JMS Journées de méthodologie statistique de l'Insee 2018

Données carroyées et confidentialité

Marc BRANCHU(*), Vianney COSTEMALLE(*), Maëlle FONTAINE(*)

(*)Insee, Département des Méthodes Statistiques, Division des Méthodes et Référentiels Géographiques

marc.branchu@insee.fr
vianney.costemalle@insee.fr
maelle.fontaine@insee.fr

Mots-clés. Carroyage, confidentialité, grilles superposées, *swapping*.

Résumé

À l'heure où de plus en plus de sources statistiques sont géolocalisées, les instituts statistiques cherchent de plus en plus à diffuser leurs statistiques sur des grilles régulières de carreaux. À condition que la maille de la grille soit suffisamment fine, l'utilisateur de données carroyées peut recréer des zones selon ses besoins, en agrégeant des carreaux. Les données carroyées ne font pas exception aux règles de secret statistique, qui imposent de ne pas diffuser une statistique dès lors qu'elle concerne un trop petit nombre d'observations. Ce seuil dépend de la source ; il est de 11 dans le cas de la source fiscale. Différentes méthodes de protection peuvent être implémentées pour gérer la confidentialité de données statistiques. Les méthodes traditionnelles ne prennent généralement pas en compte la géographie, et pourraient induire des niveaux de perturbation très élevés dans le contexte de données diffusées sur des mailles très petites. S'inspirant de la littérature existante sur les méthodes de gestion de la confidentialité prenant en compte l'information géographique, nous proposons et implémentons des versions originales de deux méthodes, avec l'objectif qu'elles soient plus spécifiquement adaptées au contexte de diffusion de données carroyées. La première méthode dite des *grilles superposées* consiste à réaliser des agrégations géographiques afin de diffuser l'information sur la plus petite maille possible respectant encore le seuil. Cette méthode peut être déclinée de différentes manières. Pour les tests, nous considérons la solution dite "niveau creusé", qui optimise la part d'information diffusée au niveau le plus fin. De surcroît, nous mettons en place un système de clés de répartition pour diffuser malgré tout une information sur tous les petits carreaux, même ceux sous le seuil, et ainsi gagner en facilité d'utilisation. La seconde méthode est celle dite de *swapping*. Elle consiste à échanger directement dans les micro-données les caractéristiques de ménages à risque, de telle sorte que toute statistique concernant un carreau sous le seuil fasse l'objet d'une perturbation. Dans l'une et l'autre de ces méthodes, l'utilisateur a l'assurance, s'il s'intéresse à un carreau sous le seuil, que l'information diffusée a subi une perturbation par rapport à l'information originale. Nous présentons les avantages et inconvénients relatifs de ces deux méthodes, et les comparons à l'aune d'un certain nombre d'indicateurs visant à mesurer la perte d'utilité des données à la suite de la perturbation.

Abstract

Cet article présente deux méthodes de gestion de la confidentialité adaptées au contexte de diffusion de données carroyées, avec un seuil minimal d'observations à respecter. Deux méthodes sont décrites et testées sur les données exhaustives de Filosofi 2014. La première, celle des grilles superposées, consiste à agréger au besoin les carreaux les plus petits dans des groupes de diffusion au-dessus du seuil. La seconde méthode, dite de *swapping*, consiste à permuter entre eux des ménages à risque avant de construire les données carroyées sur la base du fichier de micro-données perturbé. Le *swapping* induit des niveaux de perturbation légèrement plus faibles que la méthode des grilles superposées, mais pose davantage de problèmes d'affichage, puisqu'il n'est pas possible de retrouver une maille agrégée pour laquelle on assure à l'utilisateur que l'information n'a pas été perturbée.

Introduction

Le carroyage consiste à créer une grille de carreaux sur laquelle l'information géolocalisée est agrégée puis diffusée. Cela suppose par conséquent de disposer d'une source de données où chaque observations est géolocalisée, c'est-à-dire qu'on associe un point de coordonnées (x,y) à chaque observation. L'avantage d'une telle grille est de fournir de l'information finement localisée permettant à un utilisateur d'analyser *a posteriori*, par regroupement de carreaux, des zonages qui lui seraient spécifiques.

Deux enjeux principaux apparaissent lorsque l'on souhaite carroyer une source de données géolocalisées. Le premier est celui de la qualité des informations diffusées au niveau de petits carreaux, de quelques centaines de mètres de côté. Pour carroyer une source issue d'une enquête comme le recensement de la population, une étape d'estimation des variables au niveau des carreaux est nécessaire. Plus les carreaux sont petits, moins l'estimation risque d'être de bonne qualité. La qualité des informations diffusées peut aussi être impactée par la qualité du géoréférencement des données. Le deuxième enjeu des données carroyées est celui de la confidentialité : plus les carreaux sont petits plus on risque de rompre le secret statistique.

Le secret statistique vise à protéger les individus ou les entreprises : un institut comme l'Insee ne peut pas diffuser publiquement des informations qui permettent d'identifier un individu ou de révéler des caractéristiques de natures personnelles ou sensibles. On distingue généralement le secret statistique primaire qui concerne les informations directement mises à disposition des utilisateurs du secret statistique secondaire qui a trait aux informations qu'un utilisateur pourrait déduire *indirectement* à partir de l'ensemble des données diffusées par l'institut. Tout au long de l'article, ces deux notions de secret statistique primaire et secondaire vont intervenir.

Les travaux présentés ici ne concernent que la confidentialité des données carroyées et ne traite pas de leur qualité. Ils sont relatifs à la diffusion de la source Filosofi 2015 sous forme carroyée. Néanmoins, les méthodes envisagées doivent être pensées pour s'appliquer à d'autres sources et pour produire sur le plus long terme des jeux de données carroyées comparables et cohérents entre eux. Pour une approche plus générale sur la confidentialité des données spatiales et le raffinement des méthodes traditionnelles de protection par une prise en compte de l'information géographique, nous invitons le lecteur à se reporter au chapitre 14 du manuel d'analyse spatiale à paraître [3].

Nous présentons dans cet article deux méthodes spécifiquement développées et testées dans le cadre de la diffusion carroyée de la source Filosofi 2015. Cependant, ces méthodes ont été conçues dans une perspective de mobilisation pérenne pour d'autres sources. Elles reprennent et prolongent des méthodes déjà utilisées ou envisagées par le passé pour le même type de source. La première d'entre elles, dite méthode des grilles superposées, est une méthode de découpage

itératif, au même titre que la méthode des rectangles (voir annexe 3.3) utilisée par l'Insee en 2013 pour diffuser des statistiques issues de la source fiscale sur des carreaux de 200 mètres de côté. La différence est que la méthode décrite ici s'appuie sur des grilles dont la constitution est indépendante de la source à diffuser, rendant possibles les analyses multisources. La seconde méthode décrite dans cet article est une méthode dite de *swapping*, piste qui avait été envisagée au départ pour diffuser cette même source.

Dans le cadre de la gestion du secret statistique, on distingue généralement les **méthodes non perturbatives** (par exemple agrégation ou suppression de données) des **méthodes perturbatives**, qui ne donnent pas accès à l'information exacte. On distingue également les méthodes **pré-tabulées** qui interviennent sur la table initiale des observations des méthodes **post-tabulées** qui interviennent sur les tables agrégées (ou données tabulées).

Dans cet article, nous présentons les deux méthodes en question, tout d'abord en donnant une description littérale, puis en fournissant les détails de l'algorithme. Dans une deuxième partie, nous présentons les résultats des tests menés pour ces deux méthodes, en les comparant à l'aune de différents indicateurs de perturbation.

1 Deux méthodes de confidentialisation

Une fois que les observations d'une source statistique sont géolocalisées il est possible de regrouper ces observations par carreaux, selon leur position géographique. En général, l'utilisateur souhaite disposer de carreaux les plus petits possibles, pour mener des analyses les plus précises possibles. En revanche, le secret statistique primaire oblige à restreindre la taille des carreaux : s'ils sont trop petits il comporteront moins d'observations que le seuil de confidentialité [1]. De plus, pour des questions de qualité du géoréférencement de la source, il faut également se restreindre à des carreaux de taille suffisamment importante. Dans la suite, les plus petis carreaux considérés seront des carrés de 200 m (ou 250 m) de côté.

1.1 La méthode des grilles superposées

1.1.1 Généralités

La méthode des grilles superposées est une méthode agrégative non perturbative qui peut être déclinée en deux produits : les grilles superposées au niveau que l'on nomme "naturel", et celles au niveau que l'on nomme "creusé". Il est de plus possible d'ajouter une étape à chacun de ces produits, appelée "clés de répartition", qui permet d'estimer des valeurs sur tous les plus petits carreaux non diffusables. L'ajout de l'étape de clés de répartition en fait une méthode perturbative.

Le système des grilles superposées consiste en plusieurs grilles simples qui s'emboîtent les unes dans les autres, et qui sont de plus en plus fines (figure 1). Les carreaux d'un niveau donné résultent d'un découpage régulier des carreaux du niveau supérieur. Ceci permet de diffuser des informations à plusieurs niveaux de précision, selon la densité de population sous-jacente. Les deux systèmes de grilles envisagés sont :

1. système à 7 niveaux : 200 m, 1 km, 2 km, 4 km, 8 km, 16 km et 32 km ;
2. système à 8 niveaux : 250 m, 500 m, 1 km, 2 km, 4 km, 8 km, 16 km et 32 km.

Pour faciliter la suite de l'exposé, on définit un "carreau père" comme étant constitué de plusieurs sous-carreaux "fils", et pour un carreau "fils" donné, les autres carreaux ayant le même

1. En l'occurence, pour la source fiscale, le secret statistique [4] impose de ne pas diffuser de statistique dès lors que celle-ci concerne moins de 11 ménages fiscaux.

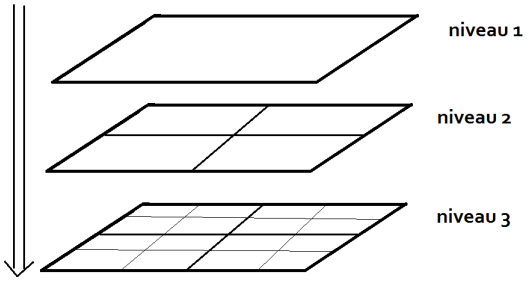

FIGURE 1 – Représentation schématique du système de grilles superposées.

"carreau père" sont appelés les carreaux "frères".

1.1.2 Niveau naturel

La première alternative, dite "niveau naturel", consiste à ne pas diffuser d'information sur les carreaux dont au moins un des carreaux "frères" n'est pas diffusable (car contenant moins d'observations que le seuil de confidentialité). Si un carreau "père" a au moins l'un de ses fils non diffusable, alors le niveau naturel correspond à ce carreau "père" et aucune information n'est diffusée à l'intérieur des carreaux-fils, qu'ils soient au-dessus du seuil ou non. Le pavage en carreaux de niveau naturel est équivalent à la méthode parfois appelée *quadtree* dans la littérature (par exemple Behnisch et al. 2013 [2] mettent en place cette méthode pour représenter des statistiques de logements). La méthode du *quadtree* consiste à raffiner le découpage de chaque carreau, en les découpant en 4 sous-carreaux [2], jusqu'à ce que le découpage fasse apparaître un carreau sous le seuil de confidentialité. Les grilles superposées au niveau naturel permettent d'obtenir un pavage complet du territoire constitué de carreaux de différentes tailles (exemple en figure 2). Au contraire, les grilles superposées au "niveau creusé" détaillées ci-après ne permettent pas d'avoir une partition du territoire en carreaux diffusables.

1.1.3 Niveau creusé

En préambule on introduit brièvement ce qu'est la différenciation géographique. Il s'agit d'une technique qui consiste à faire la différence entre une zone géographique et une autre zone contenue à l'intérieur de la première afin de déduire des informations sur la zone complémentaire de la zone intersectée. Par exemple, si on connait la valeur d'une variable sur un carreau "père" possédant 4 "fils", et qu'on connaît également la valeur de cette même variable sur les trois premiers fils, par différence entre le père et les trois premiers fils on en déduit la valeur de la variable sur le quatrième fils (voir figure 3). Cette technique ne fonctionne que si la variable en question est de nature **additive**.

La deuxième alternative, dite "niveau creusé" consiste à diffuser certains des carreaux dont au moins un des carreaux "frères" n'est pas diffusable. Certains seulement, car par différenciation entre le carreau "père" et certains carreaux "fils", on peut avoir de l'information sur les carreaux "fils" restants (voir figure 3). Il faut alors s'assurer qu'en diffusant seulement certains carreaux «fils», on ne puisse pas retrouver des informations sur des carreaux contenant moins d'observations que le seuil de confidentialité. Pour cela on choisit de blanchir en plus un des

2. Il peut aussi s'agir d'un découpage en 25 sous-carreaux, ou tout carré d'entier naturel

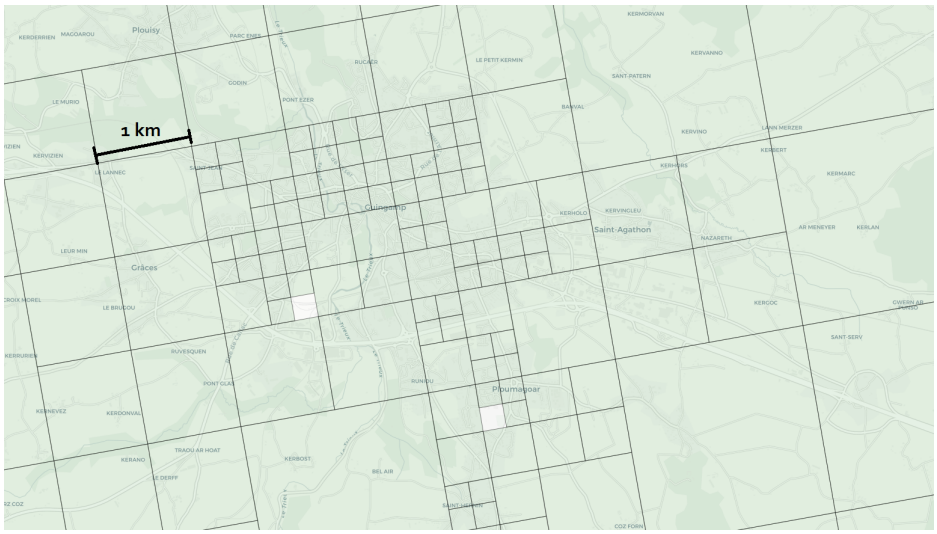

FIGURE 2 – Exemple de pavage obtenu à partir des grilles superposées au niveau naturel.
Champ : Zone rurale autour de la ville de Guingamp en Bretagne.
Note : Les plus petits carreaux représentés sur cette figure sont des carreaux de 250 mètres de côté. Tous les carreaux représentés contiennent au moins 11 ménages fiscaux. Les carreaux blancs sont les carreaux non habités.
Source : Insee, Filosofi 2014.

$$\text{père} = \text{fils1} + \text{fils2} + \text{fils3} + \text{fils4}$$

FIGURE 3 – Schéma illustrant le principe de la différenciation entre deux grilles de niveaux différents.
Note : En diffusant le père au niveau i et fils1, fils2 et fils3 au niveau i+1, un utilisateur peut en déduire fils4, par différence entre le père et fils1 + fils2 + fils3.

carreaux «fils» diffusables (celui qui contient le moins de ménages). De plus, au niveau creusé, un carreau père non diffusable peut très bien avoir certains de ses fils qui sont, eux, diffusables.

C'est le cas si un carreau père comporte plus d'observations que le seuil de confidentialité mais qu'il a néanmoins été blanchi pour protéger un des carreaux frères sous le seuil. Les fils du carreaux père au-dessus du seuil peuvent très bien être eux-memes au-dessus du seuil et donc diffusables. Il faut seulement qu'on ne puisse pas reconstituer l'intégralité du père à partir de tous les fils : pour cela, soit on blanchi un des fils au-dessus du seuil, soit certains fils sont déjà automatiquement blanchis car en-dessous du seuil. Afin de mieux visualiser ce processus on pourra se reporter à la figure 4.

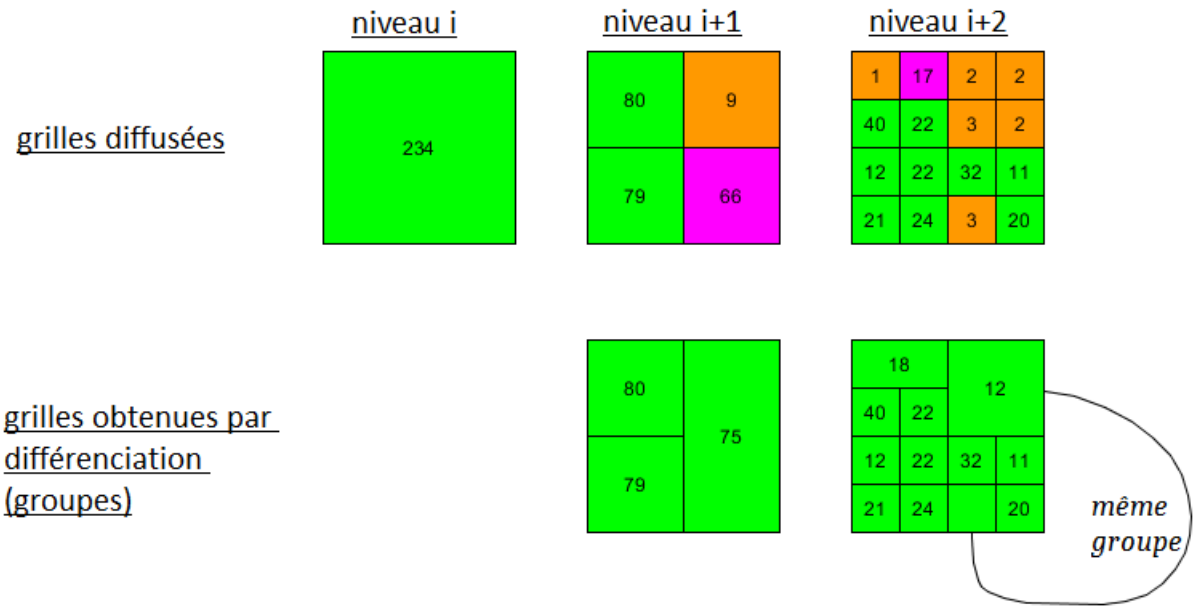

FIGURE 4 – Exemple de grilles superposées au niveau creusé

Note : On considère ici un seuil de confidentialité de 11. Les carreaux vert sont ceux au-dessus du seuil et diffusés, les carreaux violets sans ceux au-dessus du seuil mais blanchis et les carreaux oranges sont ceux en-dessous du seuil et forcément blanchis.
La première ligne indique les grilles de niveaux i, i+1 et i+2 qui sont diffusées et la deuxième ligne indique les grilles qu'on peut obtenir par différenciation entre les différents niveaux.

Description de l'algorithme utilisé

On décrit plus précisément l'algorithme relatif à la constitution des grilles superposées de niveau creusé. L'objectif principal est d'attribuer à chaque carreau un état de diffusabilité (1 s'il est diffusable ou 0 sinon - on dit alors que le carreau est blanchi). L'objectif secondaire est de regrouper les carreaux blanchis entre eux : par différenciation entre les différents niveaux des grilles superposées, on peut déduire de l'information sur un ensemble minimal de carreaux blanchis. Ceci permettra au final de connaître le maillage le plus fin pouvant être obtenu soit directement soit par des différenciations multiples.

On introduit les notations suivantes :
- il y a p niveaux (c'est-à-dire p grilles emboîtées) ;
- pour un carreau i donné, on note n_i le nombre d'observations contenues dans ce carreau ;

- pour un carreau i donné, on note E_i son état ($E_i = 1$ si le carreau est diffusé et $E_i = 0$ s'il est blanchi) ;
- pour un carreau i donné, on note F_i sa "force de protection" ($F_i = 0$ si le carreau n'est pas blanchi ou s'il est sous le seuil, et $F_i > 0$ si le carreau est blanchi mais au-dessus du seuil). Si le carreau i est *a priori* diffusable, mais qu'on a tout de même décidé de le blanchir afin de protéger certains de ces carreaux "frères", F_i donne le nombre d'observations ainsi protégées par le blanchiment de ce carreau ;
- pour un carreau i donné, on note G_i son groupe. Pour un niveau de grille donné (par exemple la grille des carrés de 1 km), le groupe d'un carré diffusé ne contient qu'un carré : lui-même. Pour les carrés non diffusés (i.e. blanchis) le groupe désigne le plus petit ensemble de carrés dont on peut déduire la valeur par différenciation entre les grilles de niveaux inférieurs (grilles de mailles plus larges). On peut se reporter à la figure 4 qui donne une visualisation des groupes obtenus par différenciation.

La démarche est alors la suivante (des optimisations sont possibles pour réduire le coût computationnel) :

1. D'abord, on implémente une fonction `DetermineEtatForceGroupe` qui détermine l'état, la force de protection et le groupe des fils à partir de l'état, la force de protection et le groupe du père (voir annexe 3.3 pour le détail du pseudo-code de cette fonction) ;
2. ensuite, on traite les carreaux du premier niveau (le niveau avec les mailles les plus grandes) : pour chaque carreau i du niveau 1 :
   - si $n_i < seuil$, $E_i = 0$; $F_i = 0$; $G_i = 1$
   - si $n_i \geq seuil$, $E_i = 1$; $F_i = 0$; $G_i = G_{i-1} + 1$
3. enfin, on détermine récursivement les états de tous les carreaux fils (de niveau supérieur ou égal à 2) :
   - pour chaque niveau n de 1 à p-1 :
     - pour chaque carreau i :
       — $ResFils$ = `DetermineEtatForceGroupe`(i, E_i, F_i, G_i, seuil) et on met à jour les informations sur les fils avec $ResFils$.

1.1.4 Clés de répartition

L'étape supplémentaire des clés de répartition permet d'obtenir des estimations sur toutes les mailles élémentaires par une simple règle de proportionnalité. Ces clés sont mises en place pour diffuser *in fine* des statistiques au niveau le plus fin. Elles correspondent au nombre de ménages par carreau ramené au nombre total de ménages de la dernière zone de niveau diffusable pour ce carreau (c'est le "groupe" au sens défini plus haut, qui peut être par exemple, le carreau du niveau supérieur, mais le plus souvent des agrégations de carreaux proches plus ou moins complexes et pas forcément contigüs). En effet, l'information relative au "groupe" n'est pas considérée comme confidentielle, puisqu'elle peut être retrouvée en théorie par différenciation.

Pour une variable donnée, on diffuse pour un carreau blanchi au niveau creusé la valeur de cette variable estimée par la clé de répartition. Cette valeur correspond au total de la variable au niveau du groupe auquel appartient ce carreau blanchi multiplié par la valeur de la clé de répartition. Cette méthode s'applique aux variables additives. Pour un ratio, on peut appliquer cette méthode séparément pour le numérateur et le dénominateur et reconstituer le ratio.

Cette étape consiste à aller jusqu'au bout de ce que l'Insee demandait aux utilisateurs des données carroyées issues de la source Revenus Fiscaux Localisés, millésime 2010. En appliquant lui-même cette méthode, l'utilisateur prend conscience que la valeur de certains carreaux n'est pas exacte mais est simplement une estimation (donc la valeur est perturbée).

Pour les variables de comptage, cela conduit la plupart du temps à diffuser des nombres non entiers, ou à adopter des règles plus complexes pour diffuser des nombres entiers sans perdre la cohérence.

Selon l'interprétation faite du seuil de 11 ménages, il peut y avoir un ou plusieurs jeux de clés de répartition. Si l'on s'autorise à diffuser n'importe quelle variable sur un carreau dès lors que celui-ci contient au moins 11 ménages, alors un seul système de clés est calculé. En revanche, si l'on considère que pour toute variable, le comptage ne peut être diffusé sans perturbation s'il ne concerne pas au moins 11 ménages, alors il y aura autant de grilles et de systèmes de clés de répartition associés, que de variables diffusées. Dans la suite, pour simplifier le message, on se restreint à la première interprétation.

1.2 La méthode du *swapping*

1.2.1 Principe

La logique du *swapping* est toute différente. On ne cherche plus ici à agréger des carreaux jusqu'à obtenir une zone assez large contenant plus de 11 ménages fiscaux. On décide dès le début de ne diffuser qu'une seule grille composée de carrés de 200m (ou 250m) et de perturber les carreaux contenant trop peu d'observations. Cette perturbation s'obtient en échangeant les positions géographiques de ménages fiscaux, d'où le nom de *swapping*. Cette manière de perturber les données assure qu'au niveau global la distribution de chaque variable est conservée ainsi que leurs corrélations (mis à part la corrélation spatiale). L'utilisation du *swapping* dans le cadre de données géolocalisées a déjà fait l'objet de travaux, dont certains sont décrits dans Buron Fontaine 2018 (chapitre 14 du manuel d'analyse spatiale à paraître [3]). Plus récemment, d'autres développements ont eu lieu pour mettre en place une méthode de *swapping* adaptée à des données carroyées. Cette nouvelle version de la méthode de *swapping*, implémentée dans le langage R par la section Analyse spatiale de la Division des Méthodes et Référentiels Géographiques de l'Insee, est présentée ici, et testée sur les données de Filosofi. L'idée principale est de repérer les ménages considérés comme «à risque» de rupture de la confidentialité, du fait de leur position géographique, et de les échanger entre eux (figure 5).

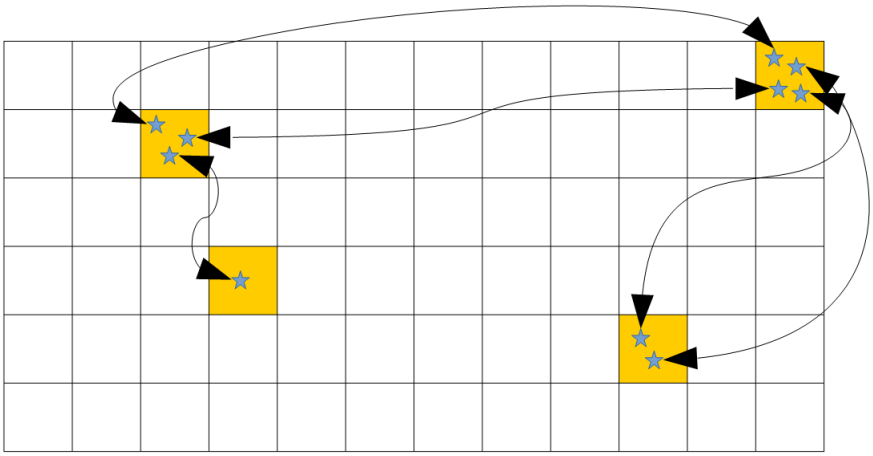

FIGURE 5 – Illustration de la méthode du swapping.
Note : Les carreaux jaunes comportent moins d'observations que le seuil de confidentialité, les ménages à l'intérieur de ces carreaux (représentés par des étoiles) sont alors échangés entre eux.

Les ménages à risque sont ici tous ceux situés dans un carreau comportant moins de 11

ménages fiscaux. On échange ces ménages entre eux selon un algorithme qui vise à :

- échanger deux ménages qui ont des caractéristiques pas trop éloignées, mais pas exactement semblables non plus ;
- minimiser la distance géographique totale des échanges, pour ne pas trop perturber les corrélations spatiales.

Il peut arriver que pour un ménage à risque donné on ne trouve aucun autre ménage à risque satisfaisant toutes les conditions d'échanges (et notamment faire partie d'un carreau différent) : dans ce cas, on s'autorise à échanger le ménage à risque en question avec un ménage qui n'a pas été identifié comme étant à risque. C'est pour cette raison qu'il peut arriver que certains carreaux au-dessus du seuil de confidentialité soient perturbés.

Contrairement à la méthode des grilles superposées, la méthode du *swapping* est largement paramétrable : les spécifications peuvent être adaptées en fonction des besoins. On peut ainsi régler :

- la définition des ménages à risque ;
- les variables sur lesquelles le rapprochement entre ménages doit être effectué, dites variables de "profil" ;
- la plus ou moins grande ressemblance des ménages échangés au regard de ces variables de profil ;
- la zone géographique maximale dans laquelle on s'autorise à déplacer un ménage.

1.2.2 Détails

Dans la méthode, on cherche un optimum global qui minimise la distance géographique entre les observations échangées, sous contrainte d'une similarité minimum des observations. Après avoir ciblé les observations à risque, la recherche des paires se fait par l'utilisation des méthodes d'*optimal matching* ([7], [8], [11]), qui permettent de résoudre un problème d'optimisation globale de façon robuste et performante. Les propriétés notables de la méthode sont les suivantes :

- on garantit que toutes les observations à risque sont échangées à la fin. Pour cela, il est possible que certaines observations qui ne sont pas à risque soient également échangées ;
- les observations à risque ne sont échangées qu'une seule fois ;
- les échanges se font entre observations de carreaux différents ;
- la méthode peut gérer à la fois des contraintes de seuil et des contraintes liées à la différenciation (à condition d'avoir repéré les problèmes en amont par une autre méthode). Le programme pourra aussi gérer d'autres types de contrainte ou des seuils différenciés par variable, modulo quelques adaptations.

On donne ici, plus précisément, les différentes étapes de la méthode mise en place.

1. **Repérage des individus à risque**

   Pour les tests réalisés sur Filosofi 2014, tous les ménages appartenant à un carreau sous le seuil de 11 ménages sont identifiés comme à risque. Des options permettent aussi d'ajouter à cette population :
   - une liste de ménages faisant l'objet de problèmes de différenciation ;
   - des seuils différents de 11 pour d'autres variables.

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 correspondent au croisement suivant :
   - taille du ménage [3] ;

3. On considère une seule catégorie pour les ménages de 5 individus ou plus.

- zone géographique (par exemple le département [4]) ;
- "profil", au sens :
  · pour une liste de variables quantitatives, on considère des quantiles suffisamment espacés pour ne pas trop contraindre (typiquement, il peut s'agir de quartiles de revenu disponible) ;
  · pour une liste de variables qualitatives, on considère des indicatrices de modalités "rares" (au sens partagées par moins de X % de la population à risque, ou à défaut, les N modalités les moins représentées dans la population à risque).

Tous ces paramètres peuvent être ajustés en fonction du niveau de contrainte souhaité pour la recherche des paires. Les quantiles et les modalités rares sont examinés au sein de la population à risque et non de la population générale, l'idée étant de partitionner la population à risque dans des strates de taille les plus homogènes possibles. Comme c'est le croisement de toutes les variables qui est considéré, ce sont les combinaisons de modalités qui sont examinées : on considère la rareté au sens multivarié, ce qui est un autre apport par rapport à la méthode de l'ONS qui considère la rareté uniquement au niveau univarié. Le choix de la liste des variables entrant dans le profil est important et doit dépendre des variables à diffuser :

- pour les variables quantitatives, il peut s'agir directement des variables quantitatives sensibles à diffuser (par exemple le revenu déclaré ou le niveau de vie), car il y a très peu de chance que les individus échangés aient exactement la même valeur ;
- pour les variables qualitatives, il doit s'agir de variables proches des variables qualitatives sensibles à diffuser, mais plus agrégées : par exemple si l'on souhaite diffuser au carreau le nombre de personnes de plus de 65 ans, on peut mettre dans le profil le nombre de personnes de plus de 50 ans : ainsi on permet d'échanger quelqu'un de 65 ans et plus avec quelqu'un de 50 à 65 ans, mais pas au-delà. On ne peut pas mettre exactement la variable à diffuser, car alors on ne confidentialise pas (le nombre de 65 ans et plus serait identique par construction avant et après perturbation).

3. **Retravail des strates ainsi obtenues**

   - Si la strate est trop grosse, on la découpe aléatoirement en deux autant de fois que nécessaire pour passer à une taille en dessous d'une taille cible maximale (des raffinements sont possibles sur cette étape pour découper non pas aléatoirement mais selon des critères géographiques par exemple – dans les tests effectués, cette taille cible est de 500 ménages) ;
   - si la strate (ainsi obtenue) a un nombre impair de ménages (en particulier de taille 1), on complète avec un individu non risqué de même taille, même zone et même profil (s'il n'existe aucun ménage ainsi défini, on relâche progressivement les contraintes sur le profil).

4. ***Matching***

   Cette étape est au cœur du processus. Chaque strate est aléatoirement coupée en deux, et on apparie deux ménages de chaque sous-population en imposant un carreau différent [5]. Pour ce faire, on minimise de façon globale la distance entre ménages échangés grâce à l'algorithme d'optimisation `fullmatch` (package R *optmatch* [6]). Cette méthode est documentée dans la littérature.

4. Il s'agit du niveau auquel on souhaite conserver les marges, mais celui-ci ne peut pas être trop fin sous peine de strates trop petites.

5. En théorie, il pourrait arriver que les strates constituées ne permettent pas de respecter cette contrainte. En pratique on n'a pas rencontré de cas où ça ne marchait pas. Il faudra sûrement envisager une solution en cas de strate entièrement incluse dans un carreau.

6. Ce package a l'avantage de mettre des alertes sur le temps d'exécution si celui-ci s'annonce long et il propose aussi un paramètre (`tol`) qui permet de choisir à quel point on permet de s'éloigner de la

Lorsque la strate est de faible effectif (le ménage présente une combinaison de caractéristiques plutôt rare), alors il faudra aller chercher plus loin géographiquement pour trouver un ménage similaire. *A contrario*, si la strate est de grande taille, l'algorithme de minimisation a toutes les chances de rapprocher deux individus proches géographiquement. La distance entre les ménages rapprochés selon cet arbitrage est donc variable, mais rarement très élevée, comme l'illustre la figure 6.

5. ***Swapping***

   Enfin, on procède à l'échange à proprement parler. Dans la table originale de niveau ménage, on échange les variables géographiques (c'est-à-dire les coordonnées x ;y ainsi que toutes les variables déterminées intégralement par les coordonnées x ;y, comme la commune ou l'appartenance à un quartier politique de la ville).

FIGURE 6 – Bretagne : exemples de permutations entre ménages avec l'algorithme proposé (100 paires choisies au hasard parmi celles effectivement échangées).
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).

2 Comparaison des méthodes

2.1 Quelques indicateurs

Dans la source Filosofi 2014, il y a 27,6 millions de ménages en France métropolitaine. On compte 1,9 millions de carreaux de 250 m de côté habités dont 77 % contiennent moins de 11 ménages fiscaux, ce qui représente 14,5 % des ménages. Pour la maille de 200 m, il y a 2,3 millions de carreaux dont 79 % sous le seuil de confidentialité, ce qui représente 18,3 % des ménages. La proportion de ménages situés dans des carreaux sous le seuil de confidentialité varie sensiblement selon les régions : elle est la plus faible pour la région Île-de-France (2,2 % pour la maille de 200 m) et la plus forte pour la région Aquitaine-Limousin-Poitou-Charentes (32,8 % pour 200 m).

2.1.1 Niveau naturel et niveau creusé

Passer du niveau naturel au niveau creusé permet d'augmenter la part de ménages diffusés à un niveau fin. Par exemple, avec la grille 200 m - 1 km-etc., 65 %[7] des ménages sont diffusés sur une maille d'un kilomètre ou une maille plus fine tandis que cette proportion est de 94 % lorsqu'on utilise le niveau creusé (voir tableau 1).

solution exacte pour gagner en temps de calcul. Il est aussi plus rapide que d'autres packages équivalents car il est codé dans un langage de bas niveau et travaille sur des objets optimisés (*InfinitySparseMatrix*).

7. Dans le tableau 1, 64,5 = 51,1 + 13,4 et 94,0 = 80,8 + 13,2.

Taille de la maille	Part de ménages diffusés au niveau le plus fin (%)				Part de ménages dans des carreaux au-dessus du seuil (%)
	Naturel		Creusé		
	200	250	200	250	Total
32 km	0,3		0,0	0,0	100,0
16 km	1,4		0,0	0,0	100,0
8 km	2,8		0,1	0,1	100,0
4 km	11,2		0,9	0,9	100,0
2 km	19,7		5,0	4,2	99,6
1 km	51,1	12,6	13,2	6,6	97,0
500 m		9,9		11,2	92,6
250 m		42,0		77,1	85,5
200 m	13,4		80,8		81,7
Total	**100,0**	**100,0**	**100,0**	**100,0**	

TABLE 1 – Part des ménages diffusés, selon le niveau le plus fin sur lequel ils sont diffusés.
Champ : France métropolitaine.
Source : Insee, Filosofi 2014.

La grille 250m-500m-1km-etc. permet de diffuser très légèrement plus de ménages à une précision d'au moins 1 km² (95 %). Néanmoins, au niveau le plus fin possible, la première grille (200m) diffuse plus de ménages (80,8 %) que la seconde (250 m), qui en diffuse 77,1 %. Il apparaît difficile de comparer les deux grilles pour savoir laquelle permet de diffuser des données plus précisément. Pour ce faire, on propose ci-après un indicateur de précision géographique qui permet de synthétiser les informations du tableau 1 en un seul chiffre.

2.1.2 Niveau creusé : indicateur de précision géographique pour différentes grilles

On considère l'indicateur suivant :

$$I = \frac{N}{\sum_{i=1}^{N} ln(m_i^2)} \tag{1}$$

où N est le nombre total de ménages et m_i est la taille du carreau le plus fin auquel le ménage i peut être diffusé sans perturbation. L'indicateur I mesure à quel point les informations des ménages sont diffusées sur une zone géographique précise. Si les ménages sont plutôt diffusés sur des carreaux de petite taille, alors I est élevé. À l'inverse, I est faible si les ménages sont plutôt diffusés sur des carreaux de grande taille.

Prendre le logarithme de la surface sur laquelle est diffusée l'information relative à un ménage permet d'atténuer la perte de précision géographique qui serait due à quelques ménages diffusés sur de très grands carreaux. On peut remarquer que si on discrétise l'espace, m_i^2 correspond au nombre d'emplacements possibles pour le ménage i et donc $\prod_i m_i^2$ est le nombre d'états possibles pour la répartition des ménages sur le territoire. Il en résulte que l'indicateur I est inversement proportionnel à la formule de l'entropie.

Pour être plus interprétable, on considère en fait un indicateur normalisé, c'est-à-dire le ratio entre I et I_{200}, qui correspond au même indicateur mais dans lequel on remplace m_i par 200, comme si tous les ménages étaient diffusés au niveau le plus fin parmi ceux considérés, à savoir

celui de 200 mètres. Ainsi, l'indicateur normalisé vaut entre 0 % et 100 % : s'il valait 100 % cela signifierait que tous les ménages sont diffusés sur des carreaux de 200 mètres de côté. La précision évaluée au sens de cet indicateur est meilleure pour le maillage allant jusqu'aux carreaux de 200 m, même s'il n'existe pas de maille intermédiaire entre le km et la maille la plus fine : l'indicateur normalisé vaut 93,7 % contre 91,3 % pour l'autre maillage envisagé (1km, 500m, 250m).

2.1.3 Grilles superposées avec clés de répartition ou *swapping*

Niveaux de perturbation

Evaluer une méthode simplement au regard de la part de carreaux perturbés donnerait une image trompeuse de la déformation induite. En effet, même si la grande majorité des carreaux sont perturbés, en réalité seuls une minorité de ménages sont contenus à l'intérieur. La carte avant / après d'une variable donnée peut donc être assez fortement modifiée dans une zone peu dense, alors même que peu de ménages ont été perturbés pour la variable en question (c'est l'effet "carte").

Sans renoncer à regarder ce genre de cartes, on considère un indicateur rendant mieux compte de la perturbation totale introduite. Il s'agit du pourcentage de la masse totale de la variable ayant fait l'objet d'une perturbation, c'est-à-dire l'indicateur :

$$\frac{\sum_{c \in carreaux} |V_c^{apres} - V_c^{avant}|}{\sum_{c \in carreaux} V_c^{avant}} \tag{2}$$

À méthode équivalente, la maille de 250 m induit légèrement moins de perturbation que la maille de 200m. À taille de maille équivalente : pour les variables quantitatives, les deux méthodes induisent des niveaux de perturbation comparables ; pour les variables qualitatives, la méthode des grilles superposées de niveau creusé avec clés de répartition induit davantage de perturbation que le *swapping*. On rappelle néanmoins que les masses de perturbation induites par le *swapping* dépendent assez fortement du jeu de paramètres choisi (dans le jeu de paramètres choisi pour les tests on impose qu'un ménage soit échangé avec un ménage de même taille ou presque et de même quartile de niveau de vie, d'où le fait qu'on ait, le plus souvent, permuté un ménage pauvre avec un autre ménage pauvre). Toujours sur cet indicateur, il convient également de noter que dans l'hypothèse où le seuil de 11 ménages serait à respecter quelle que soit la variable, les masses de perturbation seraient plus élevées. En effet, si un carreau contient 100 ménages dont 10 ménages pauvres, aucun ménage n'a eu besoin d'être perturbé pour ces tests, alors que 10 l'auraient été avec l'autre interprétation.

Distribution des perturbations par commune

On se met à la place d'un utilisateur qui souhaite connaître le total d'une variable par commune et qui ne disposerait que des données carroyées diffusées. Pour cela, il fait l'hypothèse qu'un carreau appartient à la commune à laquelle son centroïde appartient. On a fait ce calcul pour l'ensemble des 36500 communes de France métropolitaine, et on examine comment ces proxys communaux s'éloignent des totaux «réels» (tels qu'observés dans Filosofi).

Pour une maille de 250m, et pour la variable quantitative considérée en figure 6 (somme des niveaux de vie), le total est sous-estimé en moyenne de -0,2 % pour le swapping et -0,8 % pour la méthode des grilles superposées de niveau creusé avec clés de répartition («GS + CR»), en moyenne non pondérée sur l'ensemble des communes. Cet écart, faible en moyenne, est rarement élevé : pour les deux méthodes, les centiles d'ordre 1 et 99 se situent autour de -20 % et +20 %,

	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

TABLE 2 – Mesures des perturbations induites par la méthode des clés de répartition et la méthode du *swapping*.
Champ : France métropolitaine.
Source : Insee, Filosofi 2014.
Note : PR = personne de référence

et ceux d'ordre 5 % et 95 % se situent autour de -10 % et +10 %. Plus précisément, si l'on s'intéresse à la distribution de cet écart à l'intérieur de cet intervalle, l'écart semble légèrement plus élevé avec la méthode de *swapping* qu'avec GS+CR (voir figure 6).

En réalité, l'écart global peut se décomposer en deux :

1. l'écart entre le total de la commune et celui de son approximation par carreaux (commun aux deux méthodes) ;
2. l'écart entre l'approximation par carreaux sans méthode de protection et l'approximation par carreaux avec méthode de protection.

Les proportions respectives de l'un et de l'autre dans l'écart total varient selon la variable considérée et éventuellement le jeu de paramètres choisi.

Nombre de résidences principales

À moyen terme, il est envisagé de carroyer les résultats du recensement de la population. Pour que ces données carroyées soient comparables aux données carroyées de Filosofi, il faut que ces deux sources soient diffusées sur la ou les même(s) grille(s). Afin de savoir si le fait de disposer d'un niveau intermédiaire, entre la grille de maille 1km et la grille la plus fine, permettrait d'avoir de meilleures estimations locales du recensement, on a calculé le nombre de résidences principales par carreau dans les villes de plus de 10 000 habitants. Plus ce nombre est élevé, plus la qualité de l'estimation sera bonne. Dans les villes de moins de 10 000 habitants, le problème de la qualité de l'estimation ne se pose pas car le recensement y ait exhaustif.

Le tableau 3 indique que plus la taille de la maille est fine, plus la proportion de carreaux ayant plus de 1 000 logements est faible. Même pour la maille d'un kilomètre, cette proportion est assez faible puisqu'elle ne dépasse pas 20 %. La grille intermédiaire composée de carreaux de 500 m n'offre que peu de carreaux comptant plus de 1 000 logements (2 854 carreaux soit 4,4 %

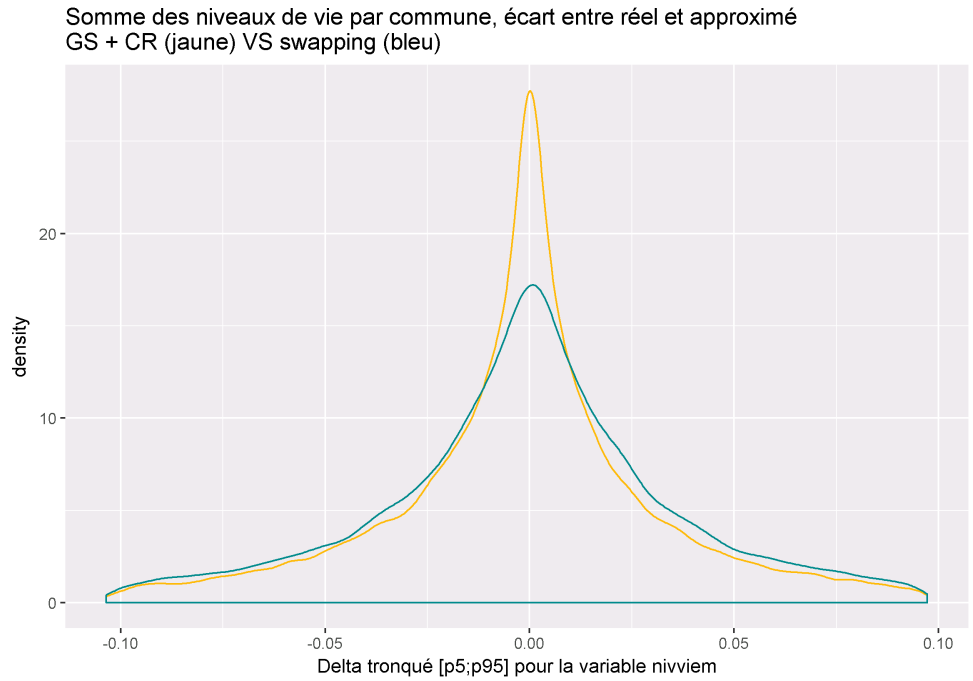

FIGURE 7 – Distribution de l'écart relatif entre les totaux réels et les totaux imputés, par commune, pour la variable indiquant le niveau de vie des ménages.
Source : Insee, Filosofi 2014
Champ : ensemble des communes de France métropolitaine (pas de pondération)

	Nombre de carreaux	**Nombre de carreaux >1000 logements**	**Proportion de carreaux >1000 logements (%)**	**Nombre de carreaux >100 logements**	**Proportion de carreaux >100 logements (%)**
200 m	248 885	256	0,1	39 246	15,8
250 m	180 240	693	0,4	40 133	22,3
500 m	65 089	2 854	4,4	25 135	38,6
1000 m	22 778	4 366	19,2	11 507	50,5

TABLE 3 – Carreaux habités présents sur les villes de plus de 10000 habitants.
Champ : carreaux dont au moins un logement est situé dans une ville de plus de 10 000 habitants, France métropolitaine.
Note : ces indicateurs varient très peu en se restreignant aux carreaux dont l'intégralité des habitants sont dans une commune de plus de 10 000 habitants.
Source : Insee, Filosofi 2014.

des carreaux de 500 m dans les villes de plus de 10 000 habitants).

2.2 Avantages et inconvénients

L'avantage de la méthode des grilles superposées est de pouvoir proposer trois produits aux utilisateurs :

- niveau naturel, sans clés de répartition : un pavage complet du territoire avec des carreaux de taille différente, selon la densité de la population ;
- niveau creusé, sans clés de répartition : plusieurs grilles de différentes mailles carré comportant des "trous" ou bien une seule grille de mailles de formes et tailles variables sans "trous" obtenue par différenciation. Le niveau creusé permet de diffuser le plus finement possible toute l'information sans perturbation ;
- niveau creusé, avec clés de répartition : une grille la plus fine possible avec perturbation des carreaux sous le seuil de confidentialité.

Le deuxième produit semble plus difficile à exploiter par la plupart des utilisateurs car chaque grille comporte des «trous» (ou alors la grille obtenue par différenciation comporte des mailles de formes variables) : il est alors difficile de reconstituer un zonage complet en dehors des zones très denses.

Au contraire, la méthode du *swapping* ne propose pas un découpage sans perturbation. Dans les zones très denses, les ménages ne seront pas échangés, car les carreaux respecteront le seuil de confidentialité, tandis que dans les zones peu denses beaucoup de ménages seront déplacés dans des carreaux voisins. L'avantage principal du *swapping* par rapport aux grilles superposées est de pouvoir mieux gérer d'autres risques que seulement celui lié au nombre d'observations dans un carreau. On peut en effet relativement facilement ajouter à la population dite «à risque» les ménages atypiques par rapport à leurs voisins du carreau, et ce même si ce carreau contient beaucoup d'observations. Les ménages ayant été échangés selon un algorithme qui fait intervenir de l'aléatoire, il semble aussi plus compliqué pour un intrus de retrouver des informations confidentielles exactes par la technique de la différenciation. Enfin, pour les deux méthodes perturbatives proposées (clés de répartition et *swapping*), se pose le problème des carreaux contenant un seul ménage (ou très peu de ménages, comme deux ou trois). Peut-on diffuser des informations sur ces carreaux même si ces informations sont perturbées ? Si oui, doit-on s'assurer que les informations perturbées sont suffisamment différentes des vraies informations ? En effet, pour les variables qualitatives, aucune des deux méthodes ne garantit que les comptages diffusés soient effectivement différents des comptages protégés. Pour les variables quantitatives, rien ne le garantit non plus pour la méthode des clés de répartition, tandis que pour la méthode du *swapping* cela peut se contrôler en amont dans la paramétrisation du processus d'échange des ménages.

3 Ce qui est fait à l'étranger

3.1 Le projet *Geostat*

Geostat est un projet d'Eurostat qui vise à diffuser les données du recensement de la population des différents pays sur des carreaux d'un kilomètre de côté. Deux jeux de données ont déjà fait l'objet d'une diffusion : *Geostat* 2006 et *Geostat* 2011 (voir rapports finaux [1] et [5]).

3.2 La Grant "*Harmonized protection of census data in the ESS*"

Au début de cette Grant, deux questionnaires ont été envoyés en 2015 et 2016 aux pays européens participant à la Grant. Il en ressort que l'Autriche, la Belgique, la France et la Suède

indiquent utiliser une forme post-tabulée de *swapping* (sans rapport direct avec les formes pré-tabulées de *swapping* évoquées plus haut, voir [10]) pour traiter la confidentialité dans les hypercubes. Seuls 21 pays déclarent produire des données carroyées pour le recensement de la population, la moitié sur une grille unique et l'autre sur plusieurs grilles de tailles différentes. Parmi ces pays, 18 n'utilisent aucune méthode spécifique.

Le cœur de cette Grant d'Eurostat consistait à identifier et tester des méthodes permettant la protection des données du recensement de la population lorsqu'ils sont diffusés sous forme d'hypercubes. Les données carroyées étaient aussi spécifiquement abordées dans ces travaux, mais le plus souvent la dimension géographique y était vue comme une variable catégorielle et donc comme une dimension de l'hypercube.

Au terme de cette Grant, Eurostat a formulé la recommandation [8] de combiner méthodes pré-tabulées prenant en compte la géographie (comme le *targeted record swapping*, similaire dans ses grands principes à la méthode testée dans le cadre de la diffusion de Filosofi) et méthodes post-tabulées (comme la *cell-key method*, confer [6]).

3.3 Sur les sites internet des INS

En parcourant les sites internet des Instituts Nationaux de Statistiques (INS) des différents pays, on s'aperçoit que plusieurs pays, en Europe ou ailleurs, diffusent des données finement localisées, soit sous forme de carreaux soit sur des zones construites pour la collecte du recensement de la population. La plupart du temps il s'agit de données issues du recensement.

De façon non exhaustive, on peut citer la Norvège, le Danemark, la Pologne, l'Autriche, la Finlande, la Lettonie, l'Estonie et l'Espagne qui produisent et diffusent (de façon libre ou payante) des données carroyées en Europe, et les Etats-Unis et le Japon parmi les pays non européens. Les grilles sont toujours des grilles régulières de carrés ; certains diffusent une seule grille et d'autres plusieurs grilles. Les grilles les plus fines sont de 50 mètres pour l'Espagne, 100 mètres pour l'Autriche, le Danemark, la Lettonie, l'Estonie, 250 mètres pour la Finlande, 1 km environ pour la Pologne, les Etats-Unis et le Japon.

Les règles de confidentialité observées sont souvent liées à un seuil minimal d'observations en-deça duquel on ne peut pas diffuser l'information. Les méthodes trouvées sur les sites internet des INS permettant de respecter ces règles sont le blanchiment de carreaux ou l'agrégation de carreaux. Plus précisément, l'Espagne utilise une méthode de découpage itératif qui ressemble à la méthode des rectangles (les carreaux obtenus en fin de processus sont des carrés ou des rectangles). La Norvège indique dans un document publié en 2009 que plusieurs méthodes ont été envisagées pour protéger les carrés en dessous de 3 observations : blanchiment, agrégation, imputation par la moyenne et floutage. Selon leur Institut, le blanchiment est gênant, car on ne peut pas retrouver les totaux sur le pays entier, l'agrégation conduit à des carreaux de tailles différentes ce qui les rend difficiles d'utilisation. Il a été décidé de ne pas indiquer le nombre exact de ménages lorsque celui-ci se trouvait entre 1 et 9, et de seulement indiquer une fourchette (entre 1 et 9 ménages).

D'autres pays diffusent des données finement localisées, il s'agit presque toujours du recensement de la population, sur les plus petites unités de collecte de l'information : les données ne sont pas carroyées mais un certain nombre d'indicateurs sont disponibles sur des zones plus ou moins petites, selon la densité de la population sous-jacente. Parmi les pays dans cette situation, on peut citer le Canada, le Mexique, le Chili, l'Australie, l'Angleterre, l'Irlande, l'Allemagne ou encore le Portugal.

8. Le livrable D3.4 est disponible à l'adresse suivante : `https://ec.europa.eu/eurostat/cros/print/book/export/html/13344_en`.

En règle générale, les zones élémentaires sont construites de façon à ce qu'il y ait suffisamment d'observations dans chacune, ce qui ne nécessite pas de traitement particulier de la confidentialité ensuite. Néanmoins, on peut citer l'Australie qui utilise une méthode perturbative pour protéger les zones géographiques (*meshblock*) contenant très peu de monde : il s'agit de la "*cell-key method*" développée par leur institut de statistique. C'est une méthode post-tabulée utilisée lorsqu'on veut protéger des variables de comptage. Elle permet de rajouter du bruit sur certaines cellules de la table à diffuser tout en gardant la cohérence des nombres totaux et marginaux d'observations.

Conclusion

La méthode des grilles superposées, à elle seule, est une méthode non-perturbative qui conduit à diffuser de l'information sur des carreaux de différentes tailles, selon la densité de population sous-jacente. Le niveau "naturel" permet un pavage en carrés du territoire, mais fournit des informations moins fines que le niveau «creusé». Le niveau creusé peut être difficile à manipuler et à comprendre pour un utilisateur standard, car il s'agit de plusieurs grilles avec des niveaux de précision différents, et chaque grille est "trouée". Pour avoir un pavage complet du territoire il faut combiner ces différentes grilles entre elles, pour obtenir les plus petites zones agrégées sur lesquelles on peut avoir de l'information. Un autre inconvénient est que ces zones au sein desquelles l'information est diffusée, ne sont pas forcément des carrés réguliers, mais peuvent être des ensembles complexes de carrés, pas forcément contigus.

Pour gagner en facilité d'utilisation, la méthode des grilles superposées au niveau creusé peut être complétée par la méthode des clés de répartitions, qui consiste à répartir, proportionnellement au nombre d'observation par petit carreau, les variables diffusées sur les zones agrégées (ensemble de carreaux contigus ou non respectant le seuil). On obtient alors en combinant ces deux méthodes une méthode perturbative : les carrés initialement blanchis au niveau "creusé" sont perturbés et on ne diffuse pas l'information exacte sur ces carrés. L'autre méthode perturbative envisagée, le *swapping*, consiste à échanger aléatoirement, tout en respectant des contraintes de proximité géographique et de ressemblance en termes de caractéristiques, les ménages situés dans les carrés contenant moins de 11 ménages fiscaux. Cette méthode a l'avantage de pouvoir être paramétrée : on peut ainsi contrôler, plus ou moins, à quel point les carreaux sous le seuil de confidentialité seront perturbés.

Dans les deux cas, avec les méthodes perturbatives, on sera amené à diffuser des informations qui ne correspondent pas à ce qui est observé originellement dans la source : on modifie volontairement la statistique, et donc l'utilité des données, en échange d'une protection du secret.

Références

[1] Backer, L. H., van Leeuwen, N. F., Lipatz, J.-L., Tammilehto-Luode, M., Tammisto, R., Makarenko-Piirsalu, D., Marsik, K., Verner, V., Bloch, H., Gundersen, G. I., Thorsdalen, B., Jablonski, R., Santos, A. M., Kuzma, I., and Kaminger, I. Geostat 1a : Representing census data in a european population grid. *Final Report* (December 2011).

[2] Behnisch, M., Meinel, G., Tramsen, S., and Diesselmann, M. Using quadtree representations in building stock visualization and analysis. *Erdkunde* (2013), 151–166.

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

[4] de la Statistique et des Etudes Economiques, I. N. Guide du secret statistique. *Documentation INSEE* (October 2010).

[5] ET AL., H. A point-based foundation for statistics : Final report from the geostat 2 project. *Eurostat, EFGS, Insee* (January 2017).

[6] FRASER, B., AND WOOTON, J. A proposed method for confidentialising tabular output to protect against differencing. *Monographs of Official Statistics. Work session on Statistical Data Confidentiality* (2005), 299–302.

[7] HANSEN, B. B. Full matching in an observational study of coaching for the sat. *Journal of the American Statistical Association 99*, 467 (2004), 609–618.

[8] 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.

[9] INSEE. Documentation complète sur les données carroyées à 200 mètres. *Online documentation* (November 2013).

[10] KOUMARIANOS, H. Traitement de la confidentialité dans la réponse au règlement européen sur les recensements de la population et du logement. *Séminaire DMS, Insee* (2014).

[11] ROSENBAUM, P. R. A characterization of optimal designs for observational studies. *Journal of the Royal Statistical Society. Series B (Methodological)* (1991), 597–610.

Annexes

Bref historique des méthodes de carroyage à l'Insee

À notre connaissance, deux sources ont déjà fait l'objet d'un carroyage à l'Insee : les Revenus Fiscaux Localisés (RFL) de 2010 et le Recensement de la Population (RP) de 2011. Le carroyage du recensement provient d'une demande d'Eurostat pour le projet Geostat dans le cadre du service Gisco, le système d'information géographique de la Commission Européenne. L'Insee a fourni en 2014 des estimations de population 2011 sur des carrés d'un kilomètre de côté. Le carroyage de la source Revenus Fiscaux Localisés (RFL) 2010 ne provient pas d'une demande internationale ; toutes les spécifications ont été choisies par l'Insee. Nous présentons brièvement ci-dessous la méthode dite "des rectangles", utilisée pour la diffusion du premier trimestre 2013.

Partition du territoire en rectangles
L'information de base (ou unité statistique) est un carreau de 200 mètres. Pour des raisons de performance informatique, La France est d'abord découpée en 36 carrés de taille égale ayant le même nombre de carreaux, qu'ils soient habités ou non. On applique alors la méthode dite des rectangles ([9]). Elle consiste à découper itérativement chaque carré initial, puis chaque sous-rectangle, en deux, soit dans la longueur soit dans la largeur, selon un critère qui permet de minimiser la dispersion géographique des ménages au sein des sous-rectangles obtenus. Chaque sous-rectangle est lui-même constitué de carreaux de 200 mètres. L'itération se termine lorsque les découpages horizontaux ou verticaux conduisent tous les deux à au moins un sous-rectangle ayant moins de 11 ménages. On obtient alors un pavage du territoire français en rectangles de tailles et formes variées.

Les rectangles sont des produits intermédiaires permettant de gérer la confidentialité. Ils ne doivent pas être utilisés en tant que tels, en particulier pour faire des cartes. Les rectangles étant de superficies variables, des cartes d'effectifs seraient en effet fallacieuses. Par ailleurs, pour les rectangles de taille relativement importante, la valeur globale masque d'éventuelles fortes disparités spatiales internes. Il convient ainsi de travailler au niveau des carreaux.

Un sous-rectangle final correspond, soit à un carreau de 200 mètres ayant plus de 11 ménages pour lequel l'information est diffusée telle quelle, soit à un regroupement de carreaux dont au moins un a moins de 11 ménages. Dans ce dernier cas, l'utilisateur est invité à reconstituer l'information au niveau carreau en répartissant le total du rectangle selon le nombre d'habitants du carreau qui est considéré comme diffusable.

Winsorisation
Enfin, avant toute agrégation, lors de cette deuxième diffusion, les revenus fiscaux par unité de consommation ont été winsorisés. Si le revenu d'un ménage est supérieur au 8ème décile de la distribution, il est rabaissé à ce seuil ; s'il est inférieur à 40 % de la médiane, il y est remonté : il s'agit d'une "winsorisation fixe". Les déciles sont définis au niveau national. Dans quelques rares cas, le revenu n'est pas rabaissé au seuil, mais à une valeur aléatoire dans un intervalle de 500 euros inférieur au seuil (respectivement remonté à une valeur aléatoire dans un intervalle de 500 euros supérieur au seuil). Pour chaque rectangle est diffusée la somme des revenus winsorisés par unité de consommation des individus. De plus, quatre autres variables considérées comme «sensibles» ont fait l'objet d'un traitement spécifique.

Différenciation et secret statistique secondaire

Diffuser des données sur une grille de carreaux d'une part, et sur des contours administratifs, comme la commune, l'Iris ou le Quartier Politique de la Ville (QPV) d'autre part, peut conduire à une rupture du secret statistique secondaire. Celui-ci correspond à la capacité qu'a un utilisateur de combiner intelligemment les données diffusées selon différentes nomenclatures géographiques afin de déduire des informations confidentielles.

La technique de la différenciation peut s'appliquer lorsque les variables diffusées sont additives (sommes ou moyennes de variables quantitatives ou de comptage) : il s'agit de déterminer une zone englobante avec la première nomenclature (par exemple les communes) et de retirer une zone englobée constituée de carreaux. On peut parfois en déduire de l'information sur une zone pour laquelle le secret statistique primaire (seuil de confidentialité de 11 ménages fiscaux) n'a pas été contrôlé (exemple en figure 8).

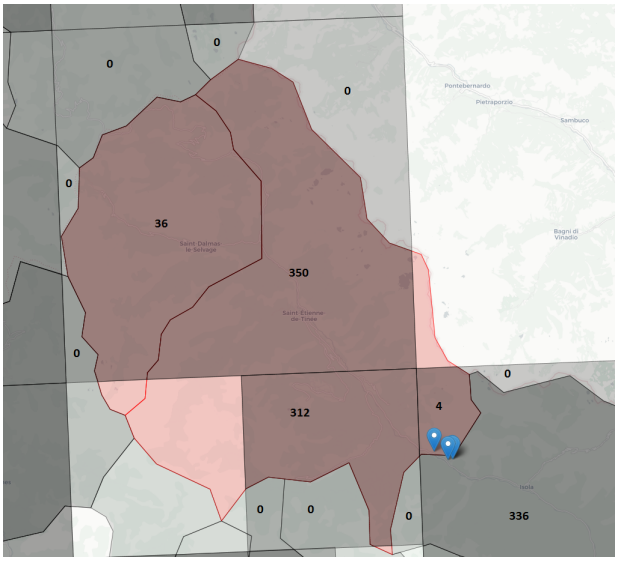

FIGURE 8 – Illustration du problème de différenciation, région PACA.
Note : En faisant la différence entre les nombres d'habitants des deux communes en rouge, et le nombre d'habitants des deux carreaux centraux grisés, on en déduit une zone de 4 habitants (correspondant aux points bleus). Les chiffres indiquent le nombre de ménages présents sur les intersections carreaux-communes.

Pour le *swapping*, les données ayant été perturbées, un utilisateur ne peut avoir la certitude que les sommes qu'il effectue soient les valeurs exactes. Toute différenciation effectuée aura une certaine probabilité d'être perturbée, et donc de ne pas fournir des informations confidentielles exactes. Mais ce seul argument ne nous assure pas qu'aucune différenciation et rupture du secret ne soit effectivement possible.

Les informations relatives à des intersections de carreaux et de communes présentant moins de 11 ménages ne sont pas toutes déductibles par différenciation. Afin de limiter la perte d'information liée à la protection du secret secondaire on cherche les croisements carreaux-communes (ou carreaux-Iris, carreaux-QPV) qui peuvent réellement être obtenus par différenciation. Par exemple, avec le carroyage des données de Filosofi 2014 au niveau creusé jusqu'à 200 m, il y a 814000 croisements carreaux-communes dont 133000 comptent moins de 11 ménages fiscaux. Tout l'enjeu est de déterminer parmi ces 133000 croisements ceux qui sont réellement atteignables

par différenciation.

Le nombre de combinaisons possibles pour réaliser des différenciations est absolument gigantesque et il est hors de portée de toutes les tester. On a néanmoins développé une méthode qui permet de s'assurer que tous les cas simples de différenciation sont pris en compte et qu'un maximum de cas problématiques soient détectés.

Pour cela, on a mis en place une procédure en deux temps :

1. simplification : on cherche à agréger des communes entre elles sachant qui si ces communes sont considérées séparément il ne peut y avoir de problème de différenciation qui apparaît ;
2. recherche exhaustive : on teste toutes les combinaisons possibles sur les données simplifiées, avec des tailles d'agrégats croissants, jusqu'à une taille d'agrégat maximale (environ 15). Au-delà, les capacités de calculs informatiques actuelles commencent à être limitantes. Cette étape est réalisée à l'aide d'un programme développé spécifiquement pour l'occasion par la Division Statistiques et Analyses Urbaines (DSAU) de l'Insee. Ce programme est écrit en C++ et peut être appelé directement à partir du logiciel R.

Pseudo-code de la fonction `DetermineEtatForceGroupe`

Algorithm 1 Détails de la fonction `DetermineEtatForceGroupe`

Require: id^{pere}, E^{pere}, F^{pere}, G^{pere}, seuil
Ensure: list((E_1^{fils}, F_1^{fils}, G_1^{fils}), ..., (E_k^{fils}, F_k^{fils}, G_k^{fils})

$Tot_B \leftarrow 0$ {Initialisation de variables utilisées par la suite}
$n_i \leftarrow$ nombre observations du fils i
on trie les n_i par ordre croissant et on réindexe : $n_1 \leq n_2 ... \leq n_k$
if $E^{pere} = 1$ **then**
  for i de 1 à k **do**
    if $n_i < seuil$ **then**
      $E_i = 0$; $F_i = 0$; $G_i = 0$
      $Tot_B = Tot_B + n_i$
    else if $n_i \geq seuil$ **then**
      if $0 < Tot_B < seuil$ **then**
        $E_i = 0$; $F_i = seuil - Tot_B$; $Tot_B = Tot_B + n_i$
      else if $Tot_B \geq seuil$ ou $Tot_B = 0$ **then**
        $E_i = 1$; $F_i = 0$
      end if
    end if
  end for
else if $E^{pere} = 0$ et $F^{pere} = 0$ **then**
  pour tout i de 1 à k, $E_i = 0$; $F_i = 0$
else if $E^{pere} = 0$ et $F^{pere} > 0$ **then**
  for i de 1 à k **do**
    if $n_i < seuil$ **then**
      $E_i = 0$; $F_i = 0$; $Tot_B = Tot_B + n_i$
    else if $n_i \geq seuil$ **then**
      if $Tot_B < F^{pere}$ then $E_i = 0$; $F_i = F^{pere} - Tot_B$; $Tot_B = Tot_B + n_i$
      if $Tot_B \geq F^{pere}$ then $E_i = 1$; $F_i = 0$
    end if
  end for
end if