Designing Origin-Destination Flow Matrices from Individual Mobile Phone Paths

The effect of spatiotemporal filtering on flow measurement

Françoise Bahoken[12], Ana-Maria Olteanu Raimond[3,4]

[1] Paris-Est University, French Institute of Science and Technology for Transport, Development and Networks (IFSTTAR),
[2]UMR 8504 Géographie-Cités, Paris, France
[3]Orange Labs, Sociology and Economics of Networks and Services department, Paris, France
[4]French Mapping Agency, Cogit Laboratory, Saint-Mandé, France

Abstract. In the past few years the mobile phone data are considered as a useful complementary source of information for human mobility research. In this paper, we focus on the computation of the Origin Destination matrix using mobile phone data. First, a new approach of OD matrix design which takes into account the spatiotemporal heterogeneities of mobile phone data is proposed. Second, some analysis allowing measuring the effect of spatiotemporal filtering on the OD matrix results are carried out.

Keywords: Human migration, Flow mapping, Origin-Destination matrix, Mobile Phone Data, Spatiotemporal Filtering

1. Introduction

Origin-Destination (OD) matrices are traditionally used in transportation, urban planning engineering and human migration studies. Matrices result from the sum of individual movements which are produced in a time interval on a given space. Their design has been estimated using a wide range of different approaches which can be grouped into three categories: directed, undirected and alternative methods. The statistics obtained using directed or undirected methods focused on the retrospective information gathering on travel, or individual questionnaires, are very sensitive to errors due to

the adopted methodology. Thus, alternative methods have been developed, focused on the markers of migrations, such as GPS or mobile phones data.

In this paper, we are interesting in designing OD flows matrix using mobile phone data as an alternative of traditional data. We assume that the individual daily movements can be captured from spatiotemporal traces of mobile phones. These traces are not the exact path of the user, but an estimation of it, derived from the mobiles' defined positions in space and time. One of the advantages of using mobile phone traces lies in the space paradigm shift : OD flows can then be generated and studied on a (pseudo) continuous basis, that is to say in fine spatial and temporal resolutions instead of being available on a discrete spatial (administrative units) and temporal ways (defined period). An OD matrix that reasonably reflects temporal distribution is often indispensable for applications ranging from short-term planning to within-day traffic control/management. Moreover, mobile phone data can be automatically collected at relatively low cost and presenting an important sample of users.

The paper is structured as follows. In the next section the state of art is presented. In Section 3 mobile phone data are briefly described. Section 4 first introduced the proposed approach to design OD flows matrix using mobile phone data and second the effects of spatiotemporal filtering are discussed. Finally, section 5 concludes and suggests some directions to future work.

2. State of art

The main issue of classical OD matrix methods (unless for registries) is that they are partial and require a complex statistical treatment in order to estimate the OD flows matrix. In fact, survey approaches implies that the OD matrix represent a snapshot of the commuting patterns over time at a selected spatial scale. Moreover, surveyed data are expensive and sometimes they are likely to be out-of-date. It is for all these reasons that alternative approaches have been developed. They focus on the use of markers of migrations in order to reconstruct individual movements and the resulting aggregated OD matrix. One way is to use the files subscriptions to a service provider like electricity, water or telecommunications furniture. Another possibility is to use mobile phone data.

In recent years, mobile phones have become one of the main sensors of human mobility at a large scale. Generally, there are two main approaches to model human mobility from mobile traces: trip-based, where aggregated data are used (Gonzalez et al. 2008, Sevtsuk & Ratti 2010) and activity-based when individual data are considered (Reades et al. 2007, Ahas et al. 2008 and Gonzalez et al. 2008, Olteanu Raimond et al. 2012).

A variety of studies having as specific goal to infer OD flows matrix using sensors data such as mobile phone and GPS were carried out in recent years. Friaz-Martinez et al. (2012) proposed a method that generates commuting OD flows matrix based on temporal variation of association rules using aggregated mobile phone data. Another approach consists on identifying moving and staying points (Byeong-Seok et al. 2005). If the mobile phone is staying on the same base station over a pre-defined threshold time, then it is consider that trip of the user is finished and an OD flow is generated. This method is very sensitive to the recording rate, and it should be applied only if the recording rate is constant. Calabrese et al. (2011) first computed stops and trips in individual trajectories and then flows for each trips of each user are extracted. Flows are aggregated by origin-destination (i.e. predefined regions) and by temporal window.

Flows can also be defined such as a path starting (the first point of the path) in the origin region and ending (the last point of the path) in the destination region (Caceres et al. 2007, Giannotti et al. 2011). A time window is thus necessary to be defined if fine temporal analysis must be carried out.

In this context, we propose a static OD flows matrix computation which means that linked flow data exist only for one time period. Such an approach is inspired from (Caceres et al. 2007, Giannotti et al. 2011). Our method, described in section 4.1, consists on taking into account the spatio-temporal heterogeneities characterizing mobile phone data.

3. Mobile phone data

Each mobile phone operator collects and store for a given period customers' mobile phones activities for billing or for technical measurements purposes. This type of collection is called "passive collection", since recordings are made automatically. They are three mainly types of mobile phone data collected using the passive collection: Call Detail Records (CDR) data, Probes data and Wi-Fi data. In this paper, only CDR data are described, since these data are used to validate our approach. For more information about mobile phone data, see (Smoreda et al. 2012).

CDR data are cell phone billing records, where location information (cell id) is automatically generated at the moment of communication: call's start (in/out-coming) and SMS (in/out). The records contains the following attributes: i) the anonymised SIM card identifier; ii) the antenna identifier; iii) the base station location; iv) the record type (call in/out, SMS in/out) and v) the time of communication activity (timestamp).

In this study the location of mobile phone users is limited to the base station location. The main advantages of CDR data are: the big mass of located data for a long period of time and for a very large spatial extends (e.g. country level). Moreover, records represents all operator' clients, not only a sample of users. The disadvantage is due to the heterogeneity of records (only when a communication occurs).

4. OD Matrix Design and the Effects of spatiotemporal filtering

In this section, we first describe the OD matrix approach. Second, the effect of spatiotemporal filters is analyzed by consider a real case study.

4.1. OD matrix design

Let's consider a spatiotemporal trajectory T_k composed by a set of n consecutive points, noted $T_k = \{p_1, p_2, ...,p_i,...,p_{n-1}, p_n\}$, where :

$p_i = (x_i, y_i, t_i)$ is a record point having a spatial location (x, y) at moment t,

$t_1 < t_2 < ...t_i<...t_{n-1} < t_n$,

i= 1..n represents the number of points composing the T_k.

Definition: For each trajectory T_k, there is a flow between the areas (i,j), noted F(i,j) if the two following conditions are satisfied :

(1) $p_1 \subset i$ *OR* ($p_2 \subset i$ *AND* $p_1 \in$ neighbors list of p_2)

AND

(2) $p_n \subset j$ *OR* ($p_{n-1} \subset j_j$ AND $p_n \in$ neighbors list of p_{n-1})

Figure 1 shows an example of our flow path definition. Let's consider two distinct trajectories (T1 and T2) belonging to two distinct users. Considering the below definition, the trajectory T1 can be consider such as an OD flow between the origin (i) and the destination (j) : at least two points of the path are related to different spatial units of the study area and the flow can be design from the crossing borders phenomena. On the opposite, the trajectory T2 cannot be consider as an OD flow between (i,j) since the first condition (1) is not met : only one point of the trajectory is included in a spatial area and crossing border phenomena do not exist.

The distinction between the two trajectories is due to spatiotemporal filters which affected the number of migrants between (i,j) and, consequently, the number of migrations or paths. It depends, on the one hand, on the tem-

poral filter -which defines position along time- and on the other, on the spatial filter -which defines the size of the areas.

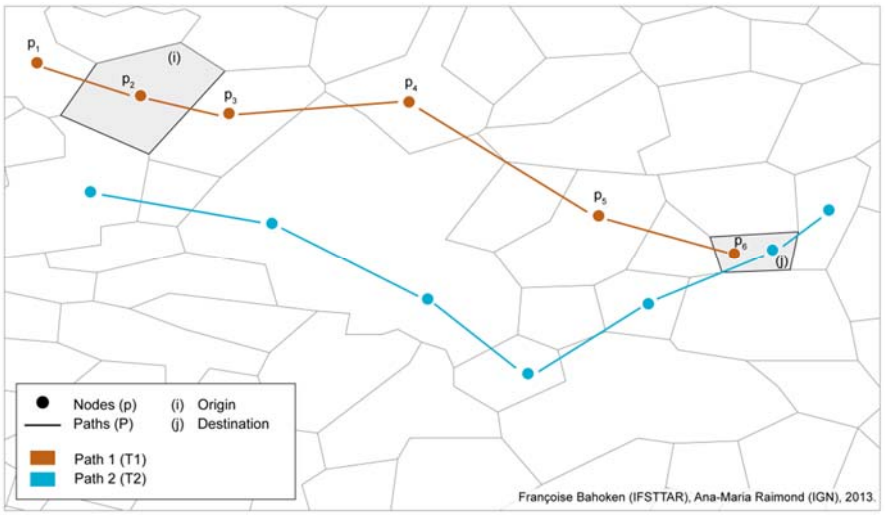

Figure 1. Spatial definition of flow path.

4.2. Data sets presentation

Three datasets are used: the CDR data, the French municipalities zoning system and the French urban nodes in *Picardie* Region.

The first dataset is the CDR data containing all mobile phone calls and SMS collected from BTS towers located in the study area during six weeks from 1st of September to 15 October 2007. BTS locations are used to build Voronoi polygons representing the antennas coverage (see Figure 2 and 4).

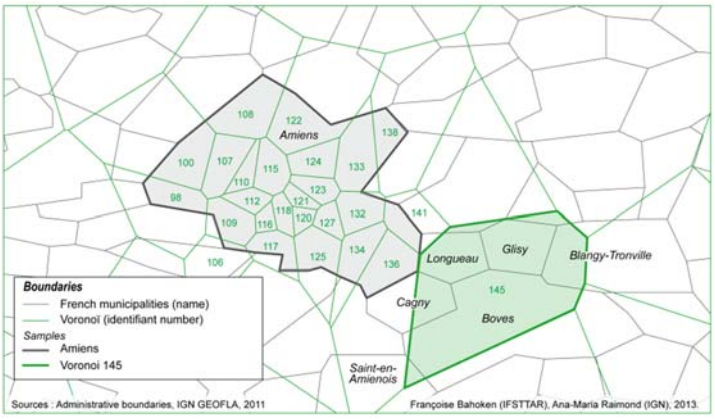

Figure 2. Voronoi polygons in the study area.

As we can see in Figure 2, the Voronoi polygons are very heterogeneous from a spatial point a view. For example, the municipality of Amien is characterized by small Voronoi areas, since the Voronoi "145" covers partially or completely five municipalities (e.g. *Boves, Longueau, Cagny and Blangy Tronville*). The dataset accounts 10,145,916 users.

Figure 3 shows the temporal distribution of mobile phone events aggregated by 60 minutes for one week. Some activity peaks can be observed at 1pm, and 7pm for weekdays and 1pm, 8pm and 9pm for weekend. These peaks are related to lunch break, commuting coherent with the French daily activity peace. We notice that the distribution of events is less important during the weekend except during the night between 1 am and 5 am.

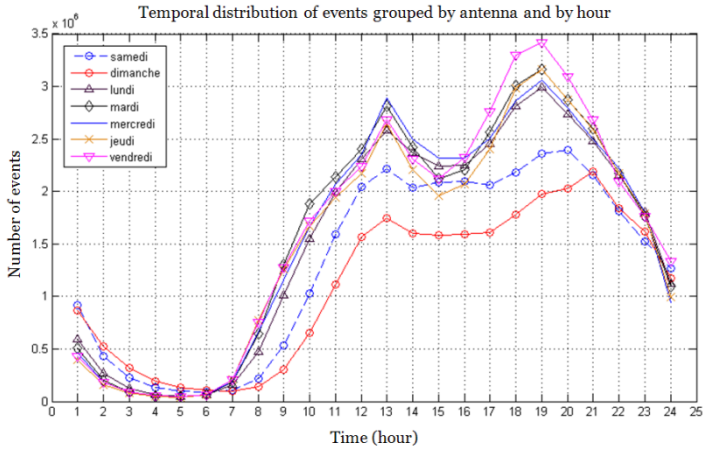

Figure 3. Temporal distribution of events during a weekday.

The second dataset contains the French municipalities zoning system areal data (polygons) representing urban areas (see Figure 4). An Urban Area (UA) represents a set of municipalities (the French *communes*) requiring several conditions: the UA must be in one piece and without enclave, composed by an urban center and must have i) more than 10,000 employments and ii) rural municipalities surrounding that have at least 40% of the resident population which is employed in the urban center (INSEE 2003).

Finally, the third dataset represents urban nodes system (polygons). A node is a group of municipalities in one piece and without enclave, consisting i) of an urban center from 5,000 to 10,000 employments and ii) rural municipalities with at 40% of the resident population which is employed in the urban center (INSEE 2003). It should be notice that the selected nodes correspond to the Central municipality of large urban cores defined in the Urban area zoning system.

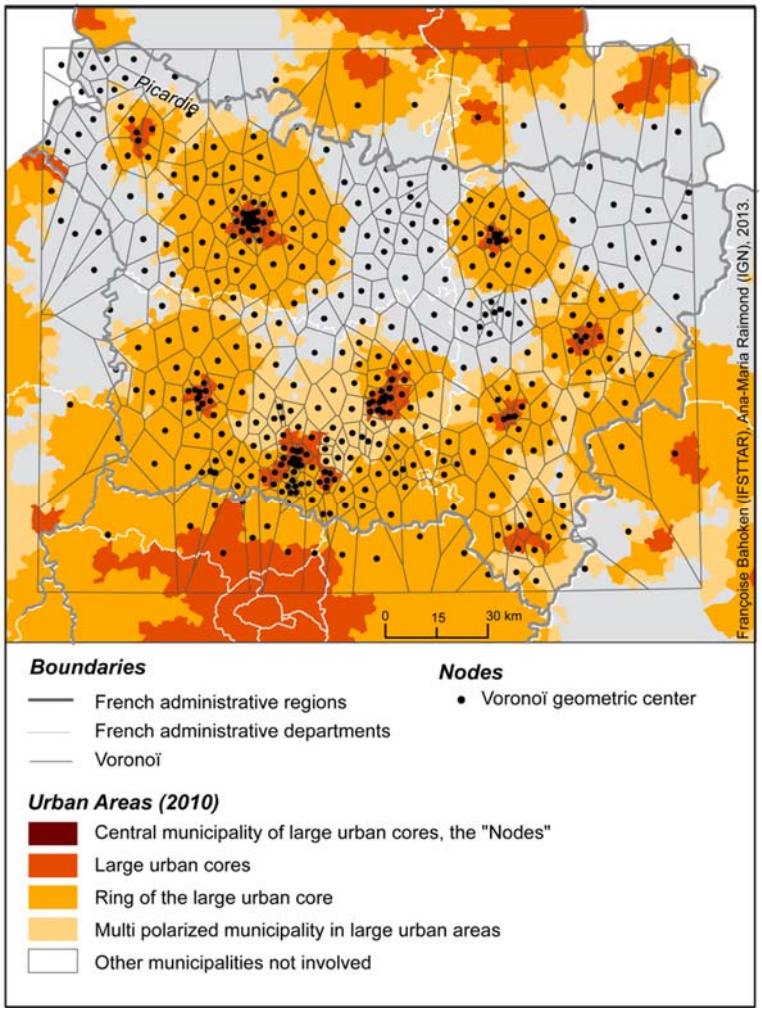

Figure 4. Urban Area zoning system (2010) and Node zoning system in The Picardie Region.

4.3. The effect of temporal filtering

As we noticed earlier, an OD matrix can be computed in defined time window interval. We are interesting to measure the quantity of links and flows which is lost when temporal filtering is applied. The number of links measures the interaction between places. The more, the number of links is the more relevant the interaction between places is. The number of flows quantifies how strong the interaction between two places is.

Generally, time window intervals are defined according to a specific goal such as flows analysis during the peaks hours, slack periods or work-days flows analysis (Lavielle 2008). According to this, different time window intervals were applied: 24h-, [6am - 10pm], [4pm - 8pm]. Thus, the temporal filtering generates three datasets. The OD matrix method was applied to each dataset at Voronoi polygons level. The results are then aggregated in four classes: Weekday class contains respectively links and flows calculated for weekdays in 24 hours time interval. Morning and Evening weekday classes represents the links and the flows computed for weekdays rush hours, respectively in [6am – 10pm] and [4pm - 8pm] time window intervals. Finally, Week-end class contains links and flows computed for week-end in 24 hours time interval. Table 1 shows the loss information (in terms of links and flows) after time filtering. Notice that *total* represents information before applying the following temporal filters (weekday, morning weekdays, evening weekdays and week-end).

As we can see in Table 1, generally the effect of temporal filtering is more important for the quantity of flows based on the total.

For example, morning and evening weekday's classes lose more than 65% of flows. However, the most effective temporal filter is the one that keeps significant values, of flows while reducing the number of small links, which help to confuse the message. From this point of view, it is then interesting to observe that when a 24 hours filter is applied for weekdays only 2% of areas interaction is lost since, 36% of flows are lost. Thus, selecting links that occur during the weekdays, allow keeping 64% of the total flux values, which is relevant and leads to a loss of 2% of the number of links that is not substantial. The 24 hours filtering is the one which keeps the part of the most relevant information (links and flows).

	Percentage of Loss Information	
	Links (%)	Flows (%)
Total	0%	0%
Weekday	2%	36%
Morning Weekdays	24%	65%
Evening Weekdays	23%	71%
Week-end	19%	64%

Table 1. Loss information account due to the temporal filtering.

4.4. The effect of spatial filtering

In order to study the effect of spatial filtering we focused on commuting flows for weekdays. Combining morning and evening flows, sub-matrices at several spatial levels allows us to generate a complete matrix representing daily flows. Daily flows are computed at three different scales (e.g. the Voronoi, the urban area and node scales) using the approach described in Section 4.1. The Voronoi scale is the most detailed scaled that it can be consider, knowing the spatial resolution of mobile phone data.

We notice that the spatial filtering has no apparent effect between the Voronoi and UA scales i.e. that no information is lost when merging Voronoi to UA. The spatial filtering effect is however relevant for node spatial scale (55% of flows are lost when passing from Voronoi scale to node scale). Nevertheless, the effect of spatial filtering can be measured by observing the distribution of inter and intra flows. Therefore to measure the change of the importance of zones, both inter zonal and intra zonal flows are analyzed. Let's remember that the elements of the OD matrix, noted Fij, represent the number of trips (flows) from one origin area i to one destination j during a determined time interval. The OD flows matrix (F) is asymmetric if i=j, then $Fij \neq Fji$.

Table 2 described the percentage of inter and intra flows considering the spatial scale. We notice that at Voronoi scale the proportion of flows is quite similar between inter and intra flows. The most relevant spatial effect is noted at node scale when 97% of flows are inside the nodes, the interaction between nodes being weak (3%).

	Inter Flows (%)	Intra Flows (%)
Voronoi	46%	54%
UA	15%	85%
Node	3%	97%

Table 2. OD matrix results at different scales.

In Figure 5 obtained flows at Voronoi scale are represented.

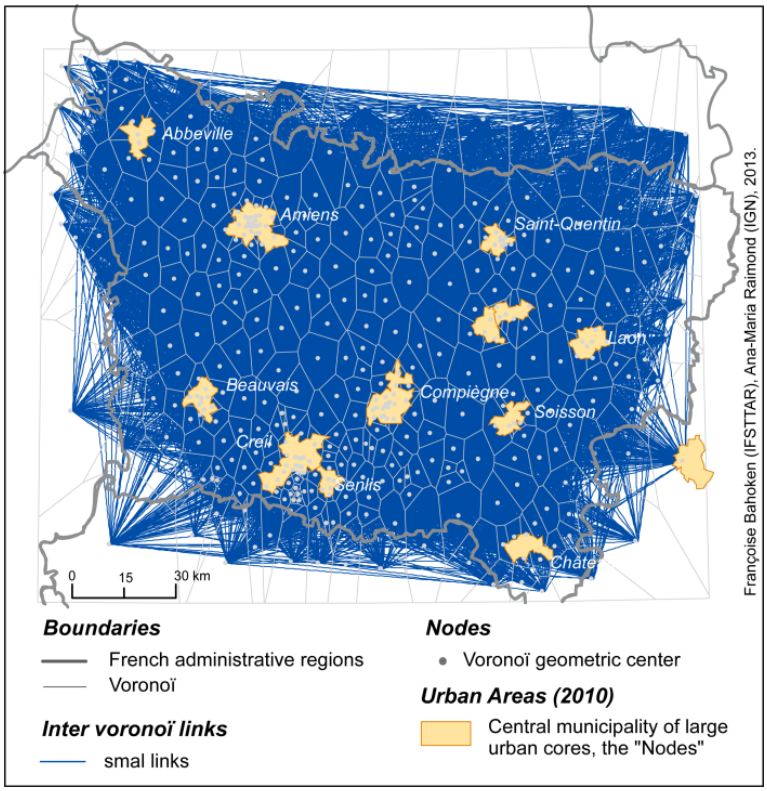

Figure 5. The total amount of flows at Voronoi Scale –example of spaghetti effect.

Mapping such flows requires the use of two different filtering procedures due to the spaghetti-effect (see Figure 5). The first one focuses on reducing the number of ties represented that we called the graphical filtering and the second one, on the aggregation procedure which is, in reality, a spatial filtering procedure based on the flow values (Bahoken, 2012).The spatial distribution of antennas determines a very specific Voronoi coverage: antennas are numerous and aggregated in urban and peri-urban areas and scattered in rural one (see Figure 5). This distribution prohibits the definition of partitions based on the relevance of

the value of flows. Therefore, we proposed an analysis based on inter Voronoï links' length, but not weighted by the number of links involved in the first place.**Fehler! Verweisquelle konnte nicht gefunden werden.** shows the normal Quantile-Quantile diagram of link's distance. It compares the observed distribution with a Gaussian. The point are not on the first bisector so that the distribution do not follow a standard Gaussian distribution law (the standard deviation is 37.6).

Figure 6. Normal distribution of links length (km).

The median value of the Voronoï links' length is 75 km, the minimum is 0.6 km and the maximum is 208 km. Thresholds can be defined to observe the distribution of flow values according to the distance. We have chosen to present, by way of illustration, the following three thresholds based on quartiles (first quartile, median, third quartile). Three classes of links are obtained: small links- distance smaller than 44 km, medium links- distance between 44 and 53 km) long links- distance greater than 99 km (see Figure 7).

The analysis of the distances between the antennas shows that areas having a relevant interaction (i.e. the number of flows is important) correspond to urban areas. We also notice that the small links are inside the urban core which corresponds to our nodes and the medium links polarized the urban areas. The *Amiens* Node has a central position in terms on in links.

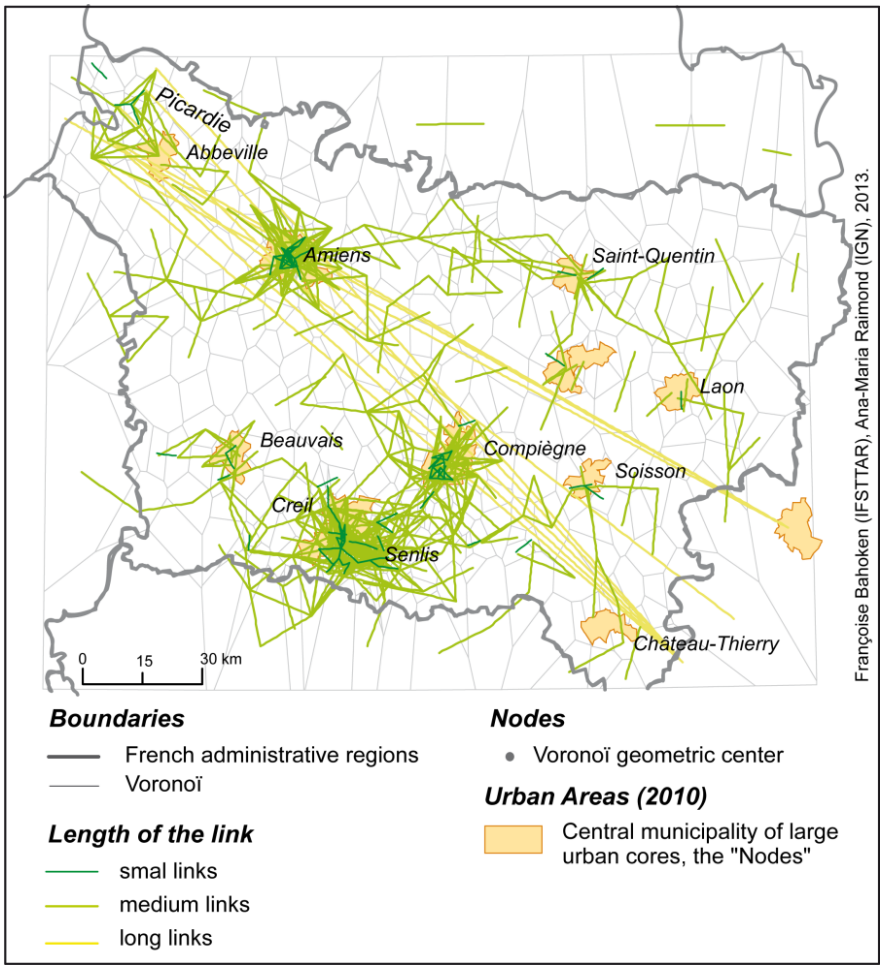

Figure 7. Flows representation according to the link's length at Node scale.

Figure 8 shows the flows representation at Node scale. It should be noted that, at this scale, there is no spaghetti effect and all bilateral flows can be represented.

The value of the flows represents the quantity which has moved between two nodes, for a week. We can note that the most important flows occurred between *Beauvais* and *Tergnier* then between *Senlis* and *Compiègne*. It is therefore at the heart of the *Picardie* region that produces the most important exchanges. Finally, it should be noted that this result obtained at the poles scales does not highlight the centrality of *Amiens* that was observed at the level of Voronoï. This assumption illustrates the instability of the statistical results when changing spatial scale, it deserves to be deep-

ened by a more analysis of the flow values, but it was not the goal of this paper.

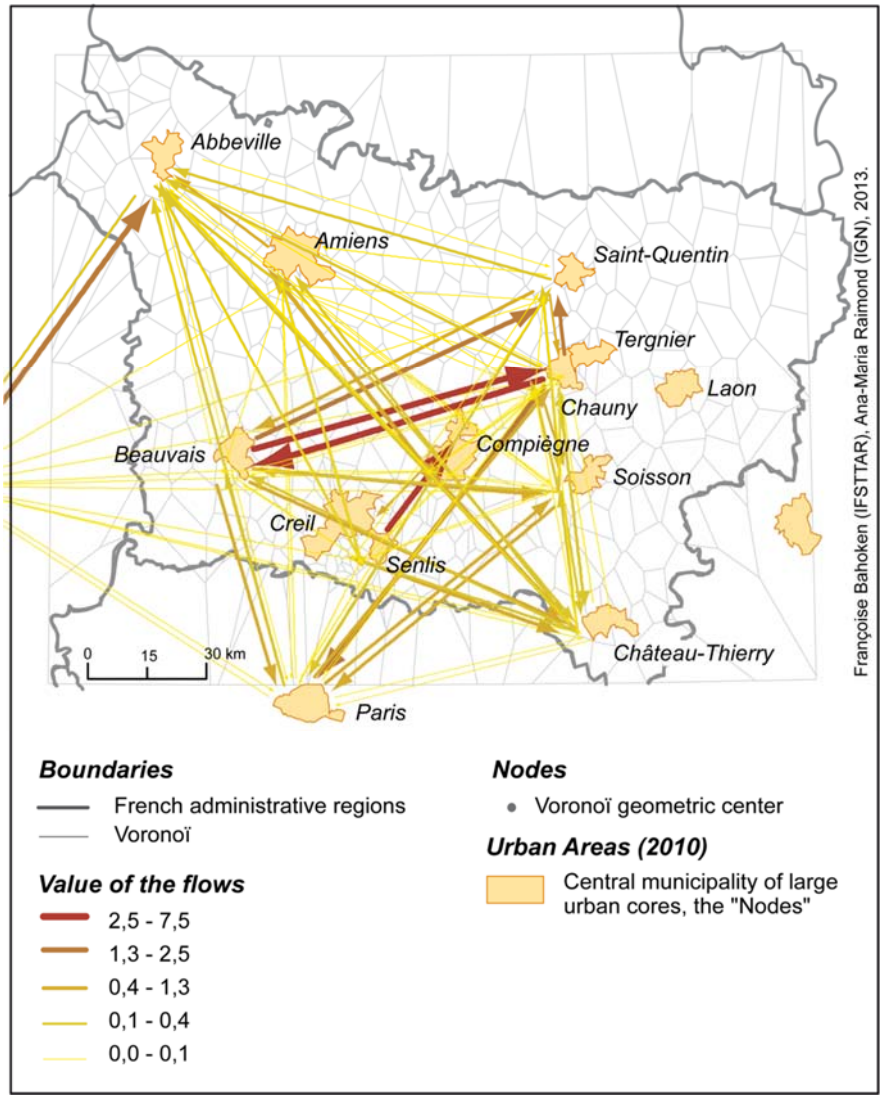

Figure 8. Flows representation at Node scale.

5. Conclusion

In this paper we focused on OD matrix design using mobile phone data and observed the loss of information associated with the different filtering procedures. A new approach allowing taking into account the mobile phone data spatiotemporal heterogeneities is then proposed. Since the computation of OD matrix depends on both temporal and spatial filtering, an analy-

sis of the effects of different spatiotemporal filtering was carried out. The results show that the more the temporal window is small the more the information in terms of links and flows is lost. Concerning the spatial filtering, flows were aggregated at three different scales: Voronoi, urban area and node. We observed that the loss of information is no relevant when flows are aggregated from Voronoi scale to UA scale and it is relevant (55%) when flows are aggregated at poles nodes. It is important to notice that the spatial filtering has relevant consequences on the results when inter and intra flows are distinguished.

In order to respond to privacy issues, few conditions were respected: all records were anonymized, no individual demographic or socioeconomic data were added to mobile phone data, the commuting flow detection were made at the spatial aggregate level and finally, the information presented is always aggregated.

References

Ahas R, Aasa A, Roose A, Mark, Silm S (2008) Evaluating passive mobile positioning data for tourism surveys: An Estonian case study. Tourism Management, 29: 469-486

Bahoken F (2011) Comparison of functional regionalization's of the world: a methodological study based on Intramax procedure, in European Colloquium on Quantitative Geography Proceedings, Athens, pp. 647-654

Bahoken F (2012) Application du raisonnement logique à la cartographie des flux, Proceedings of SAGEO, International Colloquium of Geomatics, Liège, Belgique, 7-9/11/2012

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

Caceres N, Wideberg J, Benitez F (2007) Deriving origin-destination data from a mobile phone network, Intelligent Transport Systems, IET 1(1): 15-26

Calabrese F, Di Lorenzo G, Liu L, Ratti C (2011) Estimating origin-destination flows using opportunistically collected mobile phone location data from one million users in Boston Metropolitan Area, IEEE Pervasive Computing, 10(4):36-44

Frias-Martinez V, Soguero C, Frias-Martinez E (2012) Estimation of Urban Commuting Patterns Using Cellphone Network Data, ACM SIGKDD Int. Workshop on Urban Computing, Beijin, China

Giannotti F, Nanni M, Pedreschi D, Pinelli F, Renso C, Rinzivillo S, Trasarti R (2011) Unveiling the complexity of human mobility by querying and mining massive trajectory data, The VLDB Journal — The International Journal on Very Large Data Bases, 20(5):695-719

Grasland C, Bahoken F, Beauguitte L, Pion G, Van Hamme, G (2009), Toolbox for flows and network analysis (Methodological Paper). Deliverable D.5.1. Euro-BroadMap.Vision of Europe in the World. Small or medium scale focused project FP7-SSH-2007-1

González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns, Nature 453:779–782

Holland, S.C., Plance, D.A., 2001, Methods of mapping migration flow patterns. Southeastern Geographer, 41(1): 89-104

Insee (2011) Base communale du zonage en aires urbaine 2010, URL : http://www.insee.fr/fr/methodes/default.asp?page=zonages/aires_urbaines.htm (verified, 2011, 05, 09)

Olteanu Raimond AM, Couronne T, Fen-Chong J, Smoreda Z (2012) Le Paris des visiteurs, qu'en disent les téléphones mobiles ? Inférence des pratiques spatiales et fréquentations des sites touristiques en Ile-de-France. Revue Internationale de la Géomantique, 3:413-437

Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular Census: Explorations in Urban Data Collection. IEEE Pervasive Computing 6:30-38

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):41–60

Smoreda Z, Olteanu-Raimond AM, Couronné T (2013) Spatiotemporal data from mobile phones for personal mobility assessment, In Zmud J, Lee-Gosselin M, Carrasco JA, Munizaga MA (eds), Transport Survey Methods: Best Practice for Decision Making, Emerald Group Publishing, London

Tobler, W., 1987, Experiments in migration mapping by computer, American Cartographer, 14:155-163