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

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*.

5 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.