Estimation de la précision spatiale des données de téléphonie mobile

Approche bayésienne

Milena Suarez Castillo et François Sémécurbe Marouchi

Institut National de Statistique et des Études Économiques

30 Mars 2022

Journées de Méthodologie Statistique

Spatialisation : Présentation du problème

On dispose de la population localisée au niveau des antennes que l'on souhaite répartir dans l'espace :

Antenne	Pop_{0h}	Pop_{8h}	Pop_{15h}	Pop_{20h}
1	6	837	687	9
2	107	78	32	107
3	102	101	102	104

On considère que la population est issue d'un processus présentant une double indépendance (ce qui est évidemment une grossière simplification):

- indépendance temporelle
- indépendance spatiale

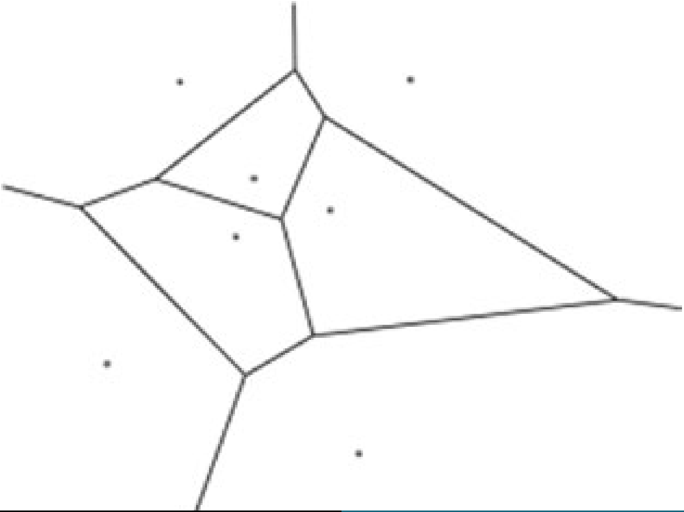
Spatialisation : Présentation du problème

- Sous les conditions précédentes d'indépendance, on peut traiter les heures et les antennes séparément.
- La **spatialisation d'une antenne** j consiste à répartir sa population à l'aide d'une nappe spatiale $S_{.j}$.
- Sans perte de généralité, on peut considérer que l'espace du géographe est discrétisé en un carroyage.
- Dans ce cas, la population dans le carreau i, Pop_i, est donnée tout simplement par la formule suivante :

$$Pop_i = \sum_{j \in J} S_{ij} \times Pop_j$$

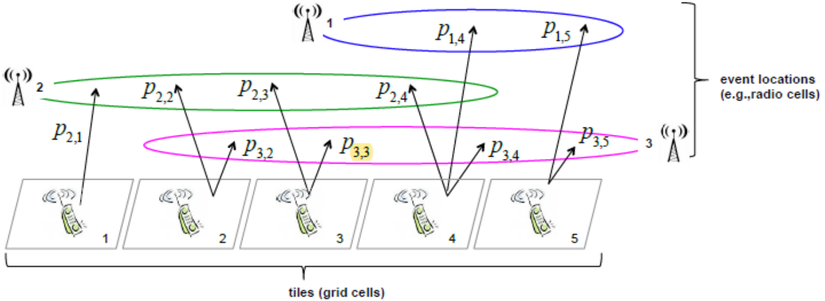
Où S_{ij} est la part de population observée à l'antenne j que l'on affecte au carreau i et Pop_j est la population de l'antenne j.

Spatialisation d'une antenne S_{ij} - Voronoi une approche *Old School*



Un peu de réseau et de probabilité mais pas trop !

En réalité, les couvertures des antennes se superposent. Dans un même carreau, un téléphone peut être rattaché à différentes antennes. Ceci se modélise sous la forme d'une carte de probabilité P_{ji}.



Overlapping event locations (Fabio Ricciato et al. 2020)

Un premier estimateur, la nappe bayésisenne

Tennekes 2018 propose d'utiliser la formule de Bayes pour obtenir une spatialisation des événements associés aux antennes :

$$Q_{ij} = \frac{P_{ji}\pi_i}{\sum_{i_0} P_{ji_0}\pi_{i_0}} \tag{1}$$

- Q_{ij} est la probabilité a posteriori de se trouver dans le carreau i sachant que l'on est observé au niveau de l'antenne j.
- Le π_i est une information auxiliaire qui décrit la répartition probable de notre population : Population Filosofi, Bâti, Routes, Camping, Bar, salles de sport...
- On peut adapter le π_i à la période (week-end, semaine) et aux heures.

Quel est le meilleur estimateur ?

- On suppose que la distribution de population est supposée connue et correspond à la répartition de l'a priori π_i.
- La localisation d'une personne est tirée aléatoirement selon la loi de probabilité de l'a priori.
- La précision de l'estimation de sa localisation $X_{g(P,j)}$ étant donné sa présence mesurée sur le réseau mobile en j est mesurée à l'aide du risque quadratique (Minimum Square Error):

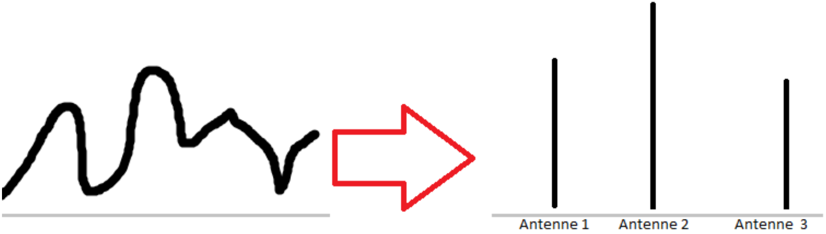
$$MSE = E_\pi(||X_{g(P,j)} - X_i||^2)$$

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

Quel est le meilleur estimateur ?

- La solution linéaire g qui minimise le risque quadratique est la moyenne a posteriori, $X_{g(P,j)} = \sum_i Q_{ij} X_i$. Un individu détecté à l'antenne j est positionné au barycentre de celle-ci.
- **Problème :** La population est répartie dans 8 millions de carreaux de $100m^2$ alors qu'en moyenne les opérateurs ont 200000 antennes.

Répartition de population et population répartie aux barycentres des antennes



Se restreindre aux estimateurs "non biaisés"

On peut restreindre les solutions linéaires R parmi celles qui suivent les deux propriétés suivantes :

- $RP\pi = \pi$, en moyenne l'estimateur reproduit la distribution de population proposée par l'a priori.
- Le barycentre d'une nappe $R_{.,j}$ correspond au barycentre de l'antenne j. Autrement dit, la nappe $R_{.,j}$ est centrée sur la solution optimale.

La nappe bayésienne respecte ces deux conditions. Dans la suite de la présentation, les résultats sont calculés pour une nappe bayesienne calculée à l'aide a carte de couverture d'Orange FluxVision dont la résolution est de $100m$ et d'un a priori uniforme.

Précision localisée

La précision locale est définit par l'espérance conditionnelle :

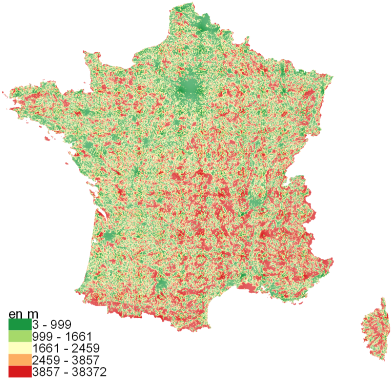
$$MSE_i = E_\pi(||X_{Q(c,i)} - X_i||^2|i)$$

D'après le théorème de Huygens, on peut décomposer le MSE en 2 termes :

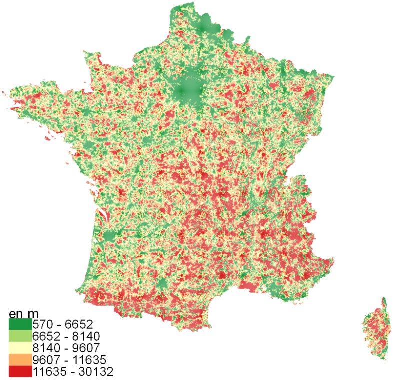
- Le biais qui décrit l'éloignement en moyenne du barycentre de la nappe X^{i_0} au carreau de référence i_0 $B_{i_0} = \|X_{i_0} - \bar{X}^{i_0}\|$
- La variance qui mesure la dispersion autour du barycentre $V_{i_0} = \sum_i N_{i_0}(i)\|X_i - \bar{X}^{i_0}\|^2$

Explorer l'espace des erreurs avec König-Huygens

Bias term

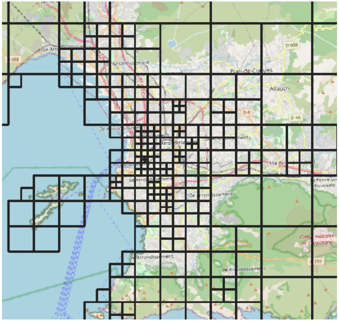


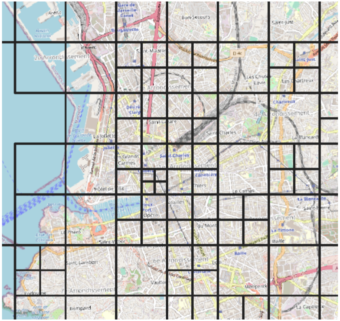
Variance term



Intégrer la précision dans la diffusion

Avec un découpage de 100 m, on a 55 000 000 carreaux, ce qui est beaucoup. On peut néanmoins utiliser nos *nappes à carreaux* N_{i_0} pour produire une grille adaptative à l'aide d'une récursion à travers les échelles !

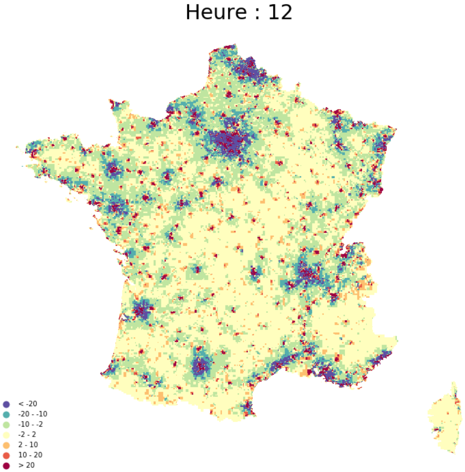




Plus les carreaux sont grands, moins la précision est bonne.
Autrement dit, on intègre la précision dans la diffusion !

Déroulée d'une journée (jeudi 28 mars)

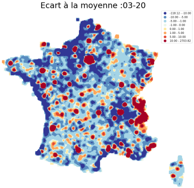
Ecart à la moyenne sur la journée (hab/km2)

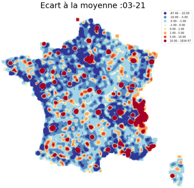


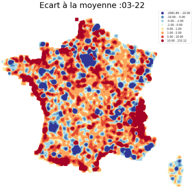
[Animations trop lourdes pour le site des archives JMS: veuillez contacter les auteurs si vous souhaitez les images]

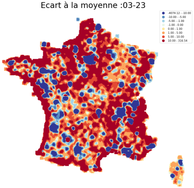
La fin de semaine, présence à 22h)

mercredi, jeudi, vendredi, samedi (écart à la moyenne sur 2 semaines)



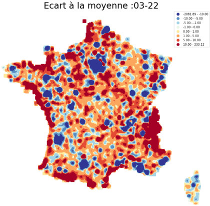


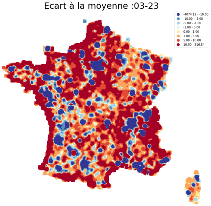


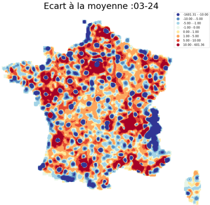


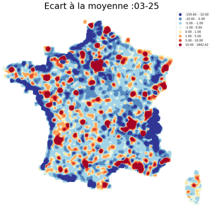
Le weekend, présence à 22h

vendredi, samedi, dimanche, lundi



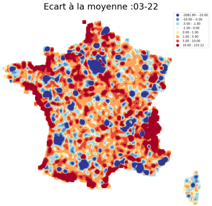


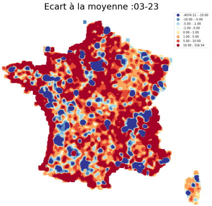


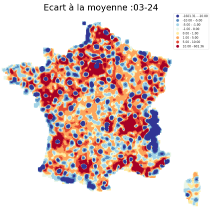


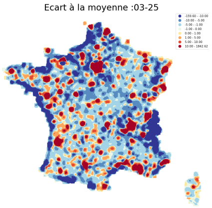
Le weekend, présence à 22h

vendredi, samedi, dimanche, lundi





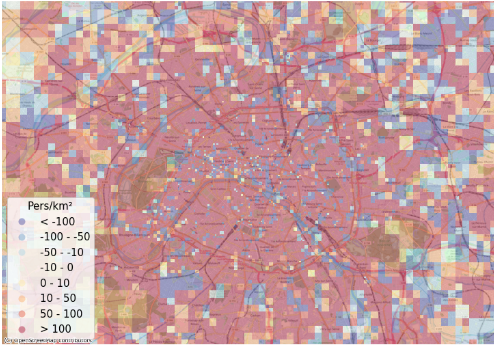




Et localement pour une journée-type

[Animations trop lourdes pour le site des archives JMS: veuillez contacter les auteurs si vous souhaitez les images]

Et localement pour une semaine-type



[Animations trop lourdes pour le site des archives JMS: veuillez contacter les auteurs si vous souhaitez les images]