Confidentialité des données spatiales

Comment diffuser des données protégées sans invalider les analyses locales ?

Maël-Luc Buron & Maëlle Fontaine

May 14, 2018

Insee

Confidentialité des données spatiales

1. Risque de divulgation et données spatiales

2. Quelques méthodes de protection prenant en compte la géographie

3. Une application à des données carroyées

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 (en France, environ 80 % des carreaux de 200 m sont sous le seuil de 11 ménages).

Introduction

- Les **données carroyées** sont présentées comme un grand apport pour des analyses locales, mais elles présentent aussi un **risque de divulgation** plus élevé ([1], [2]).
- 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**.
- D'où le besoin de réfléchir à de **nouvelles méthodes de protection qui prennent en compte l'information spatiale**, dans la recherche d'un meilleur compromis entre risque et utilité.

⇒ Comment diffuser des données à un niveau d'utilité intéressant, tout en protégeant le secret ?

Risque de divulgation et données spatiales

Données spatiales et risque de divulgation

Définition du risque de divulgation

Un **intrus** utilise des données diffusées pour obtenir des renseignements inconnus auparavant ([3], [4]) :

1. **divulgation d'identité** : l'intrus retrouve un identifiant direct ;
2. **divulgation d'attributs** : l'intrus révèle des informations sensibles, quasi-identifiants ;
3. **divulgation inférentielle** : l'intrus déduit un attribut avec un bon niveau de confiance.

$\Rightarrow$ On restreint l'accès aux données confidentielles et on applique des techniques d'anonymisation aux données diffusées.

Données spatiales et risque de divulgation

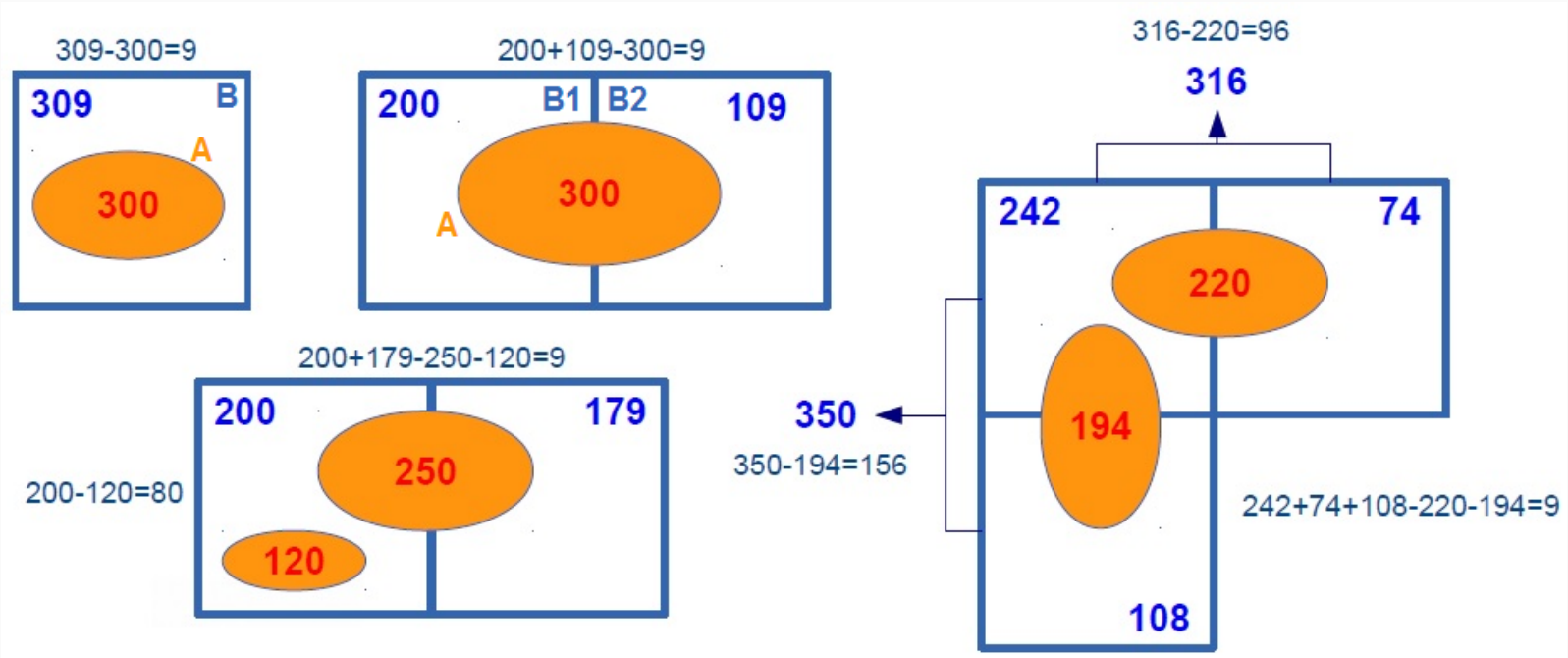
Le risque de divulgation est plus élevé en présence de données spatiales :

1. il est plus facile de mobiliser des connaissances personnelles ;
2. les observations proches se ressemblent (1ère loi de géographie de Tobler) ;
3. problème de différentiation géographique.

Différentiation géographique 1/2

- Un intrus **combine des données diffusées dans différentes géographies** et reconstitue des statistiques d'une zone plus petite ou déduit le lieu d'une observation.
- Avec des **géographies imbriquées** (exemple : régions – départements – villes), le problème se gère aisément car les opérations possibles de l'intrus sont liées à la hiérarchie des différentes géographies.
- **Sans hiérarchie**, il est nécessaire d'utiliser un algorithme long et complexe pour explorer les soustractions possibles et identifier les zones à protéger.

Différentiation géographique 2/2



Données spatiales et risque de divulgation

Comment mesurer si une observation présente un risque élevé de divulgation ?

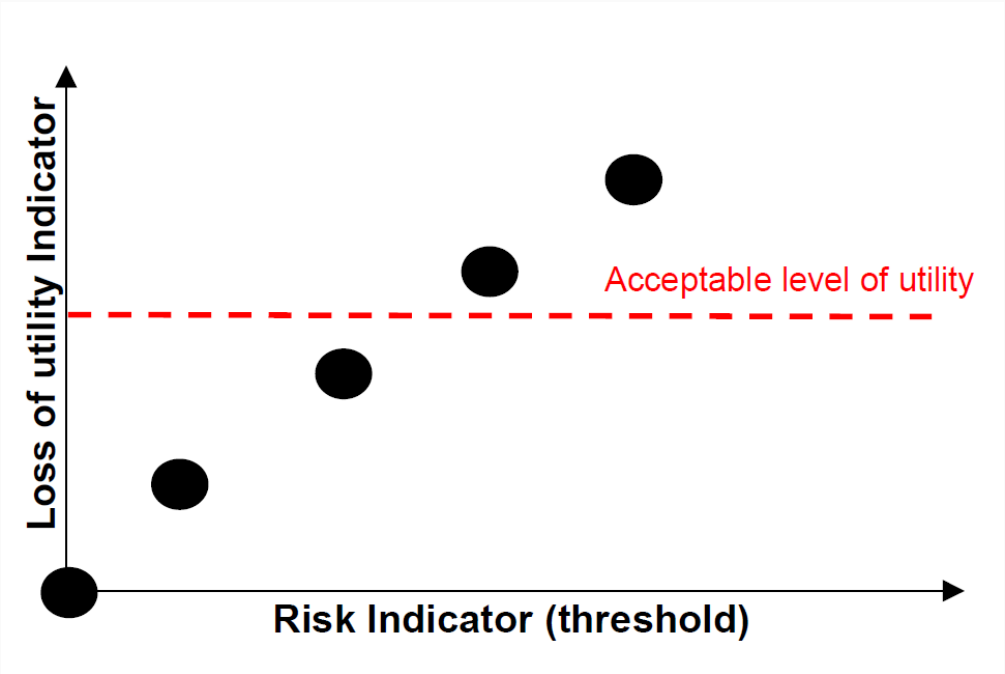
2 approches :

- **fixe** : si un comptage est sous le seuil, toutes les observations qui le constituent sont à risque ;
- **relative** : sont à risque les observations qui le sont plus que les autres. Par exemple, démarche de certaines méthodes perturbatives prétabulées :
  1. associer à chaque observation un **score individuel de rareté** qui dépend du contexte local de façon plus ou moins fine ;
  2. choisir un **seuil** au-delà duquel on considère l'observation "à risque" (un quantile par exemple) ;
  3. n'effectuer la perturbation que sur les observations ainsi **ciblées**.

Données spatiales et risque de divulgation

Comment évaluer le compromis risque-utilité d'un fichier ?

- Les ***RU-maps***, popularisées par Duncan *et al.* [14], quantifient la perte d'utilité en fonction du niveau de risque.
- Possibles **indicateurs de perte d'utilité** : déviation absolue moyenne (AAD), I de Moran, déformation des totaux par commune, etc.



Quelques méthodes de protection prenant en compte la géographie

Méthodes de protection prenant en compte la géographie

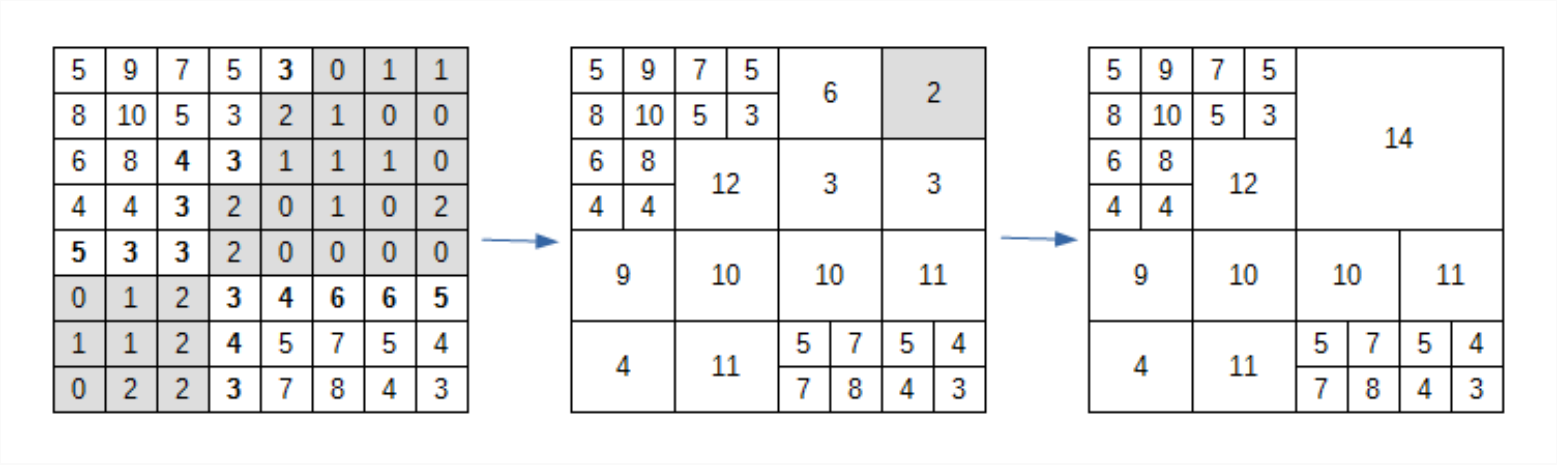
Rapide présentation de quelques méthodes :

- **agrégation géographique** : diffuser dans des mailles plus grosses (algorithmes dits *quadtree* et "des rectangles") ;
- ***geomasking*** : déplacer des individus ;
- ***targeted record swapping*** : permuter les localisations d'invidivus entre eux.

Méthodes de protection prenant en compte la géographie

Agrégation géographique : méthode *quadtree*

Petit exemple avec un seuil de 3



Agrégation géographique : méthode *quadtree*

Behnisch et al., 2013 [5]

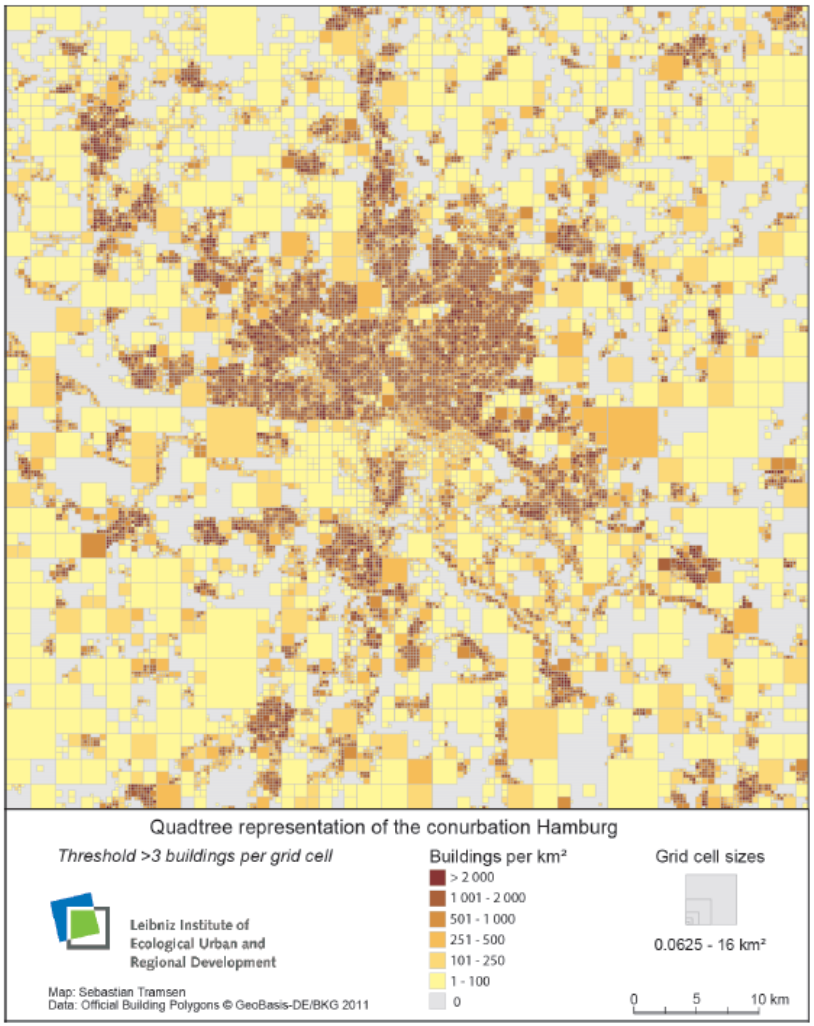


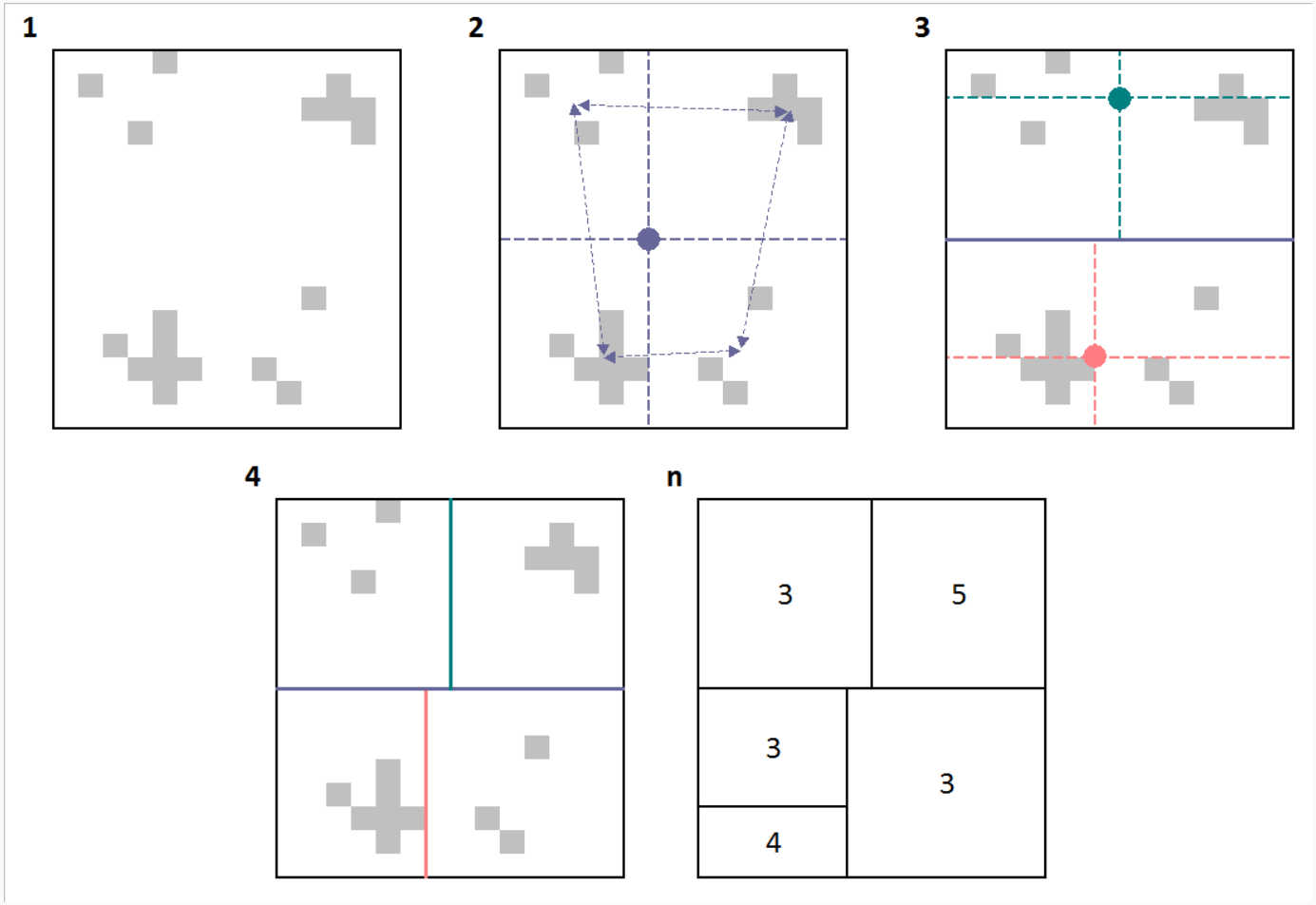
Fig. 6: Quadtree representation (threshold >3) of the conurbation Hamburg

Méthodes de protection prenant en compte la géographie

Agrégation géographique : méthode ”des rectangles”

Insee, diffusion de la source Revenus Fiscaux Localisés sur des carreaux de 200 m, 2013 [6]

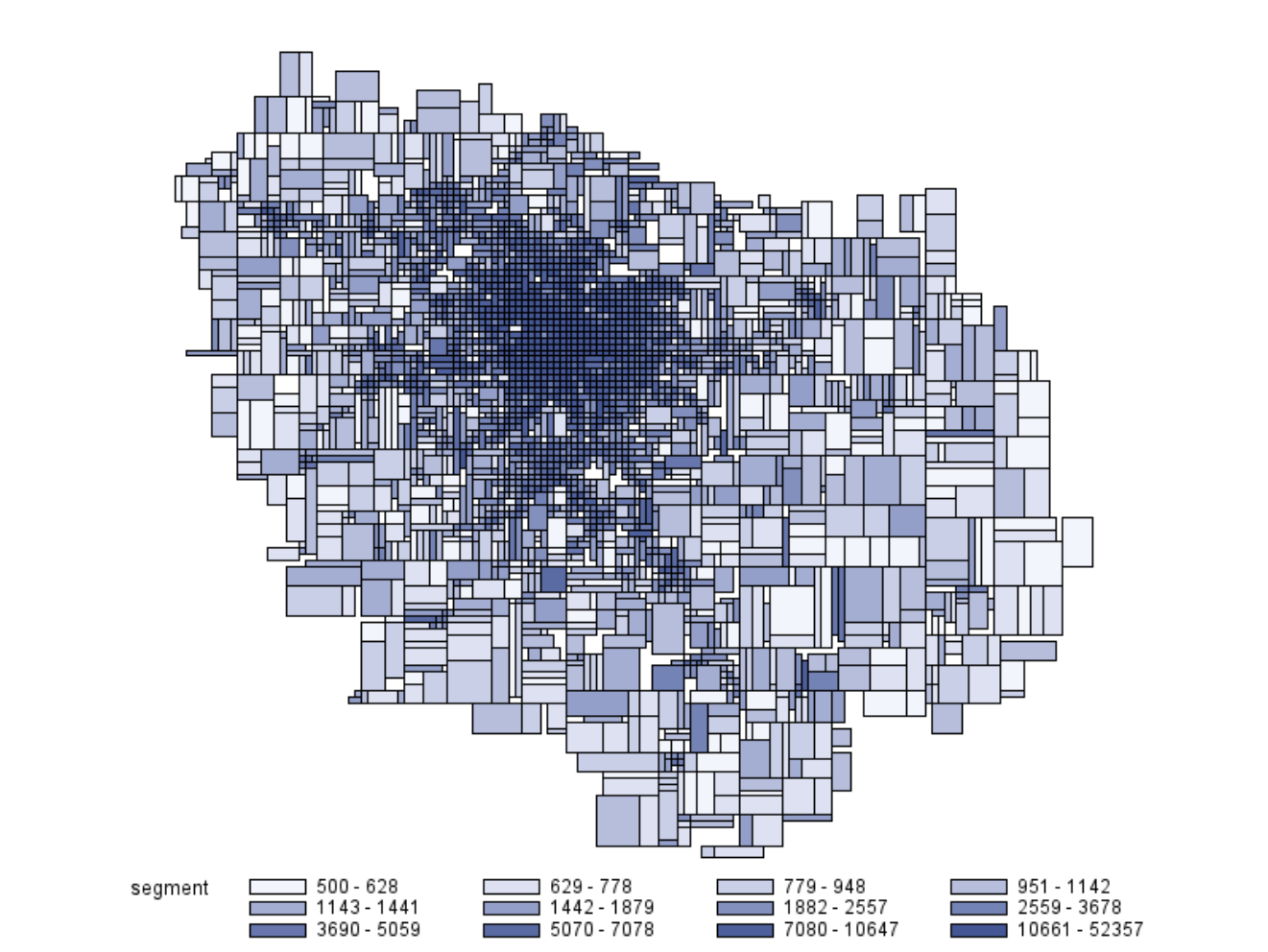
Petit exemple avec un seuil de 3



Méthodes de protection prenant en compte la géographie

Agrégation géographique : méthode ”des rectangles”

Un exemple en région Ile-de-France

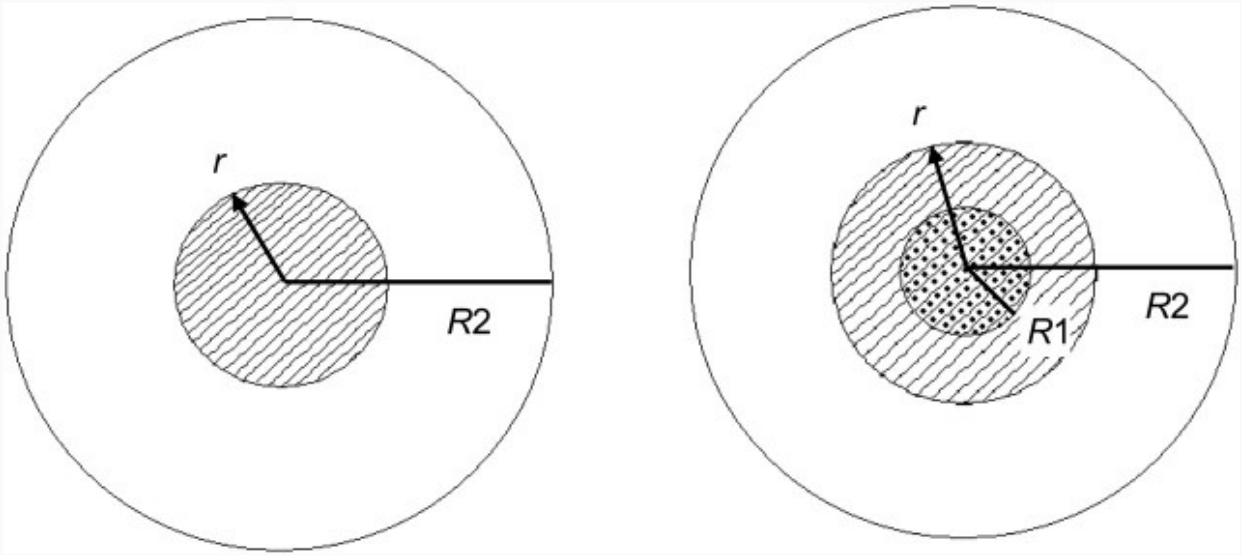


Méthodes de protection prenant en compte la géographie

Geomasking

Armstrong et al. 1999 [7], Hampton et al. 2010 [8]

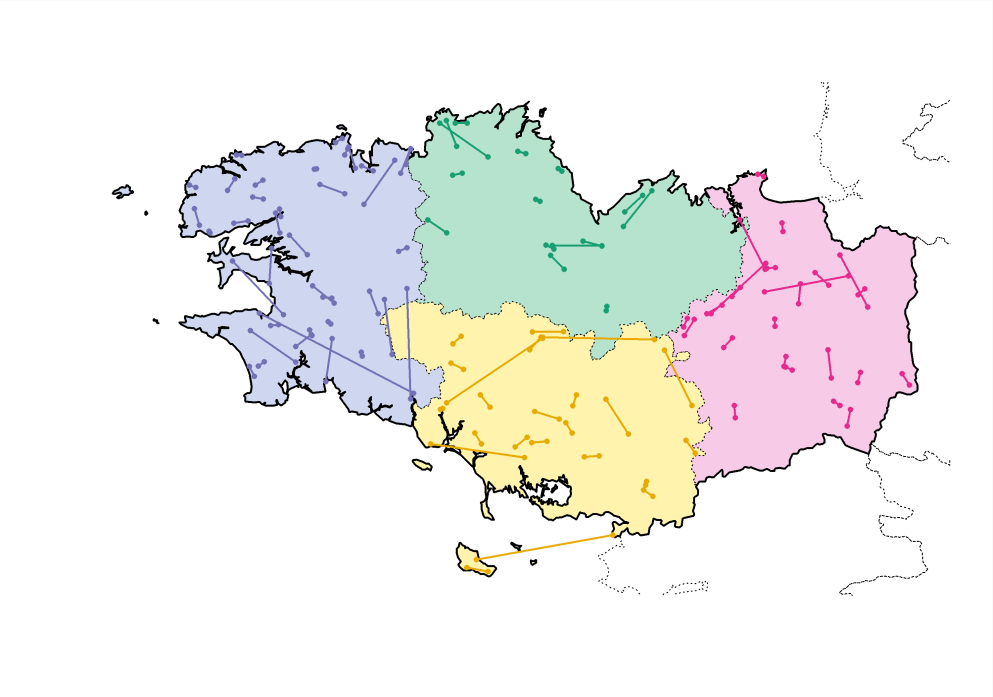
- modifier directement les localisations d'individus ;
- exemples : méthodes d'ajout de bruit aléatoire (*random perturbation* ou *donut*) ;



- souvent utilisé pour la confidentialité de données de criminalité ou d'épidémies, peu en économie (inconsistances possibles).

Targeted record swapping (TRS) : Généralités

- **Principe** : modifier les valeurs associées à des individus en permutant des ménages ;



- présenté comme un **bon compromis risque-utilité** ;
- souvent utilisé pour des données démographiques : recensements (GB [9], [10], [11], Japon [12], Hongrie [13]), données fiscales (France : travaux en cours pour Filosofi).

Targeted record swapping (TRS) : Etapes principales

1. ***Targeting*** : ciblage des ménages les plus risqués ;
2. ***Matching*** : formation de paires selon un algorithme qui rapproche des ménages différents, mais sous contrainte d'une ressemblance minimale (en termes d'attributs et de localisation géographique) ;
3. ***Swapping*** : permutation des informations géographiques des ménages d'une même paire.

Une application à des données carroyées

Application à des données carroyées

Première vague de tests

Source et champ : Fideli 2015, région Corse

Algorithme : ONS + adaptations DMRG, plusieurs jeux de paramètres

- **+** : forte conservation des relations spatiales (I de Moran) pour les variables qui sont choisies dans l'étape de matching ou qui sont fortement corrélées avec celles-ci ;
- **-** : forte déformation pour les autres variables et mesure de risque "*data-specific*" : les ménages à risque sont les α % les plus rares et non tous ceux sous le seuil.

Application à des données carroyées

Deuxième vague de tests

Source et champ : Filosofi 2014, France métropolitaine

Algorithme : DMRG, avec protection systématique des carreaux < 11

Figure 1: Niveau de vie moyen par carreau de 250 m : déformation induite par le TRS en zone peu dense (Guingamp, 22)

Application à des données carroyées

Deuxième vague de tests

Source et champ : Filosofi 2014, France métropolitaine

Algorithme : DMRG, avec protection systématique des carreaux < 11

	Carreau de 200 m	Carreau de 250 m
Part de ménage perturbés (%)	**18,4**	**14,7**
dont : déplacés d'au moins 5 km (%)	*53,0*	*55,1*
dont : déplacés d'au moins 10 km (%)	*27,0*	*28,7*
dont : déplacés d'au moins 50 km (%)	*0,9*	*1,1*
Distance moyenne entre ménages permutés (km)	**8,4**	**8,9**
Part de carreaux perturbés (%)	**81,4**	**79,9**
Part de la masse totale perturbée, pour différentes variables (%)		
Somme des revenus déclarés	4,7	3,9
Somme des niveaux de vie	3,5	2,9
Nombre de pauvres	0,3	0,3
Nombre de 65 ans et plus	14,8	12,1

Table 1: Différentes mesures de perturbations

Conclusion

Conclusion

- Gérer la confidentialité de données spatiales peut être vu comme une **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.
- Dans l'état de l'art actuel, l'information géographique est mobilisée indirectement. Une **prise en compte de l'information spatiale** plus fine serait envisageable au gré de l'augmentation des capacités de calcul ...
- ... mais le gain de précision est à mettre en regard des **difficultés à communiquer** pédagogiquement sur la méthode de protection aux utilisateurs.

Conclusion

- Il est crucial de toujours **évaluer à quel point gérer le risque dégrade l'utilité** des données.
- Dans cette idée d'un bon **compromis R/U**, Eurostat recommande de **combiner les méthodes** pré- et post-tabulées.
- Mais les **arguments scientifiques ne sont pas les seuls en jeu** : peut-on assumer la diffusion d'informations perturbées ? à quel point craint-on les erreurs d'interprétation d'un utilisateur trop pressé ?

References i

- *Confidentiality and spatially explicit data: Concerns and challenges*, VanWey et al., Proceedings of the National Academy of Sciences, Vol. 102-43, pp. 15337-15342, 2005
- *Opportunities and challenges of grid-based statistics*, Tammilehto-Luode, World Statistics Congress of the International Statistical Institute, 2011
- *Handbook on statistical disclosure control*, Hundepool et al., 2012
- *La gestion de la confidentialité pour les données individuelles*, Maxime Bergeat, Document de travail Insee M2016/07, 2016
- *Using Quadtree representations in building stock visualization and analysis*, Behnisch et al., pp. 151–166, 2013

References ii

- *Documentation complète sur les données carroyées à 200 mètres*, Online documentation Insee, 2013
- *Geographically masking health data to preserve confidentiality*, Armstrong et al., Statistics in medicine, 18-5, pp.497–525, 1999
- *Mapping health data: improved privacy protection with donut method geomasking*, Hampton et al., American journal of epidemiology, 172-9, pp.1062–1069, 2010
- *Different approaches to disclosure control problems associated with geography*, Brown et al., Proceeding of the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, 2003
- *Data Swapping for Protecting Census Tables*, Shlomo et al., Privacy in statistical databases, pp. 41–51, 2010

References iii

- *Geographically intelligent disclosure control for flexible aggregation of census data*, Young et al., International Journal of Geographical Information Science, 23-4, pp. 457–482, 2009
- *Data swapping as a more efficient tool to create anonymized census microdata in Japan*, Ito et al., Privacy in Statistical Databases, pp. 1–14, 2014
- *Targeted record swapping on grid-based statistics in Hungary*, Nagy B., Submission for the 2015 IAOS Prize for Young Statisticians, 2015
- *Disclosure risk vs. data utility: The RU confidentiality map*, Duncan, George T and Keller-McNulty, Sallie A and Stokes, S Lynne, 2001

Conclusion

Merci !