Note de synthèse

1 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*.

2 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.

3 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.

4 Données

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 Résultats

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.