Estimating the Residential Population from Mobile Phone Data, an Initial Exploration

Benjamin Sakarovitch*, Marie-Pierre de Bellefon*, Pauline Givord, and Maarten Vanhoof*****

Abstract – Many studies are focused on using data derived from mobile phones to construct statistical indicators. Mobile phone data have the advantage of providing information with both high spatial resolution and at high frequency, allowing applications such as measurements of the spatial or temporal details of population presence. Nonetheless, using mobile phone data to construct statistical indicators raises difficulties: data from a single operator are not representative of the whole population and they often lack socio-demographic detail, which limits their quality for many applications. This article is based on a database of mobile phone records from subscribers collected by a large French operator. It aims to offer a view on the potential, but also the problems posed by mobile phone data, specifically by illustrating how indicators of residential populations can or can not be estimated from them.

JEL Classification: C55, C81, R23
Keywords: mobile phones, call detail records (CDR), present population

This research was carried out within the framework of a tripartite agreement between Orange, Eurostat and the Insee's Directorate of Methodology and Statistical and International Coordination. The authors would particularly like to thank Zbigniew Zmoreda and Cezary Ziemlicki for their welcome to the Sense laboratory and their support with the use of mobile phone data, the participants of the Eurostat Big Data Task Force and especially Michail Skaliotis and Fernando Reis, as well as Elise Coudin and Vincent Loonis for their valuable advices. The authors remain solely responsible for any errors or omissions that may remain in the article.

* *Insee (marie-pierre.de-bellefon@insee.fr ; benjamin.sakarovitch@insee.fr)*
** *Insee, Crest (pauline.givord@oecd.org)*
*** *Open Lab, Newcastle University / Orange Labs (m.vanhoof1@ncl.ac.uk)*

Received on 20 February 2018, accepted after revisions on 5 October 2018
Translated from the original version: "Estimer la population résidente à partir de données de téléphonie mobile, une première exploration"

To cite this article: Sakarovitch, B., Bellefon, M. (de), Givord, P. & Vanhoof, M. (2018). Estimating the Residential Population from Mobile Phone Data, an Initial Exploration. *Economie et Statistique / Economics and Statistics*, 505-506, 109–132. https://doi.org/10.24187/ecostat.2018.505d.1968

The use of Big Data, linked to rapid advances in the capability to store and analyse huge volumes of data, has expanded significantly over the last decade. Big Data, created by the digital trails generated by the activities of individuals or companies, are often studied from the viewpoint of predictive analysis or to support decision-making. Another use is that they can also serve as source of observations useful to the construction of statistical indicators, which explains the interest shown in these data by official statistics institutes.[1] The expected opportunities for the use of Big Data in official statistics would be to reduce publication times by taking advantage of the very rapid access to useful information (e.g. in the field of economic analysis), but also to produce more detailed statistics (in particular, geographically) than the ones currently based on survey data, and finally to reduce the workload of collecting information from people and companies. As an example, automatic price gathering (from e-commerce sites or from invoicing data of major retailers) is used by several statistical institutes to construct consumer price indexes.[2] The use of alternative or additional sources to "conventional" data is subject to multiple studies, although the idea is certainly not new. Notwithstanding that official statistics have been complementing statistical surveys with government sources for decades now (e.g. for a long time Insee has used its statistical tracking of salaries on employers' social security returns), the integration of Big Data sources raises new, specific questions such as technical issues regarding data in large volumes or unstructured formats.

Data from mobile phones form part of the sources identified as particularly promising to supplement statistical information. Such data consist of regular records for the location (or at least the location of the cell tower the phone is connected with) date and time of phones belonging to the subscribers of a mobile phone operator. As such, mobile phone data can provide information on population presences at specific locations over specific time periods, and this at a very fine levels of geographical and temporal precision. Although official statistics produce information about residential population (especially by means of the census), access to the fine detail of mobile phone data would make it possible to detect the number of people who are at a given moment (which depends, for example, on tourist visits, business behaviour, etc., see Terrier, 2009), as well as the movements of people between several points. Regularly locating subscribers thus enables the mapping of population presence and the way it changes (Deville *et al*., 2014; Debusschere *et al*., 2016; Ricciato *et al*., 2015). For example, these data can be used to measure the variability in visitor numbers to certain places during the day or during the year, to improve precise knowledge of travel times using different means of transport (in particular for "small" daily journeys) and to draw up detailed mobility matrices (see Aguiléra *et al*., 2014, for evaluating performance of the Île-de-France transport network or Demissie *et al*., 2014, for Senegal). The visitor profiles of an area at different moments in time can assist the analysis of regional dynamics. Since we can expect presence (or activity) profiles to change during the day depending on the type of place (home, workplace or travel hub), Toole *et al*. (2012) were able to distinguish the main activity of areas, depending on the daily presence profiles observed mobile phone data (e.g. shops, residential, industrial or car park) across the Boston area. For France, Vanhoof *et al*. (2017) applied a similar approach at municipality scale, and revealed a correlation between aggregated activity profiles of mobile phone cell towers and the type of communities they are located in, as defined by the French statistical office's (Insee) zoning of urban areas. Ultimately, the information from mobile phone data can also be used to enhance the analysis of interpersonal networks, for example by analysing the strength of communications between subscribers or regions (Grauwin *et al*., 2017).

Nevertheless, using mobile phone data raises several questions. Firstly, it is necessary to guarantee respect for subscribers' privacy. Being able to reconstruct individual journeys using the trails left by subscribers creates a risk of "re-identification". Even by deleting all direct mentions about their identity, from a certain point onwards it is possible to attribute an observed journey to a single person with high probability (Montjoye *et al*, 2013). This requires that mobile phone data be aggregated at an adequate level to prevent individual identification, or that privacy will be protected by procedures that do not allow for practitioners to have direct access to sensitive data. The former solution has the disadvantage that it reduces information and relevance of the data, whereas the latter requires new platforms and procedures to be implemented with regard to most present-day

1. e.g. see the Scheveningen Memorandum (2013)
2. In France, the "checkout data" project is based on price records taken from invoicing data from several large retailers (see Leonard et al.*, 2017, and* Economie et Statistique / Economics and Statistics *N° 509 forthcoming).*

situations.[3] Secondly, in technical terms, mobile phone data for millions of subscribers represent huge volumes that require suitable storage and computation infrastructures.

Notwithstanding the questions raised, statistical offices are interested in the potential of this type of Big Data. For example, a Eurostat report (2014) studied the potential of mobile phone data to improve the accuracy of current tourism indicators. Additionally, several national statistical institutes have launched initial experiments in using different Big Data sources and a coordination programme was launched in 2016 to share knowledge on this subject.[4] One central element is access procedures for official statistics institutes that cover both subscribers' privacy and business confidentiality for the companies involved. For France, a CNIS report offers guidelines on reusing company data in official statistics (2016), specifically highlighting the case of mobile phone data.[5] Simultaneously, other European official statistics institutes have begun negotiations with national operators and are now engaged in experimental projects (Debusschere *et al*., 2016, for Belgium). Such experiments are needed to define what information at what level of aggregation is needed to construct relevant statistical indicators (Vanhoof *et al*., 2018).

In the case of mobile phone data, experiments have raised multiple questions. Firstly, using mobile phone data from one operator raises questions of representativeness. Access to an operator's data will only supply information about its subscribers, who only makes up part of the population. Understanding this bias requires additional information, such as the local coverage of these operators, which is necessary to obtain more detailed spatial statistics. Additionally, the level of mobile phone ownership can vary depending on population characteristics: Some people may not have a mobile phone – e.g. Wesolowski (2013) highlighted problems of the unequal distribution of telephones in different social groups in Kenya for the use of this type of data, while others may have several mobile phones.

A second difficulty in using mobile phone data relates to the grid of cell towers, which in principle does not match normal geographical grids (e.g. administrative subdivisions). Cell towers are not distributed uniformly – there are more in densely populated areas and fewer in rural areas. To use them across more traditional territorial units, translations from the cell tower grid need to be made, which introduces approximations (Ricciato *et al*., 2015).

Finally, it is essential to clarify what can be measured from mobile phone data. These data are produced "naturally" (sometimes called "organic data", as opposed to "designed data", supplied using surveys constructed with the aim of measuring the study object[6]), they simply reflect the trails left by subscribers on the mobile phone network. For a statistical indicator to have a meaning everyone can understand, it is essential first to agree a definition of what we want to measure. For example, a tourist is generally defined as a person registered "outside their usual environment". Tourist visit measurements for a place therefore require distinguishing, among people present in this place, those who do not live there but also those who do not work there regularly. To measure this information from records of subscribers' journeys require being able to identify a person's home or even their "usual" workplace (Janzen *et al*., 2018). Several studies on this question have been conducted based on mobile phone data. For example, Ahas *et al*. (2010) showed that it is possible, using trails left by an individual on the network, to reconstruct their "anchor places", i.e. places important to them, where they go repeatedly – their home and workplace being the most obvious of them (Ahas *et al*., 2010). As also emphasised by Song *et al*. (2010), the time spend by each person is generally concentrated on a limited number of places. Several algorithms have been suggested to identify a subscriber's likely home from observed journey profiles (Vanhoof *et al*., 2017; Bojic *et al*., 2015; Isaacman *et al*., 2011). This point is essential as it is a prerequisite to many other analyses (Blondel *et al*., 2015) that go beyond the simple scope of tourism.

This study is offered to, based on a practical example, illustrate the empirical questions raised by the use of mobile phone data. The study will use mobile phone data from subscribers to a French telephone operator over the course of five months in 2007. It will try to

3. For example, the Opal project (http://www.opalproject.org/about-us/) offers providing researchers a platform to run algorithms on mobile phone data to which the researcher does not have direct access: we are talking about Open Algorithm rather than Open data.

4. https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/WP5_Mobile_phone_data.

5. See Cnis-Insee report "Reuse of Company Data by the Public Statistical System".

6. In particular, this distinction has been suggested by the Census Bureau, see https://www.census.gov/newsroom/blogs/director/2011/05/designed-data-and-organic-data.html.

estimate resident population figures from this data, compare them with estimates from official statistics taken as reference data, and analyse sources of discrepancy. This methodology makes it possible to test the relevance of several home-detection algorithms and several data aggregation techniques: two essential questions when you want to use mobile phone data.

The rest of this article is organised as follows. The first section describes the different types of mobile phone records and details how to convert them to a localised population count. A second section discusses the coverage of cell towers and their translation to administrative divisions. The next section presents the different methods used to estimate resident populations. It gives details about the representativeness questions and suggests solutions that can be used to resolve them, as well as comparisons with reference sources. Finally, the last section suggests some other ideas to use mobile phone data to characterise population presence dynamics.

Data Records

A mobile phone network enables communication by transmitting radio waves between devices, repeater towers and the operator's centralised switches that direct the connection to other repeater towers for the person being called. These networks have a cellular structure, i.e. each cell tower covers a certain area and a telephone can change cell without the communication being cut off.

Principle of Mobile Phone Records

The data used here are records by "repeater towers" of the cellular network, which report the presence of subscribers' cellular telephones near these cell towers. They are mounted on towers with known coordinates. In principle it is therefore possible to construct indicators about visitor numbers to certain places, or very finely detailed geographical and temporal mobility behaviour. The frequency and regularity of these records, and therefore the level of detail (granularity) at which we will be able to construct these indicators, depends on the data type. There are several data types.

CDR (Call Detailed Records) relate to making or receiving a call or an SMS, i.e. a deliberate action by the subscriber. We therefore call them active data. These data as generally used for invoicing and operators therefore recording them "by default". In France, operators have to retain these data for six months. Besides indicating the location of subscribers, these data can be used, for example, for studies on user behaviour (call frequency, preference for text messages, etc.).

Signalling data, what we will call passive data, are generated from telecommunication and internet networks (2G, 3G, 4G), using the fact that all mobile telephones connect regularly to the nearest cell towers (with variable frequency that can range from three hours to ten minutes) without necessarily arising from the user's action on the mobile. They therefore provide more complete information than CDR, for example if you want to measure the number of visits to a place at a given moment or track people's movements. However, processing these data is more expensive. By default, these "events" are not recorded by operators: to do so requires very large storage capacities.

In terms of population coverage, the data recorded by an operator, whether active or passive, relate only to their subscribers. However, there may be "roaming" agreements that enable one operator's subscribers to use its competitors' networks when they are outside the area covered by their own operator. In France, there are few roaming agreements between the national operators, and this "roaming" situation essentially relates to foreign subscribers. In particular, this means that it is possible to identify people only passing through France, as long as they are using their telephones (for CDR data) or they at least leave them switched on (for signalling data). The SIM card makes it possible to identify the telephone operator's home country, from which the telephone subscriber's probable nationality can be inferred.[7]

Approach to convert records to population counts

A series of processing operations is needed to derive information useful for official statistics from data recorded by the mobile network (Diagram).

7. *Before June 2017, these overseas roaming costs were invoiced by the operators. Since this date, the European Commission required such invoicing to end. It is possible that this will ultimately lead to creating a more competitive European market, as the nationals from one country are more easily able to use a foreign operator and it will therefore be more difficult to identify these journeys.*

Box 1 – **Description of Mobile Phone Data Used**

The study relates to using an anonymised file of CDR data (Call Detailed Records) containing the complete record of subscribers' activities for the operator Orange across mainland France over a five month period, from mid-May to mid-October 2007[a]. The records cover about 18 million SIM cards and more than 20 billion observations. These data contain no direct information about the subscriber's name or address. However, for the study, it was possible to supplement them with certain information taken from a Customer Relationship Management (CRM) file, designated in the article as "customer file". For 12.4 million SIM cards also included in the CDR (i.e. about two-thirds), this customer file indicates the *départements* in which subscribers have stated they live. The subscriber (as identified in the customer file) is not necessarily the telephone user. Imagine for example the case of parents funding the subscription for mobile phones used by their children. Furthermore, information from the customer file can "expire", such as when moving home, as the information is not always updated.

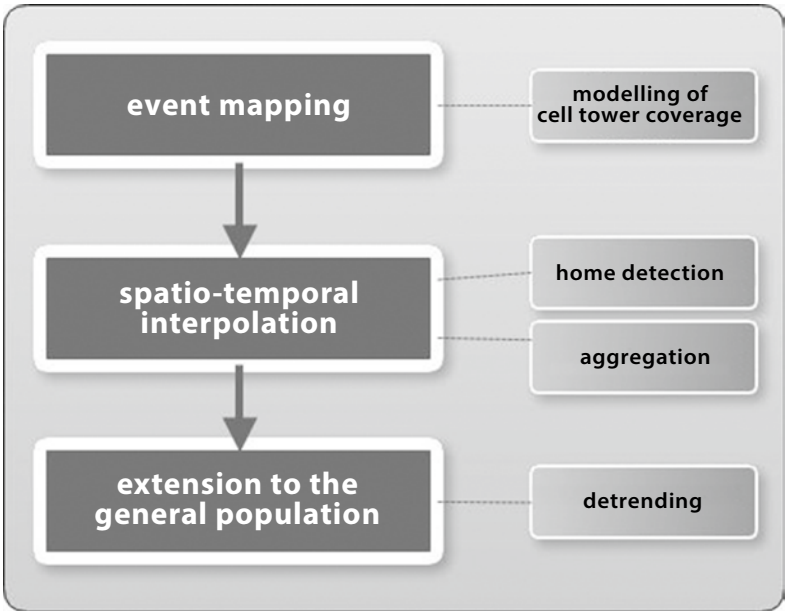
(a) These anonymised data are available to the Orange Labs SENSE laboratory, for purposes of research projects.

Table A
Structure of Call Detail Records and Essential Variables

Sending SIM card	Receiving SIM card	Event type	Sending cell tower	Receiving cell tower	Timestamp	Duration
SIM-1	SIM-2	Call	A-1	A-2	13/06/2007 -14:26:03	7m32s
SIM-1	SIM-3	SMS	A-3	-	25/08/2007 -12:04:58	-

Note: For SMS messages, we don't know the cell tower through which the message is received.

Diagram
Diagram to Show Processing of Mobile Phone Data for Oficial Statistics



The first step is mapping events recorded on the network (calls or SMS). The location of the event is inferred from information available on the location of the cell towers. To practically perform the mapping, you have to define a spatial grid on which you want to locate the different events, and secondly model the area of cell tower coverage (in particular, based on the technical characteristics of the cell towers, if they are available, see Ricciato *et al.*, 2017). As detailed in the section "A Very Non-Uniform Grid", we use the simplest way of modelling cell tower cover areas by using the Voronoï tesselation of the cell towers (see Box 2). Based on this coverage model, the event will then be located on the chosen spatial grid.

The second step is to perform a spatio-temporal aggregation to convert the record of events into aggregated data matching a predetermined definition. This consists of defining aggregation units (both temporal and spatial) to produce statistical indicators. For example, we may want to construct indicators of populations present in places based on traditional administrative subdivision (by IRIS, municipality, etc.) at specific moments of the day, or at least over fixed time periods. The grid used to convert from cell towers to places, related to their technical characteristics, does not naturally match the conventional territorial subdivision. It is therefore necessary to perform a spatial interpolation. In some case, this spatial interpolation must be coupled with a temporal interpolation, as the records from SIM cards have neither defined nor regular frequency: for example, based on call activities, we may have the location of the same telephone at 7:47 am then at 8:12 am, however the location of this same telephone at 8 am is not directly known. If the aim is to measure the population over specific times, it will be necessary to reconstruct the probable location at 8am from these available data. Finally, to estimate the resident population indicators, we must try to infer the probable home location, based on the times and locations of available in the data. The home detection algorithms that perform this step are described in a next section.

A final step seeks to obtain estimates for the reference population (the entire population of France), based on the aggregated data subscriber counts from moble phone data. This aggregation is supported by external sources (e.g. operators' market share). Several possible estimates are presented for the reference population, depending on the depth of additional information available, stressing the underlying hypotheses. These results are compared to reference statistics (resident populations, such as measured by taxation sources processed by Insee).

Approximation of Cell Tower Coverage: Simulation from Taxation Data

A Very Non-Uniform Grid of the Country

Spatial coverage across the country is uneven. For each operator, the repeater towers that supply the main information about location are sited unevenly across the country. As shown in Figure II, in 2007 the cell towers of the operator Orange were very densely distributed in urban areas but much less densely in rural areas. Furthermore, mobile infrastructures can boost the network locally to prevent it becoming saturated during events leading to large crowds – sporting events, concerts, demonstrations. In more structural terms, the development of technologies (successive releases of 2G, 3G, 4G, etc.) leads to renewing the network and therefore changes to the location of cell towers.

In practice, we can infer the probable position of a telephone from the cell towers to which it is connected. The simplest solution is to assume that it is connected to the nearest cell tower.[8] We can define subdivision of the country using a Voronoï tessellation (Box 2), which matches each cell tower to all the points in space that are nearest to it. This model of coverage is an approximation of the actual coverage of cell towers. It does not take into account that in the real world coverage areas overlap and that the load of telephones present in a given area is split among the various cell towers covering it. Still, in our simplification we consider the Voronoï polygons of all cell towers as our spatial unit of observation. Due to the unequal distribution of cell towers across the country, the areas of these polygons are very variable in size (Figure I). Figure II shows the distribution of their areas. We can see that while many Voronoï cells have quite areas small area (a few hectares), the range of areas covered is very broad and goes up to more than ten thousand hectares for some cells. These large areas do not correspond to the effective coverage of the cell towers but arise from the Voronoï tessellation in regions where the cell towers are a long way from each other and can even actually include "white" areas where no signal is received.

As confirmed by Figure VI-A, the smallest Voronoï cells are located in the most densely-populated areas.

8. This is an approximation based on the assumption that cell towers all transmit with the same power and in all directions. In reality, one mast can hold several aerials transmitting in transmission directions (all over the place) and with different ranges. Scholus (2015) or Tennekes (2015) constructed an inference model for the position of the mobile based on the detailed observation of the properties of cell towers, as well as knowledge of the distance between the telephone and the cell tower that retransmitted the signal. However, this information (properties of cell towers, distance to the telephone) is not always available in the data. Furthermore, having very frequent information can permit triangulations that make it possible to identify the position of a mobile precisely. In the ideal case, where the distances to several cell towers (at least 3) are reported, it is possible to use triangulation to deduce the exact position of the telephone.

Translating Voronoï Cells to Another Grid

The purely technical geometric partitioning of the space by Voronoï polygons obviously does not coincide with the subdivisions of the country used for circulating regional statistical data. Indeed, there is no reason for cell tower coverage to correspond to administrative boundaries of municipalities or *départements*, nor should they be contained within the finer grids used by official statistics, such as IRIS (the building blocks for circulation of infra-municipal information, interlinked within the community geography and forming uniformly-sized units in terms of population[9]). As a consequence, it requires translation to a

9. *https://www.insee.fr/fr/metadonnees/definition/c1523*

Box 2 – **Partitioning the Space, the Voronoï Tessellation**

The Voronoï tessellation is a partitioning of the space based on a set of given points: the seeds. Each point on the plane is allocated to the seed to which it is closest. The boundaries between the different areas of the plane form the sides of polygons containing exactly one seed.

This subdivision of the plane is useful for processing mobile data when you only know the locations of the various cell towers (which therefore form these seeds). We then assume that a call is transmitted using the nearest cell tower, which therefore means that the telephone is located in the Voronoï polygon associated with this cell tower.

Figure A
Example of a Tessellation Using Voronoï Polygons Derived from 7 Points

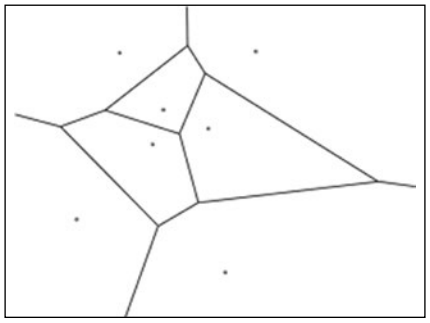
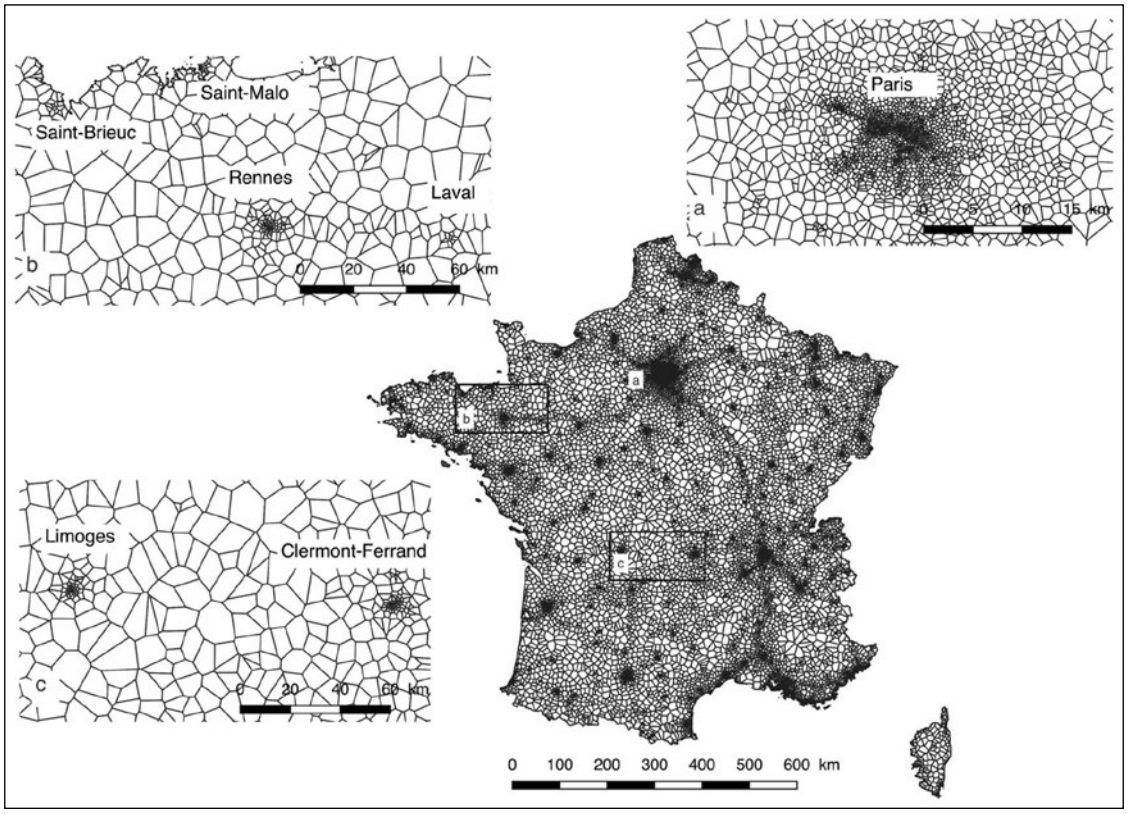
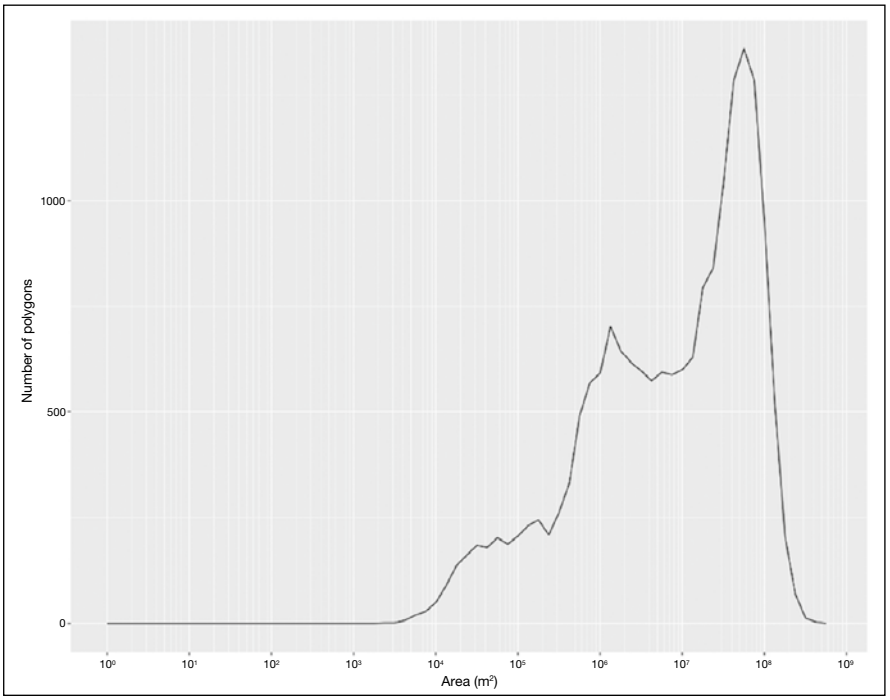


Figure I
Voronoï Polygons Associated with Cell Towers for the Operator Orange, Metropolitan France



Sources: Orange CDR.

Figure II
Area Distribution (in m²) of Voronoï Polygons Associated with Cell Towers for the Operator Orange



Reading note: The modes of the distribution are at 10^6 and 10^8 m² (the graph is plotted on a logarithmic scale); there are no polygons with area less than 10^3 m².

conventional administrative grid to estimate regional statistics from mobile telephone data and compare them with information provided at the scale of this administrative grid.

In what follows, and due to the lack of better information, this translation will be done simply weighting the areas of the polygons as they are situated in the administrative grid. The base administrative grid chosen is the municipal grid, divided into *arrondissements* (districts) for Paris, Lyon and Marseille. Subscribers' counts for the administrative grid will correspond to the sum of the estimated subscribers in the Voronoï polygons that are entirely enclosed within a unit of the administrative grid and the number of estimated subscribers weighted by the proportion of the areas of the Voronoï polygon covering an administrative unit in the case that Voronoï polygons overlap several administrative units (see also equation 1).[10]

$$N_c = \sum_{V_j} \frac{A_{V_j \cap C}}{A_{V_j}} N_{V_j} \qquad (1)$$

Where N_C represents the estimated number of subscribers in administrative unit C, N_{V_j} the number of subscribers detected in Voronoï polygon V_j, A_{V_j} the area of this Voronoï polygon, and $A_{V_j \cap C}$ the area of the intersection between the administrative unit and the Voronoï polygon.

Within equation 1, we base on the assumption that the population density present is uniform over the whole polygon. This assumption is obviously debatable, in particular in rural areas where dwellings are typically more concentrated. In the next section we evaluate the impact of the translation from Voronoï grid to the administrative grid by reproducing it for a conventional official statistics dataset, namely Tax files. Since tax files are complete (available for the entire population) and geolocated, they serve us well to investigate the effect of translating between both grids.

Simulating the Approach on Tax Data to Evaluate the Scale of the Approximation

Insee has complete information about the resident population at regional scale. The "Localised Social and Tax File" (Filosofi), which replaces and supplements the "Localised Tax Revenue File" (RFL), is made up from complete files of physical persons' tax returns

10. https://www.insee.fr/fr/metadonnees/definition/c1523

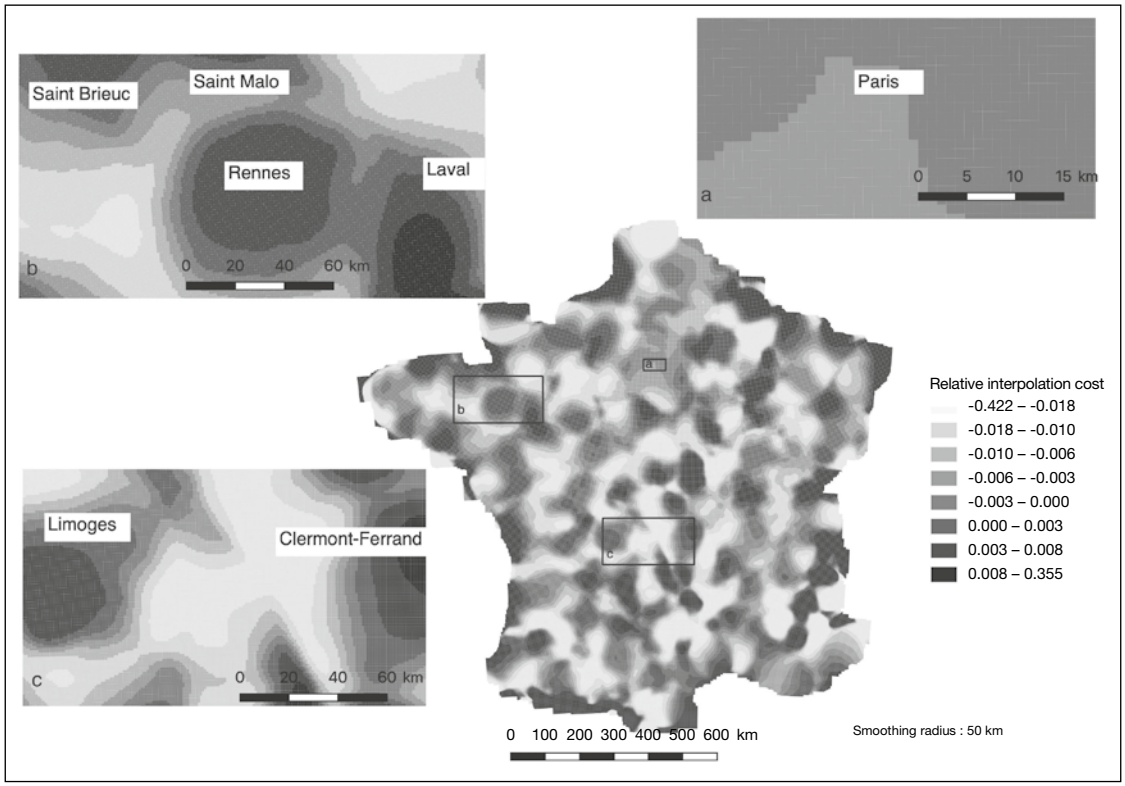
and local housing tax. This information is available in even greater detail than the mobile telephone data, since it is geolocated.[11] However, temporal accuracy is much less since this information is produced annually. Furthermore, these tax files only provide information about where people live and not about their actual presence in certain places (which can vary during the day). Nonetheless, they may constitute an interesting source of comparison to evaluate the suitability of mobile telephone data to reconstruct conventional statistical indicators, such as population density.

From the geolocated tax data, we estimate the spatial distribution of the number of inhabitants for municipalities and Voronoï polygons. Similar to previous section, we translate the information at the Voronoï polygon grid to the administrative grid and compare the outcome with the direct estimate at administrative level (municipality, in this case). The term "interpolation cost" describes the measurement error contributed by the translation between both grids. In practice, this is the difference between the number of inhabitants measured directly for the administrative unit and the estimate obtained from the translation (normalised to the reference).

Figure III illustrates the distribution across the country of the interpolation cost. For municipalities located in rural areas, this interpolation generally leads to overestimating the municipal population. The grid translation is actually based on the assumption that the population density is uniform over an entire polygon. The grid of cell towers is looser in less densely-populated areas. The corresponding

11. https://www.insee.fr/fr/statistiques/fichier/2520034/donnee-carroyees-documentation-generale.pdf

Figure III
Relative Difference between the Municipal Population and the Population Estimated by Grid Translation (Interpolation Cost)



Reading note: The interpolation cost is the difference between the municipal population obtained directly in the tax source and that estimated from spatial interpolation (using equation (1)). A negative interpolation cost corresponds to an overestimate of the municipal population, a positive interpolation cost is an underestimate).
Sources: Filosofi 2011; authors' computations.

polygons therefore cover a larger surface area, even if dwellings are more spread out – this makes the underlying assumption all the less plausible.[12] We find these differences when we estimate the effects by size of municipality. For municipalities smaller than 10,000 inhabitants, the relative difference related to spatial interpolation is an average overestimate of 53% (see Figure C1 in the Online complements[13]). Conversely, for municipalities larger than 10,000 inhabitants, spatial interpolation rather tends to underestimate the actual municipal population – nonetheless the relative differences are smaller (without ever being negligible): they are 10% on average.

The results suggest that using a grid that does not superimpose directly over the "conventional" subdivisions is a significant factor for the quality of estimates produced from these data. One solution would be to set aside the administrative subdivision by considering Voronoï polygons as the base unit, but it has the disadvantage of being based on a grid – that for the cell towers – that is neither stable over time nor uniform in space. This partitioning of the space is also based on an approximation of the true cell tower coverage, which probably affects the quality of the results obtained. In reality, antennas on cell towers are directional and only provide coverage up to a certain distance. This also explains the presence of "white areas" mapped by ARCEP since 2017.[14] In addition, their areas of coverage are very often super-imposed, unlike a tessellation. Having this information about the technical capabilities of cell towers would make it possible to refine the actual partitioning of the corresponding space. For example, exploratory work by the Dutch Central Bureau of Statistics (CBS) proposes using a Bayesian inference procedure to allocate and point in space to one or other of the nearly cell towers, based on their power and orientation.[15] Future work will be able to reveal the benefit gained in terms of accuracy and the cost in terms of complexity. But the data we are using does not contain the technical information needed for this exploration. Furthermore, as discussed below, other problems are raised by using mobile phones, which result both from the definition of a concept (how to convert the record of a telephone call in the management data to a statistical indicator?)[16] and from that of their statistical treatment (how to obtain representative estimates of the whole population from the subscribers to a single operator?).

Constructing Statistical Indicators from Data

Home Detection

The data we have correspond to the trails left by subscribers during their journeys. In principle, these recurrent journeys indicate the use of places specific to the subscriber and so it seems possible to infer subscribers' likely home, or workplace, or other places important to them. Such information is useful or even essential to construct certain statistical indicators, such as home/work journey times or tourist numbers in certain regions. Regarding tourists, for example, they are defined according to the "statistical" definition established by the United Nations World Tourism Organization Statistics Department, stating that tourism is "the activities conducted by people during their travels and stays in places located outside their usual environment for a consecutive period not exceeding one year, for leisure, business or other reasons". While the usual environment can be interpreted to vary in size, it includes at least the home and workplace. This information is rarely available in the anonymised files to which researchers or statisticians have access and thus several home detection algorithms have been proposed that try to infer them from mobile phone data.

The general principle of home detection is to define the home from criteria based on the frequency and/or times (generally the night) users are present in a place. Vanhoof *et al*. (2018) offers a review of several home detection methods. We extract five methods to be used here:

– "Maximum activity": Home is the place where most events (making and receiving calls or SMS) take place during the period under study;

---

12. Finally, it should be noted that this consists of differences related to the number of inhabitants in the municipality – numerical differences can be amplified for very small municipalities.

13. See the link to Online complements at the end of the article.

14. The site https://www.monreseaumobile.fr/ can be used to see the white areas by network and operator.

15. This work is accessible from the available mobloc R package R, the address: https://github.com/MobilePhoneESSnetBigData/mobloc. It is also described in Dutch here: https://circabc.europa.eu/webdav/CircaBC/ESTAT/ESTP/Library/2017 ESTP PROGRAMME/46. Advanced Big Data Sources - Mobile phone and other sensors, 6 – 9 November 2017 - Organiser_ EXPERTISE FRANCE/Mobile_Phone2.pdf.

16. A statistical indicator here means the quantification of a social reality (e.g. the population present), based on a convention to be defined (for Desrosières (2008), "to quantify means to agree then to measure").

– "Number of distinct days": Home is the place where activities were recorded on the largest number of distinct days during the period under study;

– "Time constraint": Home is the place where most activities between 7 pm and 9 am were recorded during the period under study;

– "Spatial constraint": Home is the place where most activities were recorded within a 1 km radius around the cell tower during the period under study;

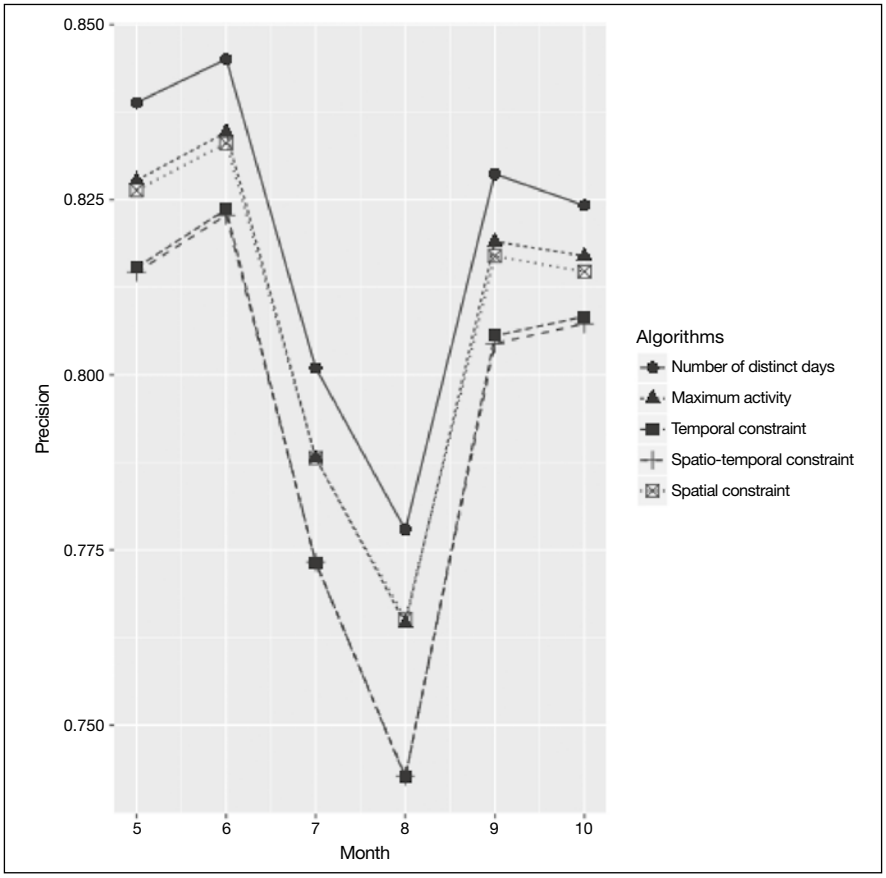
– "Spatio-temporal constraint", a combination of the previous two.

These different algorithms are all reasonable guesses. Nonetheless, they all have their limits and, for each, we can also easily think of situations where the identification of home will be incorrect. For a given subscriber, different methods can identify different places as the likely home.

To evaluate the performance of the different home detection algorithms, we have additional information supplied by the customer file, which contains the post code of the subscriber's home. This information is only available for two-thirds of the subscribers, but nonetheless it is possible to compare this post code with the estimate of home supplied by the different algorithms. Furthermore, we also have data on the resident population supplied by the localised tax files.

Figure IV shows the accuracy of the five proposed home detection algorithms, by comparing their residential population estimations with information from the customer file. The estimates are made for each month and by *département*. The accuracy corresponds to the proportion of subscribers included in the customer file for whom we have correctly identified the *département* in which they live (in the sense of it matching that in the customer file). Over the entire study period, the algorithm

Figure IV
Accuracy at *Département* Level of Home Detection Algorithms Based on the Customer File



Note: Accuracy correspond to the proportion of subscribers included in the customer file for whom the localisation algorithm determines the same *département* as the customer file.

using the number of distinct days (i.e. the place where activities were recorded on the largest number of distinct days) performed the best. Even at this aggregated regional level,[17] we see that the difference between the home *département* as identified by the algorithms and as stated in the customer file remains large (never less than 15%). The discrepancies can partly be explained by the inadequacy of the heuristic home detection methods we used. For example, accuracy falls significantly in summer and can very probably be explained by the fact that a significant proportion of the population is on holiday at that time and do not reside in their usual *département* for the whole month. This difference may also be linked to a quality problem with the customer file. Even ignoring the summer months, we see reduced accuracy over the whole period for all the algorithms (the differences observed in September-October are greater than those observed in May-June), which may be due partly to an effect of the customer file ageing (e.g. updates not made when subscribers move house). In addition, the data only contain records for the end of May (18 days) and the beginning of October (14 days), which may also explain the poorer performance than in June and September, respectively.

Additionally, it is worth noting that a user is considered as having a home in a *département* if the cell tower allocated to them by the home detection algorithm is within the *département*. There can be marginally edge effects for cell towers with Voronoï cells that overlap several *départements*. The Online complements contain maps representing the geographical distribution of this accuracy for June and August.

Adjusting Data to Obtain Estimators of Resident Population

The mobile phone data available to us relate to a single operator's subscribers only. To estimate statistics on the entire French population it is therefore necessary to perform detrending. This detrending should make it possible to convert from the subscriber population to the total population, which may differ for two reasons. The first is that the operator only covers a proportion of mobile phone subscribers. This operator's market share indicates the order of magnitude of the relative difference that we expect to find between the actual population and "raw" estimates obtained with mobile phone data. According to the *Autorité de régulation des communications électroniques et des postes* (Arcep – French electronic and postal communications regulator), the national market share of the operator Orange in 2007 was 46.7%.[18]

The second reason is that there is no simple correlation between the population of physical people and that of SIM cards. All physical people do not own a telephone (such as very young children) and conversely some have several (especially for business reasons). So we have to consider the penetration level, i.e. the ratio of the number of telephones over the reference population (the population at 1st January of year $N-1$ published by Insee). For example, in 2007, the number of portable telephones per inhabitant estimated by Arcep was 85.6% across all of mainland France. It was 81.6% for the Rhône-Alpes region but only 66.0% in Franche-Comté. In two regions, Île-de-France and the PACA region, these levels were even higher than 100% (122.3% and 104.3%, respectively).[19] Part of these differences may be linked to the characteristics of the populations. For example, the CREDOC digital barometer shows large disparities based on age in 2007: Nearly all of 18-24 year olds were equipped with a telephone while this was only true for a third of the over-70s.[20]

Formally, converting the number of subscribers N_{HD_i} identified as residing in a given spatial unit i (from the home detection algorithm – HD – corresponding to the number of separate days, the most effective according to the results above) to the resident population in this unit is supplied by the following accounting operation:

$$\widehat{N_i} == \tau_i^{-1} \cdot \alpha_i^{-1} \cdot N_{HD_i} \qquad (2)$$

where α is the local market share of the operator Orange, and τ the penetration level. These two parameters are likely to vary throughout the country, because of the Orange market share but also because of the composition of

17. In principle, the more aggregated level is less interesting, as the interest aroused by the sources derived from mobile phone data is precisely in obtaining estimators with finer spatial granularity.

18. See Ruling 07-0706 from Arcep dated 6 September 2007, https://www.arcep.fr/uploads/tx_gsavis/07-0706.pdf

19. Arcep, Le Suivi des Indicateurs Mobiles *– Figures at 31 December 2007. https://www.arcep.fr/index.php?id=9545 "Geographical distribution of mainland customers".*

20. 2015 digital barometer, available at https://www.arcep.fr/uploads/tx_gspublication/CREDOC-Rapport-enquete-diffusion-TIC-France_CGE-ARCEP_nov2015.pdf Table 2 – Proportion of individuals having a mobile telephone, p. 24.

the resident population. To obtain local estimates of residential population from mobile phone data, we therefore want to have accurate information about the variables corresponding to the detrending (at least the operator's share and penetration level) at fine geographical levels. The difficulty here is that this information is generally available at an aggregated level (national or regional). Using it uniformly over the whole country creates the risk of not being able to distinguish between actual differences in population and different market shares for different administrative units.

To investigate (and quantify) the importance of different effects on our residential population estimates, we perform detrending while increasingly adding additional information to the equation. Found estimates can then be compared to those observed in the tax source in order to understand the contribution of different effects such as market share. A first, "raw" estimate simply consists of correcting for a size effect. We simply multiply the obtained subscriber counts from mobile phone data with the ratio of the number of subscribers available in the data to the size of the residential population in mainland France (i.e. 18 million out of a total mainland population of about 62 million in 2007). A second estimate can be based on incorporating open source information on the market share and penetration rate, as was done in the previous paragraph.

A third source of information that can be added is not open-source but was available to us. It consists of the customer file, which provides an estimate of the regional distribution of subscribers. We therefore use this file to construct a detrending by *département*. This geographical level appears both sufficiently broad to reduce the problems of spatial approximation raised from using the grid supplied by Voronoï polygons, and sufficiently fine to be able to ignore the spatial heterogeneity of market shares and the penetration level mobile phones amongst the population. The number of subscribers residing in *département* k is estimated from addresses available in the customer file. As these addresses are only available for part of the file of SIM cards we have, we adjust by the size of the customer file (which comes back to supposing that the customer file's lack of coverage is uniform over the whole country). So the *département* market share simply corresponds to the ratio of this estimate of the number of subscribers residing in the *département* over the total number of inhabitants in this *département* supplied by tax sources.

$$\alpha_k \tau_k = \frac{Tot_{HD}}{Tot_{CRM}} \cdot \frac{N_{CRM_k}}{N_{Insee_k}} \quad (3)$$

Where *k* represents the index for the *département.*

A fourth and final source of information is based on *Deville et al*. (2014) who suggest estimating the municipal population densities from equivalent mobile data and a model taking account of the "superlinear effect of densely-populated areas on human activities". We therefore use this method only as a comparison with the different ways of detrending that we are suggesting.[21]

The population is then estimated using the model:

$$N_{Insee_C} = \alpha \cdot N_{HD_C^\beta} \quad (4)$$

where the parameters α and β are themselves estimated by generalised linear regression N_{Insee_C} is the number of residents in the municipality according to the tax source and N_{HD_C} is the number of people identified as resident in the municipality with mobile data.

Performing this detrending on the subscriber counts obtained from mobile phone data, we can compare the obtained residential population estimates with the aggregated tax data at different spatial scales, investigating differences in magnitude and regional distribution.

The correlation between both is measured by means of two indicators: the cosine similarity and empirical correlation coefficient (Box 3). These indicators are both independent of the size of the population involved. They therefore amount to confirming whether the estimates from mobile phone data result in residential population densities consistent with those given by the tax source.

We have measured the differences at several scales. Clearly, we will use the Voronoï polygons, which is the finest spatial scale available with mobile phone data. Because Voronoï

21. The model proposed by Deville et al. *relates to population densities. The model is estimated by least squares weighted by the population of municipalities over the logarithms of densities. The interest of official statistics focuses more towards population counts. We therefore favour a model better suited to counts and we estimate the parameters by generalised linear regression based on a Poisson family (equation 4), on which a logarithmic link function is applied.*

Box 3 – **Cosine Similarity and Empirical Correlation Coefficient**

For each geographical level, we can define the vectors of observations using data taken from the tax source ($\vec{x}$) and mobile phone data ($\vec{y}$). We therefore call the empirical correlation coefficient:

$$cor\left(\vec{x},\vec{y}\right)=\frac{\overrightarrow{(x-\bar{x})\cdot(y-\bar{y})}}{\left\|\overrightarrow{(x-\bar{x})}\right\|\cdot\left\|\overrightarrow{(y-\bar{y})}\right\|} \quad (5)$$

Where $\bar{x}$ and $\bar{y}$ are the empirical means of the sample. It is also standard to use the cosine similarity, making it possible to measure if these two vectors are similar. Formally, this is the scalar product normalised to the product of the norms of the two vectors.

$$cosim\left(\vec{x},\vec{y}\right)=\cos(\theta)=\frac{\vec{x}\cdot\vec{y}}{\left\|\vec{x}\right\|\cdot\left\|\vec{y}\right\|} \quad (6)$$

This measurement is therefore independent of the norm of each vector. In principle it is more indicated for measuring densities, while the correlation coefficient provides information about the divergences on a like-for-like basis.

polygons do not naturally superimpose on statistical or administrative subdivisions, we will also investigate subdivisions by IRIS (first intra-community level), by community, by employment area and by *département*. Figure V represents the correlation and cosine similarity between the population estimate and the population derived from geo-referenced tax data, for each level of granularity.

We note that there are at least two reasons for finding differences between the results provided by the two sources. Firstly, the measuring concepts for the home are not the same (in one case, the information is derived directly from the tax residency declaration, in the other it is only obtained very indirectly from the subscriber's call behaviour). Secondly, one of the sources is complete while the other requires detrending with only a limited amount of additional information available to enable this detrending.

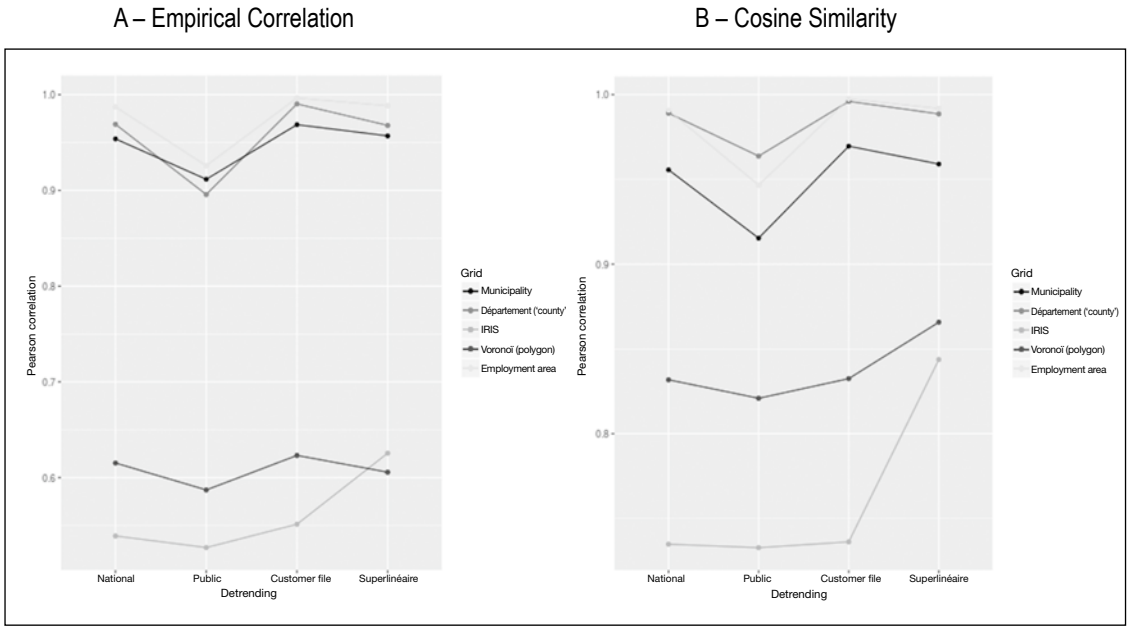
The results bring out significant divergences in the estimates obtained at very fine levels. The biggest divergences are observed in IRIS, the empirical correlation is 0.61. Using Voronoï polygons, the observations are closer compared to the IRIS grid, probably because the fact that it does not require resorting to a translation between grids, which removes one source of deviation.

The difference is smallest at the most aggregated levels. In practice, it corresponds to the accuracy of the home-detection algorithm, which can vary over the different *départements* (particularly because the cell towers are not distributed uniformly across the country). This detrending is founded exactly on the data supplied by the tax source by *département*, and it is not surprising that the estimates obtained are very similar. However, it is surprising to observe that the "loss" of accuracy by municipality is low compared to by *département*.

We have also tested the quality of our estimates for a statistical zoning that might be more relevant to mobile phone data: the zoning by employment areas. An employment area is a geographical space within which most of the working age people live and work (Aliaga, 2015). This zoning is constructed iteratively, with the aim of maximising the number of working age people who live and work in an area. In 2010, France had 322 employment areas that formed a complete partitioning of the country in similar surface areas, that hold the middle between municipalities and *départements*. Compared to other zonings, employment areas offer the best correlations between mobile phone data estimates and residential populations derived from tax data and irrespective of the detrending method. One probable explanation is that employment areas are suitable for studying the local labour market meaning that we assume most working age people that live in an employment area will also place the majority of caal in that area, at least during working days. While there is inaccuracy in the precise location of an individual's home, there is therefore a high chance that the home detection algorithms will place the individual's home in the right employment area. Our results suggest that employment areas are an appropriate geographical scale for analysing population estimates made using mobile phone data.

Figures V-A and V-B also allow comparison of the differences obtained depending on the available additional information: simple ratio of the number of subscribers, use of "public data" (the operator's national market share and regional penetration levels), use of

Figure V
Empirical Correlation and Cosine Similarity between Resident Population Estimates and the Tax Source Based on the Detrending Method and Aggregation Grid



Reading note: For employment areas, by detrending the estimates using the customer file, we find 0.99 correlation between the population estimated from mobile data and the tax resident population.
Sources: CDR, customer file for "f.client" detrending, Arcep 2007 data for "public" adjustment and Filosofi 2011; authors' calculations.

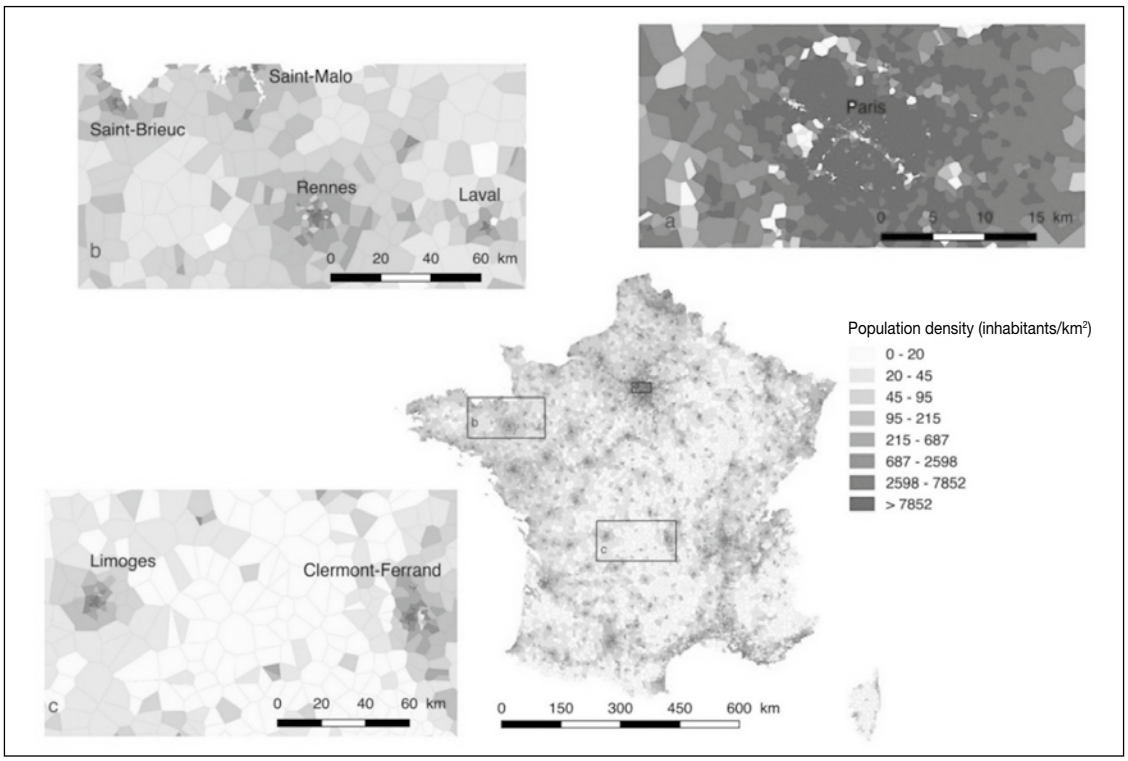
the customer file to detrend the population observed at *département* level and by estimation from the superlinear model proposed by Deville *et al*. The best estimates are obtained from customer file information. However, using additional information such as penetration levels rather tends to degrade estimates by comparison with a simple rule of three on the volume of subscribers normalised to the French mainland population. Using regional penetration levels, which can mask non-uniform intra-municipal behaviour, seems to contribute more noise instead of improving the accuracy of the estimate. Furthermore, the superlinear model estimated at national level does not yield better results in terms of the empirical correlation or cosine similarity than detrending by *départemental* market shares. It is by taking into account information about the representativeness of the operator's customers at an intermediate geographical level (the *département*) that we obtain the best results, even without considering potential non-linear effects but with simple local detrending.

Figures VI and VII provide a cartographic representation – beyond nationally-aggregated indicators – to compare the differences between population densities estimated using tax and mobile data (with *départemental* detrending by market shares). In addition, comparing these two sets of maps illustrates how many estimates based on municipality are closer to the reference than estimates at the scale of Voronoï polygons. In the close-up areas, especially around Paris, it is clear that the change of grid and aggregation by municipality or *arrondissement* provides information closer to the available references.

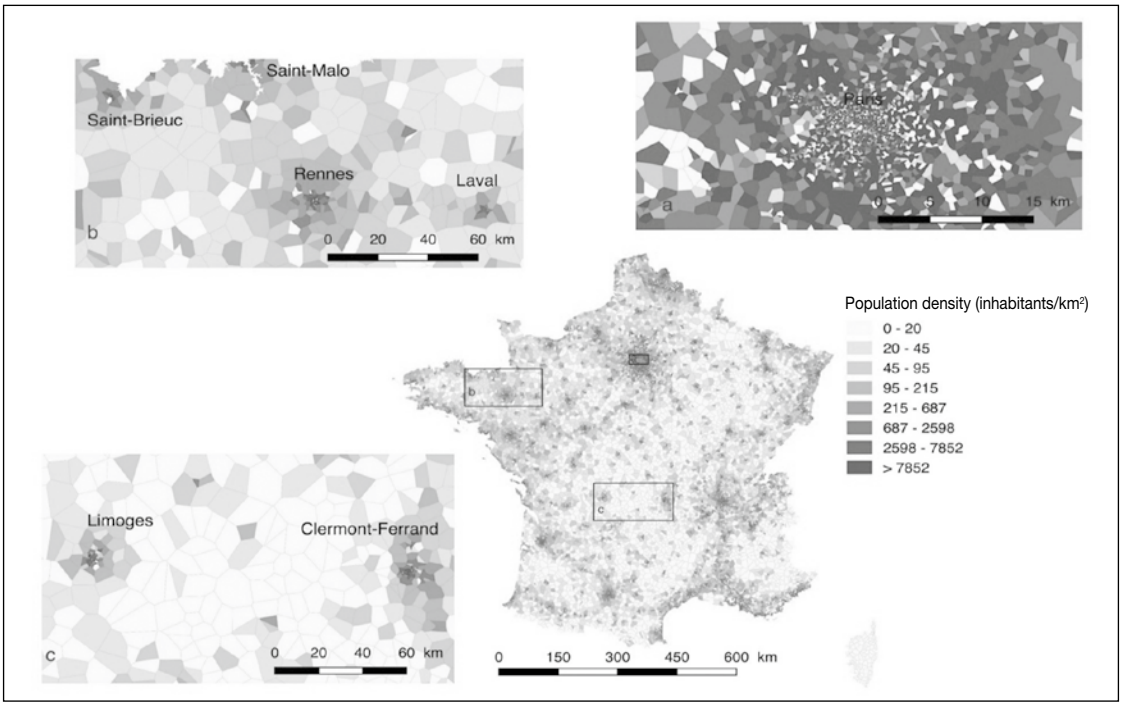
The map in Figure VIII shows the relative differences between the predictions made by municipality and detrended by *département* using the customer file, with municipal populations obtained from the tax file. The areas where the difference is highest roughly match the parts of the country where the spatial interpolation procedure creates the best outcomes (as illustrated in Figure III). We therefore remain dependent on the grid represented by Voronoï cells to produce a municipal estimate. The inaccuracy is remains highest in places where the assumption of uniform population distribution in Voronoï polygons has less chance of being confirmed (such as in areas with unevenly distributed dwellings over the region of the municipality). Sometimes the differences between estimate and reference are very large.

Figure VI
Population Density by Voronoï Polygon Calculated Using Tax Data and Mobile Phone Data

A – Tax Data



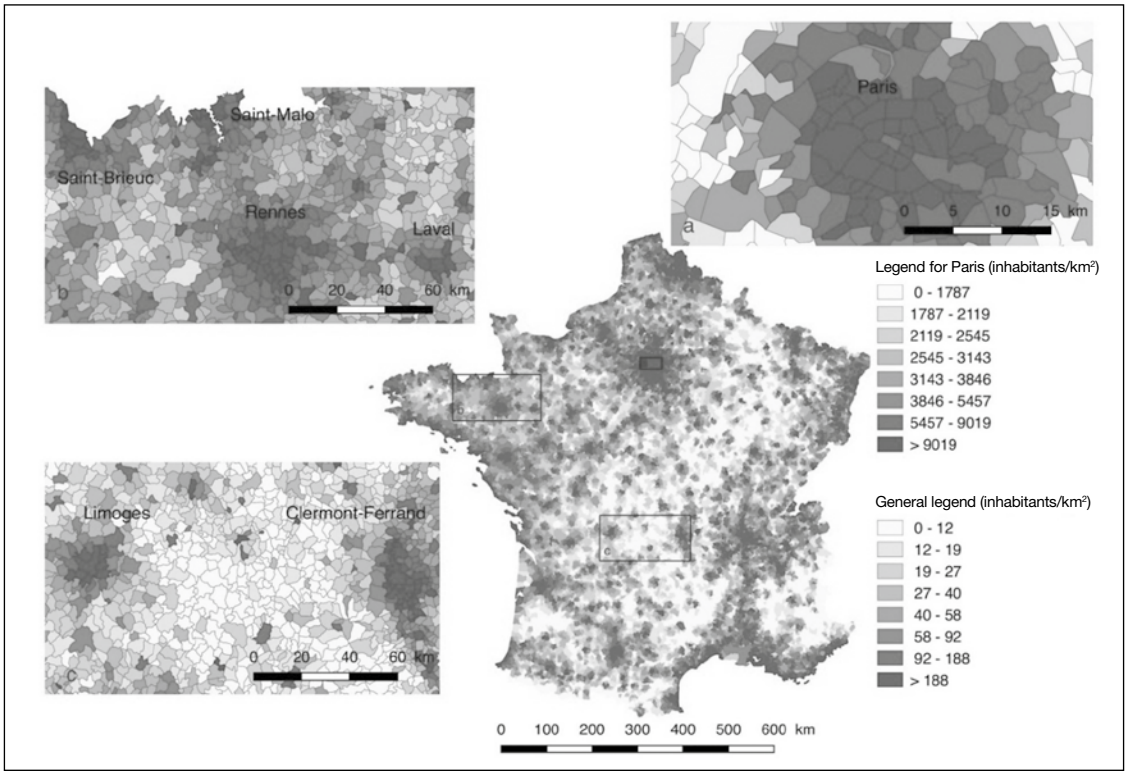
B – Mobile Phone Data



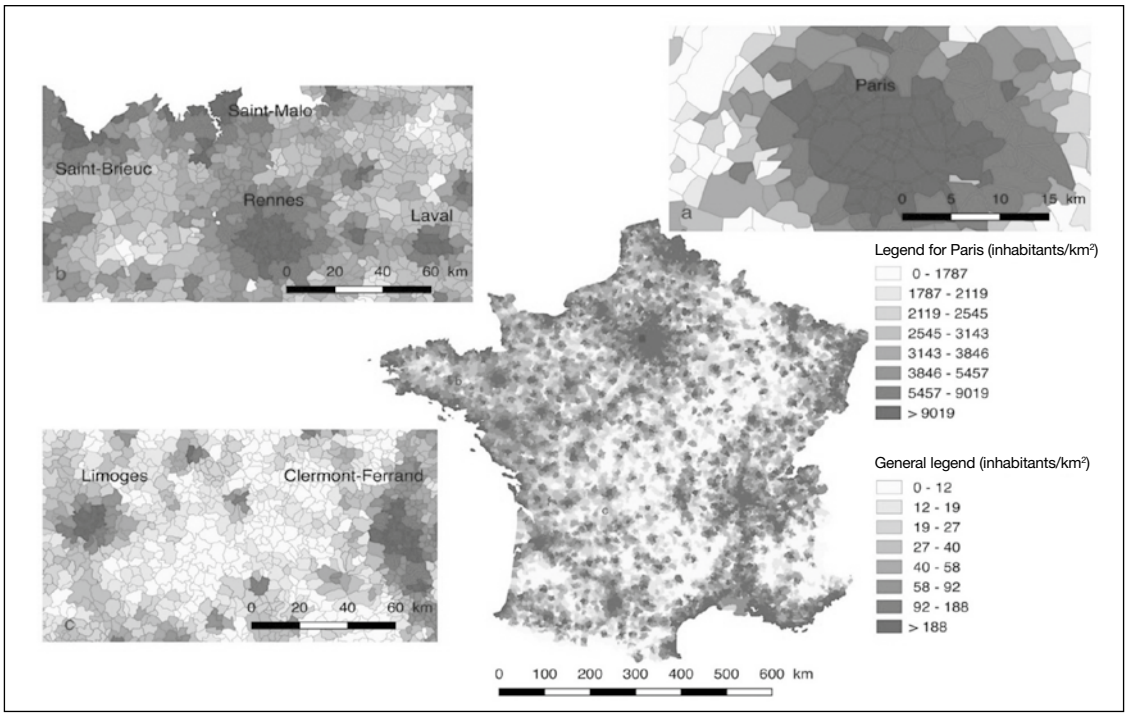
Note: The estimates are detrended by département using the customer file.
Sources: A, Filosofi; B, CDR, customer file and Filosofi; authors' calculations.

Figure VII
Population Density by Municipality Calculated Using Tax Data and Mobile Phone Data

A – Tax Data

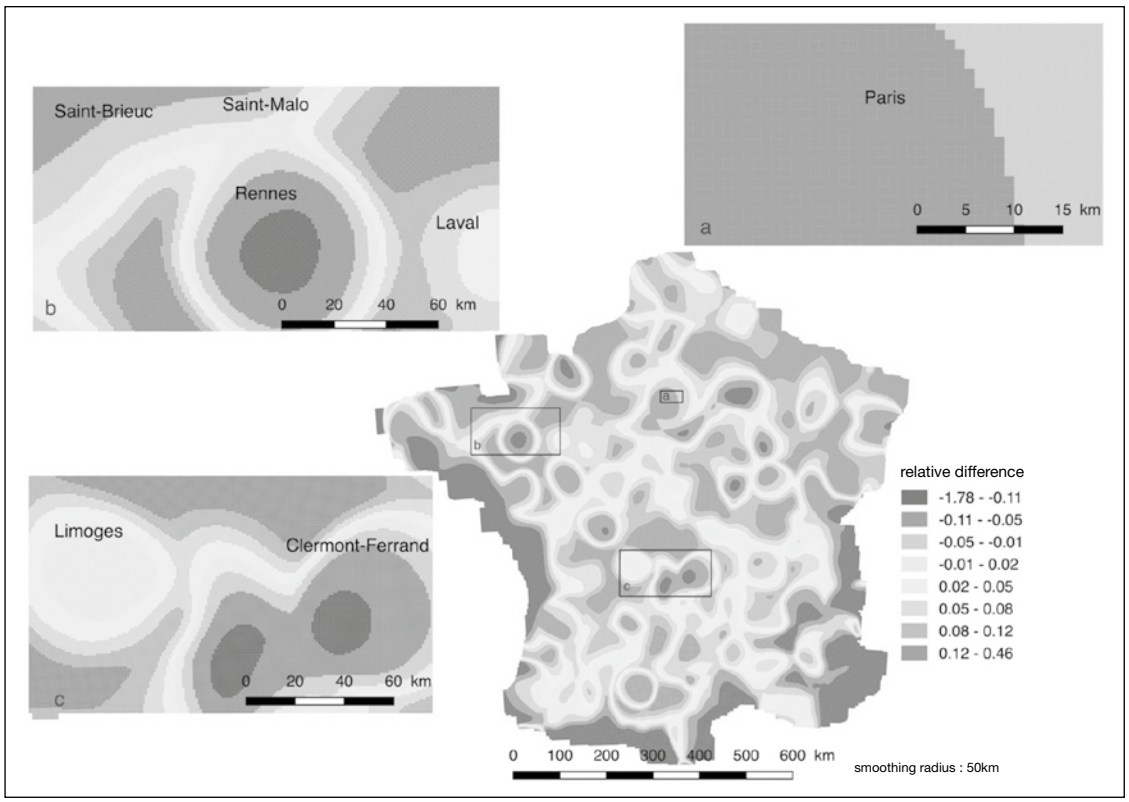


B – Mobile Phone Data



Note: The estimates are detrended by *département* using the customer file.
Sources: A, Filosofi; B, CDR, customer file and Filosofi ; authors' calculations.

Figure VIII
Map of the Relative Difference by Municipality between the Resident Population Estimate Detrended Using the Customer File and the Tax Source



Note: The differences are smoothed spatially for the representation.
Reading note: In the light grey areas the estimated populations are overestimated by a factor between 0.11 and 1.78; in the dark grey areas, it is underestimated by a factor between 0.12 and 0.46.
Sources: Orange 2007 Call Detail Records and customer file and Filosofi 2011; authors' calculations.

In some areas, the population of the municipality is underestimated by nearly half the municipal population, while in others it is overestimated by more than double (Figure VIII). These figures cover the estimates in section 2.3 on the interpolation cost in the tax source. This result is also confirmed by a more systematic analysis of the errors using statistical analysis (see Online complement C4).

Indicators such as the correlation coefficient or the cosine similarity do not take into account the spatial organisation of the points measured. However, it is plausible that the differences between the observed and predicted variables are spatially correlated, as illustrated by Figures III and VIII. For example, we may assume compensation phenomena between nearby municipalities, which are partly covered by the same cell towers and therefore by the same Voronoï polygons. Population estimates using Voronoï polygons will be distributed between these municipalities, which will create a correlation between the estimated values for these municipalities. Furthermore, as the error linked to using spatial interpolation is correlated to population density, it is likely that the differences will be similar for neighbouring municipalities. Spatial autocorrelation indicators such as Moran's *I* (Box 4) are an additional means of illustrating these phenomena.

We have calculated the value of Moran's *I* for four variables: the gross interpolation cost, the relative interpolation cost (compared to the number on inhabitants in the municipality), the gross difference and the relative difference. The four indices are significant, which confirms that these variables are not distributed randomly over the country and that there is indeed a spatial phenomenon involved.

Moran's spatial autocorrelation index for the gross interpolation cost is negative (and not significant). This is explained by the fact that when the subdivision into Voronoï polygons

Box 4 – **Moran's *I***

Spatial autocorrelation indices measure the spatial dependence between values of the same variable in different places in the space. The more the observation values are influenced by observation values that are geographically close to them, the greater the spatial correlation.

- Spatial autocorrelation is positive when similar values of the variable to be studied are grouped geographically.

- Spatial autocorrelation is negative when the dissimilar values of the variable to be studied come together geographically: nearby locations are more different than remote locations.

- In the absence of spatial autocorrelation, it can be considered that the spatial allocation of the observations is random.

The Moran index compares the way neighbouring observations co-vary with the covariance of all observations. The concept of neighbourhood is introduced using weights w_{ij} that take a value of 1 of observations y_i and y_j are similar, and 0 if not. The null hypothesis is a lack of spatial autocorrelation.

$$I_w = \frac{n}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij} (y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

$I_w > 0$ if there is positive spatial autocorrelation

leads to overestimating the population of a municipality, the population of neighbouring municipalities is underestimated, since the total population is constant. However, when the interpolation cost is normalised for the number of inhabitants, this index becomes positive – and very small, although it is significant (Table 2). Dividing by the size of the estimated population actually smooths the differences since the overestimates areas have their weight reduced relative to the underestimated areas. The gross differences and relative differences are positively correlated in space, a sign that some areas significantly concentrate municipalities having differences larger or smaller than the mean.

Table 1
Spatial Autocorrelation of Differences and the Interpolation Cost

Variable	Value of Moran's *I*
Gross difference	0.14***
Relative difference	0.13***
Gross interpolation cost	-0.11
Relative interpolation cost	0.009***

Note: *, **, *** indicate the significance at limits of 10%, 5% and 1%.

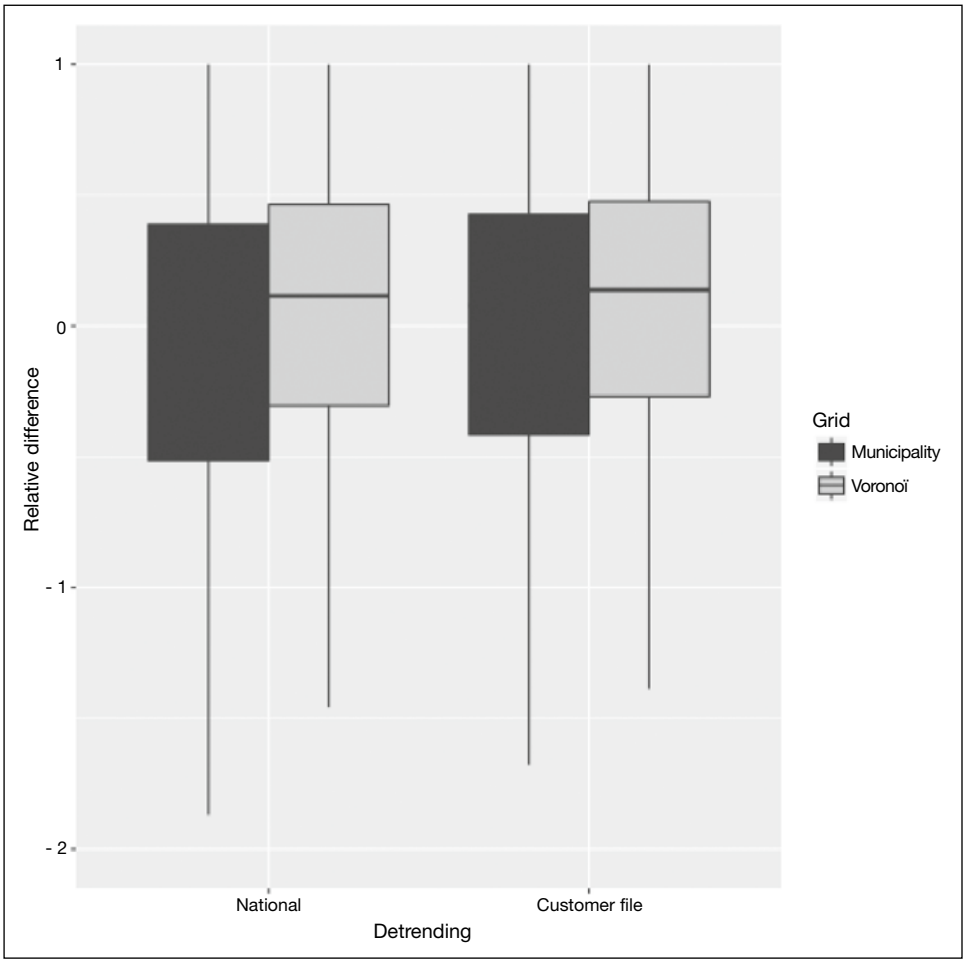
Finally, the distribution of municipal population differences, as represented in Figure IX, is narrower when detrending using the customer file by *département*. However, the median of this relative difference remains small, at the level of both the Voronoï cell and the municipality when detrending.

Using Temporal Granularity: Estimating Seasonal Variations

An important advantage of mobile phone data, other than spatial accuracy, is that it provides frequently captured data. In fact, in mobile phone data we have semi-continuous records about the presence of people, this is, when they use the network. This dimension was used indirectly in the previous estimates to identify subscribers' probable homes, but it was then used to estimate static values (the population). Using the dynamic aspects more directly can provide interesting information about the dynamic of the regions, such as variations in seasonal visitors. Such information could supplement the conventional indicators from official statistics, which only provide information about long-term changes in populations (supplied by censuses), with finer temporal information about tourist visitor numbers. Mobile phone records can be used to identify areas in which we observe large differences during the year, with great geographical precision. Looking at variations rather than at the absolute numbers of residential population estimates partly remedies the weaknesses highlighted by the previous analyses. In particular, knowing the local variability in market shares of the operator whose data we are using is less essential for investigating temporal trends then it is creating entire population estimates.

By way of illustration, we can focus on the summer months and for each month calculate the number of distinct subscribers identified in the mobile phone data for an area during

Figure IX
Distribution of Differences between Estimated Population and the Tax Source

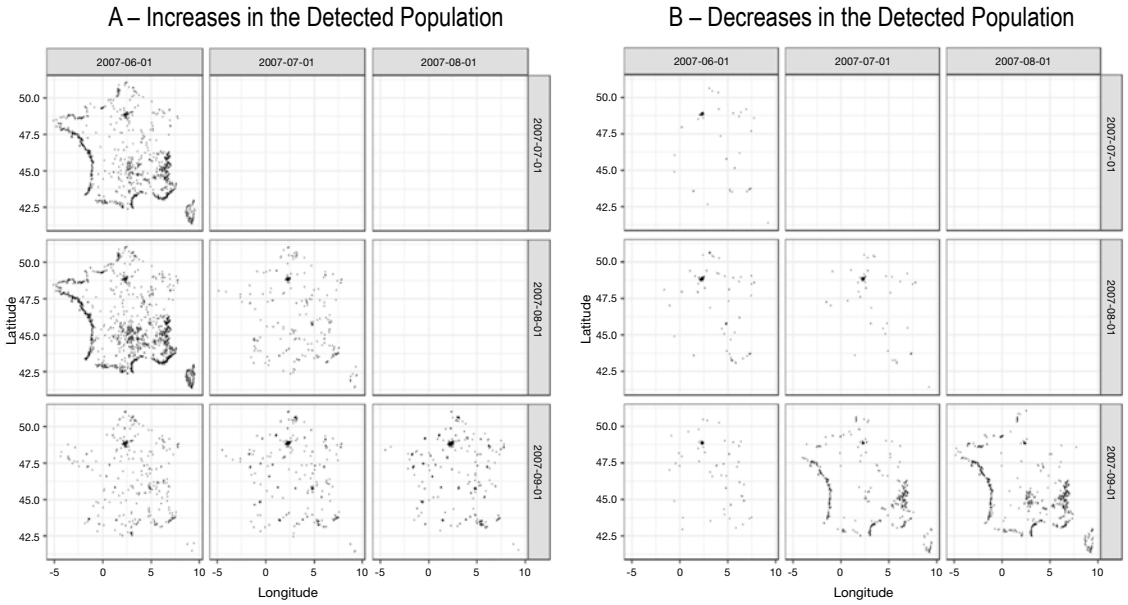


Note: For a clearer representation, outlying points are not shown. However, they represent a non-zero part of the population, for approximately 250 Voronoï cells where no tax resident is considered to live a total of nearly 60,000 users are estimated to live there.

a given month, normalised by the number of residents in previous months (here also using the most effective algorithm associated with the number of distinct days presence over a month). We work directly on the grid supplied using Voronoï polygons, to overcome difficulties linked to transposing the administrative subdivision described above. For each Voronoï cell we therefore have 6 variables corresponding to the ratios for July, August and September, normalised for the estimate of residents for June, July and August. Over all the Voronoï polygons, these variables are distributed to correspond approximately to a log-normal law centred around 1 – corresponding to a situation where the people present in a given month are identical to those identified as resident in the previous month. However, these differences can be very large, which results in a very thick tail to the distribution. To highlight the geographical distribution of these differences, Figure X shows the logarithm of these variables for the different months. To better bring out the large variations, we use different maps to show the areas where the changes are most marked. Figure X-A indicates locations where populations have increased by more than 50% between two months, and Figure X-B shows a map of the locations where they have decreased by more than 50%. The changes match the guess: in touristic areas (particularly coastal or mountain areas), we observe large population increases between June and July and between July and August, which disappear in September to return to a situation similar to that before the two holiday months.

In the rest of the country, changes are less pronounced. Still, we can also remark seasonal changes. For example, increases in populations outside the large urban centres can be observed during summer months and are reversed in September.

Figure X
Variation of the Population Present by Month



Reading note: Between June and August, the population detected as inhabitants around the dark cell towers more than doubled, essentially on the coast and in the mountains (see part A). In Online complement C4, the light blue points show cell towers around which the population fell by less than half.
Sources: CDR; authors' calculations.

* *
*

These initial analyses suggest that it would be difficult using mobile phone sources to reproduce accurate statistics for residential population counts, as produced by official statistics. This result is not surprising in itself, given the differences between the two sources (declared tax residency versus residency reconstructed by the mobile phone analyses). We may also mention the limits inherent to the "active" nature of the data used, the locations are frequent on average but not always very regular. The signalling data, which supply information about the location at a systematic frequency, may make it possible to identify homes better, for example. Even by limiting oneself to the CDR data, widespread use of unlimited text messaging packages (still not widespread in 2007) has increased their use – and therefore also the possibilities for locating subscribers more regularly. Furthermore, the availability of para-data on the coverage of cell towers seems crucial insofar as a major part of the differences found seems to come from the approximation made by modelling coverage areas using a Voronoï tessellation.

This rapid change of mobile phone usage raises a major issue for the use of this type of data by official statistics. The statistical indicators that it produces are based on clear and shared concepts – a measurement convention on the value we want to measure. To use the indicators over time, in principle it is necessary for the data (and what they relate to) to be consistent over time. A constant change of content, and the methods needed to deal with them, could complicate interpretation of the results. It therefore seems premature to target the publication of standardised indicators using mobile phone data. Furthermore, using data from a single operator raises important questions about the possibility of accessing the information needed for detrending, in particular regarding local market shares, a necessary condition for detrending at a fine level. Finally, unequal coverage of the country raises difficulties in reproducing precise analyses on grids that have meaning.

Despite these limits, records taken from mobile phones supply a rich raw material for structural studies, as they illuminate regional phenomena, by giving information about the behaviour of individuals or other variables beneficial to regional development. Thus Pucci *et al*. (2015) present an illustration of using this type of data to describe the practices and uses of urban space (in which the grid of cell towers is sufficiently small to enable

accurate analyses), and Aguilera *et al*. (2014) use them on performance measurements for urban transport networks (journey times, occupancy of trains, etc.). We can assume that these variables are less sensitive to the choice of operator and therefore that the detrending issues are raised less intensely. Galiana *et al*. (2018) were concerned with studying social and spatial segregation in urban units of Paris, Lyon and Marseille. By identifying a subscriber's probable home, and by characterising the district in which they live based on socio-economic characteristics supplied by Insee, we can calculate social segregation indicators, quantifying the tendency of people only to communicate with people living in a similar district to their own in terms of income level, and to assess if this behaviour is more or less marked depending on whether or not they live in a better-off district. This study also proposes to measure segregation in space and its change, which corresponds to the fact of meeting, during the day or the week, people coming from various districts, or conversely the fact of remaining confined to a circle similar to their own. □

Link to Online Complements: https://www.insee.fr/en/statistiques/fichier/3706217?-sommaire=3706269/505-506_Sakarovitch-de-Bellefon-Givord-Vanhoof_complement.pdf

BIBLIOGRAPHY

Aguiléra, V., Allio, S., Benezech, V., Combes, F. & Milion, C. (2014). Using cell phone data to measure quality of service and passenger flows of Paris transit system. *Transportation Research Part C: Emerging Technologies*, 43(2), 198–211. https://doi.org/10.1016/j.trc.2013.11.007

Ahas, R., Silm, S., Järv, O., Saluveer, E. & Tiru, M. (2010). Using Mobile Positioning Data to Model Locations Meaningful to Users of Mobile Phones. *Journal of Urban Technology*, 17(1), 3–27. https://doi.org/10.1080/10630731003597306

Ahas, R., Armoogum, J., Esko, S., Ilves, M., Karus, E., Madre, J.-L, Nurmi, O., Potier, F., Schmücker, D., Sonntag, U. & Tiru, M. (2014). *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics: Consolidated Report.* Luxembourg: Publications Office of the European Union. https://doi.org/10.2785/55051

Aliaga, C. (2015). Les zonages d'étude de l'Insee: une histoire des zonages supracommunaux définis à des fins statistiques. *Insee Méthodes*, 129. https://www.insee.fr/fr/information/2571258

ARCEP (2008). Le Suivi des Indicateurs Mobiles – les chiffres au 31 décembre 2007. https://archives.arcep.fr/index.php?id=9545&L=1

Blondel, V. D., Decuyper, A. & Krings, G. (2015). A survey of results on mobile phone datasets analysis. *EPJ Data Science*, 4(1), 1–55. https://doi.org/10.1140/epjds/s13688-015-0046-0

Bojic, I., Massaro, E., Belyi, A., Sobolevsky, S. & Ratti, C. (2015). Choosing the Right Home Location Definition Method for the given Dataset. In: Liu, T.-Y., Scollon C., Zhu W. (Eds.) *Social Informatics*. 7th *International Conference, SocInfo 2015, Beijing, China, December 9-12, 2015, Proceedings*, pp. 194–208. Springer International Publishing. https://doi.org/10.1007/978-3-319-27433-1_14

Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., De Meersman, F., Seynaeve, G., Wirthmann, A., Demunter, C., Reis, F. & Reuter, H. I. (2016). Big data et statistiques : un recensement tous les quarts d'heure…, *Carrefour de l'Economie*, 2016/10. https://economie.fgov.be/fr/file/801/download?token=Juj2pHbV

Debusschere, M., Sonck, J. & Skaliotis, M. (2016). Official statistics and mobile network operator partner up in Belgium, *The OECD Statistics Newsletter* N° 65, 11–14. https://issuu.com/oecd-stat-newsletter/docs/oecd-statistics-newsletter-11-2016?e=19272659/40981228

Demissie, M. G., Phithakkitnukoon, S., Sukhvibul, T., Antunes, F., Gomes, R. & Bento, C. (2016). Inferring Passenger Travel Demand to Improve Urban Mobility in Developing Countries Using Cell Phone Data: A Case Study of Senegal. *IEEE Transactions on Intelligent Transportation Systems*, 17(9), 2466–2478. https://doi.org/10.1109/TITS.2016.2521830

Deville, P., Linard, C., Martine, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D. & Tatem, A. J. (2014). Dynamic population mapping using mobile phone data, 111(45), 15888–15893. https://doi.org/10.1073/pnas.1408439111

DGINS (2013). Scheveningen Memorandum on Big Data and Official Statistics. https://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13

Desrosières, A. (2008). *Pour une sociologie historique de la quantification : L'Argument statistique I.* Paris : Presses des Mines. https://doi.org/10.4000/books.pressesmines.901

Galiana, L., Sakarovitch, B. & Smoreda, Z. (2018). *Ségrégation urbaine un éclairage par les données de téléphonie mobile*. Journées de méthodologie statistique de l'Insee, 12-14 juin 2018. http://jms-insee.fr/wp-content/uploads/S25_2_ACTEv2_GALIANA_JMS2018.pdf

Grauwin, S., Szell, M., Sobolevsky, S., Hövel, P., Simini, F., Vanhoof, M., Smoreda, Z., Barabási, A.-L. & Ratti, C. (2017). Identifying and modeling the structural discontinuities of human interactions. *Scientific Reports*, 7. https://doi.org/10.1038/srep46677

Grégoir, S., & Dupont, F. (2016). La réutilisation par le système statistique public des informations des entreprises. *Rapport du groupe de travail Insee-Cnis*. https://www.cnis.fr/wp-content/uploads/2017/10/RAP_2016_143_reutilisation_syst_stat_information_ets.pdf

Isaacman, S., Becker, R., Caceres, R., Kobourov, S., Martonosi, M., Rowland, J. & Varshavsky, A. (2011). Identifying Important Places in People's Lives from Cellular Network Data. In: Lyons, K., Hightower, J. & Huang, E. M. (Eds.), *Pervasive Computing*, vol. 6696, pp. 133–151. Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21726-5_9

Janzen, M., Vanhoof, M., Smoreda, Z. & Axhausen, K. W. (2018). Closer to the Total? Long-Distance Travel of French Mobile Phone Users. *Travel Behaviour and Society*, 11, 31–42. https://doi.org/10.1016/j.tbs.2017.12.001

Léonard, I., Sillard, P., Varlet, G. & Zoyem, J.-P. (2017). Données de caisses et ajustements qualité. Insee, *Document de travail* N° F1704. https://www.insee.fr/fr/statistiques/fichier/2912650/F1704.pdf

Montjoye, Y. A. (de), Hidalgo, C.A., Verleysen, M. & Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Science Report*, 3. https://doi.org/10.1038/srep01376

Pucci, P., Manfredini, F. & Tagliolato, P. (2015). Mobile Phone Data to Describe Urban Practices: An Overview in the Literature. In: *Mapping Urban Practices Through Mobile Phone Data,* pp. 13–35. Springer, Cham. https://doi.org/10.1007/978-3-319-14833-5_2

Ricciato, F., Widhalm, P., Craglia, M. & Pantisano, F. (2015). *Estimating Population Density Distribution from Network-based Mobile Phone Data*. Luxembourg: Publications Office. https://doi.org/10.2788/863905

Ricciato, F., Widhalm, P., Pantisano, F. & Craglia, F. (2017). Beyond the "single-operator, CDR-only" paradigm: An interoperable framework for mobile phone network data analyses and population density estimation. *Pervasive and Mobile Computing*, 35, pp. 65–82. https://doi.org/10.1016/j.pmcj.2016.04.009

Scholtus, S. (2015). Aantekeningen over het toewijzingsalgoritme voor Daytime Population. Statistics Netherlands, *Internal CBS note*.

Song, C., Qu, Z., Blumm, N. & Barabasi, A.-L., (2010). Limits of Predictability in Human Mobility. *Science* 327(5968), 1018–1021. https://doi.org/10.1126/science.1177170

Tennekes, M. (2015). Uitvoering toewijzings algoritme. Statistics Netherlands, *Internal CBS note*.

Tennekes, M. (2019). *R package for mobile location algorithms and tools: MobilePhoneESSnetBigData/mobloc*. R, Mobile Phone ESSnet Big Data. https://github.com/MobilePhoneESSnetBigData/mobloc (Original work published 2018)

Terrier, C. (2009). Distinguer la population présente de la population résidente. Insee, *Courrier des Statistiques* N° 128, 63–70. https://www.epsilon.insee.fr/jspui/bitstream/1/8564/1/cs128k.pdf

Toole, J. L., Ulm, M., González, M. C. & Bauer, D. (2012). *Inferring land use from mobile phone activity*. In: *Proceedings of the ACM SIGKDD International Workshop on Urban Computing - UrbComp '12* (p. 1). Beijing, China: ACM Press. https://doi.org/10.1145/2346496.2346498

Vanhoof, M., Combes, S., & de Bellefon, M.-P. (2017). Mining mobile phone data to detect urban

areas. In: *Proceedings of the Conference of the Italian Statistical Society*. Florence, Italy: Firenze University Press.
https://eprint.ncl.ac.uk/file_store/production/241585/32829DBE-235C-4902-A175-0A8A0BD-CAFD4.pdf

Vanhoof, M., Plotz, T. & Smoreda, Z. (2017). Geographical veracity of indicators derived from mobile phone data. In: *Netmob 2017 Book of abstracts*.
https://arxiv.org/abs/1809.09912

Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z. (2018). Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics. *Journal of Official Statistics*, 34(4), 935–960.
https://doi.org/10.2478/jos-2018-0046

Wesolowski, A., Eagle, N., Noor, A. M., Snow, R. W., & Buckee, C. O. (2013). The impact of biases in mobile phone ownership on estimates of human mobility. *Journal of The Royal Society Interface*, 10(81), 20120986–20120986.
https://doi.org/10.1098/rsif.2012.0986