Tirage spatialement équilibré d'un Échantillon-Maître

Cyril Favre-Martinoz, **Thomas Merly-Alpa (INSEE)**

8èmes journées MAASC - Nantes

8 novembre 2016

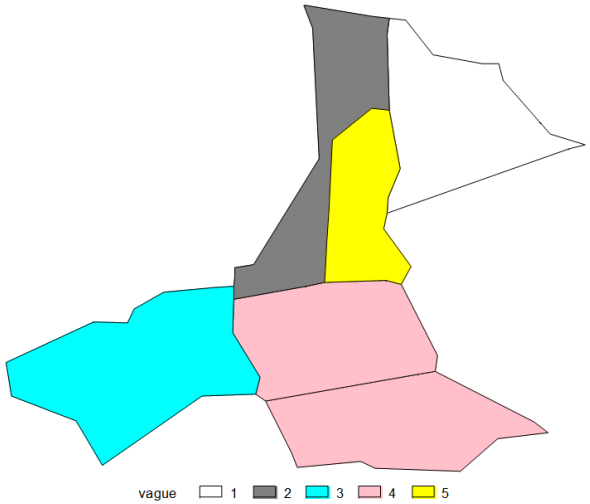
Sommaire

Sommaire

L'EM actuel

Une ZAE (Zone d'Action Enquêteur) est :

- une grande commune ;
- ou un regroupement de petites communes.



Les données proviennent du recensement rotatif (système OCTOPUSSE).

L'EM actuel

- Tirage équilibré des ZAE sur quelques variables auxiliaires
  - Nombre de résidences principales et revenu fiscal par groupe de rotation ;
  - Nombre de résidences principales par type d'espace : rural, périurbain, urbain ;
  - Davantage de variables d'équilibrage en Ile-de-France ;
  - Un total de 488 ZAE non exhaustives tirées.

Contexte

Projet NAUTILE :

- Cycle d'EM pour 10 ans : entrée en vigueur des nouvelles zones en 2020 ;
- Abandon de la base du recensement de la population ;
- Système issu des fichiers fiscaux (taxes locales, déclarations de revenus) ;
- Plus de liberté pour mettre en oeuvre des techniques de sondage poussées.

Les objectifs

- Tirage d'un échantillon d'unités primaires selon trois critères de qualité
  - Une précision en termes d'EQM raisonnable ;
  - Une proximité avec le réseau d'enquêteurs raisonnable, afin de limiter les coûts de collecte ;
  - Un tirage spatialement réparti pour...
    - limiter la corrélation spatiale, synonyme de perte de précision au niveau global ;
    - une meilleure couverture de l'espace national ;
    - faciliter l'affectation des enquêteurs.

Sommaire

Méthodes de tirage spatialement réparties

- Références :
  - *Grafström, A. (2012). Spatially correlated Poisson sampling. Journal of Statistical Planning and Inference, 142(1), 139-147.*
  - *Grafström, A., Lundström, N. L., & Schelin, L. (2012). Spatially balanced sampling through the pivotal method. Biometrics, 68(2),514-520.*
  - *Grafström, A., & Tillé, Y. (2013). Doubly balanced spatial sampling with spreading and restitution of auxiliary totals.Environmetrics, 24(2), 120-131.*

Principe du tirage spatialement équilibré

- En deux étapes
- p le nombre de contraintes d'équilibrage
- On constitue des clusters de $p+1$ unités
- On lance une phase de décollage du cube sur ce cluster, on met à jour les probabilités d'inclusion localement puis re-clustering...
- Lorsqu'il reste p unités à tirer, on applique la phase d'atterrissage.
- Implémenté en R par Grafström dans le package *BalancedSampling*
- Contrairement à la macro Cube, la phase d'atterrissage ne peut se faire qu'en supprimant successivement les contraintes d'équilibrage (peut être modifié)

Sommaire

Type de plans

- Plusieurs tailles d'échantillon : $n = 488, 427, 366$
- Niveau du tirage et de l'équilibrage : grandes régions actuelles
- Comparaison entre le tirage équilibré spatialement et non équilibré spatialement (cube classique)
- La marge pour la comparaison : la précision sur l'EM actuel.
- Alternative (non testée) : Tirage avec une phase d'équilibrage au niveau régional suivi d'une phase d'atterrissage regroupée au niveau national

Variables d'équilibrage

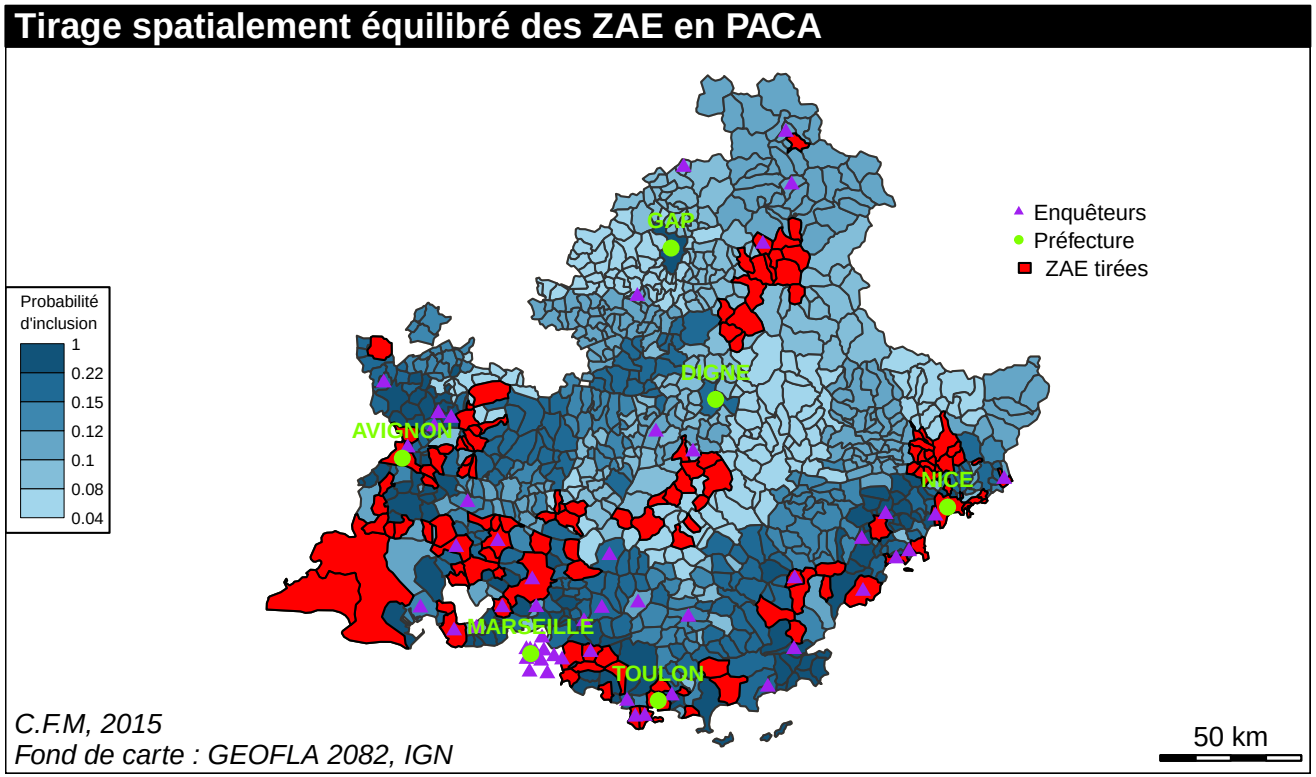
- Nombre de résidences principales
- Type de zonage urbain : rural, périurbain et urbain
- 3 classes d'âge : moins de 20 ans, de 20 à 59 ans et 60 ans et plus
- Nombre d'étrangers, de familles monoparentales, de familles de grande taille et de propriétaires
- Nombre de logements HLM

Résultats en termes de répartition spatiale

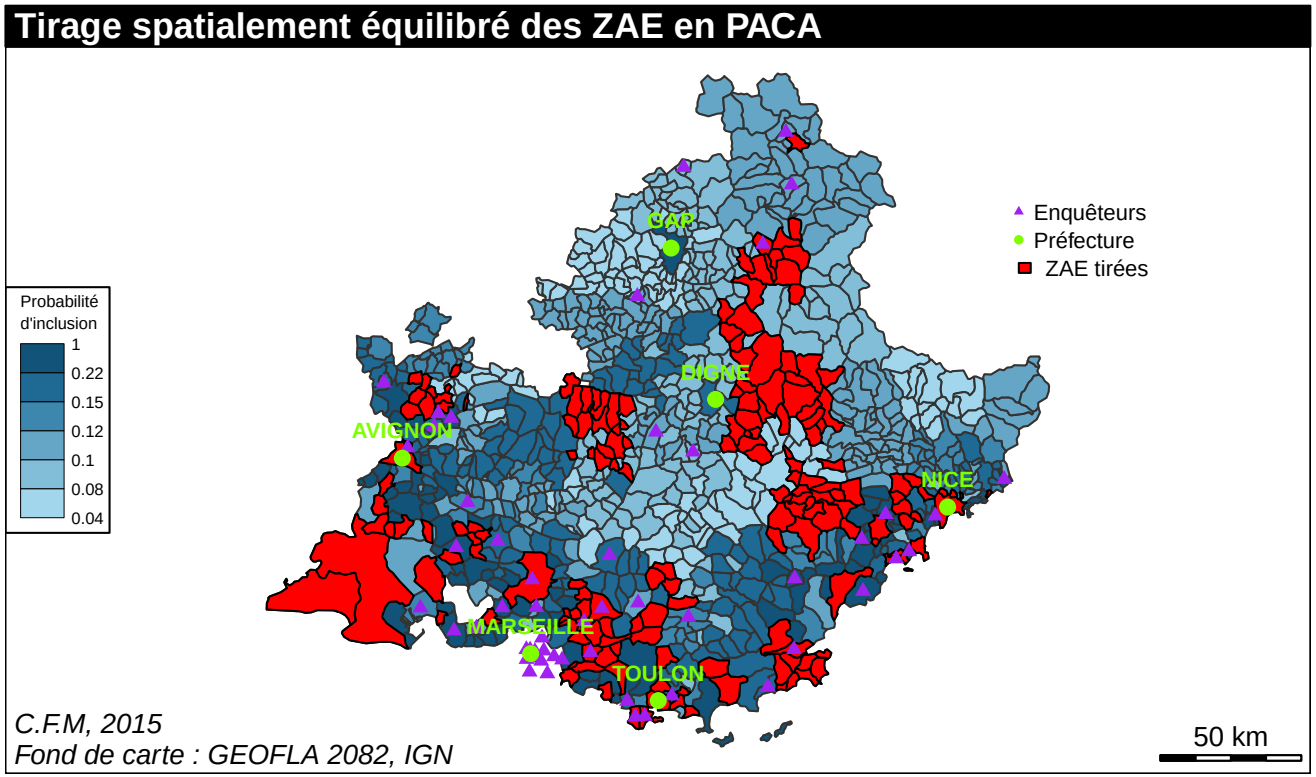
Critères de répartition spatiale :

- A l'oeil nu, en essayant de regarder la présence de clusters à l'issue du tirage ;
- Polygones de Voronoi ;
- D'autres critères : Inertie sur les coordonnées géographiques ?

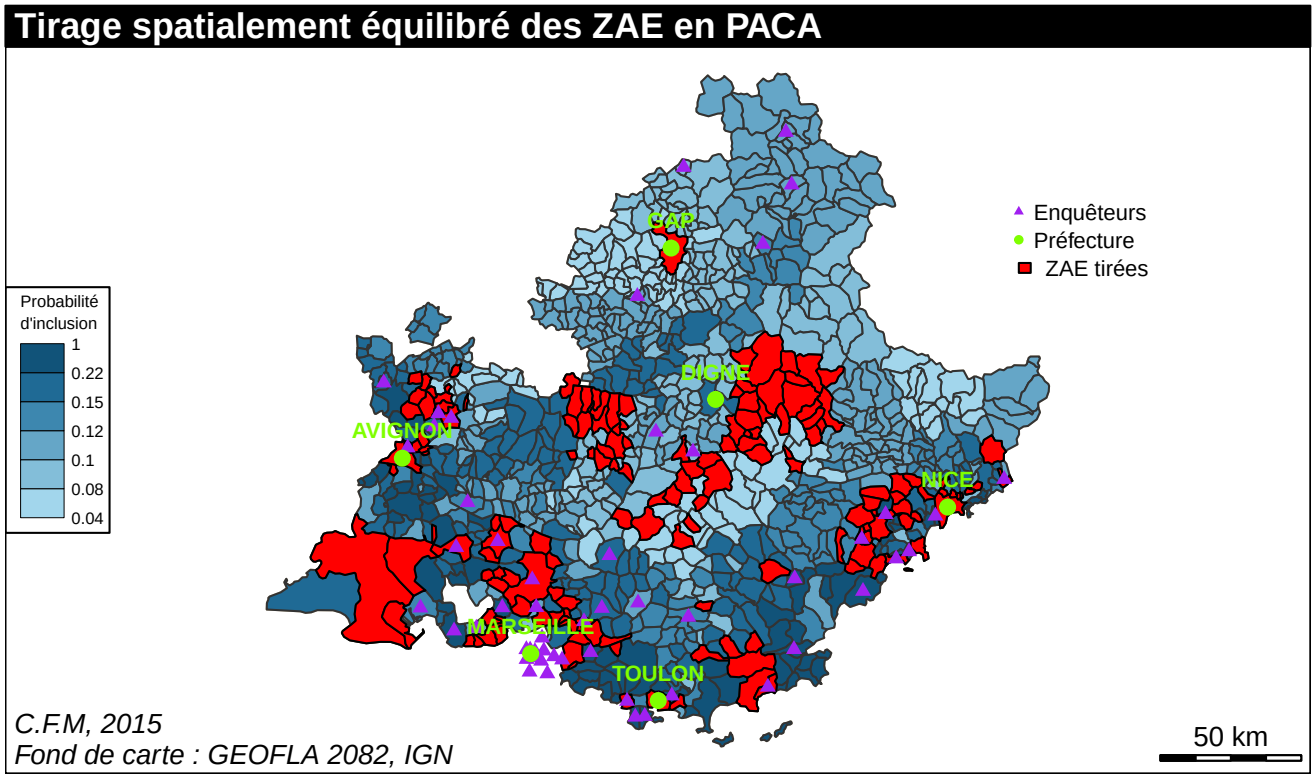
Répérage à "l'oeil nu" des clusters d'UP



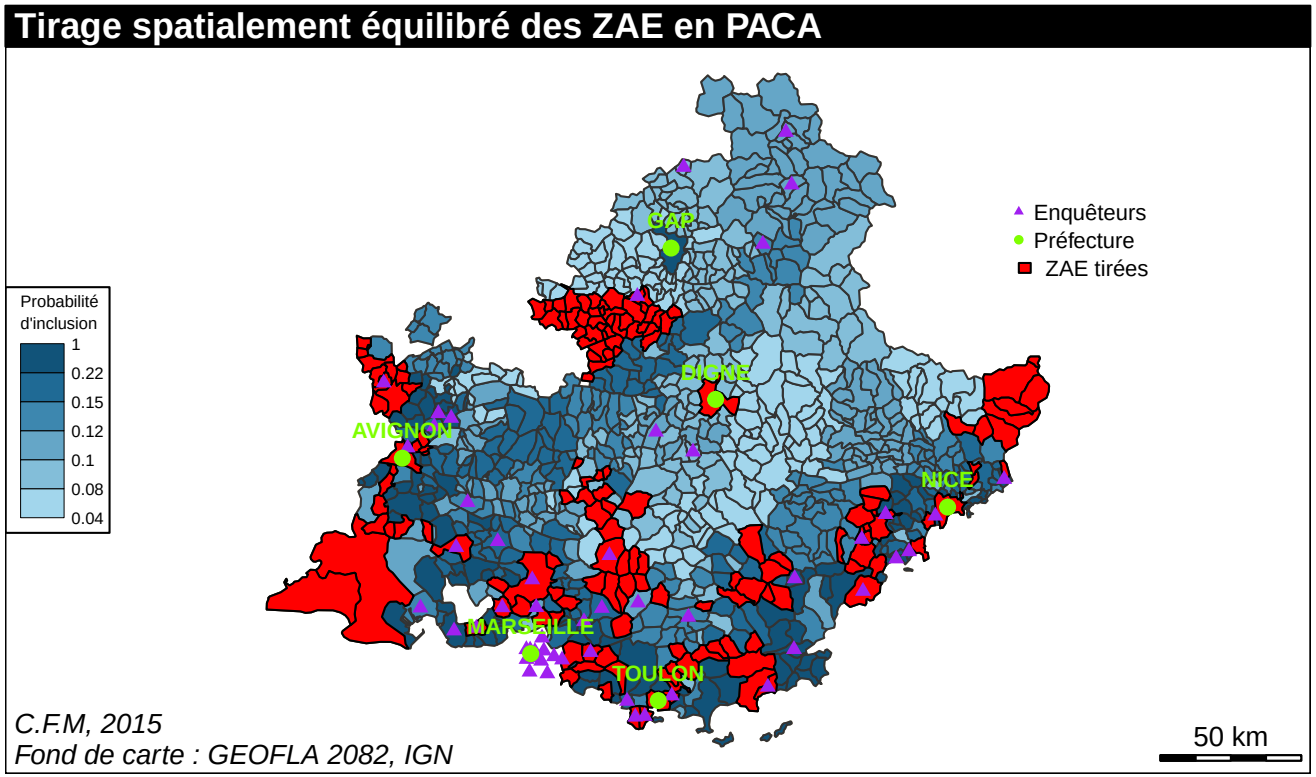
Répérage à "l'oeil nu" des clusters d'UP



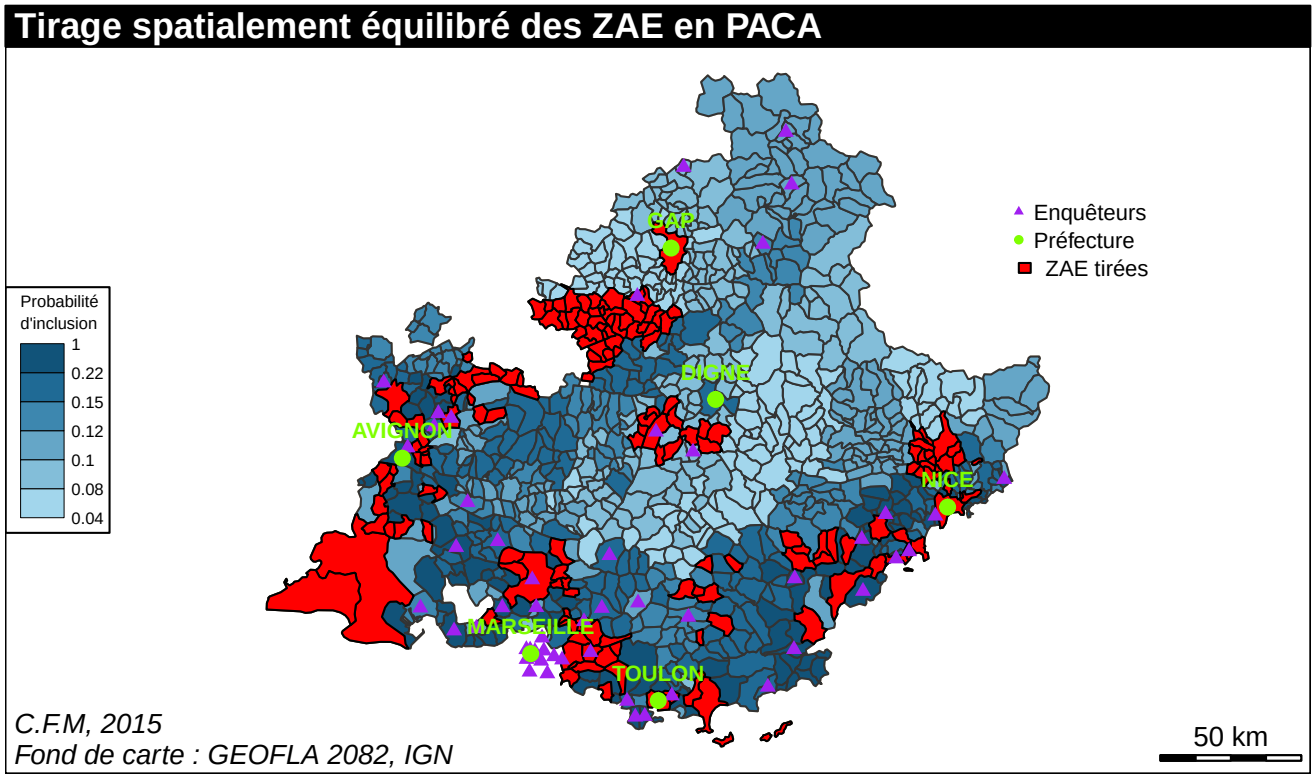
Répérage à "l'oeil nu" des clusters d'UP



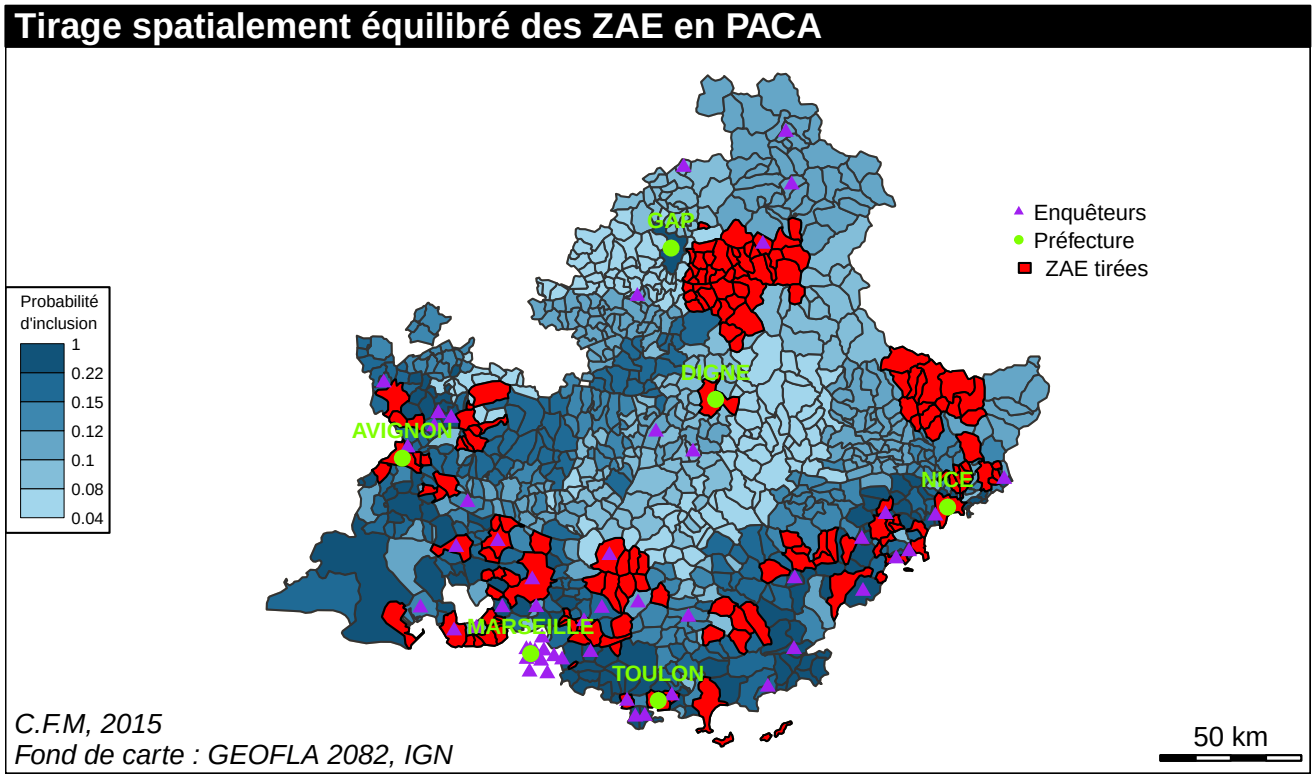
Répérage à "l'oeil nu" des clusters d'UP



Répérage à "l'oeil nu" des clusters d'UP



Répérage à "l'oeil nu" des clusters d'UP

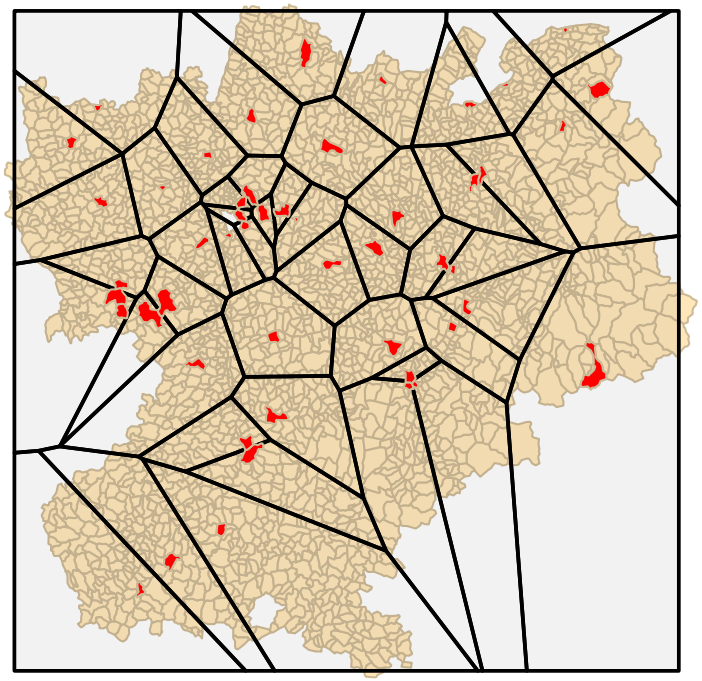


Les polygones de Voronoi

- Découpage de l'espace qui regroupe l'ensemble des points du plan plus proches d'une unité primaire tirée que de toutes les autres unités primaires tirées.
- On définit δ_i comme la somme des probabilités d'inclusion des UP incluses dans le polygone i.
- Si le plan est spatialement équilibré, on devrait avoir $\delta_i \approx 1 \ \forall i$
- On définit l'indicateur de Voronoi Δ par :

$$\Delta = \frac{1}{n} \sum_{i}^{n} (\delta_i - 1)^2$$

Les polygones de Voronoi : région Rhône-Alpes-Auvergne



Résultats en termes d'équilibrage spatial

	$n = 488$		$n = 427$		$n = 366$	
	Cube	Cube spatial	Cube	Cube spatial	Cube	Cube spatial
Δ	0.35	0.27	0.48	0.38	0.73	0.59

TABLE 1 – Valeur de l'indicateur de Voronoi pour différentes tailles d'échantillon

Résultats en termes de précision

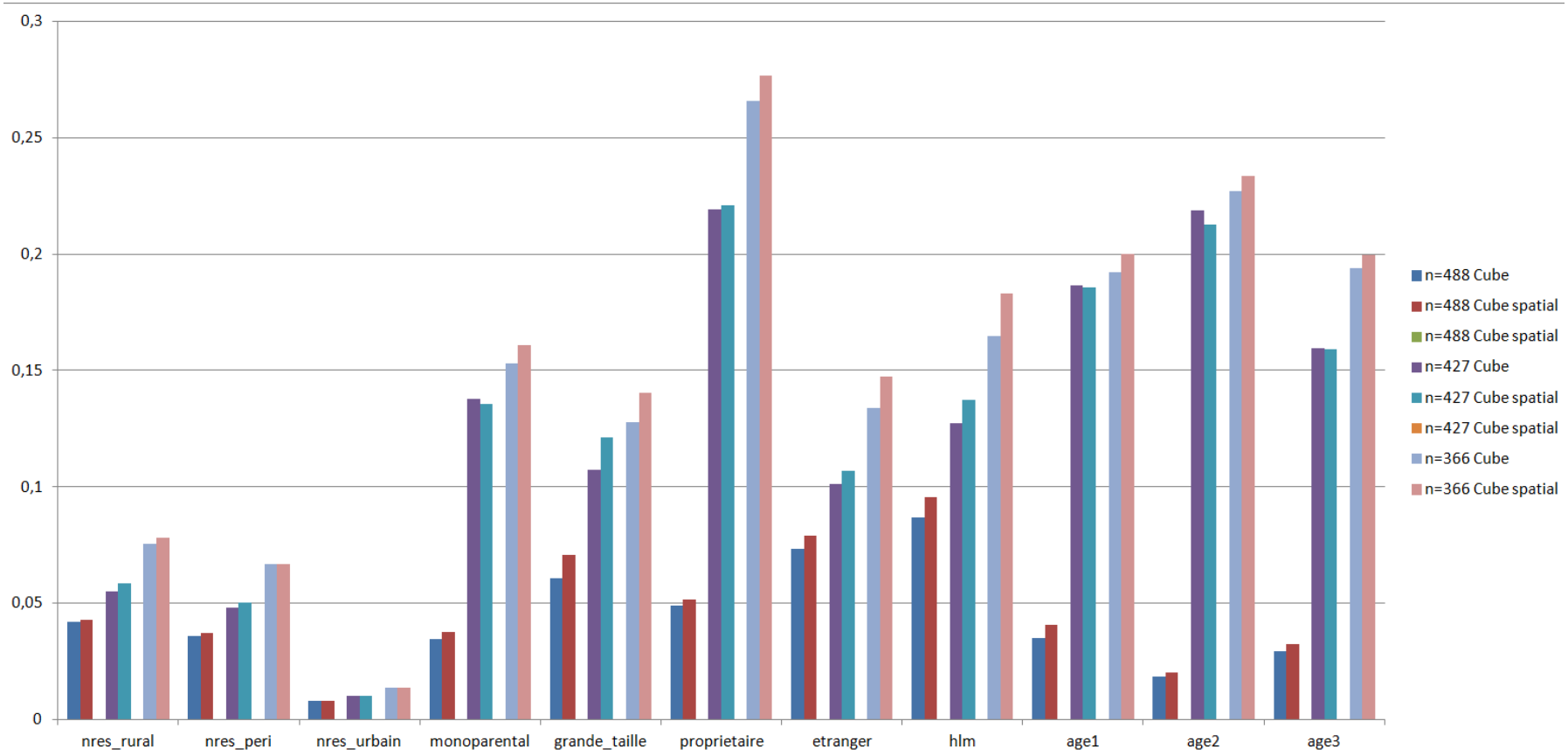


FIGURE 1 – Ratio de l'EQM pour le tirage considéré par rapport à l'EQM de l'EM actuel pour les variables auxiliaires

Résultats en termes de précision

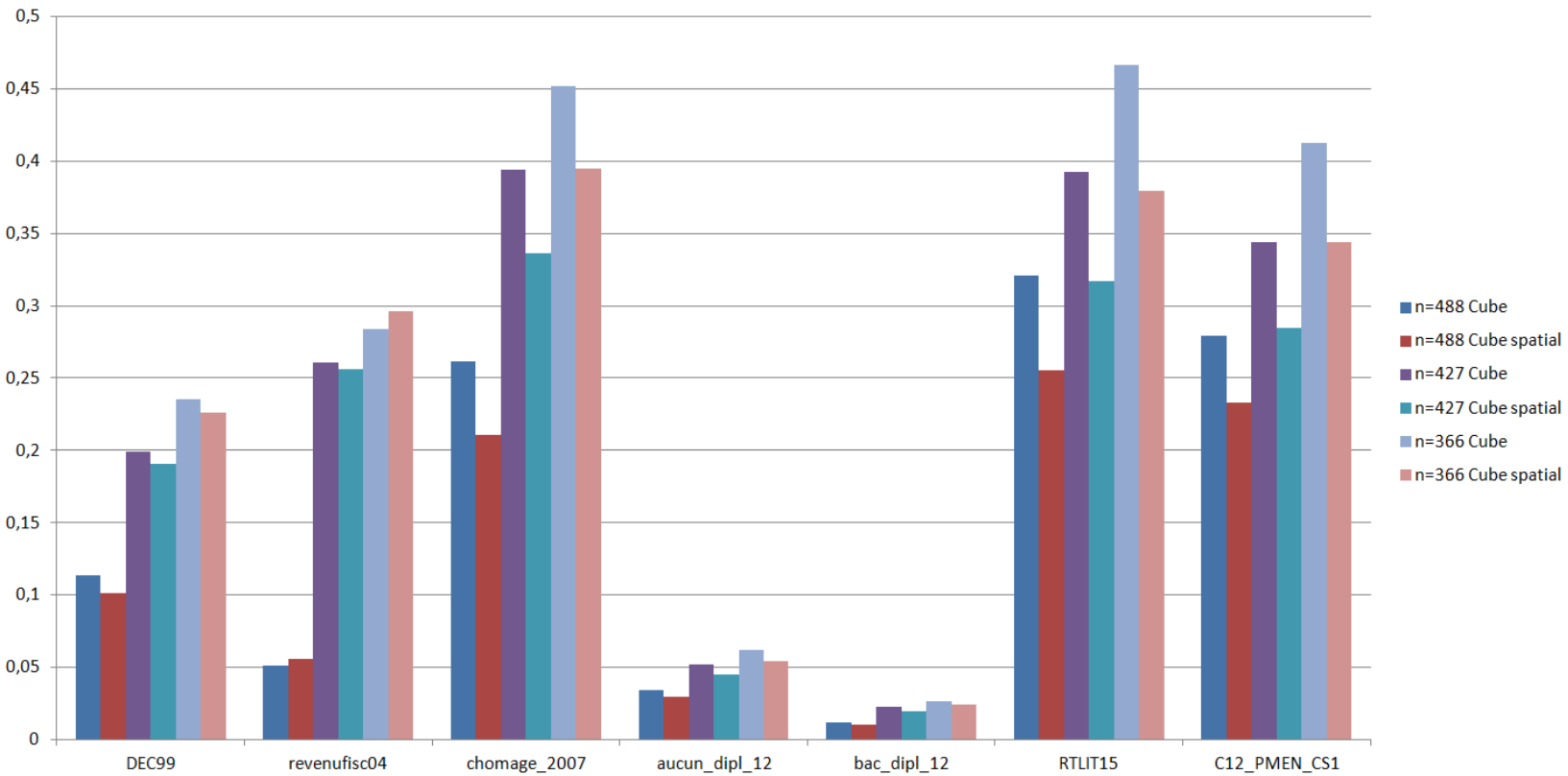


FIGURE 2 – Ratio de l'EQM pour le tirage considéré par rapport à l'EQM de l'EM actuel pour différentes variables d'intérêt

Sommaire

Calcul de précision d'un EM

- On souhaite connaître la variance liée au tirage de premier degré d'un échantillon maître, afin d'évaluer plus tard la précision des enquêtes réalisées.
- On veut calculer la variance issue d'un tirage spatialement équilibré.
- Cette problématique est centrale pour :
  - évaluer l'efficacité de la méthode de tirage ;
  - permettre la comparabilité avec les autres pays ;
  - fournir des estimations de précision robustes.

Calcul des $\pi_{i,j}$

- La méthode de tirage spatialement équilibré respecte les π_i ;
- Pour estimer la variance, il nous faut également les $\pi_{i,j}$:

$$\widehat{Var}(\hat{t}_{y\pi}) = \sum_{i \in S} \sum_{j \in S} \frac{y_i}{\pi_i} \frac{y_j}{\pi_j} \frac{\Delta_{ij}}{\pi_{ij}}. \quad (1)$$

  où $\Delta_{ij} = \pi_{ij} - \pi_i \pi_j$.
- On les estime par simulations ;
- Certains couples (i, j) sont trop rarement conjointement tirés et donc $\pi_{i,j} \approx 0$;
- Cela crée de la variabilité dans l'estimation de la variance ;
- On propose différents seuils en dessous desquels $\pi_{i,j}$ est supposé égal à 0 : $10^{-3}, 10^{-4}, 10^{-5}$.

Un autre estimateur

- Une autre piste est proposée dans l'article de Grafström et Tillé
- On combine deux estimateurs de variance :
  - l'estimateur de variance proposé par Stevens et Olsen (2003) dans le cas de tirages spatialement dispersés ;
  - l'estimateur de variance de Deville et Tillé (2005) utilisé dans le cas de plans de sondage équilibrés.
- L'estimateur s'écrit :

$$\hat{V}_{SB} = \frac{p+1}{p} \sum_{h=1}^{H} \frac{n_h}{n_h - p} \sum_{i \in S_h} (1 - \pi_i) \left(\frac{e_i}{\pi_i} - \bar{e_i} \right)^2$$

où e_i est le résidu et $\bar{e_i}$ la moyenne locale de e_j autour de i.

Résultats

- On calcule de façon empirique une variance Monte Carlo comme référence ;
- Les estimations via le calcul des $\pi_{i,j}$ sont :
  - plus stables quand le seuil est plus grand...
  - ...mais plus biaisées également !
- L'estimation $\hat{V}_{SB}$ a de bonnes performances en termes de biais et de robustesse, hormis pour les variables parfaitement expliquées par les variables d'équilibrage.

Sommaire

Résultats en termes de précision

- Pourquoi ces gains importants par rapport à l'EM actuel :
  - Une information disponible sur toute l'UP et non plus seulement sur les communes recensées ;
  - On introduit cinq fois moins de contraintes d'équilibrage ;
  - Cela permet d'ajouter davantage de variables d'équilibrage.
- Gain tirage équilibré vs non équilibré spatialement :
  - On s'affranchit de la corrélation spatiale ;
  - (x, y) : variable explicative de nombreux phénomènes (tourisme, etc) ;
  - On se prémunit du vieillissement de l'équilibrage.

Les questions en suspens

- Quel niveau de tirage envisagé : national, régional, départemental ?
- Combien d'unités primaires ?
- Des nouvelles unités primaires,
  - qui s'affranchissent de la contrainte des groupes de rotation du RP ;
  - plus homogènes entre elles et plus hétérogènes en intra ;
  - qui faciliteront le tirage équilibré.
- Tester d'autres algorithmes de tirage spatial : pivot spatial ?