Location Precision in Mobile Phone Data
Application to present population in France

Keywords: Mobile Phone Data; Precision; Present Population Statistics; Quadtree; Signaling Data

1 Introduction

At an increasingly large scale and with increasing details, mobile phone data are shedding a new light on population dynamics. Yet, spatial resolution of cellular-based statistics is scarcely documented, highly dependent on the network topology and very heterogeneous across territories. In this paper, we propose a general framework to evaluate the location estimation precision of cellular network events. This evaluation combined with a quadtree algorithm enable us to build an adaptive spatial grid featuring small tiles for high accuracy areas and large tiles for low accuracy areas. The spatial precision is embedded within the dissemination grid. Our proposals are directly tested on producing a new present population statistics for metropolitan France, building from three main input datasets: raw signalling data, network coverage data and a highly precise geography of French residents.

2 Methods

Spatial mapping. The geolocation of mobile devices - from events recorded at the network cell level - can be described as solving an inverse problem [1]. Given a tesselation of interest, with I tiles, let us denote $u \in \{0,1\}^I$ the vector encoding the true tile of presence of the device at the origin of the event. When the device is in tile i_0, $u = 1_{i_0}$ the vector such that $u_{i_0} = 1$ while $u_i = 0 \ \forall \ i \neq i_0$. Given a cellular grid with J cells, we define the random variable $c \in \{0,1\}^J$, encoding the cell recording the event of this device: when the recording cell is j_0, $c = 1_{j_0}$. Telecom network coverage data gives us a probability matrix P, such that P_{ji} represents the probability of being detected at cell j while being in tile i:

$$P_{j,i} = \mathbb{P}\{\text{device detected in cell } j | \text{device in tile } i\}$$

Realistic models are used by mobile network operators to instanciate the matrix P in particular to be able to fulfill legal obligation.[1] On average, we expect to observe on network cells $\mathbb{E}[c] = Pu$ translating the presence of the device.[2] The estimate $\hat{u}$ can be written in general as $\hat{u} = g(P, c)$ where g is a chosen *spatial mapping*. In this paper we focus on a linear estimator $\hat{u} = Qc$. Q distributes presence over the cells in the tiles:[3]

$$Q_{i,j} = \mathbb{P}\{\text{device mapped to tile } i | \text{device detected in antenna } j\}$$

[1] Such as providing emergency call location.

[2] Note that u could also encode the presence of many devices, possibly with individual weights.

[3] Although our estimation of spatial accuracy may be applied to any spatial mapping, for our empirical results we follow [2] who suggest to deduce Q from P Bayes' rule by introducing a prior that reflects where the population is most likely located (e.g. based on land-use). In the results presented here, we use a uniform prior.

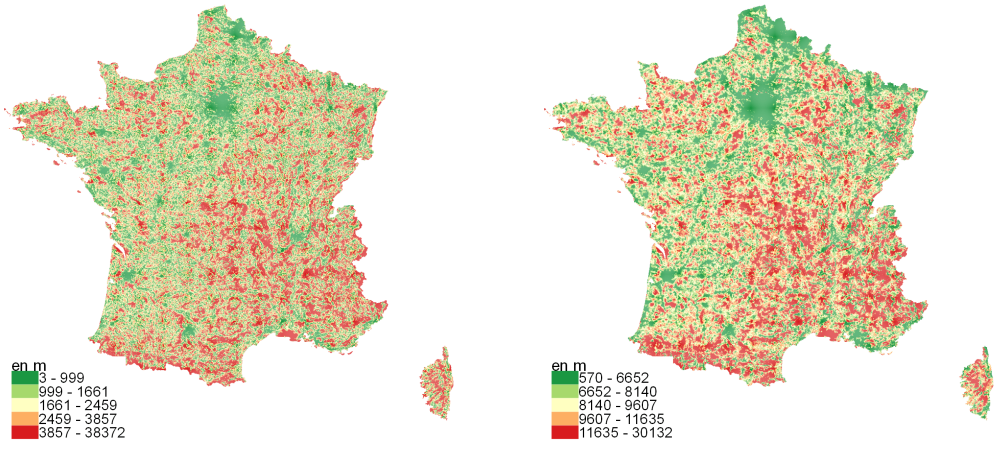
Estimating accuracy locally. The accuracy of this linear estimator can be approached locally by defining the probability to localise in i a device who is in i_0 and connects to the network probabilistically through P.

$$N_{i,i_0} = \mathbb{P}\{\text{device mapped to tile } i | \text{device in tile } i_0\}$$

Formally, $N = QP$. A good estimator Q should lead to a high N_{i_0,i_0} probability (correct mapping), or at least a high probability of tiles i in the neighborhood of i_0. With previous notations, $N1_{i_0} = \mathbb{E}[\hat{u}|\text{device in tile } i_0]$. Given x the tile coordinate, such that x_i is the coordinate of tile i, and $\hat{x}$ denotes the inferred location coordinates from $\hat{u}$, the average inferred location of the device located in i_0 is obtained from $\hat{x}_{i_0} = \mathbb{E}[\hat{x}|\text{device in tile } i_0] = \sum_i N_{i,i_0} x_i$. The spatial error integrates the uncertainty from P and Q, and can be evaluated with the mean squared error of the inferred location:

$$MSE_{i_0} = \mathbb{E}[\|\hat{x} - x_{i_0}\|^2| \text{ device in tile } i_0] = \underbrace{\mathbb{E}[\|\hat{x}_{i_o} - x_{i_0}\|^2 | i_0]}_{\text{Bias}} + \underbrace{\mathbb{E}[\|\hat{x}_{i_o} - \hat{x}\|^2 | i_0]}_{\text{Variance}}$$

The bias term describes the distance between the average inferred location and true location i_0 while the variance term measures the uncertainty around the average inferred location. This method allows the representation of the spatial accuracy of the estimates (Figure 1) - and notably shows how accuracy greatly varies in space.



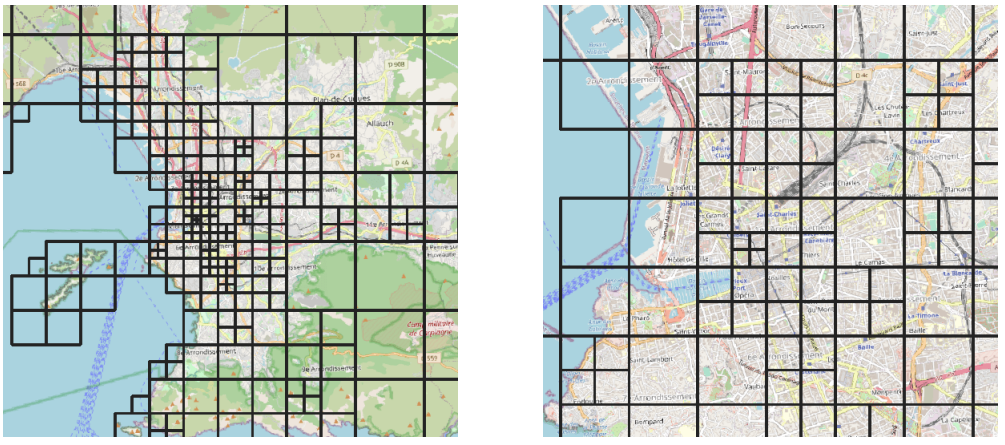
(a) Bias term (b) Variance term

Figure 1: Square root of the Biais and Variance terms (Unit: meter). *Note:* For this computation, matrix P was obtained from Orange Fluxvision and the prior was considered uniform.

Embedding precision within dissemination. We build a quadtree which directly embeds the calculated spatial precision by gathering tiles until the probability of correct location in the macro tile I_0 (group of tiles) is higher than a threshold: $N_{I_0}(I_0) > s$. We derive present population estimates within this reduced spatial grid, which visually provide a clear idea of the achievable precision (Figure 2.).

Present Population Statistics. We build hourly present population statistics over metropolitan France using 3 months of signaling data from Orange network and the geography of French residents from INSEE.[4] As our goal is to count individuals - and not active mobile phones, we added two steps: device temporal interpolation and

[4]Our approach only requires exchanges of anonymous aggregates between INSEE and ORANGE. Processing of individual data was performed by each data owner.



(a) Around Marseilles (b) Focus on the Old Port of Marseilles

Figure 2: Reduced grid with a threshold $s = 1\%$. The larger the cells, the less accurate the precision. *Note:* The grid was based on Orange Fluxvision matrix P and uniform prior.

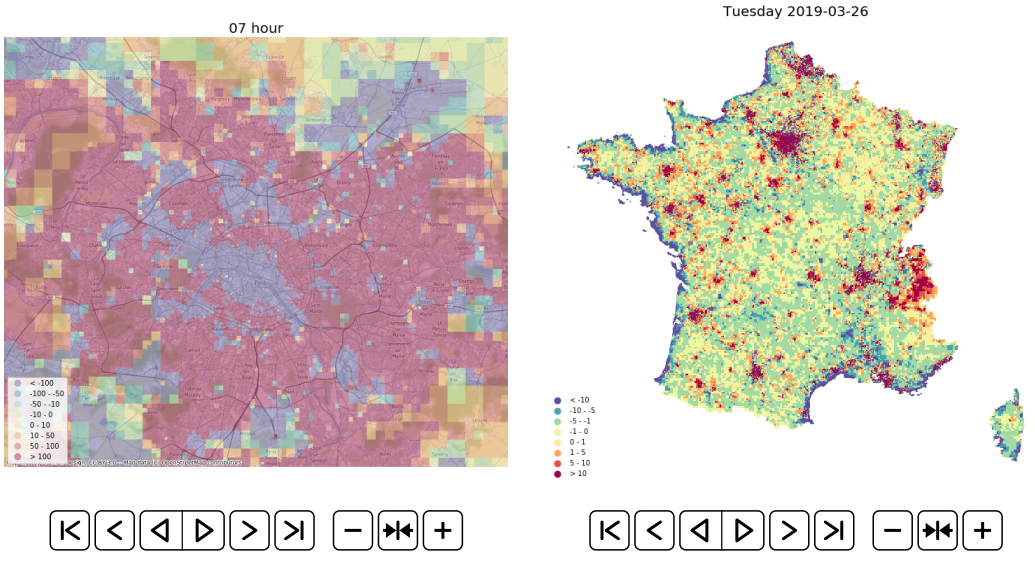
device residency-based weighting. To bypass the temporal sporadic presence of users over the network, we use device trajectory interpolation to rely on the closest-in-time location for each hour, before any aggregation. To be representative of the French resident population, we build device residency-based weights in several steps.[5] For each device, we estimate a home cell (a cell covering the device owner home). We then map French residents over home cells using realistic information on the coverage of each tile of 100 meters by Orange cells as provided by Orange Fluxvision (matrix P). We define weights with the ratio of actual residents divided by the network-detected residents at the level of contiguous groups of home cells which contain at least twenty detected resident devices.[6] Importantly with this approach, each detected device is assumed to be representative in its presence patterns of its neighbourhood. Finally, we build weighted sums of devices at the cell $\times$ hour level. These final aggregates may be mapped according to any spatial mapping. Given a spatial mapping Q, device residency-based weights w, and one record c per device i, we can reconstruct the present population vector as $\sum_i w_i Q c_i$. In the following empirical application, we use a bayesian inversion of matrix P with uniform prior to build Q.

3 Results

This new statistics allows the study of within day and across days variations of population. Around the Paris urban area (Figure 3.a.), we can see commuting movements in and out of the business centers. The suburbs are emptying while the center is filling up. In the course of a month (Figure 3.b.), tourism mobilities are predominant. On weekdays, the population is concentrated in cities. On Friday evenings, large segments of the population leave cities and move to the countryside and coastal areas, the return movements take place on sunday evenings. Through these two examples, we illustrate the useful visual properties of the adaptive grid. It allows us to describe the urban and rural areas at the same time while communicating the relative precision of the estimates in space.

[5] We filter devices which are identified as mobile phones (to filter M2M) and retain devices which are present at least 30 days out of the three months so as to ensure a relative stability of our scope (e.g. to filter movement due to client churning - irrelevant to inform on total counts).

[6] Therefore, at each hour, the weighted sum of network-detected residents equals the actual residents number at a very local level, and not only at the national level.



(a) hour-by-hour the 2019-03-08

(b) Day-by-Day from March to June 2019

Figure 3: Variation of Present Population at the hourly and daily level. *Note:* The same grid is use on the maps. On the Day-by-day, we keep only the present population at 22 o'clock.

4 Conclusion

We suggest a method to evaluate locally the probability that an event happening in a given region of space can be effectively mapped to this region. This method can be used for any spatial mappings and any assumption on the modelling of device-network interactions (e.g. voronoi-like tessellation, simplified radio propagation model...). We suggest to embed this information into the dissemination of the results - through a quadtree-derived adaptive grid. Our results are substanciated with an empirical application on French data. We build a new present population statistics over the French territory drawing particular attention to its consistency with official statistics and to the spatial mapping of the results.

References

[1] Fabio Ricciato, Giampaolo Lanzieri, Albrecht Wirthmann, and Gerdy Seynaeve. Towards a methodological framework for estimating present population density from mobile network operator data. *Pervasive and Mobile Computing*, page 101263, 2020.

[2] Martijn Tennekes, Y Gootzen, and Shan H Shah. A bayesian approach to location estimation of mobile devices from mobile network operator data. In *CBDS Working Paper 06-20*. 2020.