Diffusion de données finement localisées

Ne laisser personne sur le carreau

Marc Branchu*, Vianney Costemalle*, Maëlle Fontaine*

Journées de méthodologie statistique, 14 juin 2018

Insee, Division des Méthodes et Référentiels Géographiques

Introduction

Introduction

- Respecter le **secret statistique**, c'est garantir l'impossibilité pour un intrus de deviner les données personnelles d'un individu "enquêté" (ménage, entreprise...).
- En pratique, cela prend souvent la forme d'un **seuil** à respecter.
- En général, les mailles de diffusion (Iris, communes) contiennent un nombre minimum d'observations ; le cas de **comptages faibles ou nuls** n'apparaît que lorsque l'on croise des variables.
- La diffusion de **données carroyées** modifie ce contexte.

Introduction

- Les **données carroyées** sont présentées comme un grand apport pour des analyses locales ([9], [8]) : zonages à façon possibles, maillage stable dans le temps, multisource possible.
- Mais elles présentent aussi un **risque de divulgation** plus élevé : en France, environ 80 % des carreaux de 200 m sont sous le seuil de 11 ménages.
- Pour autant elles sont logées à la même enseigne que les autres données tabulées concernant le secret.
- Le risque à utiliser les méthodes traditionnelles de confidentialité est de trop perturber les données et donc de leur enlever toute **utilité**.

⇒ Présentation et tests de deux méthodes spécifiquement développées par l'Insee en vue d'une diffusion carroyée de Filosofi

Diffusion de données finement localisées

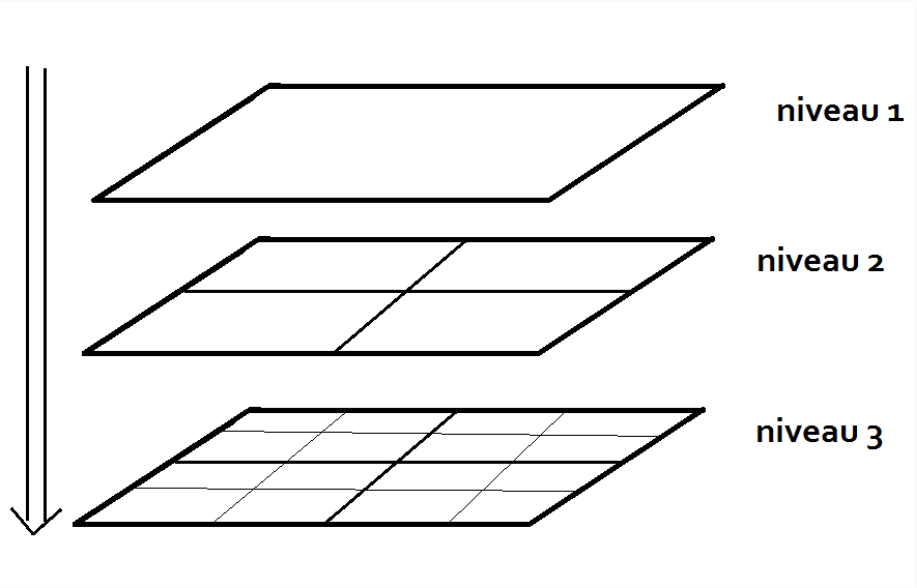
1. Méthode des grilles superposées
2. Méthode de swapping
3. Résultats

Méthode des grilles superposées

Méthode des grilles superposées

Principe général

- si le carreau contient trop peu d'observations, il est agrégé avec ses carreaux voisins ;
- plusieurs grilles simples emboîtées, de plus en plus fines, par exemple : 250m - 500m - 1km - 2km - 4km - 8km - 16km - 32 km ;



- l'information est alors diffusée avec plusieurs niveaux de précision.

Méthode des grilles superposées

Plus la maille est fine, moins on a de carreaux diffusables

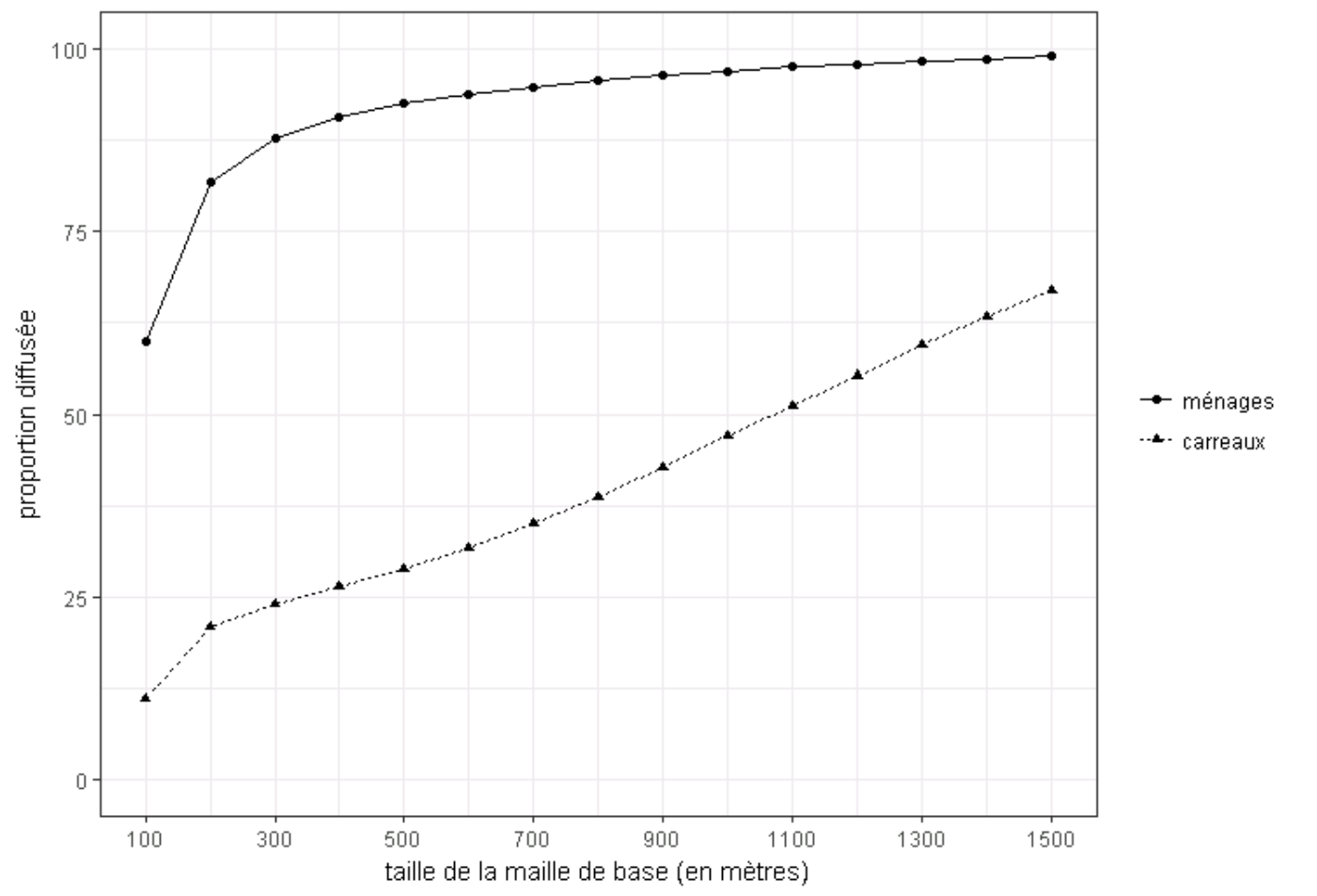
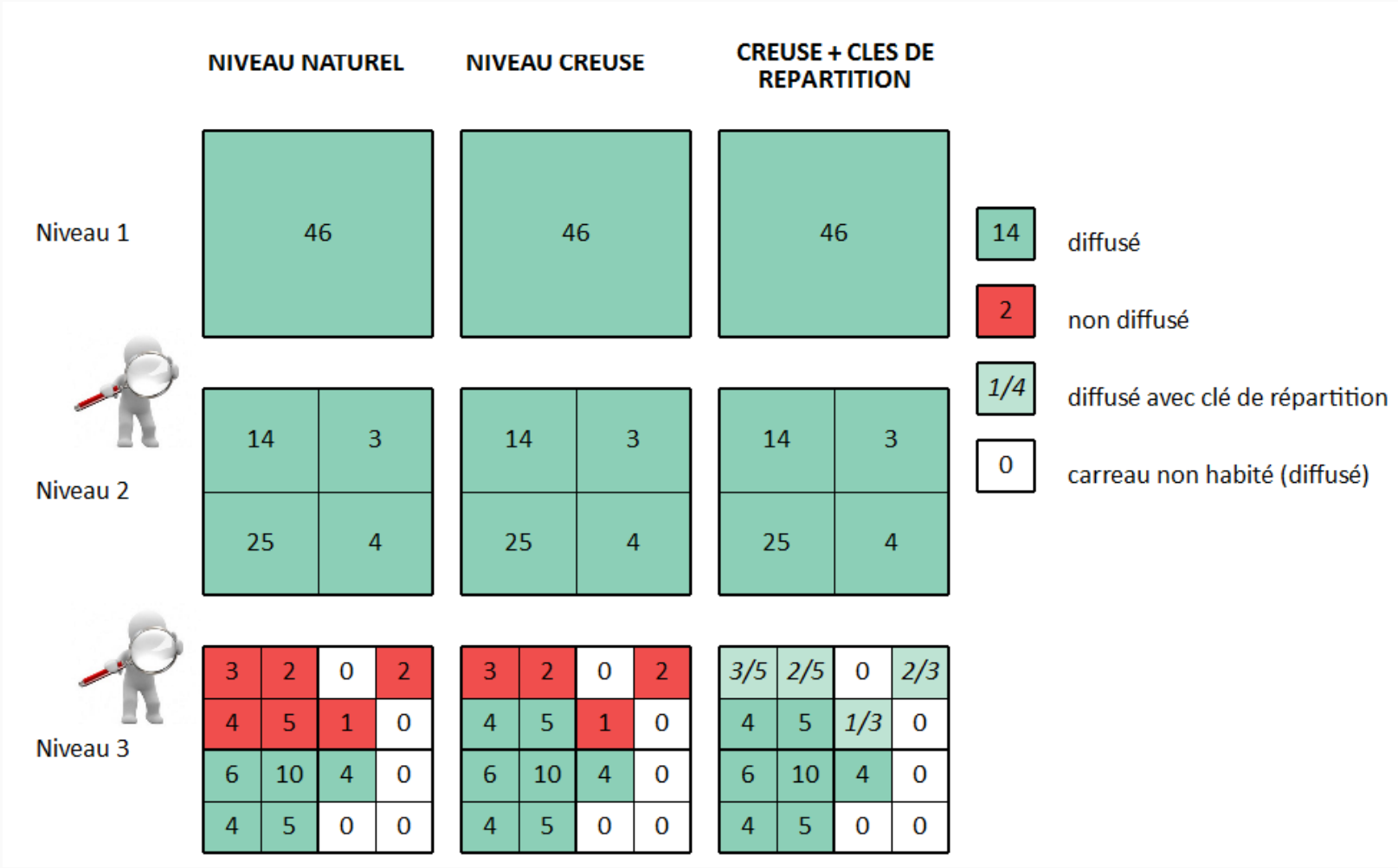


Figure 1: Part de carreaux et de ménages diffusés en fonction de la taille du carreau

Méthode des grilles superposées

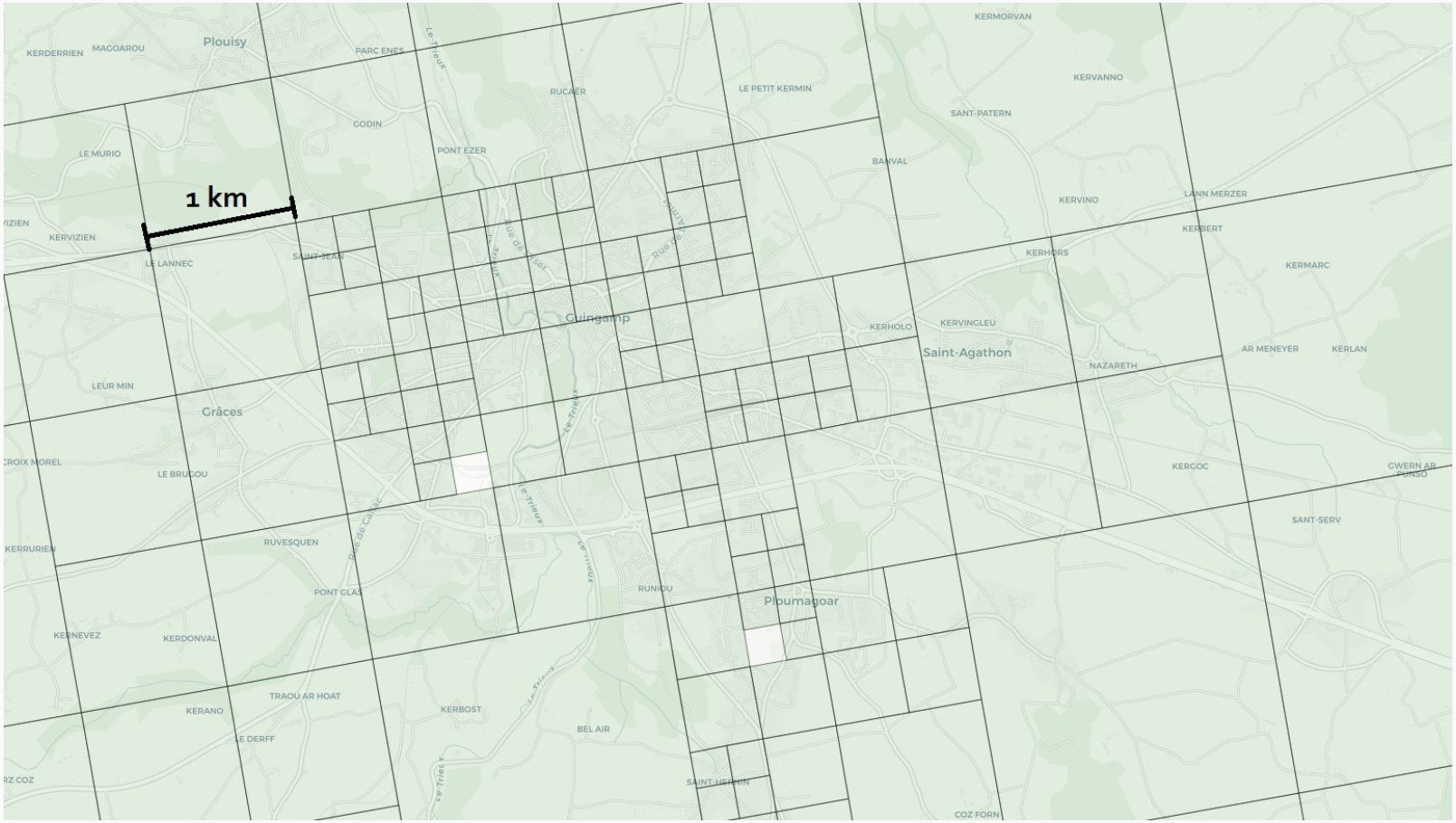
Une méthode qui peut se décliner en plusieurs produits différents

Exemple avec un seuil de 3



Méthode des grilles superposées

Un exemple : niveau naturel, région de Guingamp (22)



Méthode des grilles superposées

Les clés de répartition

- diffuser malgré tout des statistiques au niveau le plus fin ;
- Pour chaque carreau c non diffusable : $cle_c = \frac{\#\{menage \in c\}}{\#\{menage \in G_c\}}$;
- pour chaque variable X à diffuser : $X_c^* = cle_c * X_c$;
- la méthode devient **perturbative**.

Méthode des grilles superposées

Avantages

- facilité d'utilisation (à condition de mettre en place les clés de répartition) ;
- il existe un pavage pour lequel l'information n'est pas perturbée (niveau naturel).

Inconvénients

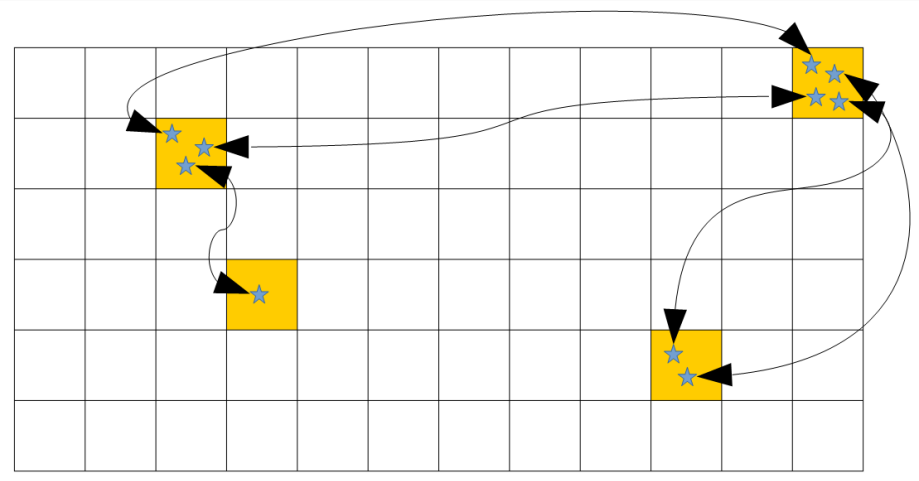
- comptages non entiers ;
- autant de jeux de clés que de variables à diffuser ?

Méthode de swapping

Méthode de *swapping*

Principe général

- Perturber l'information associée à une maille en permutant les ménages situés à l'intérieur avec d'autres ménages situés ailleurs ;
- on parle plus précisément de *targeted record swapping* ;



Note : Les carreaux jaunes comportent moins d'observations que le seuil. Les ménages à l'intérieur seront échangés entre eux.

- présenté comme un **bon compromis risque-utilité** ;
- souvent utilisé pour des données démographiques : recensements (GB [1], [7], [10], Japon [4], Hongrie [5]).

Méthode de *swapping* i

L'algorithme

1. **Repérage des individus à risque**
   - tous les ménages dans des carreaux sous un seuil ;
   - auxquels on peut ajouter si on le souhaite une liste de ménages.

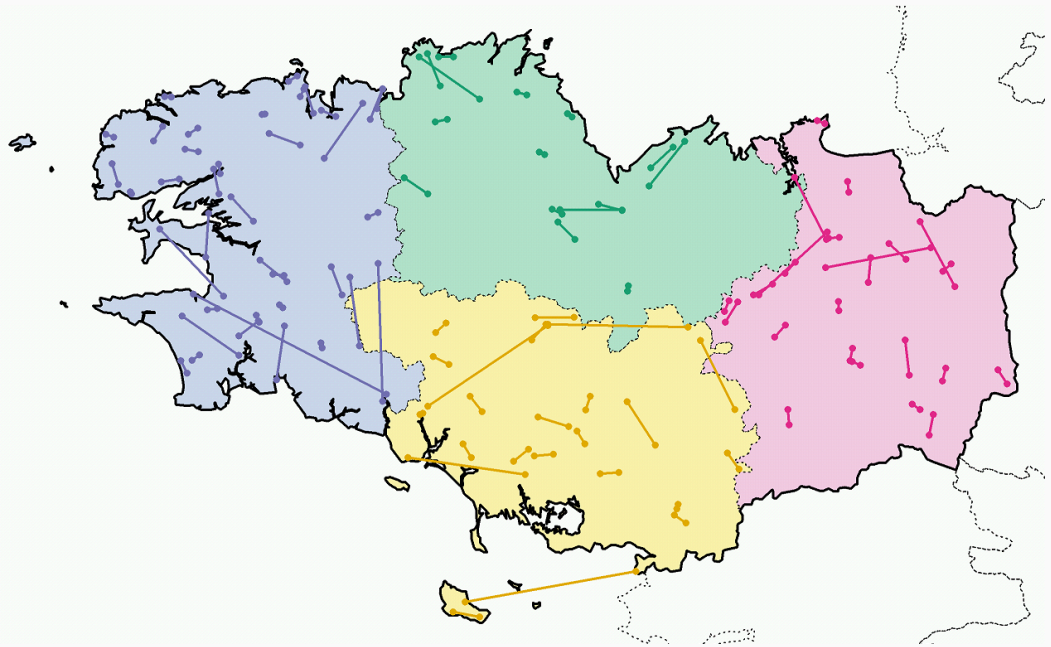
2. **Caractérisation des ménages à risque**
   On constitue différentes strates parmi les ménages à risque. Les échanges ne pourront avoir lieu que pour des ménages d'une même strate. Les strates doivent être de taille suffisante mais les plus homogènes possibles. Elles correspondent au croisement : **taille * zone * profil**, où :
   - taille = nombre d'individus (1, 2, 3, 4, ≥ 5) ;
   - zone = zone géographique agrégée (/ex. département) ;
   - profil = quantiles (var. quanti.) ou $\mathbb{1}_{modaliterare}$ (var. quali.).

Méthode de *swapping* ii

3. Retravail des strates
   pour que son effectif soit pair et sous une taille cible.
4. *Matching*
   Dans chaque strate, on constitue des paires de ménages en **minimisant de manière globale les distances géographiques entre les observations échangées** : package *R* `optmatch` (inspiration des méthodes d'*optimal matching*, [3], [6])
5. *Swapping*
   échange des variables géographiques

Méthode de *swapping*

Bretagne : exemples de permutations entre ménages (100 paires choisies au hasard parmi celles effectivement échangées avec l'algorithme proposé)



Note : Pour la maille de 200 mètres, la part de ménages échangés en Bretagne s'élève à 30 %. Selon la région, elle varie entre 2 % (Ile-de-France) et 33 % (Nouvelle-Aquitaine).

Méthode de *swapping*

Avantages

- facilité d'utilisation du produit final ;
- méthode fortement paramétrable (contrôle du niveau de perturbation) ;
- gestion possible des autres risques (ménages atypiques...).

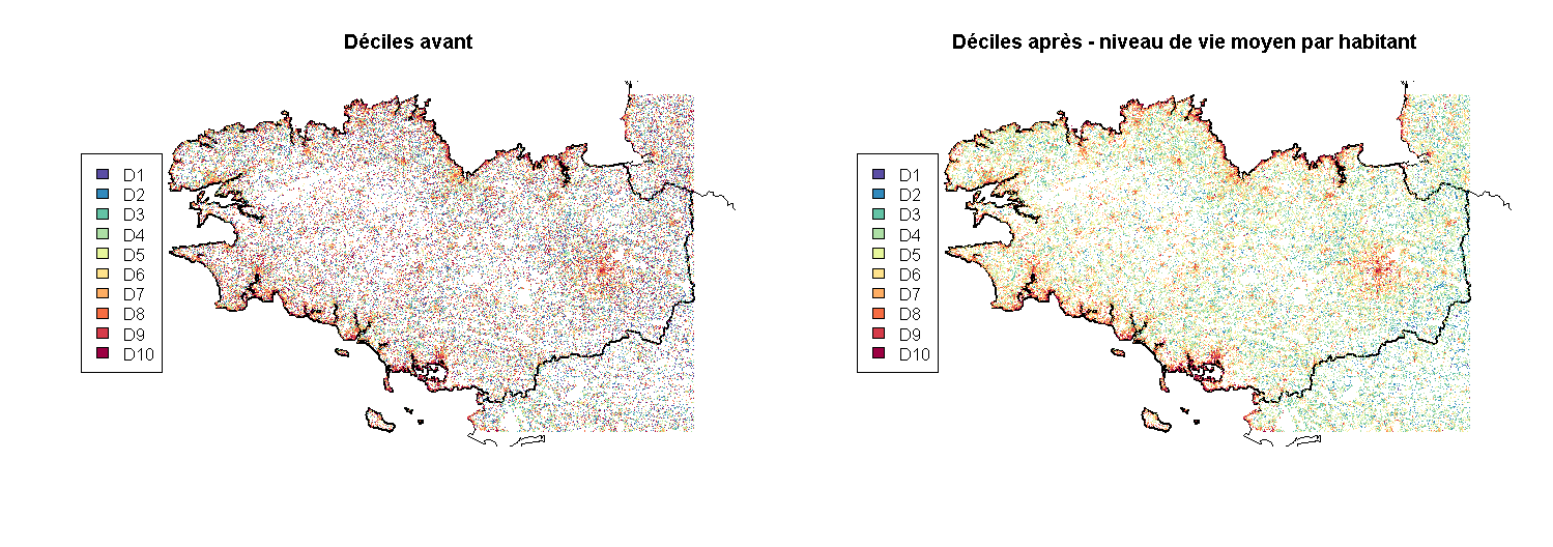
Inconvénients

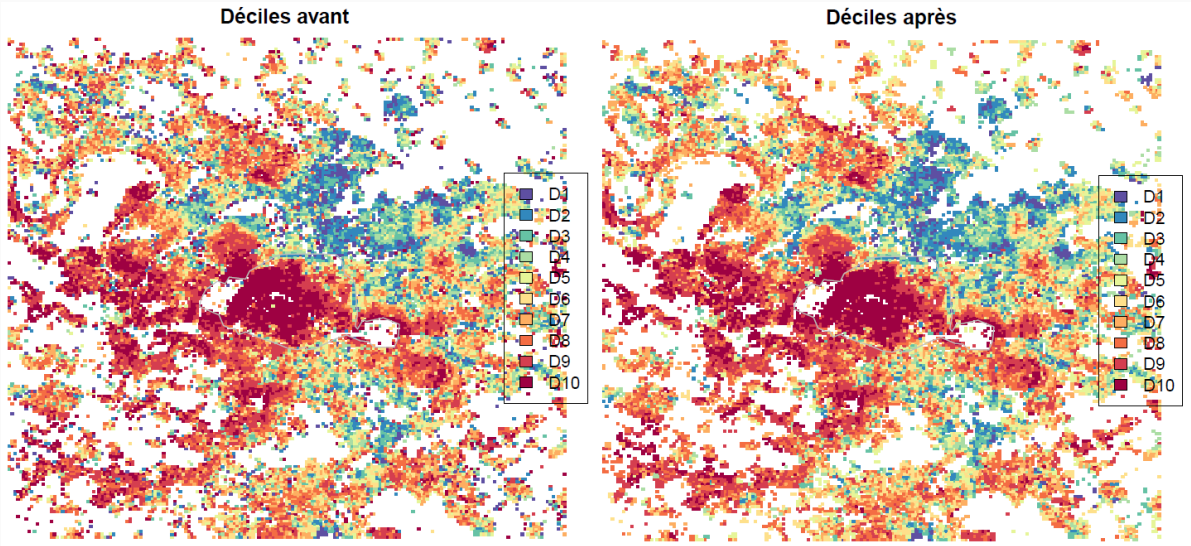
- il n'existe pas de niveau pour lequel on garantit que l'information n'est pas perturbée ;
- peut donner à l'utilisateur l'impression que rien n'a été fait.

Résultats

Résultats

Cartes avant/après : Attention à l'"effet-carte" !





Résultats

Niveaux de perturbation

$$\frac{\sum_{c \in carreaux} |X_c^{apres} - X_c^{avant}|}{\sum_{c \in carreaux} X_c^{avant}} \quad (1)$$

	Grilles superposées de niveau creusé + clés de répartition		*Swapping*	
Indicateur	200 m	250 m	200 m	250 m
Part de carreaux perturbés (en %)				
Variable 'nombre de personnes'	78,5	81,4	2,2	2,3
Variable 'nb. ménages dont PR >= 65 ans'	78,3	80,8	34,6	34,3
Variable 'nombre de pauvres'	72,4	70,9	0,7	0,7
Variable 'somme des niveaux de vie'	80,5	82,7	81,4	79,9
Masse de perturbation, tous carreaux (en %)				
Variable 'nombre de personnes'	4,4	3,6	0,1	0,1
Variable 'nb. ménages dont PR >= 65 ans'	13,3	10,9	14,8	12,1
Variable 'nombre de pauvres'	16,4	13,7	0,3	0,3
Variable 'somme des niveaux de vie'	4,2	3,5	3,5	2,9

Champ : France métropolitaine.

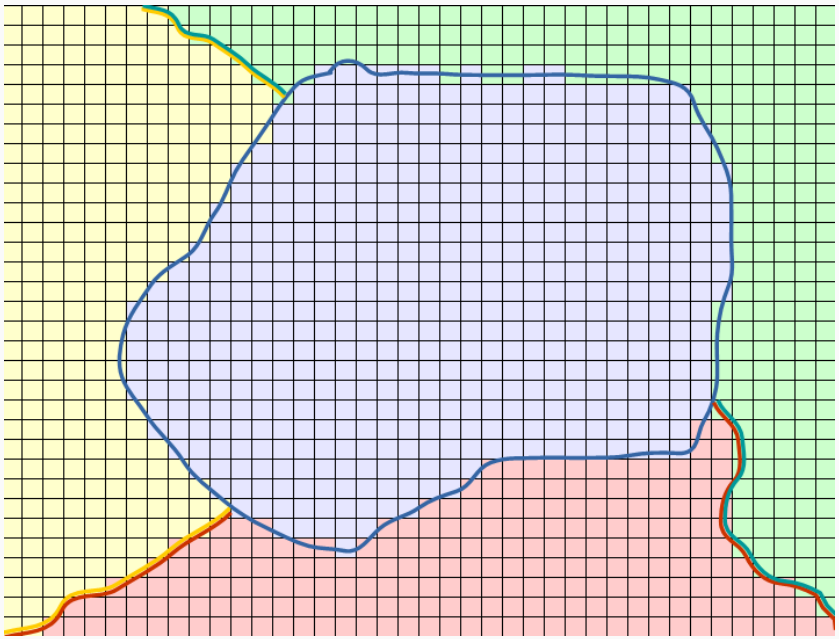
Source : Insee, Filosofi 2014.

Note : PR = personne de référence

Résultats

Perturbations par commune

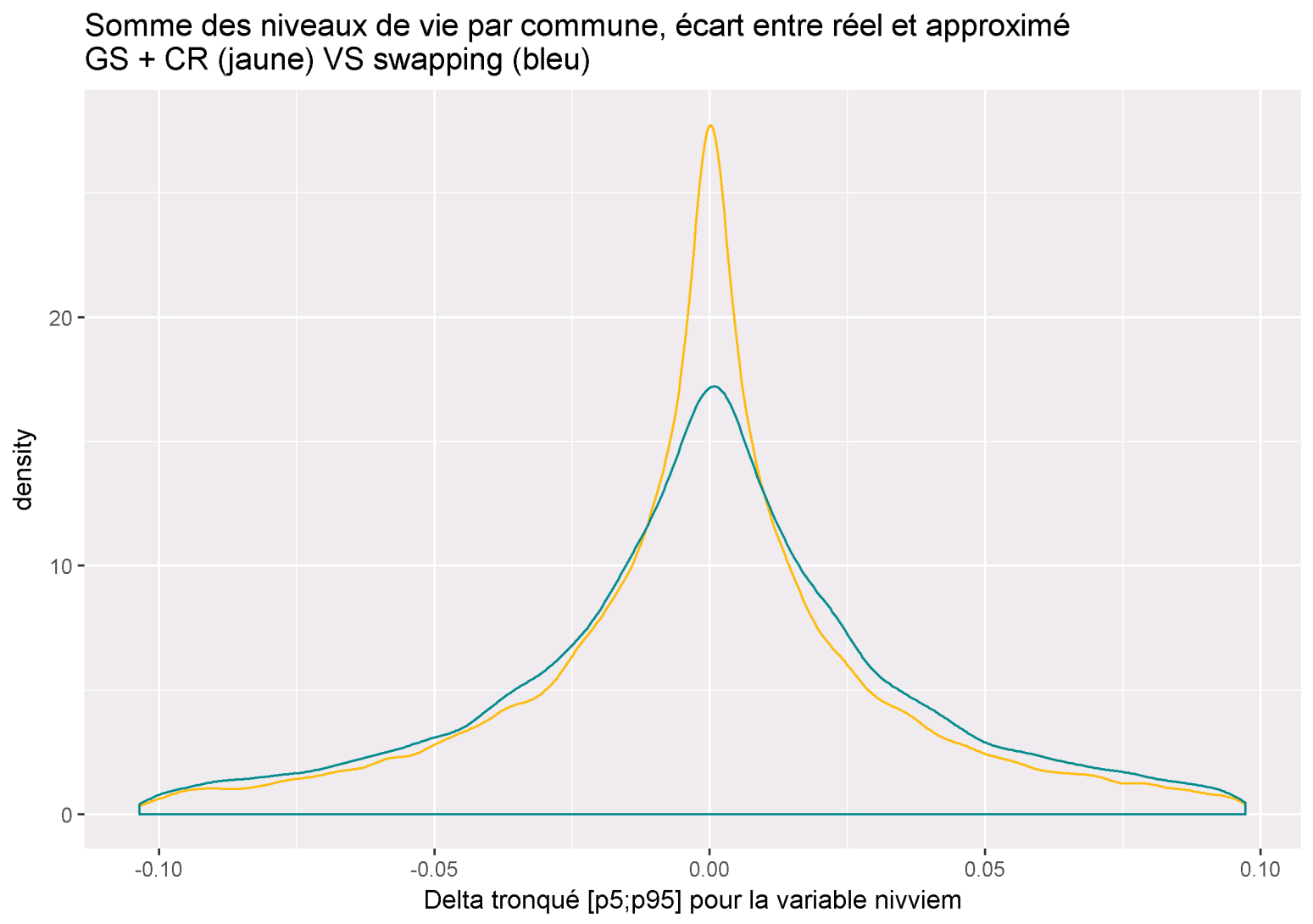
- un utilisateur souhaite connaître le total d'une variable par commune à partir des données carroyées ;
- hypothèse qu'un carreau appartient à la commune à laquelle son centroïde appartient ;



- pour chaque commune, différence entre ce proxy et le total "réel".

Résultats

Perturbations par commune



Conclusion

Conclusion

- **Contexte** = diffuser des données carroyées avec un seuil à respecter.
- Gérer la confidentialité de données finement localisées = **opportunité d'affiner les méthodes**, car le risque de divulgation dépend fortement de la densité de population et de la ressemblance d'un individu avec ses voisins.
- Parmi les nombreuses méthodes de confidentialité possibles [2], on en a présenté 2 (grilles superposées et *targeted record swapping*).
- La **facilité d'utilisation** se gagne **au prix** d'une introduction de **perturbation** ou de la **perte d'exhaustivité** de l'information.
- Les **arguments scientifiques ne sont pas les seuls en jeu**.
- Autres enjeux = cohérence entre sources, communication aux utilisateurs, capacité d'assumer la diffusion d'informations perturbées, crainte des erreurs d'interprétation d'utilisateurs trop pressés ...

Conclusion

Merci !

Références i

Brown, D.
Different approaches to disclosure control problems associated with geography.
Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality (2003).

Buron, M.-L., and Fontaine, M.
Manuel d'analyse spatiale.
Eurostat, EFGS, Insee (2018).

Hansen, B. B., and Klopfer, S. O.
Optimal full matching and related designs via network flows.
Journal of computational and Graphical Statistics 15, 3 (2006), 609–627.

Références ii

Ito, S., and Hoshino, N.

Data swapping as a more efficient tool to create anonymized census microdata in japan.

In *Privacy in Statistical Databases* (2014), pp. 1–14.

Nagy, B.

Targeted record swapping on grid-based statistics in hungary.

Rosenbaum, P. R.

A characterization of optimal designs for observational studies.

Journal of the Royal Statistical Society. Series B (Methodological) (1991), 597–610.

Shlomo, N., Tudor, C., and Groom, P.

Data swapping for protecting census tables.

In *Privacy in statistical databases* (2010), Springer, pp. 41–51.

Références iii

Tammilehto-Luode, M.

Opportunities and challenges of grid-based statistics.

In *World Statistics Congress of the International Statistical Institute* (2011).

VanWey, L. K., Rindfuss, R. R., Gutmann, M. P., Entwisle, B., and Balk, D. L.

Confidentiality and spatially explicit data: Concerns and challenges.

Proceedings of the National Academy of Sciences 102, 43 (2005), 15337–15342.

Références iv

Young, C., Martin, D., and Skinner, C.
Geographically intelligent disclosure control for flexible aggregation of census data.
International Journal of Geographical Information Science 23, 4 (2009), 457–482.