Journées de méthodologie statistique de l'Insee 2022

Estimation de la précision spatiale des données de téléphonie mobile

François Sémécurbe (*), Milena Suarez Castillo (*)

(*) Insee, Direction de la méthodologie et de la coordination statistique et internationale

francois.semecurbe@insee.fr milena.suarez-castillo@insee.fr

Mots-clés. : Téléphonie mobile, estimation de précision, inférence bayesienne.

Domaines. 7. 11.

Motivation

L'exploitation des données de téléphonie mobile collectées au niveau des antennes relais pour produire des populations présentes est un enjeu majeur de la statistique publique. L'abondance des antennes en France métropolitaine (plus de 130000 pour les trois opérateurs historiques selon l'ARCEP en septembre 2021) masque des situations locales contrastées. Une antenne peut couvrir de larges pans de l'espace rural et réciproquement un quartier urbain peut être couvert par plusieurs dizaines d'antennes, de sorte que la précision spatiale des populations présentes est très variable entre les territoires.

L'estimation de cette précision, rarement documentée, est l'objet de ce papier. La transformation d'une information connue au niveau des antennes en une information spatiale revient à résoudre un problème inverse [1]. La population est supposée être contenue dans une grille spatiale régulière composée de I carreaux. Le vecteur $u = \{0,1\}^I$ décrit la localisation d'une personne connectée au réseau. Lorsque la personne est dans le carreau i_0, $u_{i_0} = 1$ et $u_i = 0$ pour $i \neq i_0$. Le vecteur aléatoire $c = \{0,1\}^J$ décrit l'antenne de connexion du téléphone de la personne parmi les J antennes du réseau : pour l'antenne de connexion j_0, $c_{j_0} = 1$. P_{ji} est la probabilité d'être connecté à l'antenne j sachant que la personne est localisée dans le carreau i. Si P est la matrice formée par les P_{ji}, en moyenne $E(c) = Pu$. L'estimation de u s'écrit sous la forme $\hat{u} = g(P; c)$. g désagrège l'information connue à l'antenne en une distribution spatiale.

Dans le cas ou l'estimateur g est linéaire $\hat{u} = Qc$, Q est une nappe spatiale qui répartit en probabilité de présence la localisation d'une personne. [2] propose une estimation bayésienne : $Q_{ij} = \frac{P_{ji}\pi_i}{\sum_{i_0} P_{ji_0}\pi_{i_0}}$ où π est un vecteur de probabilité reflétant la connaissance a priori sur la répartition de la population dans l'espace. Après avoir justifié celle-ci en l'abordant sous la forme d'une solution à un problème de minimisation quadratique sous contrainte, une exploration de sa précision spatiale est proposée.

Précision des estimations bayésiennes

L'objectif, ici, n'est pas d'estimer une distribution de population mais de déterminer la précision spatiale des approches bayésiennes. La distribution de population est supposée connue et

correspond à la répartition de l'a priori. La localisation d'une personne (son carreau d'appartenance i codé par le vecteur u) est tirée aléatoirement selon la loi de probabilité de l'a priori. La précision de l'estimation de sa localisation étant donné sa présence mesurée sur le réseau mobile c est mesurée à l'aide du risque quadratique (Minimum Square Error) :

$$MSE = E_\pi(||X_{g(P,c)} - X_i||^2)$$

où $X_\cdot$ désigne les coordonnées spatiales des localisations (pour les carreaux leur centre).

La solution g qui minimise le risque quadratique est la moyenne a posteriori, $X_{g(P,c)} = \sum_i Q_{ij} X_i$. Un individu détecté à l'antenne c est positionné au barycentre pondéré par la loi a posteriori de celle-ci. Cette solution bien qu'optimale a un défaut : la distribution estimée (un ensemble discret de points moyens) est parcellaire par rapport à une distribution de population qui recouvre continûment l'espace. Le recours à la totalité de l'information spatiale contenue dans la nappe Q pour désagréger le lieu de présence d'une personne permet de dépasser cette limite. La nappe Q est optimale parmi les nappes R qui respectent les deux contraintes suivantes :

$$RP\pi = \pi \tag{1}$$

$$\sum_i R_{ij} X_i = \sum_i Q_{ij} X_i, \, \forall j \tag{2}$$

La première condition garantie l'absence de biais globale. En moyenne, l'estimation reproduit la distribution de l'a priori. La deuxième implique que les nappes sont centrées sur les barycentres des distributions à posteriori et par la même sur la solution optimale sans contrainte.

Le MSE associé à la nappe Q a pour valeur :

$$MSE = \sum_i \pi_i \sum_j P_{ji} \sum_{i'} Q_{i'j} ||X_i - X'_i||^2$$

Précision locale

$\sum_j P_{ji} \sum_{i'} Q_{i'j} ||X_i - X'_i||^2$, s'interprète comme l'erreur quadratique moyenne de localisation d'une personne située dans le carreau i et localiser via Q. Elle correspond à l'inertie de la nappe moyenne $N_{u_i} = QPu_i$ associée au carreau i. Cette erreur de localisation se décompose en deux termes de précision. Le biais décrit la distance entre le carreau i et le centre de gravité de la nappe N_{u_i} ; la variance l'erreur de localisation autour de cette position moyenne. Cette décomposition permet de représenter cartographiquement la précision de localisation d'une personne à l'aide du réseau téléphonique (Figure 1).

Intégrer la précision dans la diffusion

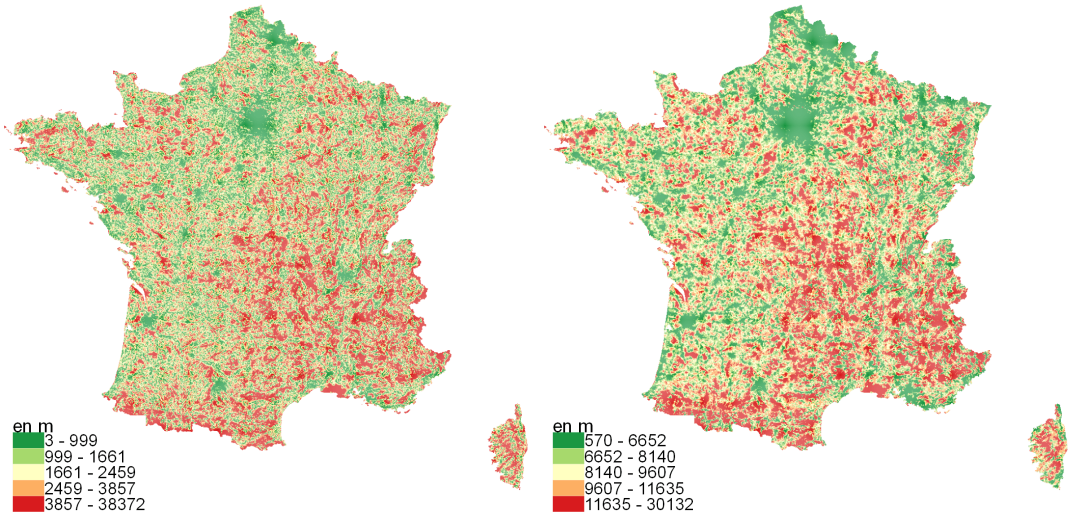
Les nappes N_{u_i} ont été utilisées pour créer des grilles adaptatives intégrant directement la précision spatiale à l'aide d'un algorithme de type quadtree. Les carreaux sont regroupés ensemble tant que $N_{I_0}(I_0) > s$ n'est pas supérieur au seuil s, $I0$ représente l'agrégation des carreaux. Plus la précision de localisation est bonne et plus les carreaux sont petits (Figure 2).

Bibliographie

[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.

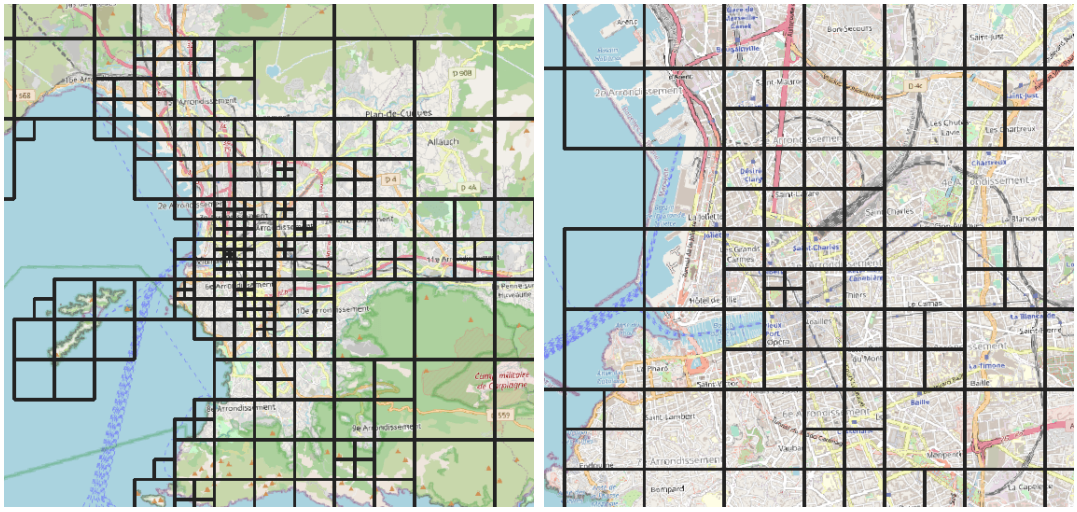
[2] Martijn Tennekes, Y Gootzen, and Shan H Shah. A bayesian approach to location estimation of mobile devices from mobile network operator data. In CBDS Working Paper 06-20. 2020

Annexes



(a) Biais (b) Variance

FIGURE 1 – Racine carrée du biais et de la variance. (L'unité est le mètre). Les cartes ont été élaborées à l'aide d'informations fournies par Orange décrivant l'organisation de leur réseau en 2019.



(a) Autour de Marseille (b) Focus autour du Vieux-Port

FIGURE 2 – Exemple de grille adaptative. Les cartes ont été élaborées à l'aide d'informations fournies par Orange décrivant l'organisation de leur réseau en 2019.