Échantillonnage spatial via des distances socio-économiques

Comparaison de méthodes pour le tirage d'un Échantillon-Maître

Samuel Givois, Thomas Merly-Alpa

Journées de Méthodologie Statistique 2018

Mercredi 13 juin 2018

1. Contexte

2. Méthodes envisagées

3. Comparaison des méthodes

4. Résultats

5. Conclusion

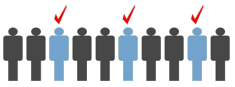
Contexte

Un nouvel échantillon-maître (EM)

- 10 ans de l'EM actuel, nouvelles sources de données
  - un nouveau projet, un nouvel EM
- Dernier EM : tirage équilibré région par région
- Essais avec cube local, gains a priori [7]
- Explorer plus profondément les méthodes spatiales

Méthodes envisagées

Attentes théoriques



Estimateur Narain-Horvitz-Thompson du total : $\hat{t_y} = \sum_{k \in S} \frac{y_k}{\pi_k}$

Équilibrage : $\sum_{k \in S} \frac{x_k}{\pi_k} = \sum_{k \in U} x_k$

Existence d'un modèle : $y = {}^t x.\beta + \epsilon$

où $\mathrm{cov}(\epsilon_k, \epsilon_l) = \sigma_{kl}$

Décomposition de la variance anticipée :

$$E_p E_M[(\hat{t_y} - t_y)^2] = \sum_{k \in U} \sum_{l \in U} \sigma_{kl} \frac{\pi_{kl} - \pi_k \pi_l}{\pi_k \pi_l}$$

Méthodes

- Cube : équilibrage région par région [6, 13]
- Cube « stratifié » : équilibrage national [2]
- Pivot local : répartition spatiale [11]
- Cube local : équilibrage et répartition spatiale [12]
- Répartition de l'information auxiliaire [10]

Distances

- Distance euclidienne :

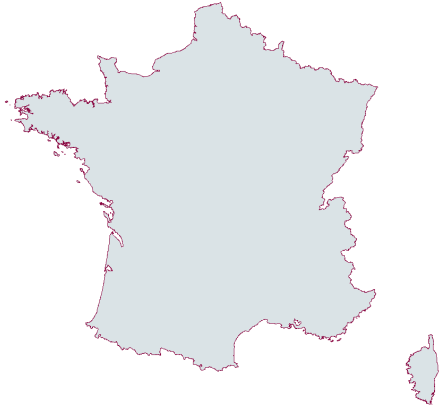
$$d_{kl} = |x_k - x_l|_2$$

- Distance dans un hyperplan factoriel issu d'une ACP [15]
- Distance de Mahalanobis [16] :

$$d_{kl} = \sqrt{{}^t(x_k - x_l)\Sigma^{-1}(x_k - x_l)}$$

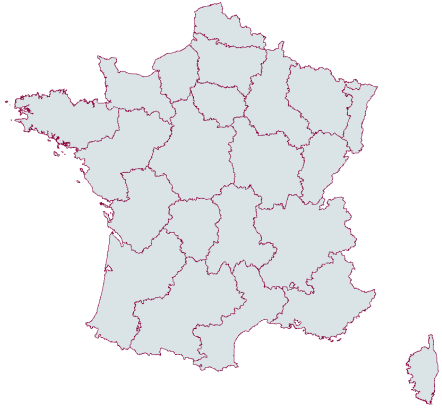
Comparaison des méthodes

Unités primaires de test



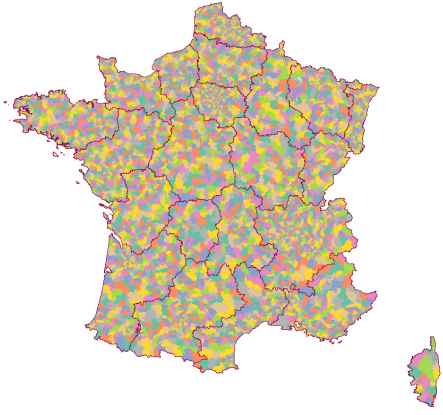
- France métropolitaine
- 22 anciennes régions
- 5213 unités primaires [8]

Unités primaires de test



- France métropolitaine
- 22 anciennes régions
- 5213 unités primaires [8]

Unités primaires de test

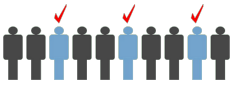


- France métropolitaine
- 22 anciennes régions
- 5213 unités primaires [8]

Variables auxiliaires [13, 7]

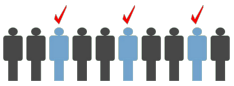
- Probabilités d'inclusion
- Territoire (rural, urbain, ...)
- Revenu
- Âge
- Composition du ménage
- Propriétaires de leur logement
- Nombre de HLM

Variables de « distance »



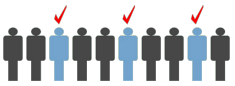
- Résidences principales, secondaires, ...
- Niveau de diplôme
- Catégorie socioprofessionnelle

Variables d'intérêt



- Logement détaillé
- Revenu détaillé
- Genre
- Élections présidentielles
- Naissances et décès

Critères de comparaison



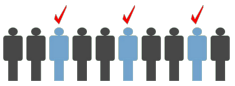
Estimation du total :

- Biais
- Coefficient de variation
- Erreur quadratique moyenne

Au-delà :

- Quantiles [14]
- Variance de l'indicateur de Voronoï

Estimations des performances



5 méthodes $\times$ 36 distances $+$ 2 méthodes

4 tailles d'échantillons

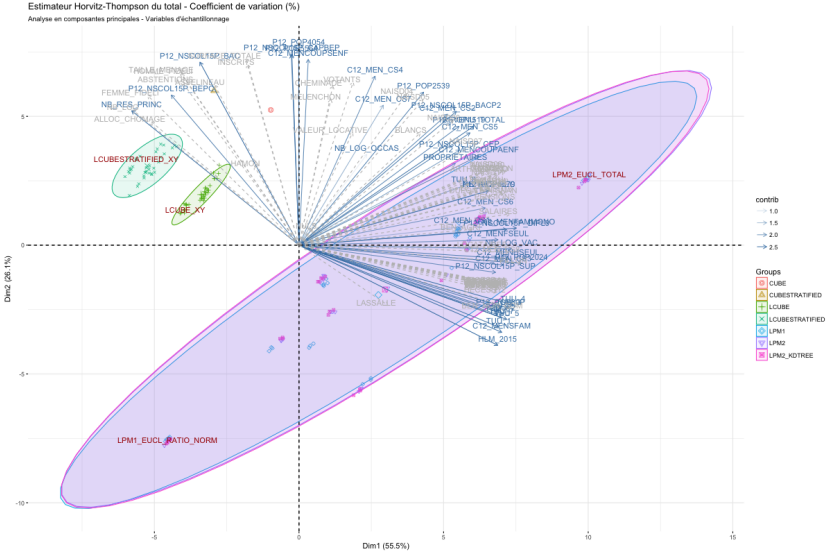
- $\mathbf{n} = 567$ (100 %)
- $\mathbf{n} = 508$ (87,5 %), $\mathbf{n} = 439$ (75 %), $\mathbf{n} = 306$ (50 %)

Estimation par Monte-Carlo

- 10000 simulations

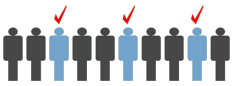
Résultats

Comparaison globale



Conclusion

Principaux résultats



- Indépendant de la taille d'échantillon
- Cube local meilleur que le cube « simple »
- Pivot local plus variable
- Équilibrage national paraît préférable

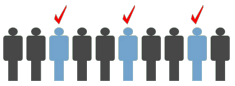
Préconisation : cube local, équilibrage national (« stratifié »), distance géographique (X, Y)

Aller plus loin

- Choix des variables
  - Beaucoup de variables
  - Territoire difficile à capter
  - Disponibilité des données
- Évolution démographique

Merci de votre attention

Bibliographie I



Pascal Ardilly et Yves Tillé :
Multi-stage sampling.
Sampling Methods : Exercises and Solutions, pages 159–207, 2006.

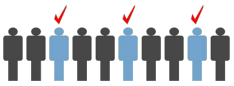
Guillaume Chauvet :
Stratified balanced sampling.
Survey Methodology, 35(1):115–119, 2009.

Guillaume Chauvet et Yves Tillé :
A fast algorithm for balanced sampling.
Computational Statistics, 21(1):53–62, 2006.

Marc Christine et Sébastien Faivre :
Octopusse : un système d'échantillon-maître pour le tirage des échantillons dans la dernière enquête annuelle de recensement.
Actes des Journées de Méthodologie Statistique de 2009, 2009.

Jean-Claude Deville et Yves Tille :
Unequal probability sampling without replacement through a splitting method.
Biometrika, 85(1):89–101, 1998.

Bibliographie II



Jean-Claude Deville et Yves Tillé :
Efficient balanced sampling : the cube method.
Biometrika, 91(4):893–912, 2004.

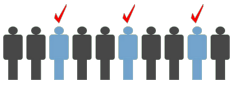
Cyril Favre-Martinoz et Thomas Merly-Alpa :
Utilisation des méthodes d'échantillonnage spatialement équilibré pour le tirage des unités primaires des enquêtes ménages de l'insee.
SFDS - 9ème Colloque Francophone sur les Sondages, 2016.

Cyril Favre-Martinoz et Thomas Merly-Alpa :
Constitution et tirage d'unités primaires pour des sondages en mobilisant de l'information spatiale.
SFDS - 49èmes Journées de Statistique, 2017.

A Grafström et J Lisic :
Balancedsampling : balanced and spatially balanced sampling. r package version 1.5.2, 2016.

Anton Grafström et Niklas LP Lundström :
Why well spread probability samples are balanced.
Open Journal of Statistics, 3(1):36–41, 2013.

Bibliographie III



Anton Grafström, Niklas LP Lundström et Lina Schelin :
Spatially balanced sampling through the pivotal method.
Biometrics, 68(2):514–520, 2012.

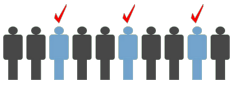
Anton Grafström et Yves Tillé :
Doubly balanced spatial sampling with spreading and restitution of auxiliary totals.
Environmetrics, 24(2):120–131, 2013.

Fabien Guggemos :
Simulations de tirages de zones d'action pour les enquêtes de l'insee.
Actes des Journées de Méthodologie Statistique de 2009, 2009.

Rob J Hyndman et Yanan Fan :
Sample quantiles in statistical packages.
The American Statistician, 50(4):361–365, 1996.

Ronan Le Gleut :
Analyse factorielle et sondage - utilisation de méthodes d'échantillonnage spatial.
SFDS - 49èmes Journées de Statistique, 2017.

Bibliographie IV

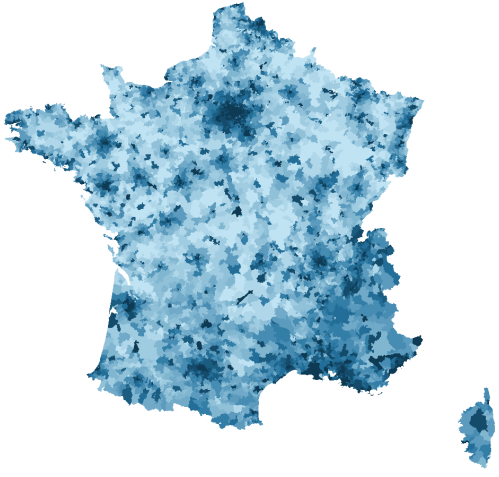


Goeffrey J McLachlan :
Mahalanobis distance.
Resonance, 4(6):20–26, 1999.

Yves Tillé et Matthieu Wilhelm :
Probability sampling designs : Principles for choice of design and balancing.
Statistical Science, 32(2):176–189, 2017.

Autocorrélation spatiale

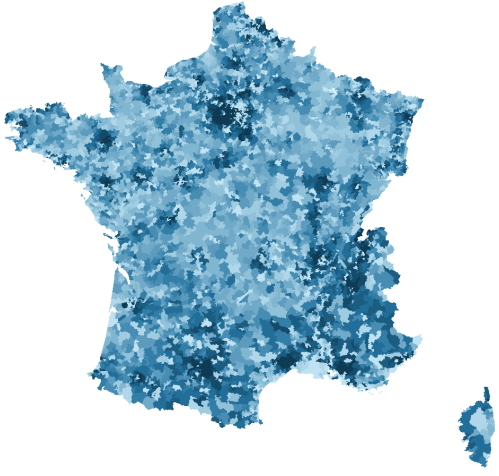
Autocorrélation spatiale



Nombre de diplômés du supérieur long :

- y_k
- $\hat{\epsilon}_k = y_k - \hat{\beta}_0 - \hat{\beta}_1 . \pi_k$
- $\hat{\epsilon}_k = y_k - {}^t x_k . \hat{\beta}$

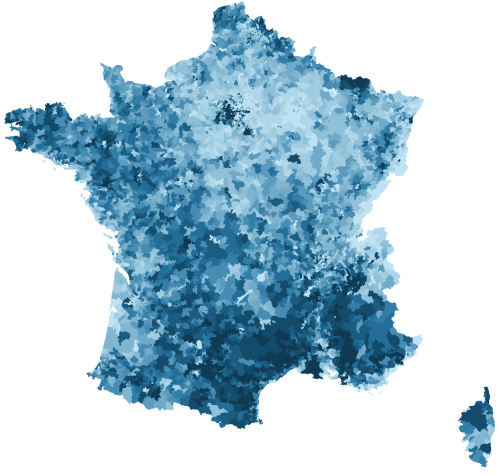
Autocorrélation spatiale



Nombre de diplômés du supérieur long :

- y_k
- $\hat{\epsilon}_k = y_k - \hat{\beta}_0 - \hat{\beta}_1.\pi_k$
- $\hat{\epsilon}_k = y_k - {}^t x_k.\hat{\beta}$

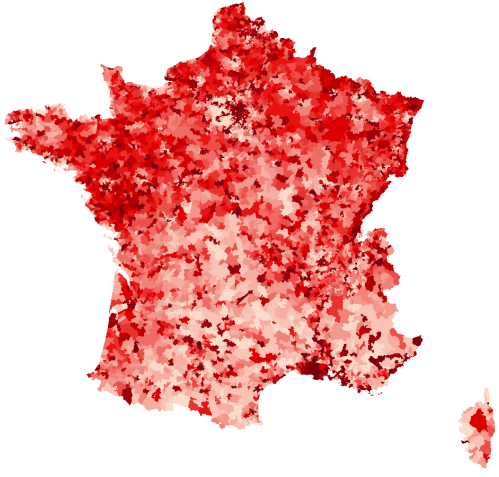
Autocorrélation spatiale



Nombre de diplômés du supérieur long :

- y_k
- $\hat{\epsilon}_k = y_k - \hat{\beta}_0 - \hat{\beta}_1 . \pi_k$
- $\hat{\epsilon}_k = y_k - {}^t x_k . \hat{\beta}$

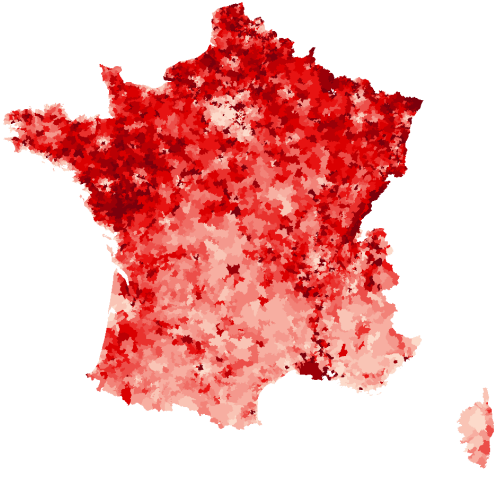
Autocorrélation spatiale



Nombre d'ouvriers :

- y_k
- $\hat{\epsilon}_k = y_k - \hat{\beta}_0 - \hat{\beta}_1.\pi_k$
- $\hat{\epsilon}_k = y_k - {}^t x_k.\hat{\beta}$

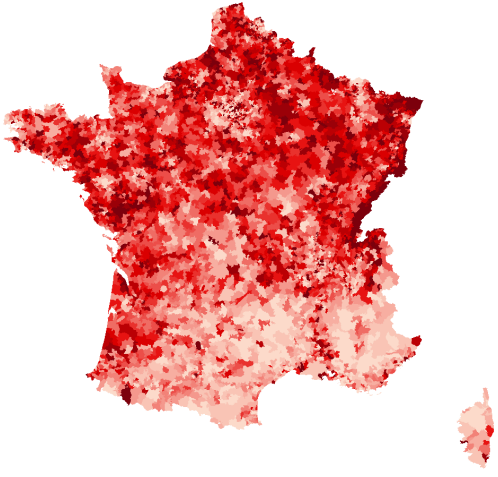
Autocorrélation spatiale



Nombre d'ouvriers :

- y_k
- $\hat{\epsilon}_k = y_k - \hat{\beta}_0 - \hat{\beta}_1.\pi_k$
- $\hat{\epsilon}_k = y_k - {}^t x_k.\hat{\beta}$

Autocorrélation spatiale



Nombre d'ouvriers :

- y_k
- $\hat{\epsilon}_k = y_k - \hat{\beta}_0 - \hat{\beta}_1.\pi_k$
- $\hat{\epsilon}_k = y_k - {}^t x_k.\hat{\beta}$