Summary note

1 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.

2 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.

3 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.

4 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.

5 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.