AVOUAC Romain

ENSAE 2ème année

Stage long

Année scolaire 2018-2019

Estimating present population using mobile phone data

Insee

Montrouge

Maître de stage :

Benjamin Sakarovitch

03/09/18 - 01/03/19

Abstract

Due to their ubiquitous use in general population, mobile phone data appear as a high potential source for official statistics. They provide information on human behaviour with fine-level spatial and temporal resolution, allowing applications such as the measurement of present population. Yet, using mobile phone data for population estimation raises an important difficulty : when a phone event is recorded, only the location of the antenna to which the device has connected is available, leaving the exact geographical location of the user unknown. As a result, researchers using mobile phone data to study human behaviour generally model antennae coverage by a Voronoi tessellation. However, there are many downsides to relying on this approximation, which can substantially skew population estimation. In this study, we implement a bayesian model to improve the estimation of population. Prior information based on land use data is added, which allows the localization of phone users to depart from the Voronoi based one. The model is estimated using a dataset of phone events from around 18 million customers of a French major operator. Validation is performed by comparing population estimates to French localized tax data. Using multiple metrics, we provide evidence that the bayesian model substantially improves population counts as compared to previous estimates. Yet, we show that the gain in quality mainly occurs in rural and peri-urban areas, whereas densely populated areas remain poorly estimated. Overall, our results highlight the great potential of combining mobile phone data with sources of official statistics to complement traditional methods of population estimation.

Contents

1 Introduction

In recent years, there has been a growing interest in the use of "big data" sources in official statistics (Daas et al., 2015). The ever-growing number of sensors in electronic devices generate large amounts of individual digital tracks, both with a high frequency and at a fine spatial resolution. As a result, there is a high potential in combining these data with traditional official statistics sources, like surveys or administrative data (Cheung, 2012). Such combination could reduce delays in access to relevant information (e.g. short-term economic circumstances), enable the production of information at a more disaggregated level, and reduce data collection costs.

Due to their ubiquitous use in general population, data generated by mobile communication technologies appear as a high potential source for official statistics. For each mobile event, a timestamp and a location are

collected. The aggregation of all these events provides information on population distribution, with fine-level spatial and temporal resolution. Consequently, mobile phone data have been used with success in various applications. In particular, they appear as a major new data source for traditional fields of study of human and social behavior, such as mobility patterns (Calabrese et al., 2013), modes of transport (Demissie et al., 2016), and the analysis of social and spatial segregation (Galiana et al., 2018).

Due to their dynamic nature, these data provide the opportunity to proceed to present population counts (in a given venue, city, region...) whereas official statistics sources, notably census data, mostly produce information on resident population (Terrier, 2009). Yet, information on the evolution of present population evolution over time has potential use for a wide variety of actors. It provides knowledge on changes in population densities which are due to tourism (Vanhoof et al., 2017) and national or international migration flows (Hugues et al., 2016). It it used to study attendance patterns of particular areas, which are useful to analyze territorial dynamics (Toole et al., 2012). Such temporal variations also appear very important for local and regional actors, who seek to adapt supply of local services (transport, emergency, etc.) to demand.

Multiple challenges arise with the use of mobile phone data for official statistics. The most obvious one is the necessity to ensure confidentiality of collected data by avoiding possibility to identify individuals. Another difficulty is to proceed the huge amounts of data collected by mobile phone, which requires specific and robust IT infrastructures. In this study, we are tackling another important challenge : the geographical location of events. In most applications involving mobile phone data, only the location of the antenna to which a device has connected to is available, leaving the exact geographical location of mobile devices unknown (Ricciato et al., 2015). As a result, researchers working on population estimation using mobile phone data often rely on the Voronoi tessellation approximation (Okabe et al., 2009) : when a phone event is observed, it is assumed that the device connects to the nearest antenna (see for instance Deville et al., 2014 ; Vanhoof et al., 2018). However, there are many downsides to using this hypothesis to estimate the geographical location of events (Ricciato et al., 2017 ; Sakarovitch et al., 2019). These shortcomings have the potential to cause important bias in studies relying on this hypothesis.

In this paper, we implement a probabilistic method for the detection of devices location, based on a bayesian model developed by Tennekes (2018). The main advantage of this method is that it is highly flexible, in the sense that it can accomodate a wide range of situations as regards availability of technical information on antennae. In this study, we focus on the case mostly found in the literature in which no technical information on antennae are available. The bayesian framework makes it possible to map events observed at Voronoi level on a regular grid in a probabilized way. This transition enables us to make use of available prior information on the grid, which allows localization of phone users to depart from a purely Voronoi based one. We implement this model on an extensively studied data set consisting in call detail records (CDR) of French MNO Orange customers in 2007. Land use data is mobilized as prior information. Validation is performed by comparing population estimates to French localized tax data, from which we get the number of people

legally residing in each tile of the grid. Using multiple metrics, we provide evidence that the bayesian model substantially improves population estimates as compared to baseline. Besides, we use spatial analysis tools to demonstrate that the spatial structure of residential population is better replicated. Yet, we show that the gain in quality mainly occurs in rural and peri-urban areas, whereas densely populated areas remain poorly estimated. We discuss these limitations and explain how technical information on antennae as well was signalling data could be included in the bayesian model to further improve population counts.

2 A bayesian framework to improve geographical location of mobile phone events

In this section, we present the methodological framework on which our population estimates are based. First, we present the research problem : mapping the unknown location of mobile phone users. Then, we present the bayesian model used to perform this mapping in a probabilistic way. Specifically, we highlight one of the main advantage of this framework : it is highly flexible, in the sense that it can accommodate a wide range of situations as regards available information.

2.1 The spatial mapping problem

Each phone event, defined as a connection of a mobile device to a network, leaves a digital track. As such, these tracks represent a major opportunity to get a very fine view of spatial footprints of mobile phone users. However, these tracks generally do not contain information on the geographical location of the device. Data collected by mobile network operators (MNO) are primarily devoted to customers billing and network analysis. But these tasks only require the identification of the antenna that the mobile device has connected to. Although geographical coordinates of the antenna tower are known precisely, the exact location of the user is not. Consequently, when using mobile phone data to describe human behavior, the first step is to map events to a likely location.

Various methods can be used to proceed to this mapping. The main factor in choosing between them is the amount of available technical information on the MNO network. However, such information is rarely available, and may be considered as sensitive by MNOs when it is. Thus, many authors in the literature resort to a common approximation : coverage areas are modelled by a Voronoi tessellation. Each point of the space is mapped to its closest antenna, which produces a partition of the space in convex polygons (figure 1). Coverage of a given antenna is then assumed to correspond to the polygon in which it is located.

There are several downsides to relying on Voronoi tessellation to approximate the geographical location of phone users. First, many technical factors influence the choice of an antenna by a particular device, such

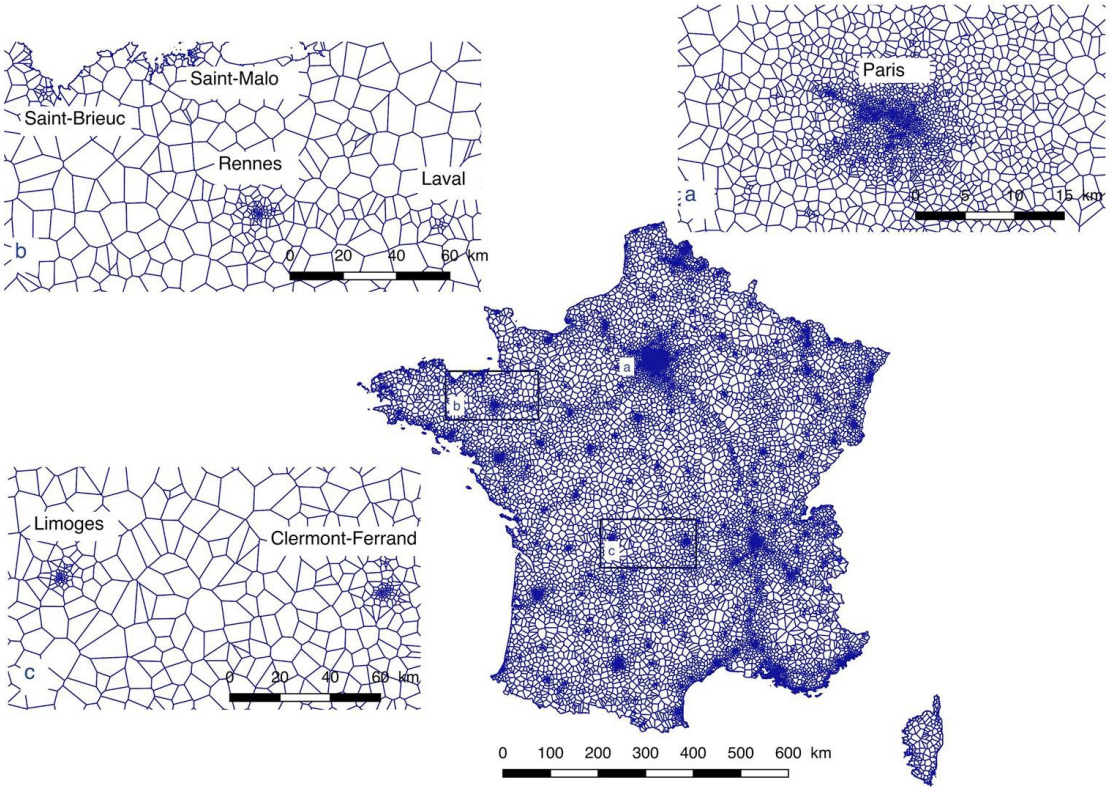


Figure 1: Voronoi tessellation of France based on the location of Orange antennae in 2007 (source : Sakarovitch et al., 2019)

as the type of the antenna (omnidirectional or directional), its coverage range, the generation of wireless technology used by the device (2G, 3G, 4G) and even its brand (different brands use different device-cell matching algorithms). As a result, it is often the case that a given phone connects to an antenna which is not the closest one. Besides, Voronoi tessellation constitutes a partition of the space, and thus fails to take into account the fact that antennae coverage areas actually overlap, so that one antenna can take over if another one is overloaded. In fact, mobile phones connections switch from one antenna to another even when standing still. These shortcomings explain why counting present population based on this hypothesis can generate significant bias.

Another important downside to using Voronoi tessellation is that the shape of its polygons are entirely dependent on the spatial distribution of antennae. This poses two problems. First, there is no reason for Voronoi polygons to match any of the spatial units used in official statistics. Yet, if we want to combine present population estimates with other sources, we have to be able to produce these statistics on a common spatial unit, such as Eurostat INSPIRE grid, French municipalities, etc. Second, as evident from figure 1, the size of the Voronoi polygons depends on the local density of antennae : in rural areas, antennae distribution is sparse and thus polygons are large, and conversely for urban areas. This is not directly an issue for present

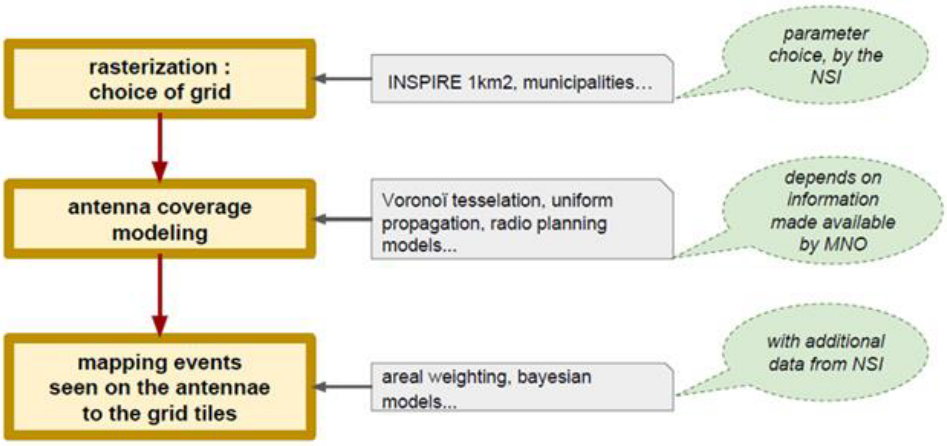


Figure 2: Spatial mapping procedure

population estimates, as this heterogeneity rightfully reflects the fact that estimates can't have the same precision everywhere. However, it can become problematic when these estimates are used in combination with other data sources, such as localized fiscal data to study social and spatial segregation, as it can cause modifiable areal unit problem (Openshaw, 1984) and so introduce bias.

Our research problem is thus to map events which are observed on the mobile network to a given input grid. Furthermore, we want to be able to incorporate any available information (MNOs data on antennae coverage, official statistics sources) that can improve the spatial accuracy of the mapping. Figure 2 sums up this procedure.

2.2 A bayesian model to probabilize mobile phone events location

All along this report, we use a localization method of mobile phone users based on a framework developed by Tennekes (2018). It fundamentally relies on Bayes' formula :

$$\mathbb{P}(tile_i|cell_j) \propto \mathbb{P}(tile_i)\mathbb{P}(cell_j|tile_i)$$

Each phone event is observed at cell level, i.e. the actual antenna transmitting the signal[1]. The problem is then to infer in which tile[2] of the grid the mobile phone that generated this record was located. We want this

[1]Up until now, we used the term "antenna" loosely. Yet it is important to distinguish the actual transmitter of the signal (the cell) and the antenna mast, on which multiple cells are generally located. When a phone event is observed, we know the specific cell to which the device has connected. Geographically speaking, this only gives us the coordinates of the mast on which it is located, next to some other cells. However, this is particularly important if we have access to information on cell technologies, because we can take it into account to compute probabilities based on theoretical coverage. In this report, we use the terms "antenna" and "cell" interchangeably, keeping in mind that geographical coordinates are those of the mast.

[2]Tiles are the spatial units composing a grid. In spatial analysis, they are often also referred to as "cells", but we choose to call them "tiles" so as to distinguish them from antennae cells.

location to be probabilized over the grid : as the only geographical information we have is the coordinates of the antenna to which the device has connected, we can never be certain of the actual location of the user. Bayes' rule states that this probability is proportional to any prior information we might have on each tile multiplied by the probability that the signal comes from a given cell knowing that the phone was on a particular grid tile. This latter probability — the likelihood — corresponds to the way we model antennae coverage. It should be as faithful as possible a reproduction of the actual coverage areas of antennae. The way we model it is thus entirely dependent on available information, from no information except antennae coordinates (Voronoi modelisation) to the most detailed information (best service areas, radio propagation models. . .).

This bayesian framework appears as a satisfactory solution to the research problem stated before. First, it can be computed on any given input grid. Second, it can accomodate a wide range of situations as regards information availability on antennae coverage from the MNO. Finally, various data sources — notably official statistics ones — can be mobilized to compute prior distributions on the chosen grid, so as to take into account the unequal probabilities of human presence on the French territory. In section 3, we provide an application of this model to a mobile phone dataset for which no information on antennae coverage was available.

2.3 Implementation as an R package : mobicountR

The analysis presented in this paper was implemented using R programming language. To improve reproducibility of results, all functions used were gathered inside an R package called `mobicountR`. A first group of functions consists in tools to build and manage the spatial units used in the analysis. A second group of function is devoted to prior distributions computation for the bayesian model using various statistical sources. A third group of functions serves to implement the bayesian framework, by computing posterior distribution over a given input grid. Finally, we provide a fourth group of functions devoted to diagnostic analysis and quality evaluation by comparison with a given validation data source. The documentation of the package is presented in appendix A.

3 Application to home detection in the absence of technical information on antennae

In this section, we present an application of the bayesian framework to resident population estimation using a 2007 dataset from French MNO Orange. Since no technical information on antennae are available for these data, we rely on the Voronoi approximation to model coverage areas, similarly to previous studies based on this dataset. The bayesian model is used to map events observed at Voronoi level on a regular grid in a

probabilized way. This interpolation enables us to make use of available prior information on grid tiles. We provide multiple evidence that the bayesian model substantially improves population estimates as compared to baseline from previous works.

3.1 Data

3.1.1 Mobile phone data

This analysis is based on a 2007 pseudonymised call detail records (CDR) dataset from major French MNO Orange. CDR keep tracks of mobile phone users activity, such as sending or receiving a call or a text message. As shown in table 1, we know for each event the unique identifiers of the caller and the callee, event type, duration, and the unique identifier of the cell. Five months of exhaustive activity of more than 18 million Orange customers is available, representing around 15 billion events.

As detailed below, we validate our population estimates using localized tax data. These data give information on legal residence of French households, and so characterize resident population. For the comparison to be relevant, we focus on home detection, i.e. predicting people likely place of residence. This is done by subsetting mobile phone data at two temporal levels. First, we limit the analysis to September 2007 data, so as to measure a concept of mensual residence. Besides, September is the month with less holiday of the time period (June-October) we have access to. Finally, another reason is purely technical : the probabilistic method of localization we implement is much more computationally intensive than previous methods in the literature, which relied on deterministic home detection. This restriction thus also serves to limit computational load. Similarly, we restrict daily tracks to the 19:00-09:00 time span. This choice is based on results presented by Vanhoof et al. (2018), who find that this particular time interval maximizes correlation between population estimates and validation data.

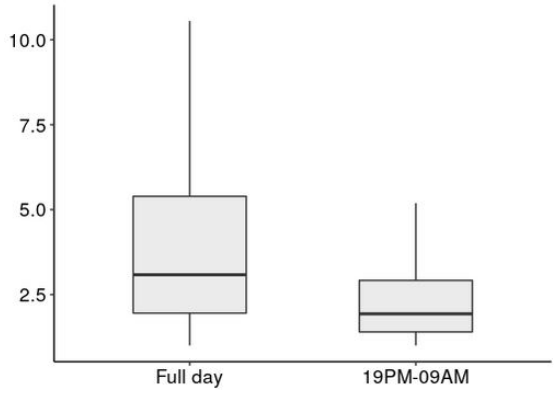
We shall already outline an important limitation of this dataset. In 2007, mobile phone use wasn't nearly as ubiquitous as it is today. Besides, as these data come from CDR, only users actions are recorded, but not automatic connections of mobile phones to the network — as opposed to signalling data. A direct result of these shortcomings is temporal scarcity of individual tracks (figure 3). Half of Orange customers use their phone at least three times a day, with a high heterogeneity between them in this regard. Yet, most users appear regularly in the dataset over the month, which is useful as we seek to analyse regularity of their location. Not surprisingly, these statistics drop when we restrict the time range to 19:00-09:00. Despite this obvious limitation, we are still able to observe most users at least 7 times during the month for an average of two events per night.

Table 1: Orange 2007 CDR : example

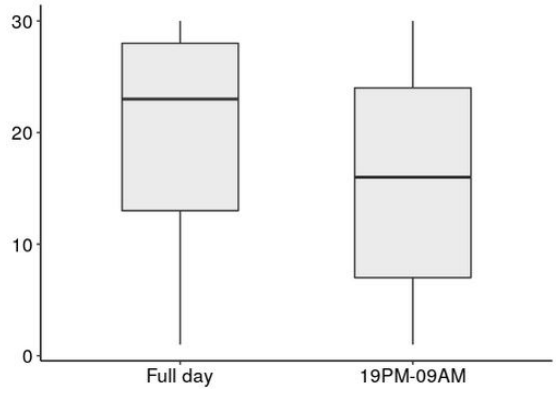
Timestamp	Caller	Callee	Event	Duration (sec.)	Area_id	Cell_id
2007/10/01 12:09:55	HJT22RR1	R482GH001	VO	365	1548	530012
2007/10/01 12:10:32	TR001BB	25GG2477	SMS	12	32110	255337

Figure 3: Orange 2007 CDR : summary statistics of users activity in September

(a) Average number of daily events per user



(b) Number of distinct days users appear



3.1.2 Prior information on grid tiles

Various data sources can be mobilized to add prior information on tiles of the grid. Relevant source choice for prior depends on the exact concept of present population we seek to measure. As we choose to focus on nightly present population, we want a prior which gives more weight in the localization of mobile phone users to residential areas. We can't choose geolocated tax data for this purpose as we already use it as validation dataset.

Against that background, we resort to the BD TOPO, a dataset on land cover and land use maintained by the French mapping agency (IGN). The BD TOPO provides a three-dimensional vectorial description of French territory elements (forest, hydrography...) and infrastructures (transport infrastructures, energy networks...). Notably, it provides an exhaustive vectorial description of buildings, classified in various types based on their use. As we seek to measure nightly present population, we focus on the "other building" type (*bâti indifférencié*), described as buildings with a minimal area of 20 m^2 which does not possess any attributed function. This type of building mainly consists in housing, but also encompasses public buildings which can't be classified in other types, such as schools, hospitals, museums, etc. In order to get the best approximation of the number of people residing in each tile of the grid, we take advantage of the presence of building heights in the dataset to compute a building volume per tile[3]. This volume is then normalized over the grid to get a

[3]Height is missing for about 10 % of buildings. We impute them using closest neighbour rule inside each department.

probability distribution which can be used as prior in the bayesian model.

3.1.3 Validation data

In order to be able to estimate potential quality gain of population counts due to the bayesian framework, we need to compare our estimates to a validation dataset. However, in most applications involving mobile phone data, no ground truth data are available. As a result, most authors in the literature resort to "high-level" validation data by comparing their estimates with population counts from census data (Deville et al., 2014). This solution is problematic in the sense that it doesn't enable the creation of performance metrics at individual level to optimize location algorithms. However, in absence of such ground truth data, "high-level" validation data still appear as a satisfying solution to validate population estimates.

Localized tax data can be useful as a validation tool for population estimates (Vanhoof et al., 2018). Here, we use Insee (French National Institute of Statistics and Economic Studies) localized tax income device (RFL). This dataset gives information on the fiscal income at residence level for more than 26 million French households. For our purpose, its main interest is that it is geocoded, so that we have access to the number of people in each household as well as geographic coordinates of their (tax) residence. Using these coordinates, we attribute each household of the dataset to the grid tile in which its residence is located. As a result, we obtain resident population counts over the same grid on which we map mobile phone events, which can then be uses as a benchmark for our population estimates.

One important limitation of this dataset is that we use 2011 income tax year, since it is the oldest available year for which tax data were exhaustive. We thus make the assumption that population patterns did not change much during this time span.

3.2 Methodology

3.2.1 Implementation of the bayesian framework in the absence of technical data on antennae

The Orange 2007 CDR dataset has been extensively studied in the litterature. As no technical information on antennae is available for these data, most of these works rely on the Voronoi tessellation approximation (Vanhoof et al., 2018 ; Galiana et al., 2018 ; Sakarovitch et al., 2019). Besides, home detection is generally performed in a deterministic way, in the sense that a single most likely place of residence is assigned to each phone user. This allocation is performed using various heuristics, such as maximum activity ("home is the cell tower where most activities of the user occurred during a specific observation period") or distinct days ("home is the cell tower where the maximum active days of a user were observed during a specific observation period") [see Vanhoof et al., 2018 for a detailed presentation of possible heuristics and their respective performance].

Table 2: Descriptive statistics on spatial units

	mean	s.d.	min	P10	P25	median	P75	P90	max
Number of tiles per Voronoi polygon	142	179	1	4	12	62	224	390	1,788
Number of Voronoi polygons per tile	1	0	1	1	1	1	1	2	28
Number of Voronoi polygons:	18,084								
Number of tiles:	2,166,965								

Since no technical information on antennae is available, we can't fully abstract ourselves from the Voronoi approximation : for each phone event, the only geographical information we have are coordinates of the antenna which processed the action. However, for the reasons explained in section 2, we want to be able to map events on a given input grid rather than Voronoi tessellation, which sizes only depend on the spatial distribution of antennae. Similarly to Galiana et al. (2018), we interpolate events from the Voronoi tessellation to a regular grid composed of tiles of 500x500 meters, covering metropolitan France. Such a grid offers a good compromise between fine spatial resolution of the mapping and computational load. Table 2 provides descriptive statistics on the different spatial units used and their relations.

The bayesian framework makes it possible to transition from Voronoi polygons $(v_j)_j$ to tiles $(t_i)_i$ of the grid in a probabilistic way, while at the same time incorporating prior information on the grid. We adjust the general framework presented in section 2 to this specific use case :

$$\mathbb{P}(t_i|v_j) \propto \mathbb{P}(t_i)\mathbb{P}(v_j|t_i)$$

Given that a phone event has been observed in a Voronoi polygon j — which is equivalent to say that it has been observed by the associated antenna (cell) j — we are interested in computing the posterior distribution $(\mathbb{P}(t_i|v_j))_i$ in order to map this event to the grid. As no information on antennae coverage areas is available, we use simple areal weighting to model the likelihood :

$$\mathbb{P}(v_j|t_i) = \frac{s(t_i \cap v_j)}{s(v_j)}$$

where $s(t_i \cap v_j)$ is the area of the intersection between tile i and Voronoi polygon j, and $s(v_j)$ is the total area of Voronoi polygon j. Any probability distribution over tiles $(\mathbb{P}(t_i))_i$ can be used as prior information.

At the level of a given individual, we thus transition from a track made of individual events $(v_{j_1}, ..., v_{j_m})$ observed with certainty at Voronoi level to a larger sequence of events $((c_{i_1}, ..., c_{i_n}), ..., (c_{i_1}, ..., c_{i_n}))$ observed at grid level with probabilities $((p_1^{j_1}, ..., p_I^{j_1}), ..., (p_1^{j_m}, ..., p_I^{j_m})$. These probabilities are normalized so that they

add up to one for each individual. Thus, by summing probability sequences of all individuals, we get present population counts at grid level, which add up to the number of distinct Orange customers in the 2007 dataset. A simplified example of this process is presented in figure 4.

The ability to add any available prior information on grid tiles is a distinctive feature of the bayesian model of localization. It proves especially useful when no technical information on antennae is available. Indeed, if we choose an uninformative prior equal to one for all observations, events are mapped on grid tiles based solely on their relative share in the Voronoi polygon which they intersect. In this case, the model basically boils down to an estimation directly at Voronoi level[4], as events are assumed to be uniformly distributed over the tiles wich intersect a given Voronoi polygon. Sakarovitch et al. (2019) show that this uniform interpolation — which they perform at city level — is actually the main cause of estimation error, with an almost one for one ratio. Since adding prior information is what fundamentally allows localization of phone users to depart from a purely Voronoi based one, it has important potential for mitigating such interpolation error.

3.2.2 Statistical adjustement of the sample

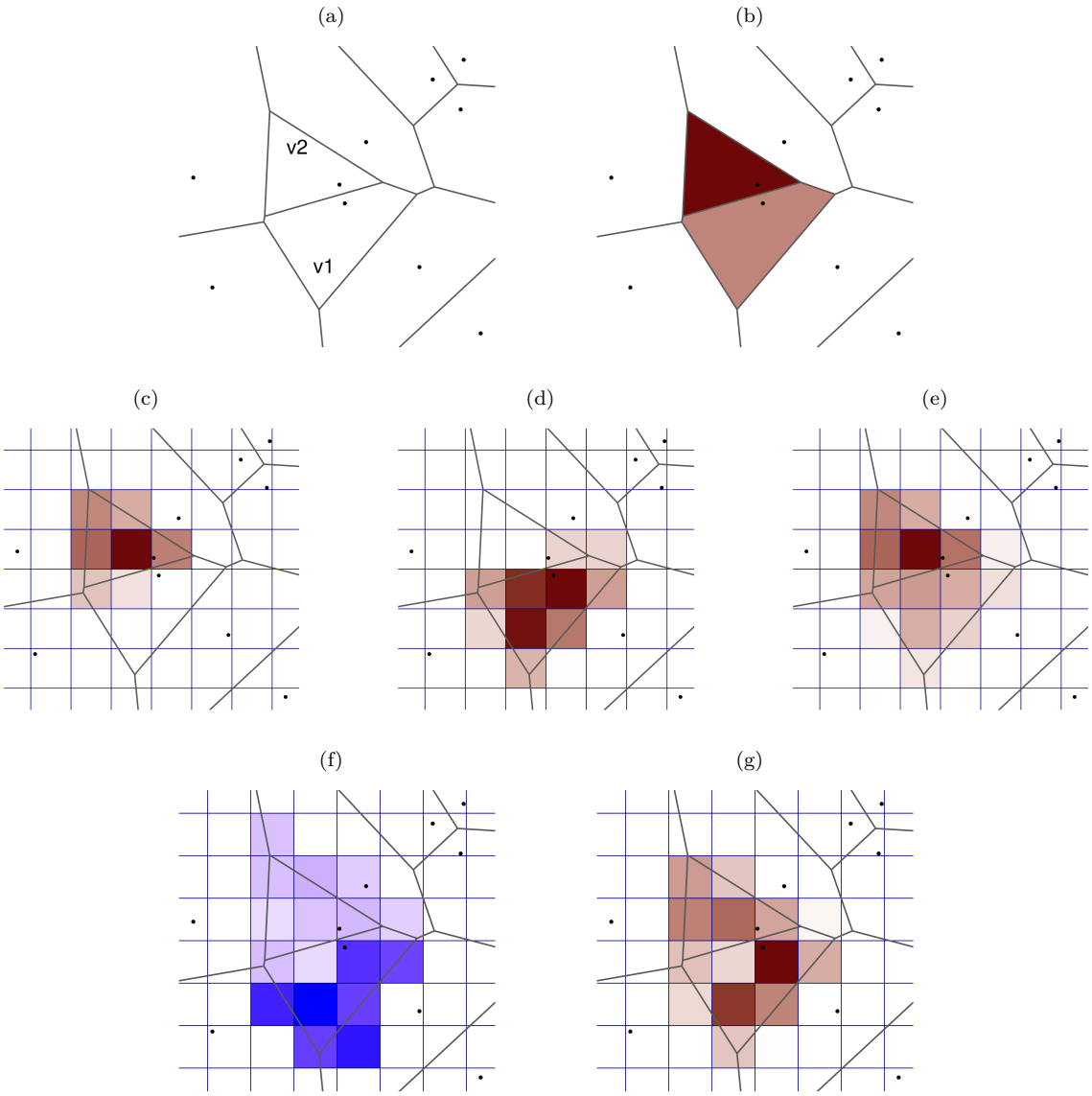
The final step of the detection process is to make resulting estimates comparable with general population. They can diverge for two main reasons. First, population estimates are computed using data from customers of a single MNO (Orange), who represent only a given share of the population. Second, we must take into account the penetration rate of mobile phones, i.e. the fact that some people don't own any phone whereas some people own several ones. Furthermore, there rates are known to be quite heterogeneous over the territory. To overcome these issues, we must adjust our sample in the following way :

$$\widehat{N}_i = \frac{N_{PHD_i}}{\tau_i.\alpha_i}$$

where N_{PHD_i} represents population count of spatial unit i resulting from the probabilistic home detection, τ_i the penetration rate in spatial unit i, and α_i Orange market share in spatial unit i. Similarly to Sakarovitch et al. (2019), we adjust our sample to make it representative of general population using Orange customer relationship management (CRM) in 2007 at department level. The number of Orange customers living in department k is estimated using customers addresses in this list. Since addresses are not available for all customers, we adjust by the size of this file. Finally, we obtain department k market share by dividing the estimated number of customers living there by the total number of people living in this departement

[4]Actually, some changes can occur for tiles which intersect two or more Voronoi polygons, since events are mapped on tiles based on their relative share in the polygon(s) they intersect. Figure 4 illustrates such a case. In practice, such edge effects produce limited change in population estimates at tile level for two reasons. First, most Voronoi polygons over the territory have substantial size and thus contain many tiles (see figure 4), which limits edge effects occurrence. Second, because population distributions generally exhibit high level of spatial autocorrelation (see section 3.2.4 for a detailed discussion over this concept). Thus, nearby Voronoi polygons tend to be similar in terms of population density, which limits the magnitude of edge effects.

Figure 4: Example of spatial mapping using the bayesian framework



Lecture : this plot illustrates home detection process for a given phone user with a fictive example. We focus on two specific Voronoi polygons v_1 and v_2 (a). Let's assume that, over the month, 2/3 of the nightly phone events of the user are located in v_2 — i.e. processed by the corresponding antenna — and 1/3 are located in v_1 (b). We allocate these events to the grid based on conditional probabilities $(\mathbb{P}(t_i|v_j))_{i,j}$, computed via Bayes' formula. If allocation is performed using only relative shares of tiles in the Voronoi polygon they intersect (see (c) and (d)) — i.e. if we use an uninformative prior equal to 1 for all tiles — then probabilistic home detection is given by (e). But let's assume that housing volumes are higher in tiles located at the South of the grid (f). In this case, probabilistic home detection is modified when the prior is incorporated (g).

according to tax data :

$$\widehat{\tau_k.\alpha_k} = \frac{\frac{Tot_{PHD}}{Tot_{CL}}.N_{CL_k}}{N_{TAX_k}}$$

These adjustments place our population estimates in a same order of magnitude as localized tax data, and thus enables to compare distributions over the territory.

3.2.3 Validation metrics

In order to evaluate the performance of the bayesian model of localization, we rely on several metrics to compare population estimates with validation data. First, we estimate Pearson's r correlation coefficient, which is the most used indicator in similar works (Deville et al., 2014 ; Vanhoof et al., 2018 ; Sakarovitch et al., 2019). For a vector of population estimates $\vec{x}$ and a vector of validation population data $\vec{y}$ of the same length n — the number of grid tiles — it is defined as :

$$r(\vec{x}, \vec{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Pearson's r indicates both direction and strength of the linear relationship between estimates and validation vectors. However, it is known to be quite sensitive to outliers in either or both compared distributions (Kim et al., 2015). Spatial distributions of population tend to be heavily skewed to the right because of high heterogeneity between rural and urban areas and are prone to heteroscedasticity issues, which makes comparison between correlation coefficients difficult (Wilcox, 2009). Furthermore, they generally exhibit high degree of spatial autocorrelation — i.e. spatially close observations are similar to each other — which requires specific testing procedure to avoid spurious correlation (Clifford et al., 1989). To deal with these limitations, we also estimate Spearman's ρ rank correlation coefficient. If we denote by $\overrightarrow{rg_x}$ and $\overrightarrow{rg_y}$ the respective rank ordering vectors of $\vec{x}$ and $\vec{y}$, then Spearman's ρ is obtained by computing Pearson's correlation coefficient over the rank vectors :

$$\rho(\vec{x}, \vec{y}) = r(\overrightarrow{rg_x}, \overrightarrow{rg_y})$$

Spearman's correlation test is less restrictive than Pearson's in the sense that it assesses any monotonic relationship between estimates and validation vectors, not only a linear one. This property makes it useful for comparison of spatial distributions of population, which often exhibit nonlinearities. Besides, the rank transformation makes it robust to the presence of outliers.

The statistical adjustment we make using Orange client list enables us to directly quantify differences between estimates and validation counts. Again, we use two different metrics to do so in order to take into accounts specificities of population distributions. First, we estimate the root mean squared error (RMSE) — also often called root mean squared deviation (RMSD). Using the same notations as above, it is defined formally as :

$$RMSE(\vec{x}, \vec{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2}$$

We also estimate the mean absolute error (MAE), defined as :

$$MAE(\vec{x}, \vec{y}) = \frac{1}{n} \sum_{i=1}^{n} |y_i - x_i|$$

These two measures are both frequently employed as accuracy measures of statistical models. They share many properties : they both express *average* error in units of the variable of interest, here population counts. Besides, they are both defined on $\mathbb{R}^+$ and negatively-oriented (the closer to 0, the better the quality of estimates). But there is also a very important difference between these two metrics, due to the way they are defined : taking the square root of errors before avering in the RMSE gives more weight to large errors of prediction. In our case, this can be either desirable or not : on the one hand we would like to penalize large under- and over- estimates of population, on the other the potential presence of very important outliers can significantly affect the metric. We thus choose to also rely on MAE, which is much less prone to this issue since all errors are weighted equally (see Willmott & Matsuura, 2005 and Chai & Draxler, 2014 for a detailed discussion, and a case for using and comparing both metrics in statistical applications in the latter).

3.2.4 Accouting for the spatial dimension of population repartition

The validation metrics we presented above summarize the statistical link between estimates and validation distributions, and enable an evaluation of mean estimation error. However, they are aspatial in nature, in the sense that they would give identical results for any random permutation of grid tiles on the map. Yet population repartition patterns are obviously far from being randomly distributed, with high population densities in cities and their suburbs, and conversely low population densities in rural areas. Furthemore, transitions from low density to high density areas tend to happen in a smooth way rather than abruptly. This property, which characterizes most spatial distributions, is called spatial autocorrelation : neighbouring spatial units tend to exhibit similar values of a given variable. Against that background, another important way to evaluate the quality gain of our estimates is thus to check whether they are able to reproduce the underlying spatial structure of resident population repartition. We resort to spatial analysis tools to explore this demension.

First, we can determine whether the distributions we analyze are characterized by the presence of spatial autocorrelation. Several indices have been developped for this purpose. Generally speaking, they consist in computing the correlation between spatially close observations of a given phenomenon. "Closeness" is assessed by a spatial weights matrix W, whose elements $(w_{ij})_{i,j}$ are such that $w_{ij} = 1$ if grid tiles i and j are contiguous[5]. If we denote the observed variable as $\vec{x}$, the general form of spatial autocorrelation indices is

[5]There are many other possible ways to define neighbourhood structure : based on a distance, taking a given number of nearest neighbours, etc. We choose to rely on contiguity for simplicity purpose : computing a contiguity matrix involves much less operations than computing distances between pairs of tiles. This is particularly important in our case due to the high number of grid tiles.

given by :

$$Corr(Y, WY) = \frac{Cov(Y, WY)}{\sqrt{V(Y)V(WY)}}$$

In this study, we choose to rely on Moran's I index (Moran, 1948), which is often preferred to other spatial autocorrelation indices in the literature beacause of greater stability (Upton et al., 1985). Keeping the same notations as above, it is defined as :

$$I_W = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \quad \text{for } i \neq j$$

Moran's I enables to test whether neighbouring observations co-vary relatively to the variance of the distribution. The null hypothesis is that neighbours do not co-vary in any particular way, in which case $I_W = 0$. A strictly positive Moran's I indicates positive spatial autocorrelation, meaning that similar values of the observed variable tend to cluster geographically. The test is performed using the z-score :

$$\frac{I_W - E(I_W)}{\sqrt{V(I_W)}} \overset{H_0}{\sim} \mathcal{N}(0, 1)$$

where :

$$E(I_W) = -\frac{1}{n-1}$$

and :

$$V(I_W) = \frac{n[(n^2 - 3n + 3)S_1 - nS_2 + 3S_0^2] - k[(n^2 - n)S_1 - 2nS_2 + 6S_0^2]}{(n-1)(n-2)(n-3)S_0^2}$$

with $S_0 = \sum_i \sum_j W_{ij}$, $S_1 = 2\sum_i \sum_j W_{ij}$, $S_2 = 4\sum_i W_{i.}^2$ and $k = \frac{\sum_i (x_i - \bar{x})^4 / n}{(\sum_i (x_i - \bar{x})^2 / n)^2}$.

Moran's I makes it possible to test for the presence of spatial autocorrelation in the distributions we analyse. In particular, it is useful to check whether the distributions of estimation errors exhibit spatial autocorrelation, which can highlight possible systematic bias of our model. Yet, Moran's I doesn't provide any information on the specific structure of autocorrelation when its presence is acknowledged. As such, it is thus not sufficient to determine which of our estimates distributions better replicate the actual spatial structure of resident population repartition. Surprisingly, the literature dealing with the problem of statistically comparing two maps is rather sparse, and no fully satisfactory solution emerges. Lee (2001) develops a bivariate spatial association measure by integrating Pearsons'r and Moran's I in a single statistic. However, this so-called L statistic has undesirable properties. First, the value of the L statistic between a given spatial distribution and itself is not equal to one but rather to a measure of the spatial autocorrelation level exhibited by the distribution, which renders difficult its interpretation. Second, the measure remains based on Pearson's r and thus can still be significantly altered when outliers are present. Levine et al. (2009) propose a statistical procedure for comparing outputs from ecological niche models. But it relies on the hypothesis that mean pixel difference scores have a normal distribution, which appears unlikely in our case due to the aforementioned peculiarities of population distributions. Finally, another possibility we explored was to use local spatial autocorrelation measures, such as Moran's local I (Anselin, 1995). These measures quantify the degree of

spatial autocorrelation exhibited by each spatial unit — in our case, each tile of the grid — based on its value and the values of its neighbours. Such an approach has been used by Wulder et al. (2007) to compare outputs from a forest growth model. An important shortcoming of these measures is that they are very computationally expensive to compute as they are not easily vectorized, due to their local nature. As a result, we could only compute them in areas of limited size.

Against these limitations of spatial analysis tools to compare the spatial structures of two distributions, we also provide an analysis inspired by fractal geometry, similar to Sémécurbe et al. (2018). Basically, it consists in comparing the way population is spatially distributed at several nested spatial resolutions. We consecutively aggregate our 500x500 meters grid into 1km, 2km, 4km, 8km and 16km grids respectively. For each population distribution (home detection with uniform prior, home detection with land use prior and localized tax data), we compute an entropy measure at each resolution to summarise the amount of information contained in the distribution. Then, again for each population distribution, we compute the differences in entropy between subsequent scales — these differences correspond to the fractal dimensions — which intuitively represent the amount of information required to be able to transition from the aggregated level to the disaggregated level. Equivalently, it quantifies the degree of randomness of the spatial structure of the distribution : a fractal dimension close to 2 indicates a uniform distribution, whereas a fractal dimension close to 0 characterizes a highly concentrated distribution. Many entropy measures can be used to perfom the analysis, which are all based on Rényi's general formula of entropy (Rényi, 1961).

$$H_\alpha(X) = \frac{1}{1-\alpha}\log\sum_i P(X = x_i)^\alpha$$

where X is a random variable representing population repartition over the grid, normalized so that$\sim\sum_i P(X = x_i) = 1$. In particular, the limiting value of H_α as $\alpha \to 1$ is the Shannon entropy (Shannon, 1948), widely used in information theory.

3.2.5 Computation of results

Let's briefly detail how the steps of the analysis are carried out in practice to produce the results presented below. Computation of priors and posterior distributions is performed using the R functions presented in the previous section. The output of this first step is a table which gives for each Voronoi-grid intersection the value of the posterior, for each available prior. This table is then used as an input for the home detection step. Because of the massive amount of data that CDR represent — several terabytes in total — this second step is performed using big data framework Spark on an HDFS (Hadoop Distributed File System — distributed storage platforms specifically designed for big data tasks) infrastructure, at the MNO office. The output of the second step is a table which gives for each prior estimated population counts over the grid. Finally, computation of statistics on this table are performed using R again, and spatial representations are produced using QGIS (an open source geographic information system (GIS) application optimized for viewing spatial

data), apart from kernel smoothing which is performed using the R package *btb*.

3.3 Results

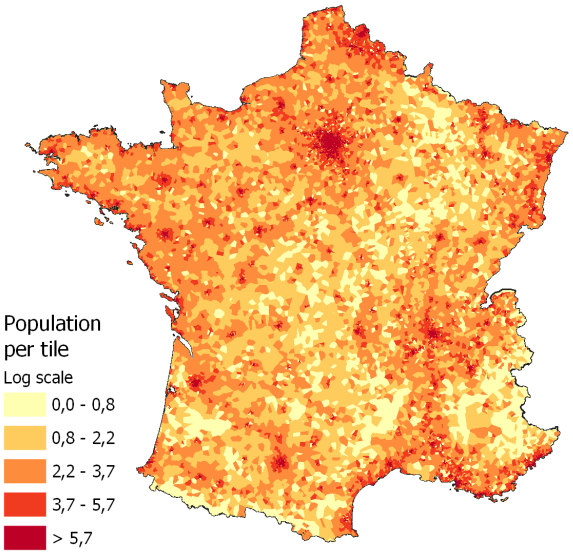
Two configurations of the bayesian model of phone users localization are estimated. In the first one, we use an uninformative prior equal to one. We call it "uniform prior" because it assumes that probabilities of a given phone event are uniformly distributed over the tiles which intersect the Voronoi polygon in which it is detected. This configuration is very close to what has been used in the literature using Orange 2007 CDR dataset (notably Galiana et al., 2018), and thus constitues our baseline estimates. In the second configuration, we estimate the bayesian model with a land use prior, which we also call "bdtopo prior" in reference to the statistical source from which it was computed. Below, we present several analyses which aim at quantifying the gain in accuracy of population counts due to the land use prior. To do so, we compare population distributions computed with each prior to the reference one, computed from localized tax data.

First, we plot the two estimated population distributions as well as the reference one over the grid at France level (maps (a), (b) and (c) of figure 5). These maps enable a direct visual comparison of spatial repartition of the population in each configuration. Since distributions are heavily right-tailed, we use log transformation for better representation. Tiles with no inhabitant are classified in the lowest density range. Map (a) shows that we are still highly reliant on the Voronoi approximation, as shapes of Voronoi polygons appear quite clearly[6]. It also highlights one of the main limitation of this hyptohesis : when comparing with reference (c), it appears that the population of many areas is overestimated. Most specifically, it seems that a lot of rural areas — characterized by large size Voronoi polygons — are slightly to severely overestimated. Urban areas appear too big in size as compared to reference, leading to visible overestimates in their suburbs. Comparatively, the distribution of population estimated using land use prior appears much more similar to the reference one. Adding the prior dramatically improves the replication of sparsity of population distribution over French territory. This first analysis can be confirmed by plotting in each tile difference to reference data regarding population counts, for each prior (maps (d) and (e)). Again, we observe that the main effect of adding the land use prior is to "clean" layers of population overestimation, particulary around important cities. However, at first glance, the land use prior doesn't seem to solve issues of either under- and over-estimation in large cities, particularly noticeable for Paris and its suburbs.

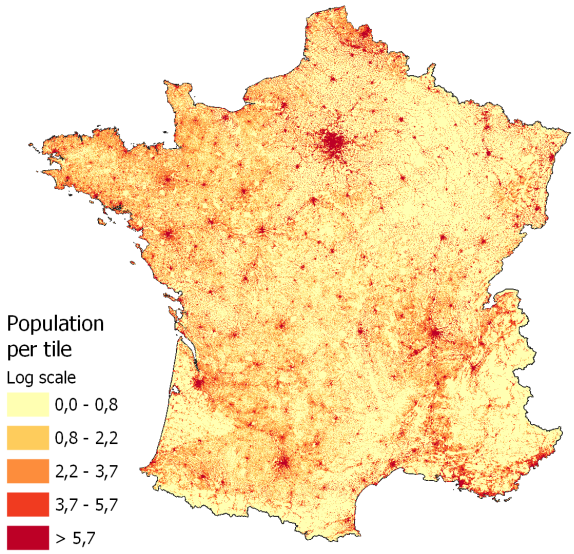
[6]We shall make emphasize that in map (a), we are not representing population counts per Voronoi polygon. Such a map would be semantically incorrect, since Voronoi polygons are heterogenous in size. Here, population counts are represented over (regular) grid tiles. The fact that Voronoi polygon shapes appear is a direct consequence of the Voronoi approximation we rely on for the estimation and the fact that using an uniform prior doesn't modify the distribution.

Figure 5: Population counts from probabilistic home detection (PHD) at France level

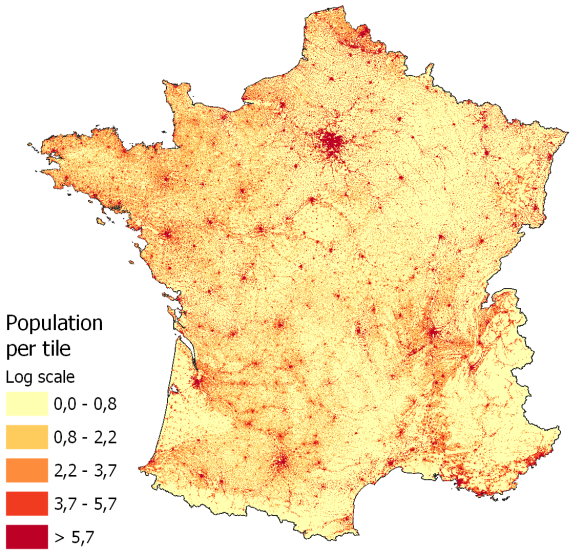
(a) PHD with uniform prior



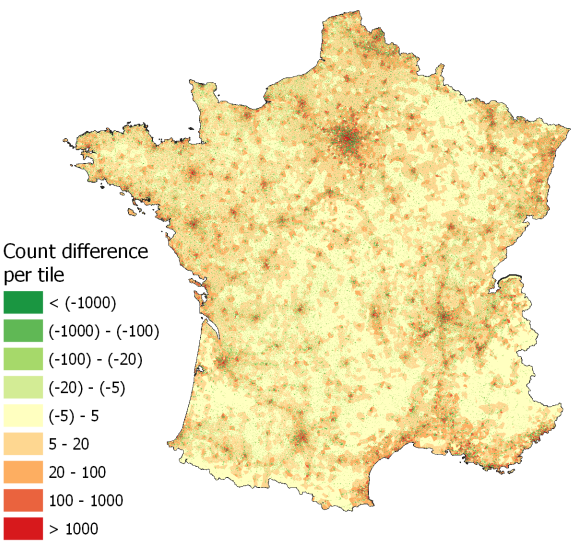
(b) PHD with land use prior



(c) Reference : localized tax data



(d) Difference between PHD with uniform prior and reference



(e) Difference between PHD with land use prior and reference

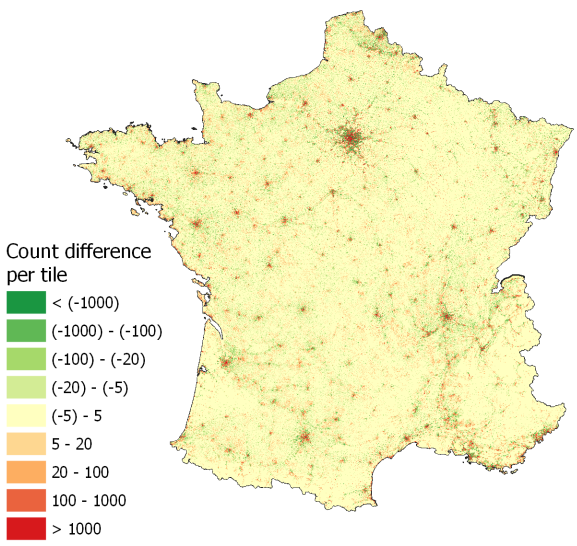


Table 3: Comparison for each prior of population counts with reference (tax data) using various metrics

Prior	Metric			
	Pearson's r	Spearman's ρ	Root mean squared error	Mean absolute error
Uniform	0.69	0.3	159.5	32.6
Bdtopo	0.79	0.87	149.7	16.8

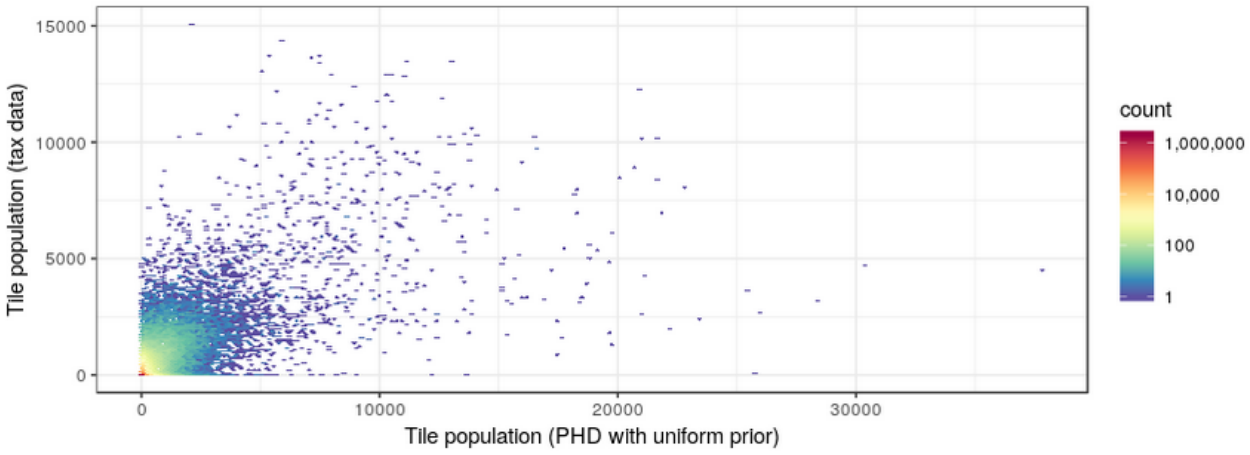
This graphical assessment of the quality gain of population estimates due to the land use prior can be confirmed by a statistical comparison of the two estimated distributions with the reference one (table 3). We use the various validation metrics presented in section 3.2.3. First, we compare Pearson's r linear correlation coefficient. Using this metric, the gain from using the land use prior appears rather small (ten percentage points), thus contrasting sharply with graphical comparison. However, as previously discussed, multiple issues arise when comparing correlation coefficients with distributions such as population counts. As evident from figure 6, population distributions are heavily concentrated around low values, but exhibit very long right tales. It is thus likely that outliers artificially increase correlation coefficients : very high values of population estimates co-occur with very high values of validation population. Furthermore, these distributions exhibit obvious heteroscedasticity[7]. Comparison of Spearman's ρ coefficients appears to confirm this analysis. The coefficient with uniform prior is seriously reduced (39 percentage points) while the coefficient with land use prior improves (8 percentage points). Correlation analysis performed using Spearman's coefficient thus suggests a dramatic improvement of the shape of population estimates distribution due to the land use prior which was not captured using Pearson's coefficient.

We come to similar conclusions when we assess average estimation errors relatively to validation data. Using root mean squared error (RMSE), we find that including the land use prior reduces average error by only 6 %. However, using mean absolute error (MAE) gives very different results. First, errors of estimation are much less important (PHD misallocates between 16 and 32 people by tile on average, against 150 to 160 using RMSE). Besides, the improvement due to the land use prior is much more pronounced : the average error rate is halved. As discussed before, this difference in conclusions between the two metrics is likely due to the way they are defined : RMSE gives much more weight to important errors, whereas MAE weights all errors equally. This discrepancy thus tends to confirm the analysis we made by comparing maps of distributions in figure 5 : the land use prior removes small to moderates errors, but it doesn't solve issues of important under- and over-estimation which are observed in big cities and their suburbs (also clearly visible in figure 6).

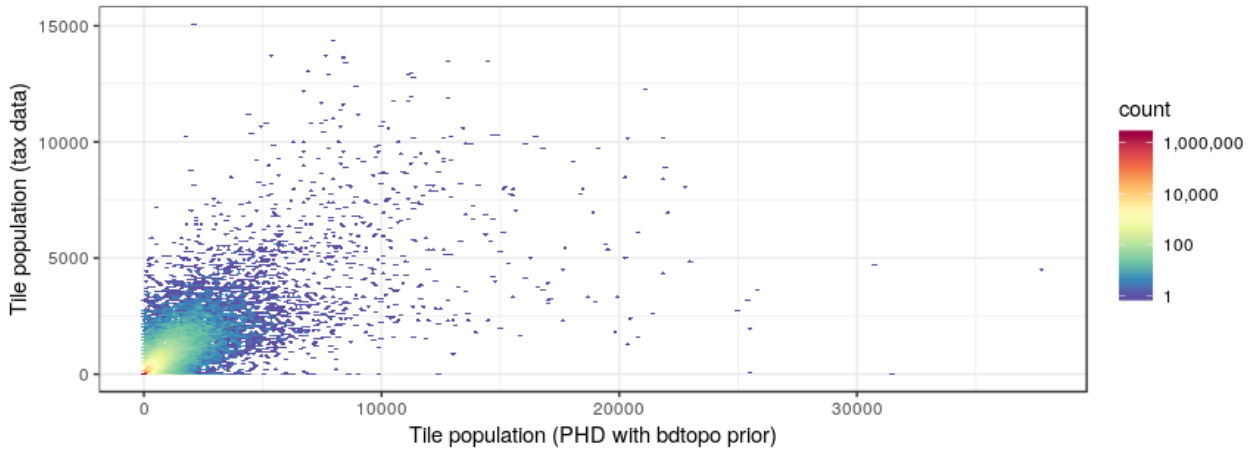
[7] A method which is often used to mitigate the effects of outliers and heteroscedasticity in such distributions is to apply log transformation to variables of interest. Here, this method is problematic since population distributions over the French territory exhibit high proportion of uninhabited tiles whereas the log function is not defined in zero. We use it for representation purpose in figure 5 by classifying uninhabited tiles with lowest density tiles. However, it wouldn't be rigorous to perform correlation analysis using this simplification.

Figure 6: Scatter plots of tax data population counts against probabilistic home detection (PHD) estimates

(a) PHD with uniform prior



(b) PHD with land use prior

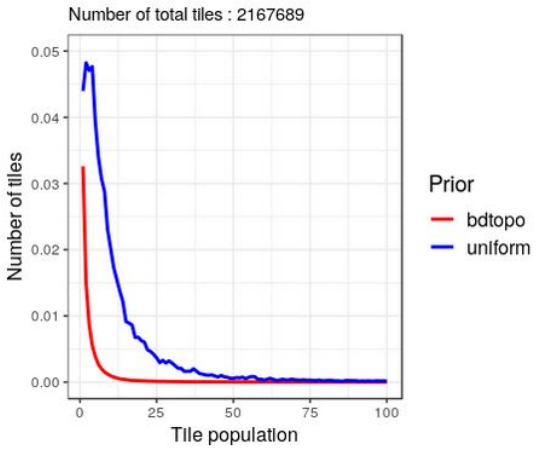


This last claim is further confirmed when we compare estimates with each prior depending on tile's "true" population according to validation data. 50 % of empty tiles are rightly declared as not populated with land use prior, against only 2 % with uniform prior. In figure 7, we extend this analysis by comparing misallocations with each prior, depending on tile's population. Graph (a) confirms that the land use prior greatly reduces the number of grid tiles in which some people are detected whereas there should be none according to tax data. Conversely, graph (b) points out that the land use prior tends to result in too many tiles declared as empty whereas they shouldn't be. However, this latter effect is of very limited magnitude, which strongly supports the use of land use prior in the bayesian model of localization.

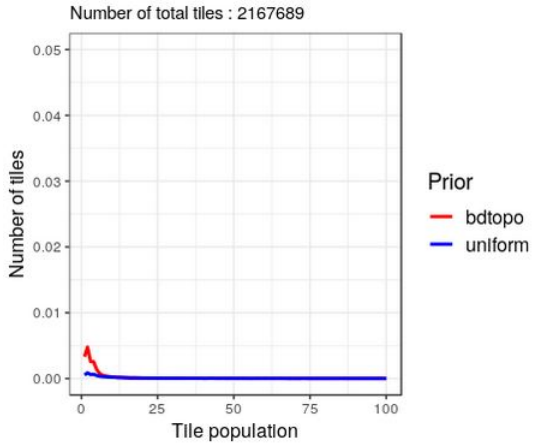
Statistical analysis substantiates the improvement in population estimates quality due to the land use prior. But apart from the visual comparison of maps done with figure 5, we didn't take into account the spatial dimension in the analysis. Indeed, all the results presented above would hold true for any random permutation

Figure 7: Advantage and shortcoming of land use prior for probabilistic home detection (PHD)

(a) Share of grid tiles wrongly declared as populated by PHD



(b) Share of grid tiles wrongly declared as empty by PHD



Lecture : the first graph shows the share of grid tiles (y-axis) in which probabilistic home detection (PHD) allocates n people (x-axis) whereas there should be none according to reference (tax data). The second graph shows the share of grid tiles in which PHD allocates no people whereas there should be n people according to reference.

of the grid on the map. However, spatial autocorrelation — or lack thereof — is a major property of spatial distributions, which we must take into account for comparisons to be relevant. We thus resort to spatial analysis tools in order to describe this dimension of population counts.

Table 4 displays the results of the testing procedure for the presence of spatial autocorrelation, based on global Moran's I statistic. All three population distributions, the two estimated ones and the validation one, exhibit significant and substantial levels of spatial autocorrelation. Besides, the ordering of the coefficients appear consistent. PHD with uniform prior exhibits the highest level of autocorrelation : since population is distributed uniformly in each Voronoi polygon, most neighbouring tiles display the same population count. Conversely, PHD with land use prior exhibits the lowest level of autocorrelation. This might be explained by its tendency to declare too much tiles as having no building volume (see figure 7), which mechanically set them as empty, whatever the number of phone calls detected in the corresponding Voronoi polygon. We also find that estimation errors, computed as absolute difference between home detection with each prior and localized tax data, are characterized by significant spatial autocorrelation. This result is consistent with the findings obtained by visual inspection of maps (d) and (e) of figure 5 : with both priors, there are evidence of substantial clustering of overestimation in urban centers and of either over- or under-estimation in city suburbs. More surprising is the fact that they exhibit a similar level of clustering : for the same reasons as explained above, errors tend to be the same for whole Voronoi polygons when using uniform prior, so that we could have expected higher spatial autocorrelation. However, the main effect of incorporating the land use

Table 4: Spatial autocorrelation of population distributions and estimation errors

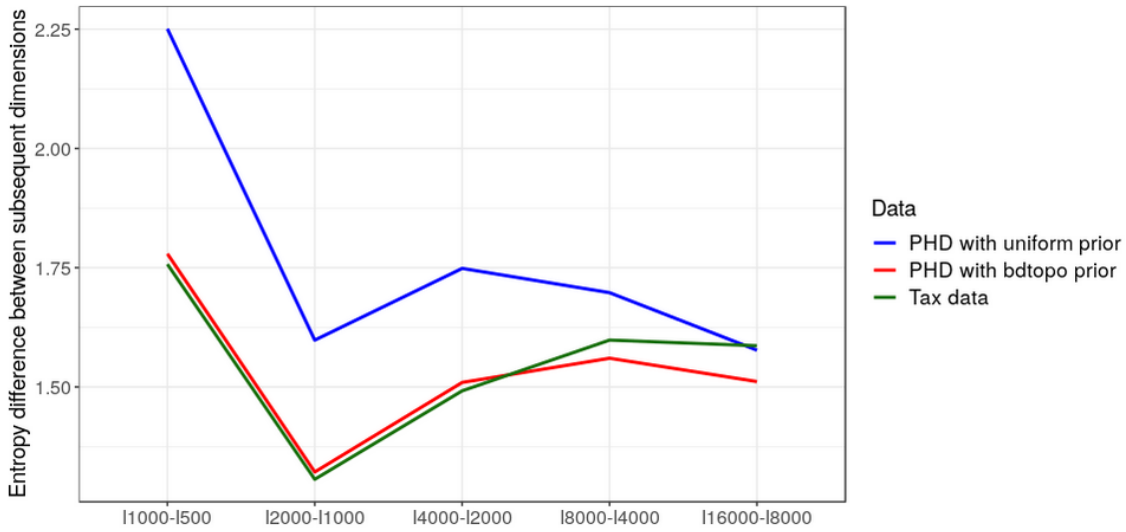
Variable	Moran's I	p-value
PHD with uniform prior	0.78	0.00
PHD with land use prior	0.69	0.00
Local tax data	0.74	0.00
Estimation error with uniform prior	0.37	0.00
Estimation error with land use prior	0.36	0.00

prior is precisely to remove this type of errors in rural and peri-urban areas. We thus end up with a similar level of clustering, since tiles with no more error are mechanically grouped together.

While spatial analysis tools are useful to detect the presence of autocorrelation in the distributions, it is yet not sufficient to determine whether we better reproduce the underlying spatial structure of resident population repartition. We thus propose a method inspired by fractal analysis to try to quantify the additional information on the spatial structure of population provided by the use of the bdtopo prior (figure 8). The comparison of entropy evolution distributions suggests that adding the land use prior substantially improves the replication of changes in information over scales, as evident from the proximity of the curves corresponding to localized tax data and PHD with the bdtopo prior. Besides, the ordering of the curves appears coherent : the values of information differences over scales using the uniform prior are consistently above the two others, confirming the much higher uniformity of this distribution observed in figure 5. Not surprisingly, the distributions converge for the higher scales : at these levels of aggregation, differences in population counts between spatial units are much less heterogeneous, so that no distribution stands out in terms of information structure.

Another way to take into acount the spatial dimension of population repartition is to simply decline the statistical analysis at disaggregated level so as to highlight spatial disparities in quality of home detection. Since the statistical adjustment of the sample was done at department level, it makes sense to replicate the analysis at that level. In figure 9, we plotted MAE and RMSE relative evolutions between PHD with uniform and land use prior for each French department. Again, important differences between the two metrics are observed, due to the way errors are weighted in each case. Although fine interpretation of divergence between the two metrics proves difficult, a particular pattern emerges : important quality improvements of estimates with the land use prior seems to be very often localized in rural areas, whereas departments where major French cities are located appear to benefit the less from the prior addition. This pattern is consistent with our previous results, which show that the land use prior doesn't solve substantial under- and over-estimation which happen in important cities and their suburbs.

Figure 8: Effect of scale change on Shannon's entropy measure

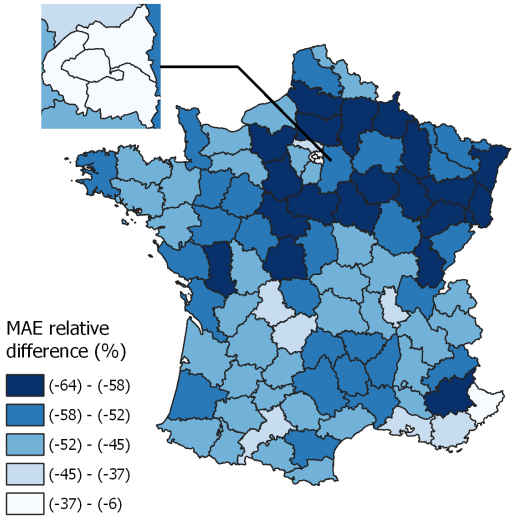


As an illustration of this finding, we focus on two examples at opposite ends of the spectrum as regards quality evolution of estimates (figure 10). Meuse belongs to the departments which benefit the most from adding the land use prior, according to both performance metrics (MAE is reduced by 64 % and RMSE by 31 %.). In subplots (a) and (b), we plotted absolute differences between counts from PHD and those from validation data in Meuse, for each prior. Subplot (a) highlights quite clearly the limitations of the Voronoi hypothesis : since people likely place of residence are allocated uniformly in the Voronoi polygon in which they are detected, estimation errors tend to affect all the tiles intersecting the Voronoi polygon. This major source of bias is greatly reduced when we estimate PHD using the land use prior, as it eliminates tiles with low volume of housing in which it is unlikely that people spend the night. This effect of the land use prior is observed in all rural areas, as well as nearby moderate size cities. On the contrary, its effect appears very limited in crowded urban areas. This is for instance blatant in Paris and its immediate suburbs, which belong to areas exhibiting lowest RMSE and MAE reductions with the land use prior. Differences between estimates and validation data are again plotted for comparison (subplots (c) and (d)). As nearby tiles can exhibit substantial heterogeneity in this area, we resort to kernel smoothing so as to facilitate visualization[8]. Although some local differences arise, the two distributions appear almost identical, confirming the inability of the land use prior to mitigate estimation issues which arise in important cities and their vicinity. This result holds true in other major French cities, although to a lesser extent due to the specificities of Paris.

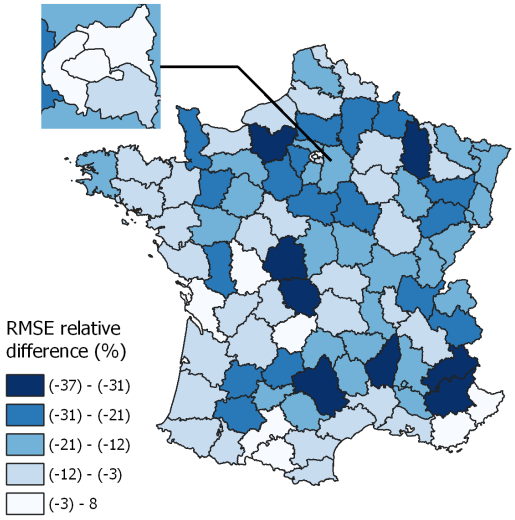
[8]More specifically, distributions are smoothed using a bisquare kernel over a 100x100 meters grid, with bandwith parameter ("size" of the neighbourhood on which computations are carried out) set to 2km. See Genebes, Renaud & Sémécurbe in Loonis & De Bellefon (2018) for a detailed presentation of kernel smoothing.

Figure 9: Accuracy gain of population counts due to the use of land use prior at department level

(a) Relative change in mean absolute error (MAE) due to adding bdtopo prior



(b) Relative change in root mean squared error (RMSE) due to adding bdtopo prior



3.4 Limits

Our results indicate substantial improvement of the quality of population estimates due to adding prior information on land use in the bayesian model of localization. Specifically, the main effect of the land use prior appears to be the removal of small to moderates overestimation errors all over French territory. However, it doesn't solve issues of major under- and over-estimation which characterize important cities and their suburbs. Various explanations can be put forward to explain the existence of these errors and the fact they are not mitigated by the addition of prior information on tiles.

First, we still rely on the Voronoi approximation to estimate population counts. When a phone event is localized in a given Voronoi polygon, the use of prior information on tiles will just ensure that it is projected on the grid in a relevant way, i.e. by taking into account unequal probabilities of nightly presence on the territory. However, the event remains bound to be allocated to tiles which belong to or intersect the Voronoi polygon, whereas it is entirely possible that the user is actually localized in another Voronoi polygon. This shortcoming alone could constitutes a major source of the misallocations observed in urban areas. In rural areas, where antennae distribution is very sparse and thus Voronoi polygons much wider than average, the Voronoi approximation seems more justified : in such configuration, antenna cells tend to be macro cells (i.e. cells covering very wide areas, often in a circular way) and thus the approximation can actually fit technical reality. On the contrary, it doesn't appear justified in urban areas : as evident from figure 1, Voronoi polygons are much smaller in urban reas, and are even often smaller than grid tiles in centers of major cities. In this case, the Voronoi approximation produces substantial bias since it doesn't account for overlapping of

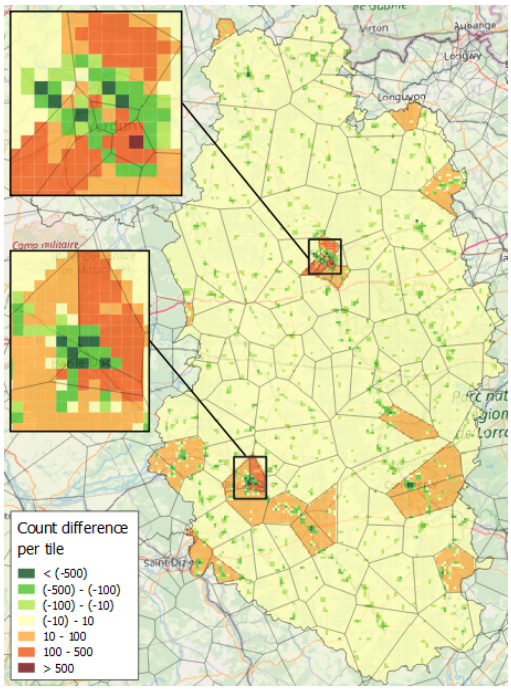
antennae coverage areas. Besides, adding prior information on land use doesn't make any change because of low number of tiles per Voronoi polygons, and the fact that housing tends to be uniformly distributed in dense areas. Availability of antennae coverage maps based on waves propagation models from the MNO could highly mitigate this source of error.

Another potential explanation for these errors could be that the concept of present population we measure is not the same over the territory. Urban areas are characterized by distinctive behaviours : tourism, the importance of social activities (restaurants, movies theaters, ...), different working hours, etc. As a result, the time window to which we restrict the sample of mobile phone events (19:00 to 09:00) — chosen so as to provide information regarding places of residence, while at the same time ensuring sufficient nightly observations because of tracks temporal scarcity — might capture more events which interfere with home detection than in rural areas. A careful examination of overestimation areas in major French cities gives weight to this hypothesis : overestimation particularly occurs in places known to be highly touristic (historical city centers, commercial areas, places with a lot of hotels...) as well as in transport key points (railway stations, airports). The land use prior fails to mitigate this type of errors because these areas are generally residential, and transport key points are classified in BD Topo "Bâti indifférencié" type because of their public nature. This analysis suggests that some of the differences between estimates and validation data in urban areas may not be errors, but simply due to measuring temporarily present population rather than nightly residing one. The much greater temporal resolution of tracks from signalling data could help optimize time span restriction.

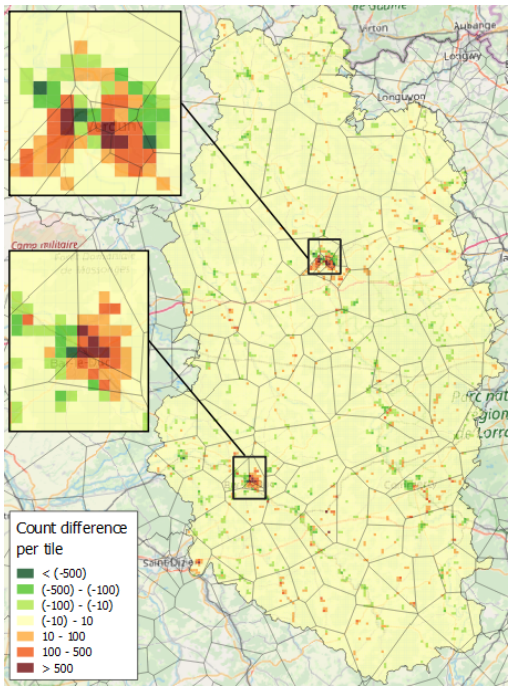
Finally, some limitations associated to the data sources we exploit might explain observed differences in a more general way. For availability reasons, we are constrained to deal with data sources observed at different temporality — 2007 for mobile phone data, 2011 for localized tax data, and 2015 for BD Topo. We thus make the hypothesis that evolutions are not meaningful enough to produce significant bias. Besides, some differences might be produced by the statistical adjustment of the sample. It relies on Orange client list in 2007, which quality is uncertain. It is also possible that market shares and penetration rate vary at finer levels than the departmental one.

Figure 10: Comparison of prior effect in rural and urban areas

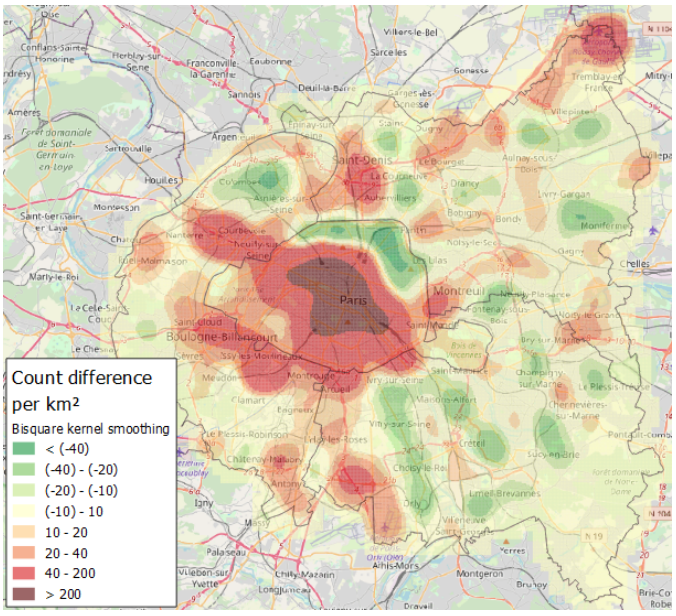
(a) Difference between PHD with uniform prior and reference : zoom on Meuse department



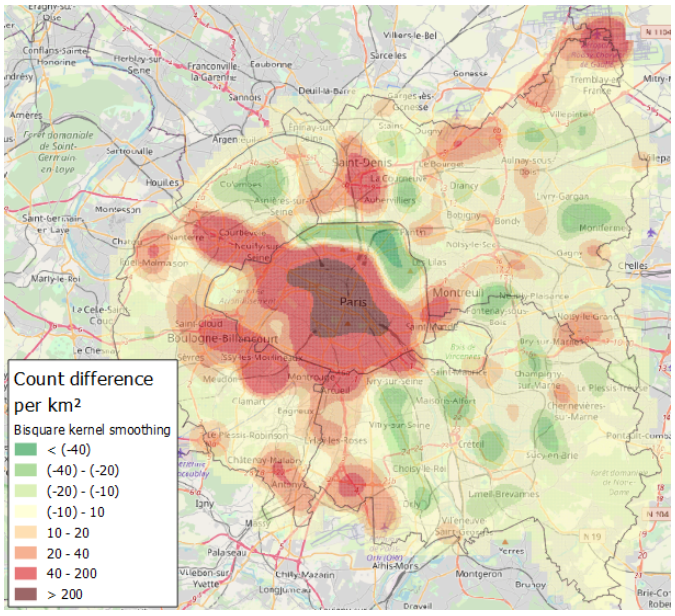
(b) Difference between PHD with land use prior and reference : zoom on Meuse department



(c) Difference between PHD with uniform prior and reference : zoom on Paris and immediate suburb



(d) Difference between PHD with land use prior and reference : zoom on Paris and immediate suburb



4 Conclusion

Mobile phone data appear as a high potential source to complement traditional ways of counting population. However, the lack of technical information on antennae has led researchers to rely on Voronoi approximation in most applications, which can introduce major bias. Our main contribution is to show that a simple bayesian model of localization can substantially improve population estimation by enabling the combination of mobile phone data and other statistical sources. We also point out that the indicators of estimation quality which have been used in previous studies — mainly Pearson's r correlation coefficient — can be misleading due to the pecularities of population distributions. Finally, we show that tools from spatial analysis and fractal geometry theory can be used to evaluate the quality gain of new estimates as regards replication of the spatial structure of validation data.

In this paper, we focused on *resident* population estimation in order to validate new estimates against localized tax data. But it is our opinion that the main potential of mobile phone data is to enable the estimation of *present* population, which sources from official statistics fail to measure satisfactorily. A main advantage of the bayesian framework is that it can accomodate a wide range of situations as regards information availability. As such, we would suggest future research to take advantage of this flexibility by estimating new configurations of the model, for instance with additional technical information on antennae (best service areas, radio propagation models) or different prior information sources (transport infrastructure, points of interest, etc.). Further investigation is also needed to determine how to validate present population estimates as sources based on census are not sufficient anymore for this purpose.

References

Anselin, Luc. 1995. "Local indicators of spatial association—LISA." *Geographical Analysis* 27 (2). Wiley Online Library: 93–115.

Calabrese, Francesco, Mi Diao, Giusy Di Lorenzo, Joseph Ferreira Jr, and Carlo Ratti. 2013. "Understanding individual mobility patterns from urban sensing data: A mobile phone trace example." *Transportation Research Part C: Emerging Technologies* 26. Elsevier: 301–13.

Chai, Tianfeng, and Roland R Draxler. 2014. "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature." *Geoscientific Model Development* 7 (3). Copernicus GmbH: 1247–50.

Cheung, Paul. 2012. "Big data, official statistics and social science research: Emerging data challenges." In *World Bank Meeting, Washington.*

Clifford, Peter, Sylvia Richardson, and Denis Hémon. 1989. "Assessing the significance of the correlation

between two spatial processes." *Biometrics*. JSTOR, 123–34.

Daas, Piet JH, Marco J Puts, Bart Buelens, and Paul AM van den Hurk. 2015. "Big data as a source for official statistics." *Journal of Official Statistics* 31 (2). De Gruyter Open: 249–62.

Demissie, Merkebe Getachew, Santi Phithakkitnukoon, Titipat Sukhvibul, Francisco Antunes, Rui Gomes, and Carlos Bento. 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). IEEE: 2466–78.

Deville, Pierre, Catherine Linard, Samuel Martin, Marius Gilbert, Forrest R Stevens, Andrea E Gaughan, Vincent D Blondel, and Andrew J Tatem. 2014. "Dynamic population mapping using mobile phone data." *Proceedings of the National Academy of Sciences* 111 (45). National Academy of Sciences: 15888–93.

Galiana, Lino, Benjamin Sakarovitch, and Zbiegniew Smoreda. 2018. "Ségrégation urbaine : un éclairage par les données de téléphonie mobile."

Hughes, C, E Zagheni, GJ Abel, A Wisniowski, A Sorichetta, I Weber, and AJ Tatem. 2016. "Inferring migrations: Traditional methods and new approaches based on mobile phone, social media, and other big data." *Brussels: European Commission*.

Kim, Yunmi, Tae-Hwan Kim, and Tolga Ergün. 2015. "The instability of the Pearson correlation coefficient in the presence of coincidental outliers." *Finance Research Letters* 13. Elsevier: 243–57.

Lee, Sang-Il. 2001. "Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I." *Journal of Geographical Systems* 3 (4). Springer: 369–85.

Levine, Rebecca S, Krista L Yorita, Matthew C Walsh, and Mary G Reynolds. 2009. "A method for statistically comparing spatial distribution maps." *International Journal of Health Geographics* 8 (1). BioMed Central: 7.

Loonis, Vincent, Marie-Pierre De Bellefon, and others. 2018. "Handbook of Spatial Analysis." Insee.

Moran, Patrick AP. 1948. "The interpretation of statistical maps." *Journal of the Royal Statistical Society. Series B (Methodological)* 10 (2). JSTOR: 243–51.

Okabe, Atsuyuki, Barry Boots, Kokichi Sugihara, and Sung Nok Chiu. 2009. *Spatial tessellations: concepts and applications of Voronoi diagrams*. Vol. 501. John Wiley & Sons.

Openshaw, Stan. 1984. "The modifiable areal unit problem." *Concepts and Techniques in Modern Geography*. GeoBooks.

Rényi, Alfréd, and others. 1961. "On measures of entropy and information." In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Ricciato, Fabio, Peter Widhalm, Massimo Craglia, and Francesco Pantisano. 2015. *Estimating Population Density Distribution from Network-Based Mobile Phone Data*. Publications Office of the European Union.

Ricciato, Fabio, Peter Widhalm, Francesco Pantisano, and Massimo Craglia. 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. Elsevier: 65–82.

Sakarovitch, Benjamin, Marie-Pierre De Bellefon, Pauline Givord, and Maarten Vanhoof. 2019. "Allô, où es-tu ? Estimer la population résidente à partir de données de téléphonie mobile, une première exploration." *Economie et Statistique*.

Sémécurbe, François, Cécile Tannier, and Stephane G Roux. 2018. "Exploring the deviations from scale-invariance of spatial distributions of buildings using a Geographically Weighted Fractal Analysis. An application to twenty French middle-size metropolitan areas."

Tennekes, Martijn. 2018. "Geographic location of events." https://github.com/MobilePhoneESSnetBigData/mobloc.

Terrier, Christophe. 2009. "Distinguer la population présente de la population résidente." *Courrier Des Statistiques* 128: 63–70.

Toole, Jameson L, Michael Ulm, Marta C González, and Dietmar Bauer. 2012. "Inferring land use from mobile phone activity." In *Proceedings of the Acm Sigkdd International Workshop on Urban Computing*, 1–8. ACM.

Upton, Graham, Bernard Fingleton, and others. 1985. *Spatial data analysis by example. Volume 1: Point pattern and quantitative data*. John Wiley & Sons Ltd.

Vanhoof, Maarten, Liane Hendrickx, Aare Puussaar, Gert Verstraeten, Thomas Ploetz, and Zbigniew Smoreda. 2017. "Exploring the use of mobile phone data for domestic tourism trip analysis." *Netcom. Réseaux, Communication et Territoires*, nos. 31-3/4. Netcom Association: 335–72.

Vanhoof, Maarten, Fernando Reis, Thomas Ploetz, and Zbigniew Smoreda. 2018. "Assessing the quality of home detection from mobile phone data for official statistics." *Journal of Official Statistics* 34 (4). Sciendo: 935–60.

Wilcox, Rand R. 2009. "Comparing Pearson correlations: Dealing with heteroscedasticity and nonnormality." *Communications in Statistics-Simulation and Computation* 38 (10). Taylor & Francis: 2220–34.

Willmott, Cort J, and Kenji Matsuura. 2005. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." *Climate Research* 30 (1): 79–82.

Wulder, Michael A, Joanne C White, Nicholas C Coops, Trisalyn Nelson, and Barry Boots. 2007. "Using local spatial autocorrelation to compare outputs from a forest growth model." *Ecological Modelling* 209 (2-4).

Appendices

Appendix A : R package documentation

Package 'mobicountR'

October 30, 2019

Type Package

Title Improving present population counts using mobile phone data

Version 0.1.0

Description
a flexible bayesian framework to improve present population counts using mobile phone data.

License CC0

Encoding UTF-8

LazyData true

Imports minior,
dplyr,
rlang,
foreach,
sf,
readr,
stringr,
tidyr,
data.table,
leaflet,
RColorBrewer

RoxygenNote 6.1.1

R topics documented:

bay_coverage *Computes the bayesian posterior distribution using a coverage map*

Description

This function implements the bayesian model for mobile events detection, assuming that a coverage map

Usage

```
bay_coverage(df_inter, prior)
```

Arguments

priors	a list of data.frames, one for each prior. See 'Details' for required structure
weights	numeric vector of same length as priors; weights used to compute the linear combination of priors. Must sum to one for the output to be a probability distribution.

Details

data.frames in priors should have the same structure as those produced by prior_building and prior_res_pop :
a first column indicating tiles IDs, and a second column giving the distribution of the prior on the grid.

Value

a data.frame with two columns : tiles ID and a prior distribution over tiles.

Examples

```
## Not run:
# Import priors
prior_bdtopo <- prior_building(paths_shp = "~/bdtopo_shp", grid = grid_500_france)
prior_rfl <- prior_res_pop(rfl = rfl11)
# Include inputs in a list
list_priors <- list(prior_bdtopo, prior_rfl)
# Define weights to be used (same order as the list of priors)
weights = c(0.3, 0.7)
# Combine priors
prior_comb <- prior_combine(priors = list_priors, weights = weights)

## End(Not run)
```

bay_voronoi *Computes the bayesian posterior distribution using a voronoi tessela-tion*

Description

This function implements the bayesian model for mobile events detection, assuming that no information on the MNO's antennas coverage is available. A voronoi tesselation is thus used to approximate the area covered by each antenna. If available, prior information over tiles can be used to improve the quality of the detection.

Usage

```
bay_voronoi(df_inter, prior = NULL, var_prior = NULL,
  var_final = "proba_final", filter_null = TRUE)
```

Arguments

df_inter	a data.frame which indicates for each voronoi x grid intersection the relative area of the intersection in the corresponding voronoi; should be either produced by or have the same format as the output of inter_grid_voronoi.
prior	a data.frame; should be either produced by or have the same format as those produced by the prior computation functions of mobicount. Defauts to NULL, in which case a prior equal to one is used (uniform).
var_prior	character; name of the variable which contains the prior distribution; ignored if prior = NULL.
var_final	character; name of the variable which should contain the posterior distribution.
filter_null	logical; should voronoi x grid intersections with a null posterior probability be deleted ? Defauts to TRUE.

Value

a data.frame with two columns : an ID for each voronoi x grid intersection formatted as "voronoi_ID:grid_id", and the posterior distribution.

Examples

```
## Not run:
# Import prior
prior_bdtopo <- readr::read_csv(~/bdtopo_france.csv)
# Import voronoi x grid intersections table
inter_df <- readr::read_csv(~/inter_grid_voro.csv)
# Compute bayesian posterior distribution
bay_df_bdtopo <- bay_voronoi(inter_df, prior_bdtopo, "proba_bdtopo")

## End(Not run)
```

`create_grid` *Creates a regular grid over a spatial unit*

Description

Creates a regular square grid over the bounding box of an sf object (e.g. a country limits shapefile imported with `st_read`).

Usage

```
create_grid(x, tile_size, crs = 2154)
```

Arguments

`x`	data.frame of class `sf`.
`tile_size`	integer; length of the side of a square (tile).
`crs`	integer; desired projected coordinate system; defauts to 2154 (Lambert 93).

Details

The output data.frame contains a variable grid_id which provides an unique identifier for each tile. Identifiers are constructed as
x_centroid/100:y_centroid/100 so that tile centroids are easy to compute back when needed.

Value

the created grid as a data.frame of class `sf`.

Examples

```
## Not run:
# Import shapefile of country limits
fr_limits <- sf::st_read("~/france_shp/francemetro_2015.shp")
# Create grid
grid <- create_grid(x = fr_limits, tile_size = 500)

## End(Not run)
```

`inter_grid_voronoi` *Computes intersections between grid and voronoi tesselation*

Description

Computes intersections between a grid (e.g. created with `create_grid`) and a voronoi tesselation computed over an MNO antennas map. Intersections probabilities, defined as the relative area of an intersection in the correspond voronoi, are computed during the process.

Usage

```
inter_grid_voronoi(grid, voronoi, grid_id = "grid_id",
  voronoi_id = "NIDT", sf = FALSE, crs = 2154)
```

Arguments

`grid`	data.frame of class `sf`.
`voronoi`	data.frame of class `sf`.
`sf`	logical; if `TRUE`, output is an sf object; defauts to `FALSE`.
`crs`	integer; desired projected coordinate system; defauts to 2154 (Lambert 93).

Details

The output data.frame contains a variable `proba_inter` which provides, for each intersection between a tile and a voronoi, the relative area of the tile in the corresponding voronoi. This value is comprised between 0 and 1 and thus assimilated to a probability of intersection.

Value

a simple data.frame if `sf=TRUE`. A data.frame of class `sf` if `sf=FALSE`.

Examples

```
## Not run:
# Create grid
grid_sf <- create_grid(x = fr_limits, tile_size = 500, export = FALSE)
# Import shapefile of voronoi tesselation
voronoi_sf <- sf::st_read("~/Antenne_voronoi_rev.shp", crs = 2154)
# Compute grid-voronoi intersections
table_prob <- inter_grid_voronoi(grid = grid_sf,
voronoi = voronoi_sf, prob = TRUE, sf = FALSE)

## End(Not run)
```

mobicount *Improving present population counts using mobile phone data*

Description

mobicount provides a flexible bayesian framework to improve present population counts using mobile phone data.

Author(s)

Romain Avouac

prior_building *Computes prior from building registers*

Description

Computes a prior distribution over a grid using data from building registers (e.g. shapefiles of buildings over the territory).

Usage

```
prior_building(paths_shp, grid, grid_id_var = "grid_id",
  dir_inter = paste0(tempdir(), "/dir_inter"), area_min_max = NULL,
  height_var = "HAUTEUR", impute = TRUE, height_min_max = c(NA, 130),
  parallel = TRUE, n_workers = 5, crs = 2154)
```

Arguments

paths_shp	character vector; complete paths to building shapefiles.
grid	a data.frame of class sf.
grid_id_var	character; name of the variable indicating tiles unique identifier in grid.
dir_inter	character; path to the desired output directory for prior files.
area_min_max	numeric vector of length 2; minimum and/or maximum areas allowed for buildings, outside of which buildings are filtered. To provide only one value, indicate the other as NA.
height_var	character; name of the variable indicating building heights. If set to NULL, prior is computed using only building areas.
impute	logical; should heights of zero height buildings be imputed ?
height_min_max	numeric vector of length 2; minimum and/or maximum heights allowed for buildings, outside of which buildings are filtered. To provide only one value, indicate the other as NA. Minimum height is ignored if impute = TRUE.

parallel	logical; if multiple building shapefiles are provided, should the computations on each of these files be run in parallel ?
n_workers	number of worker processes that doParallel will use to execute tasks in parallel. Ignored if parallel = FALSE.
crs	integer; desired projected coordinate system; defauts to 2154 (Lambert 93).

Details

If multiple shapefiles are provided (e.g. shapefiles at an infra-national level), prior computation can be computed in parallel using foreach.

If the provided grid was not computed using create_grid, tiles ID should be formatted the same way : "x_centroid/100:y_centroid/100".

Value

A data.frame of class data.table with two columns : tiles ID and a prior distribution over tiles.

Examples

```
## Not run:
# Import building shapefiles paths
bdtopo_shp <- list.files("~/bdtopo")
# Compute grid
grid_500_fr <- create_grid(x = fr_limits, tile_size = 500)
# Compute prior
build_prior_dt <- prior_building(paths_shp = bdtopo_shp, grid = grid_500_fr)

## End(Not run)
```

---

prior_combine *Combines multiple priors*

Description

Combines multiple prior distributions over a same grid by computing a new distribution as a linear combination

Usage

```
prior_combine(priors, weights)
```

Arguments

priors	a list of data.frames, one for each prior. See 'Details' for required structure
weights	numeric vector of same length as priors; weights used to compute the linear combination of priors. Must sum to one for the output to be a probability distribution.

Details

data.frames in priors should have the same structure as those produced by prior_building and prior_res_pop :
a first column indicating tiles IDs, and a second column giving the distribution of the prior on the grid.

Value

a data.frame with two columns : tiles ID and a prior distribution over tiles.

Examples

```
## Not run:
# Import priors
prior_bdtopo <- prior_building(paths_shp = "~/bdtopo_shp", grid = grid_500_france)
prior_rfl <- prior_res_pop(rfl = rfl11)
# Include inputs in a list
list_priors <- list(prior_bdtopo, prior_rfl)
# Define weights to be used (same order as the list of priors)
weights = c(0.3, 0.7)
# Combine priors
prior_comb <- prior_combine(priors = list_priors, weights = weights)

## End(Not run)
```

---

prior_res_pop *Computes prior from resident population*

Description

Computes a prior distribution over a grid using resident population data from fiscal localized income sources (e.g. RFL or FiLoSoFi data in France).

Usage

```
prior_res_pop(rfl, indiv_id_var = "DIRNOSEQ", n_indiv_var = "nbpersm",
  x_y_var = c("x", "y"))
```

Arguments

rfl	a data.frame of class data.table containing fiscal localized incomes data. If a simple data.frame is provided, it is automatically converted to a data.table.
indiv_id_var	character; name of the variable indicating household unique identifier in rfl.
n_indiv_var	character; name of the variable indicating household sizes (number of individuals) in rfl.
x_y_var	character vector of length 2; names of the variables indicating x and y coordinates of the household in rfl.

Details

If the output `data.frame` of this function is to be used as in , x and y coordinates in input data (`rfl`) should be converted beforehand in Lambert-93 (ESPG:2154).

In the output `data.frame`, tiles ID are formatted as "x_centroid/100:y_centroid/100" in order to match the formatting of grids created by `create_grid`.

Value

a `data.frame` of class `data.table` with two columns : tiles ID and a prior distribution over tiles.

Examples

```
## Not run:
library(sas7bdat)
# Import data from fiscal localized data in 2014
rfl_df <- read.sas7bdat("path_to_filosofi_2014/menages14.sas7bdat")
# Compute prior using resident population
res_prop_prior <- prior_res_pop(rfl = rfl_df, indiv_id_var = "IDENTIFIANT")

## End(Not run)
```

---

sfc_as_cols *Transforms a sf POINT geometry to x,y columns*

Description

Transforms a data.frame of class `sf` with a POINT geometry (e.g. a grid of which centroids has been computed using `st_centroid`) to a conventional data.frame with coordinates as two numeric columns.

Usage

```
sfc_as_cols(x, names = c("x", "y"))
```

Arguments

`x`	data.frame of class `sf` with a sfc_POINT geometry.
`names`	character vector of length 2; names of the two new columns with x,y coordinates; defauts to `c("x","y")`.

Details

This transformation is useful for efficient filtering of coordinates, e.g. to subset a grid to a smaller area bounding box.

Source of the function : https://github.com/r-spatial/sf/issues/231

Value

A data.frame with two new columns corresponding to x,y coordinates of spatial units. The initial geometry column is removed.

Examples

```
## Not run:
# Import grid and compute centroids
grid <- sf::st_read("~/grid_500_france.shp")
grid <- grid %>% st_centroid()
# Transform to a data.frame with centroid coordinates as new columns
grid_nogeo <- grid %>% sfc_as_cols()

## End(Not run)
```

Note de synthèse

Contexte

Ce rapport présente un stage d'une durée de 6 mois réalisé conjointement à l'INSEE et à Orange. Ce travail a été réalisé sous la supervision de Benjamin Sakarovitch, data scientist au SSP-Lab de l'INSEE. Le SSP-Lab est une unité récente qui vise à stimuler et diffuser l'innovation au sein du service statistique public. Une des missions majeures de cette unité consiste à exploiter les nouvelles sources de données, notamment volumineuses, en complément des sources traditionnelles. Du fait de l'usage quasi-généralisé des téléphones portables, les données mobiles apparaissent comme une source prometteuse pour la statistique publique. En particulier, elles permettent l'estimation de la population *présente* là où les sources usuelles de la statistique publique décrivent la population *résidente*.

Problème de recherche

Cette étude s'intéresse à un des problèmes majeur qui caractérise l'exploitation des données mobiles : la localisation géographique des évènements. Dans la plupart des cas, la position géographique exacte d'un mobile lors d'un évènement est inconnue, seule celle de l'antenne à laquelle il se connecte est connue. Une étape de *mapping* spatial est donc généralement nécessaire avant de pouvoir procéder à des estimations de population présente. Dans la littérature, ce *mapping* est traditionnellement réalisé par une tesselation de Voronoï : on fait l'hypothèse qu'un mobile se connectera toujours à l'antenne la plus proche. En généralisant à tous les points possibles, on obtient une partition de l'espace en polygones convexes nommés "polygones de Voronoï", auxquels on attribue les évènements mobiles. Cette méthode de *mapping* présente cependant de nombreuses limites, aussi bien sur le plan technique que statistique, et peut ainsi générer des biais importants dans les estimations de population.

Méthodologie

Nous mobilisons un modèle de localisation des évènements mobiles proposé par Tennekes (2018). Ce modèle repose sur la formule de Bayes :

$$\mathbb{P}(carreau_i|antenne_j) \propto \mathbb{P}(carreau_i)\mathbb{P}(antenne_j|carreau_i)$$

Chaque évènement est observé au niveau de l'antenne auquel s'est connecté le mobile. Le problème consiste alors à inférer la position de l'utilisateur sur une grille spatiale de notre choix. Il est souhaitable que cette localisation soit probabilisée : comme seules les coordonnées de l'antenne sont connues, la position de l'utilisateur ne peut jamais être estimée avec certitude. Selon la formule de Bayes, cette probabilité est proportionnelle à une information *a priori* que nous pourrions avoir sur les carreaux de la grille, multipliée par

la probabilité que le mobile se soit connectée à une antenne donnée sachant le carreau dans lequel l'utilisateur se trouvait.

Ce modèle de localisation a de multiples avantages en pratique. Dans cette étude, nous exploitons principalement la possibilité d'ajouter de l'information *a priori* sur la grille, et montrons que les estimations de population peuvent être fortement améliorées par ce biais, sans même disposer d'informations annexes sur la couverture théorique des antennes.

Données

Nous exploitons un jeu de données CDR datant de 2007, fourni par l'opérateur Orange. Il contient la trace numérique de l'ensemble des appels et SMS passés par les 18 millions de clients d'Orange durant 5 mois en 2007, soit au total 15 milliards d'évènements. Afin de tenir compte de l'inégale répartition de la population sur le territoire, nous utilisons les données sur le bâti indifférencié issues de la BD TOPO comme information *a priori*. Nous validons nos estimations de population en les comparant aux données fiscales localisées (RFL), à partir desquelles nous calculons les densités de population sur la grille. Le fait de disposer de ces données exhaustives comme validation nous permet de quantifier de manière précise le gain de qualité permis par l'ajout d'information *a priori*.

Résultats

La comparaison visuelle des cartes nationales de densités de population permet d'observer un gain de qualité considérable de l'estimation de population. Une analyse statistique confirme ce premier constat : à partir de métriques robustes à la présence de valeurs extrêmes, nos résultats indiquent une forte hausse de la corrélation entre les distributions de population, ainsi qu'une réduction de moitié de l'erreur absolue d'estimation en moyenne. Enfin, des outils issus de la statistique spatiale nous permettent de mettre en évidence une meilleur réplication de la structure d'autocorrélation de la distribution de validation.

Malgré ce gain de qualité, les distributions estimées restent caractérisées par de fortes erreurs d'estimation dans les grandes villes et leur alentours. Nous avançons deux explications à la persistence de telles erreurs : une qualité insuffisante de la modélisation de la couverture théorique des antennes, et une trop faible profondeur temporelle des données CDR. L'exploitation de données de *signalling* (données passives) devrait permettre de fortement limiter ces erreurs.

Summary note

Context

This report details a 6-months internship undergone jointly at the National Institute of Statistics and Economic Studies (INSEE) and at Orange, a major French mobile network operator (MNO). This work was done under the supervision of Benjamin Sakarovitch, a data scientist at INSEE's SSP-Lab. The SSP-Lab is a recent unit aimed at fostering innovation in the official statistical system. A major mission of this unit is to leverage new data sources, notably big data sources, to complement traditional statistical sources. Mobile phone data, due to their ubiquitous use in general population, appear as a very promising new data source for official statistics. In particular, it provides the opportunity to proceed to present population counts (in a given venue, city, region...) whereas official statistics sources, notably census data, mostly produce information on resident population.

Research problem

This study tackles a major challenge that arises when dealing with mobile phone data : the geographical location of events. In most applications involving such data, only the location of the antenna to which a device has connected to is available, leaving the exact geographical location of mobile devices unknown. A spatial mapping of phone events is thus required in order to produce accurate information on present population. In the literature, this mapping is usually performed through a Voronoi tessellation approximation : when a phone event is observed, it is assumed that the device connects to the nearest antenna. This produces a partition of the space in convex polygons named Voronoïs, on which phone events are mapped. However, there are many downsides to using this hypothesis, both of technical and statistical nature, which can introduce substantial bias in population estimates.

Methodology

We leverage a localization method of mobile phone users based on a framework developed by Tennekes (2018). It fundamentally relies on Bayes' formula :

$$\mathbb{P}(tile_i|cell_j) \propto \mathbb{P}(tile_i)\mathbb{P}(cell_j|tile_i)$$

Each phone event is observed at cell level, i.e. the actual antenna transmitting the signal. The problem is then to infer in which tile of a given input spatial grid the mobile phone that generated this record was located. We want this location to be probabilized over the grid : as the only geographical information we have is the coordinates of the antenna to which the device has connected, we can never be certain of the actual location of the user. Bayes' rule states that this probability is proportional to any prior information we

might have on tiles of the grid multiplied by the probability that the signal comes from a given cell knowing that the phone was on a particular grid tile.

This localization model has multiple advantages. In this study, we leverage the possibility to add prior information on the grid to show that population estimates can be substantially improved even if no technical information on antennae theoretical coverage are available.

Data

This analysis is based on a 2007 pseudonymised call detail records (CDR) dataset from major French MNO Orange. Five months of exhaustive activity of more than 18 million Orange customers is available, representing around 15 billion events. In order to take into account the unequal probabilities of presence on the territory, we exploit land use data from the BD TOPO as prior information. We validate our population estimates by comparing them to localized tax data, from which we get the number of people residing in each tile of the grid. Having this ground truth enables us to quantify the gain in quality due to using relevant prior information.

Results

The visual comparison of population density maps at France level highlights a dramatic improvement in the replication of the ground truth population distribution. This conjecture is confirmed by statistical analysis : using outliers-robust metrics, we find both a substantial increase in correlations between population distributions and a reduction by half of absolute estimation error on average. We also use tools from spatial statistics to show that we better replicate the autocorrelation structure of the ground truth distribution.

Yet, we find that the estimated distribution remains characterized by either under- or over-estimation of population in important cities and their suburbs. We hypothesize that insufficient quality of antennae coverage modelization and temporal scarcity of the data are the main factors to explain this limit. Reproducing this study on more recent data from signalling sources could greatly reduce these shortcomings.

Résumé des stages longs

Stage à la DSDS de l'INSEE, 11/06/18 - 31/08/18

Contexte

Ce stage a été réalisé à la section Revenu des Ménages de la Direction des Statistiques Démographiques et Sociales de l'INSEE. Ce stage était l'occasion d'acquérir une expérience de recherche en sociologie quantitative. La thématique de recherche a été proposé par Louis-André Vallet, et la supervision du stage a été réalisée conjointement par Céline Goffette (CREST) et Jérôme Accardo (INSEE).

Problème de recherche

Pour un certain nombre de pays et particulièrement dans le cas de la France, les conclusions des différentes littératures sur la transmission du statut socio-économique ne sont donc pas convergentes : là où les sociologues mettent en évidence une hausse de la fluidité sociale, i.e. un affaiblissement du lien entre origine et position sociales, les économistes mettent au contraire en évidence une transmission substantielle de la capacité à générer des revenus qui semble se perpétuer au fil des générations. L'objectif de cette étude est de proposer une analyse intermédiaire entre les approches sociologique et économique de la mobilité en étudiant dans quelle mesure la capacité à générer des revenus se transmet au fil des générations selon l'origine sociale des individus.

Méthodologie

Cette étude a pour objet la comparaison des distributions de revenus conditionnelles à l'origine sociale. Nous nous inscrivons dans la lignée du cadre économique et de la procédure statistique proposés par Lefranc, Pistolesi et Trannoy (2004) pour analyser l'évolution de l'inégalité des chances entre 1979 et 2000. Les auteurs proposent de définir l'inégalité des chances en recourant à une expérience de pensée : supposons que les individus ont la possibilité de choisir leur milieu social d'origine. L'inégalité des chances prévaut dès lors qu'un individu rationnel préfère toujours une distribution de revenus à une autre. Et comme le choix hypothétique de l'individu s'apparente à un choix risqué – on compare des distributions de probabilités sur les différents niveaux de revenu – la comparaison des distributions se fait à partir d'un critère de dominance stochastique.

Données

Nous appliquons ce cadre statistique à la série des enquêtes Revenus Fiscaux (ERF) et Revenus Fiscaux et Sociaux (ERFS) de l'Insee. Ces enquêtes couvrent la période 1996-2015, et permettent en cela de déterminer si

la réduction de l'inégalité des chances mise en évidence par les auteurs entre 1979 et 2000 s'observe également au cours des deux dernières décennies.

Résultats

Nous présentons tout d'abord des statistiques descriptives qui mettent en évidence une inégalité des chances substantielle au niveau statique. Puis nous exposons les résultats issus de l'application de la procédure de Lefranc et al aux ERFS, qui met en lumière une stabilité de l'inégalité des chances entre 1996 et 2015, contrastant avec la forte réduction de l'inégalité à laquelle concluent les auteurs pour la période 1979-2000.

Stage à Orange, 11/03/19 - 24/05/19

Contexte

Ce stage a été ralisé au laboratoire SENSE d'Orange Labs. Il constitue le prolongement de mon stage d'application réalisé conjointement à l'INSEE et à Orange Labs. La raison de ce prolongement a été la possibilité de travailler sur des données mobiles très récentes, permettant d'envisager des estimations de population présente de haute précision spatio-temporelle. Comme le stage d'application, il a été réalisé conjointement sous la supervision de Benjamin Sakarovitch, *data scientist* au SSP-Lab de l'INSEE, et de Zbigniew Smoreda, sociologue au laboratoire SENSE d'Orange.

Problème de recherche

Le problème de recherche est identique à celui du stage d'application : la localisation géographique des évènements. Cependant, contrairement à l'étude effectuée lors du stage précédent, l'enjeu est ici de s'affranchir complètement de la modélisation par polygônes de Voronoï dans le cadre du *mapping* spatial des évènements. Cette extension est rendue possible par la disponibilité de données précises sur la couverture théorique des antennes.

Méthodologie

Nous mobilisons à nouveau le modèle de localisation des évènements mobiles proposé par Tennekes (2018) :

$$\mathbb{P}(carreau_i|antenne_j) \propto \mathbb{P}(carreau_i)\mathbb{P}(antenne_j|carreau_i)$$

Nous modifions cependant le terme de vraisemblance $\mathbb{P}(antenne_j|carreau_i)$, qui correspond de fait à la modélisation de la couverture des antennes, afin de tenir compte de l'information technique additionnelle dont nous disposons sur la couverture théorique des antennes. Cette modification témoigne d'une force majeur du modèle bayésien de localisation : il permet de couvrir un large éventail de situations en termes de données

disponibles, aussi bien sur la couverture des antennes que pour incorporer de l'information *a priori* sur la grille.

Données

Nous disposons d'un jeu de données de *signalling* (i.e. non plus seulement des données actives – appels et SMS – mais également passives – connexions fréquentes du mobile aux antennes sans action de l'usager) couvrant l'intégralité des clients d'Orange. Ces données présentent une profondeur temporelle largement plus importante que celles utilisées lors du stage d'application. Les cartes de couverture théorique des antennes sont également fournies par Orange.

Résultats

La courte durée du stage n'a pas permis de produire de résultats finaux concernant la population présente. De nombreux problèmes computationnels se sont posés : du fait de leur profondeur temporelle considérable, les données de *signalling* utilisées représentent un volume d'environ 1To par jour, ce qui rend très délicat leur exploitation, même en utilisant des outils de calcul distribué adaptés (Spark sur une infrastructure HDFS). Le stage a cependant permis de décrire statistiquement ces données – qui n'ont jamais été exploitées auparavant – ainsi que de coder les scripts informatiques nécessaires à de futures estimations de population présente.

Stage à Foundamental (Berlin), 11/06/19 - 06/09/19

Contexte

Ce stage a été réalisé à Foundamental (Berlin), un fonds de *venture capital* qui investit dans des *startups* innovantes dans le domaine de la construction. Il a été supervisé par Clemens Meyer zu Rheda, qui est *data scientist* à Foundamental. Ce stage a été l'occasion pour moi de réaliser d'une part une expérience dans une équipe internationale, et d'autre part de découvrir un milieu professionnel très différent de ceux de mes précédentes expériences (statistique publique ou laboratoires de recherche).

Problème de recherche

L'équipe chargée de la *data science* vise à développer des processus permettant de simplifier et de rendre plus efficient le travail des investisseurs, à toutes les étapes du processus de décision de l'entreprise. Dans ce cadre, le principal projet sur lequel j'ai travaillé consistait à détecter la similarité entre *startups* sur la seule base de leur description. L'objectif de ce projet est de pouvoir détecter les potentiels concurrents d'une entreprise dans laquelle Foundamental investit, ce qui est un facteur essentiel dans la décision d'investissement.

Méthodologie

Comme indiqué précédemment, les seules données disponibles pour procéder à la détection de similarité étaient des descriptions plus ou moins détaillées de *startups*. Le problème se ramenait donc à une application classique de *Natural Language Processing* (NLP) visant à détecter la similarité sémantique entre des textes de longueur inégale par le biais d'outils statistiques. Une des questions majeures a été de trouver le meilleur *embedding*, i.e. la meilleure manière de représenter vectoriellement le texte afin de procéder à leur comparaison. De nombreux *embeddings* ont été implémentés, des plus classiques (comptage binaire, TF-IDF, LDA) aux plus récents (word2vec, doc2vec, réseaux pré-entrainés). Une autre question importante a été de trouver la métrique appropriée pour comparer les textes une fois représentés vectoriellement. Là encore, plusieurs métriques ont été implémentées (similarisé cosine, *clustering* de textes, *word's mover distance*).

Données

Les descriptions des *startups* provenaient de plusieurs sources : principelement de bases de données commerciales (Crunchbase, Tracxn), mais certaines étaient également issues des recherches des investisseurs.

Résultats

Les résultats de la recherche par similarité étaient mitigés. En moyenne, les meilleurs algorithmes permettaient de détecter des entreprises candidates de bonne qualité, mais le nombre de faux-positifs s'est avéré important. La principale raison de ces performances contrastées est l'importante hétérogénéité de la qualité des descriptions, qui ne permet pas de détecter systématiquement des entreprises pertinentes. Face à ce constat, j'ai proposé et implémenté une application dédiée à la détection de similarité, qui permet de visualiser rapidement à la fois la description initiale et les résultats, de modifier à la volée les paramètres importants des modèles, et donc de déterminer rapidement si l'algorithme de détection de la similarité peut s'avérer pertinent pour l'entreprise considérée.