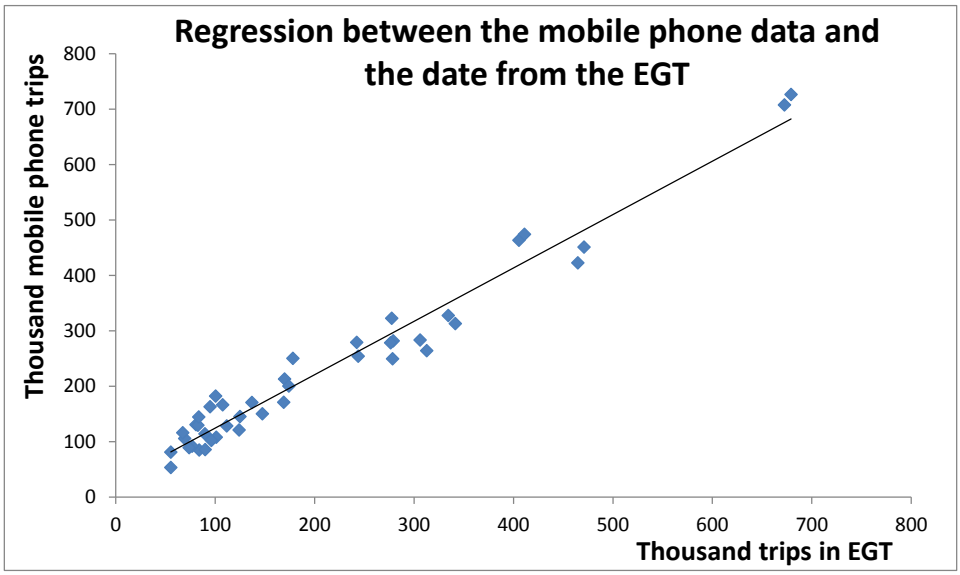


Fig. 9. Regression plot between the mobile phone data (working days only) and the data from the EGT. Source of data: Orange, STIF

We are not really able to explain them in view of the fact that there are many possible sources of error, beginning with the hypotheses which we have been forced to make throughout this study. In addition, the boundaries of the LA do not coincide with the administrative boundaries of Ile-de-France. It is therefore possible that some of the trips that have been assigned to Ile-de-France also involved neighbouring regions. This may go some way to explaining our over-estimates, but it certainly cannot be the sole cause of the disparities.

Table 3: Origin-destination matrix for the percentage difference between the mobile phone data (working days only) and the data from the EGT, Source of data: Orange, STIF

	1	2	3	4	5	6	7
1		5%	15%	4%	1%	7%	1%
2	7%		-2%	-9%	7%	15%	-2%
3	14%	-8%		-16%	-4%	-5%	1%
4	15%	-4%	-7%		15%	-10%	2%
5	62%	81%	46%	41%		73%	72%
6	72%	55%	27%	1%	57%		21%
7	26%	16%	16%	25%	53%	20%	

6. Discussion and conclusion

As many studies have already shown, mobile phone data allow us to construct origin-destination matrices. These matrices were generated from signalling data collected in the spring of 2009 on the mobile phone network operated by Orange in the Ile-de-France 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 have never been validated at the scale of a region like Ile-de-France (12 million inhabitants with an area of 12,000 km^2). More generally, world-wide, very few studies have been undertaken to validate the travel or traffic data obtained from mobile phone data (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).

This work is therefore among the first studies in this area. Chen et al. (2014) have attempted to validate the mobility data produced from mobile phone data by constructing ad hoc data that provided an accurate portrayal of "reality". Most other studies have attempted to compare mobile phone data with external sources of mobility data. These take quite a large variety of forms, for example road traffic counts in the case of Caceres et al. (2007), odometer data obtained from annual vehicle inspections in the USA in the case of Calabrese et al. (2013), but most commonly, the comparison has been with travel survey data. This is what we have done with the commuting data and the data obtained from the *Enquête Globale Transport* (EGT) which is a household travel survey carried out in 2010 of a representative sample of the population of Ile-de-France.

The comparison between the matrices obtained from commuting and mobile phone data provide quite limited results. It is reasonable for the estimated number of trips to differ, as commuting data only relates to trips from the individual's home to their place of work or study, while mobile phone data cover all trip purposes. But the analysis we have performed also show that the structures of the matrices have little in common. There is thus a high degree of dispersion in the rates of variation between the two sources of origin-destination data. This preliminary work was conducted on the basis of fairly strong hypotheses. It is therefore quite possible that more detailed analysis would make it possible to moderate some of the strongest hypotheses and improve comparability. Nevertheless, the only trip purposes covered by commuting data are work and study. It would therefore be necessary to be able to estimate the location of the home and place of work or study of the individuals from mobile phone data in order to significantly improve the matrices produced from this source. Even if a number of algorithms have been described in the literature (Chen et al., 2014; Calabrese et al., 2013; Phithakkitnukoon et al., 2012), our identification of these locations is bound to be uncertain unless we have a large number of events for each mobile phone, which is not the case with the data that we have used. It seems certain that the analysis of the mobile phone data from smartphones which are frequently connected to web-based applications would make it possible to attempt this type of analysis.

Comparison with the data from the *Enquête Globale Transport* (EGT) performed by STIF in 2010 is much more promising. This database has the advantage of covering all trip purposes, not just those related to work and study. However, it only contains individuals who live in the Ile-de-France Region. Using this database we were able construct a matrix that set out to reproduce the assumption of minimum stationary time within a location area which is necessary in order to produce origin-destination matrices from mobile phone data. In view of the size of the survey (43,000 individuals and 150,000 trips), zones were aggregated into seven zones in order to reduce the confidence interval for each element in the EGT matrix. This meant that 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 is very similar with an R^2 value of 0.96 and a slope that is very close to 1. But these very encouraging results should not distract us from the fact that the results are less satisfactory in the case of some origin-destination pairs for which the disparities can attain 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.

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 Ile-de-France Region. However, the data from the EGT only covered Ile-de-France residents. It would therefore be interesting to attempt to identify where the mobile phone owners in the database live. As we have already mentioned, this has not been done because of the short period covered by our data. However, it would be worthwhile to carry out a similar analysis on smartphone data for which there are many more events for each mobile phone, and possibly on data that covers a longer period. This would make it possible to extract solely the residents of the Ile-de-France Region in order to improve the validity of the comparison with household travel survey data. This of course assumes that the travel of smartphone owners is representative of that of the population as a whole, which remains to be verified;
- We have applied a uniform assumption of minimum stationary time of one hour for all the location areas. It would certainly be possible to refine this and vary it according to the characteristics of each LA in terms of surface area and travel speeds. Moreover, the duration of one hour is necessarily somewhat arbitrary, even if it was based on an analysis of the data from the EGT. We have therefore tested the impact of this threshold on the number of trips generated. This analysis was conducted on a single day (Figure 10). It shows that the results are highly sensitive to this assumption. It would therefore be of interest to be able to refine this threshold for each LA and also test the sensitivity of the number of trips to the selected threshold as it is by no means certain that changing the threshold would lead to a proportional change in all the matrix elements.
- The data obtained contain spatial information whose resolution corresponds to the location areas. However, the events in the Orange database pass through base stations. This information is potentially useful. Initially, the zoning would not be changed, as only the database which contains changes in LA gives the mobile traces the spatial continuity which our trip generation method requires. It would however enable us to refine our analyses, in particular as regards ping-pong effects, or alternatively refine our assumption as regards the minimum stationary time within an LA;
- After this, it would be interesting to analyse the data from 3G probes which monitor the exchange of data in addition to phone calls, SMS messages, LA updates and handovers. These databases contain more events, allowing us to make novel hypotheses for trip generation (Chen et al., 2014);
- The boundaries of the location areas are identified by analysing Voronoi polygons. This means there is a high degree of uncertainty about the boundaries. Moreover the actual limits to the coverage of the base stations varies according to mobile phone traffic, the weather and the local topography. It would be interesting both to refine the base station boundaries and hence those of the LA and also study the impact of uncertainty about LA boundaries on the construction of origin-destination matrices;
- Each database uses its own zoning. This means we have to construct conversion matrices to convert one zoning system into another. The analysis we have conducted is based on the surface area of the built up part of each zone. New databases are now available that contain not only the built-up surface area, but also the built volume. INSEE has recently started to make census data available which is much more fine-grained than the IRIS database (Aggregated Units for Statistical Information - *Ilots Regroupés pour l'Information Statistique*). This

makes it possible to assign individuals and jobs to each block of buildings, considerably increasing the accuracy of the conversion matrices that are used for the transformation from one zoning system to another (Manout, 2014);

- The expansion of matrices obtained from mobile phone data is based on the very simple assumption that the population of mobile phone owners for whom we have been able to constitute at least one trip is representative. 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;
- Finally, 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.

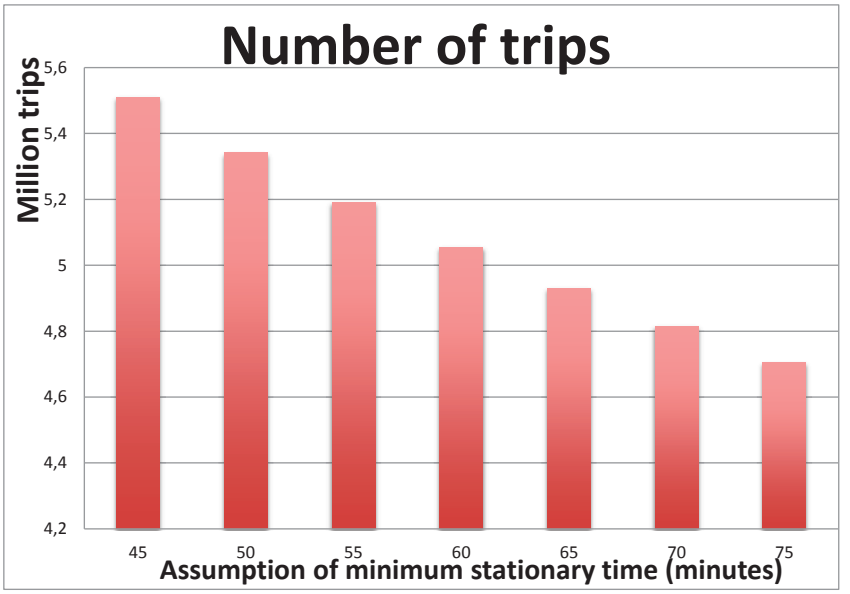


Fig. 10. Number of trips made according to the minimum stationary time assumption. Source of data: Orange

Mobile phone data therefore seem very 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 Ile-de-France Region and STIF for making the EGT 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.

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 (2014) Autorité de Régulation des Communications Electroniques et des Postes, *Observatoires / Services Mobiles*.

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, Laboratoire d'Economie des Transports, Orange Labs, 133p.

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.

Enquête Globale Transport de l'Ile-de-France : http://www.stif.org/IMG/pdf/%20Enquete_globale_transport_BD-2.pdf

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.

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

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.

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.

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.

Manout O (2014), *Codification des connecteurs de zones pour les transports en commun*, mémoire de master TER, Université Lumière Lyon2 – ENTPE, Lyon

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, Italie.

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, *Transportation Research Part C,* 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.

Pierdomenico F, Valerio D, Ricciato F, Hummel K (2011), Steps towards the Extraction of Vehicular Mobility Patterns from 3G Signaling Data, In: *Traffic Monitoring and Analysis*, Springer Berlin Heidelberg, pp. 66-80.

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

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

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.

STIF (2010), *Notice méthodologique de l'Enquête Globale Transport du Syndicat des Transports d'Île de France*, rapport du Syndicat des Transports de l'Ile-de-France, Paris.

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

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.

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.

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.

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

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

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