JMS Journées de méthodologie statistique de l'Insee 2018

Identification des problèmes de différenciation géographique

Vianney COSTEMALLE (*)

(*) Insee, Département des méthodes statistiques, Division des méthodes et référentiels géographiques

vianney.costemalle@insee.fr

Mots-clés. Secret statistique, confidentialité, différenciation, données géographiques, carroyage, graphes.

Résumé

Lorsqu'on diffuse des données on s'assure généralement qu'elles ne concernent pas moins d'un nombre minimal d'observations (seuil de confidentialité) afin de respecter le secret statistique. C'est en particulier le cas lorsqu'on diffuse des données selon un zonage géographique : chaque zone doit être suffisamment grande pour comporter plus d'observations que le seuil de confidentialité. Un problème supplémentaire se pose néanmoins si une même source de données est diffusée selon plusieurs zonages géographiques simultanément : par combinaison et recoupement des diverses zones, un utilisateur pourrait déduire des informations sur des zones plus petites que celles qui avaient été prévues pour garantir le secret statistique au départ. Par conséquent, rien n'assure dans ce cas qu'une rupture du secret ne soit pas possible. La différenciation géographique est une technique qui consiste à faire la différence entre une zone englobante (composée par exemple de communes) et une zone englobée (composée par exemple de carreaux) et déduire des informations sur une plus petite zone. Il y a rupture de la confidentialité s'il est possible de déduire des informations concernant un nombre d'observations inférieur au seuil de confidentialité. Comment repérer l'ensemble des lieux géographiques où un problème de différenciation apparaît ? En effet, la zone englobante peut être particulièrement complexe du fait qu'il y a environ 36 000 communes en France. Il est impossible de tester une à une l'ensemble des combinaisons possibles de communes. Les travaux présentés visent à donner une représentation du problème de la différenciation géographique sous forme de graphe orienté. Tout l'enjeu est alors de repérer les sous-graphes connexes conduisant à un problème de différenciation. Pour cela on effectue une simplification du graphe de départ en fusionnant certains sommets selon deux méthodes. On rétrécit la taille du graphe, jusqu'à converger vers les zones les plus problématiques. Si le graphe obtenu après fusion des sommets est assez petit, on peut tester exhaustivement toutes les combinaisons. Sinon, on étudie la structure du graphe, pour adopter une stratégie adaptée. On présentera les résultats de ces méthodes appliquées aux données carroyées de Filosofi 2014 pour lesquelles le seuil de confidentialité est de 11 ménages fiscaux.

Abstract

La différenciation géographique est une technique qui consiste à faire la différence, pour une variable de nature additive, entre deux zones géographiques afin de déduire de l'information sur une zone géographique plus petite. Cette technique peut mener à une rupture du secret statistique dans le cas où des données sont diffusées selon deux zonages différents. L'objet de cet article est de fournir une représentation du problème sous forme de graphe et de proposer des méthodes pour simplifier et explorer ce graphe. On applique cette démarche aux données de Filosofi 2014 diffusées au niveau des communes et de carreaux. On trouve alors qu'il y a environ 10 000 ménages à risque, c'est-à-dire pour lesquels on peut déduire de l'information concernant une zone contenant un de ces ménages et au plus 9 autres ménages.

Remerciements

Je tiens à remercier Marc Branchu avec qui j'ai eu de nombreuses discussions stimulantes sur le sujet qui m'ont aidé à clarifier mes idées et à considérer de nombreuses situations que je n'avais pas imaginées. Je remercie également Arlindo Dos Santos et François Sémécurbe pour leurs précieux conseils et pour avoir développé un programme spécifique permettant de tester exhaustivement toutes les différenciations possibles concernant un agrégat de taille fixé. Enfin, merci à Maëlle Fontaine et Vincent Loonis pour leur soutient permanent et leurs remarques pertinentes.

Introduction

Le secret statistique

Le secret statistique concerne les informations diffusées par un institut comme l'Insee : ces informations ne doivent pas révéler l'identité d'un individu, ni dévoiler des caractéristiques d'individus, de ménages ou d'entreprises qui seraient confidentielles ou personnelles. Le plus souvent en pratique, la méthode de protection des données consiste à ne pas diffuser d'information résultant de l'agrégation d'un nombre trop petit d'observations. Pour cela, on fixe un seuil de confidentialité. Ce seuil varie selon la source et la variable considérées.

Lorsqu'on diffuse des données avec une composante géographique (c'est le cas des cartes thématiques par exemple), on doit aussi veiller à ce que le secret statistique soit respecté, c'est-à-dire à ce qu'aucune des zones géographiques sur lesquelles on diffuse l'information, ne comporte moins d'observations qu'un seuil minimal. Dès problèmes peuvent apparaître dès lors que les zones sont très petites où lorsque la densité des observations est faible.

De plus, on distingue généralement le secret statistique primaire du secret statistique secondaire. Le secret statistique primaire concerne les informations directement diffusées tandis que le secret statistique secondaire concerne les informations qu'un utilisateur pourrait déduire *indirectement*, en recombinant et recoupant ces informations entre elles. Les informations déduites des informations diffusées doivent également respecter le secret statistique et ne pas conduire à révéler des données confidentielles ou personnelles. Dans le cas de données diffusées selon un ou plusieurs zonages géographiques, des manipulations sont possibles pour combiner et recouper ces zones entre elles afin de déduire des informations sur de nouvelles zones, souvent plus petites que les zones d'origines. La différenciation est une technique qui permet de déduire ainsi de nouvelles informations.

On s'intéresse à la différenciation géographique car on peut l'appliquer au cas de données carroyées : bien souvent, les données carroyées sont également diffusées selon un zonage administratif en plus. Ce sera le cas en France avec les données de la source fiscale qui font l'objet d'une diffusion au niveau des communes d'une part et qui seront prochainement diffusées au niveau d'une grille de carreaux d'autre part. On s'intéresse ici aux problèmes de confidentialité, du point de vue du secret secondaire, que peut soulever cette double diffusion. Le secret primaire est déjà pris en compte séparement dans la diffusion au niveau des communes et au niveau des carreaux.

La technique de la différenciation

La différenciation géographique est une technique qui permet de déduire de l'information sur des zones géographiques par combinaison et recoupement d'autres zones géographiques plus grandes. Cette technique peut être utilisée lorsque les données diffusées sont de nature **additive**, c'est-à-dire qu'on peut additioner ou soustraire ces données entre elles. C'est le cas par exemple lorsqu'on diffuse les revenus totaux d'une population, mais ce n'est pas le cas si on diffuse la médiane des revenus. Le principe de la technique de différenciation géographique est très simple : il s'agit de faire la différence, sur des données additives, entre l'information diffusée au niveau d'une zone englobante et celle diffusée au niveau d'une zone englobée. On en déduit alors la valeur de la variable additive sur la périphérie de la zone englobante, hors zone englobée (voir figure 1). Les zones englobante et englobée peuvent être quant à elles relativement complexes en étant composées de multiples zones plus petites.

Mise en péril du secret statistique secondaire

La différenciation, si elle est utilisée par une personne mal intentionnée, peut conduire à une rupture du secret statistique. En effet, lorsqu'un institut comme l'Insee diffuse des mêmes données selon différents zonages, en s'assurant que pour chacun de ces zonages le secret statistique soit

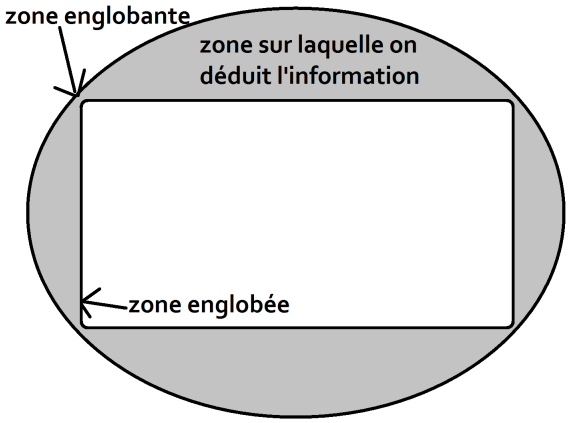


FIGURE 1 – Schéma illustrant le principe de la différenciation géographique sur données additives.

Note : Connaissant la valeur de la variable additive sur l'ovale et sur le rectangle, on en déduit la valeur de cette même variable sur la zone frontière représentée en grisée.

respecté, rien n'assure qu'on ne puisse combiner les différentes informations diffusées pour en déduire des informations confidentielles à l'aide de la technique de la différenciation. En effet, la différenciation permet de déduire des informations sur des zones plus petites que celles qui avaient été prévues au départ.

Protéger les données c'est alors s'assurer qu'aucune différenciation entres les fichiers mis à disposition du public ne puisse aboutir à connaître des informations concernant un nombre d'observations strictement inférieur au seuil de confidentialité en vigueur. Or, les possibilités de rupture du secret par différenciation grandissent, dans le contexte de diffusion de plus en plus fréquente de données carroyées. La diffusion de données sous forme d'une grille régulière de carreaux de petite taille est en effet aujourd'hui rendue possible par le géoréférencement d'un grand nombre de sources. Du fait de la multiplicité des zonages d'une part, et de la taille de plus en plus petite de la maille, d'autre part, le repérage même des problèmes de différenciation revêt donc un caractère complexe sur le plan computationnel. Tout l'enjeu des travaux présentés ici est d'essayer de repérer l'ensemble des problèmes de différenciation qui pourraient apparaître lorsqu'on diffuse des données additives selon plusieurs zonages géographiques différents. Pour protéger des données à risque il faut accepter de perdre en utilité des données diffusées : on va en effet masquer, agréger ou brouiller certaines données. Afin de déformer le moins possible les données diffusées, il est important de connaître précisément quelles sont les observations à risque de différenciation, pour concentrer les méthodes de protections sur ces seules observations.

Cas général

La technique de la différenciation peut être appliquée dans un cadre plus général du moment qu'on dispose de données additives diffusées selon plusieurs nomenclatures. Un zonage géographique, comme le partitionnement du territoire en communes, est une nomenclature particulière, mais on peut penser à des nomenclatures non géographiques comme par exemple la catégorie sociale pour des individus ou le secteur d'activité pour des entreprises. On peut donc généraliser la technique de la différenciation à des données non géographiques.

De plus, on peut aussi généraliser cette technique dans le cas où il y a trois nomenclatures ou plus. Dans ce cas, on peut effectuer plusieurs différenciations à la suite entre les différentes nomenclatures, jusqu'à déduire des informations sur des zones de plus en plus petites.

Objectifs de l'article

L'article a pour ambition de proposer une représentation des données qui permette de mieux appréhender le problème de la différenciation géographique. Il fournira de plus des algorithmes de simplification et de recherche exhaustive pour permettre de détecter un maximum de problèmes de différenciation. On ne s'intéresse qu'à la détection des observations à risque de différenciation, c'est-à-dire des observations pour lesquelles on peut déduire par différenciation des informations sur des agrégats plus petits que le seuil de confidentialité de la source. L'article n'a pas pour objet de mettre en place un traitement de protection particulier sur les observations ainsi détectées. Pour cet aspect, le lecteur pourra se référer à la littérature des méthodes de confidentialité (voir par exemple le chapitre "Confidentialité des données spatiales" du Manuel d'analyse spatiale de l'Insee [2]).

1 Le cas de deux zonages

Par la suite, on se restreint au cas où une variable additive est diffusée sur deux, et uniquement deux, zonages qu'on nommera $z_A = \{z_A^1, ..., z_A^{N_A}\}$ et $z_B = \{z_B^1, ..., z_B^{N_B}\}$. Le but sera alors d'identifier les observations "à risque de différenciation" de la base de données, c'est-à-dire les observations pour lesquelles il est possible, par différenciation, de retrouver la valeur de la variable additive sur des agrégats petits, i.e. concernant moins d'observations que le seuil de confidentialité. Pour mieux visualiser les explications qui suivent, on pourra par la suite penser au zonage A comme étant le découpage de la France en communes (et donc $N_A \sim 36\ 000$) et le zonage B comme étant un découpage en carreaux.

1.1 Plusieurs façons de représenter les données

Pour commencer, on peut remarquer qu'il existe plusieurs façons de représenter les données géographiques.

- **La carte** : cela consiste à tracer les contours des zones sur lesquelles sont diffusées les données et à indiquer où sont situées les observations au sein de ces zones. La figure 2 en est un exemple : 13 observations sont réparties sur quatre communes (z_A^1, z_A^2, z_A^3 et z_A^4) et sur six carreaux (z_B^1, z_B^2, z_B^3, z_B^4, z_B^5 et z_B^6).
- **La table individuelle** : elle comprend une ligne par observation et deux colonnes, l'une indiquant la zone du premier découpage géographique à laquelle appartient l'observation et l'autre la zone du deuxième découpage. La carte de la figure 2 peut ainsi être représentée en table individuelle comme dans la table 1.
- **La table croisée** : c'est la table des fréquences de la table individuelle lorsqu'on croise les deux zonages. Un exemple est donnée en table 2. Il y a autant de lignes que d'intersections non vides entre les deux zonages et une troisième colonne indique le nombre d'observations situées sur ces intersections.
- **La matrice de croisement** : c'est une matrice comportant autant de lignes que de zones possibles pour le zonage A et autant de colonnes que de zones possibles pour le zonage B, le coefficient de la matrice située sur la i^{eme} ligne et la j^{eme} colonne donne le nombre d'observations à l'intersection entre les zones z_A^i et z_B^j.
- **Le graphe** : De façon générale, un graphe est la donnée d'un ensemble de sommets (ou *noeuds*) et d'un ensemble d'*arêtes*. Une arête est un lien entre deux sommets. Elle est orientée si l'ordre du sommet de départ et du sommet d'arrivé a une importance. La représentation sous forme de graphe est un outil efficace pour représenter les liens entre les entités modélisées et ainsi mettre en évidence des *clusters* ou des structures particulières. On définit la *taille* d'un graphe comme étant son nombre de sommets. Un *chemin* est une liste de sommets où chaque sommet est relié à son successeur par une

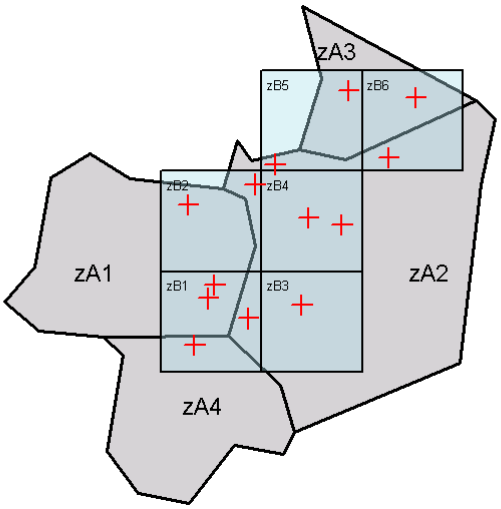


FIGURE 2 – Exemple de diffusion de données géolocalisées selon deux nomenclatures géographiques (ou zonages) différents.
Note : Les croix représentent la localisation des observations.

arête. Un *sous-graphe* d'un graphe donné est composé d'un sous-ensemble de sommets du graphe ainsi que des arêtes reliant entre eux ces sommets. Un sous-graphe est dit *connexe* s'il existe un chemin connectant chaque paire de sommets entre eux. Si on assimile la liaison entre deux sommets du graphe par une arête à une relation d'équivalence, alors l'ensemble des classes d'équivalence obtenues sont les *composantes connexes* du graphe. Pour plus d'informations sur la théorie générale des graphes le lecteur pourra se reporter au chapitre du Manuel d'analyse spatiale correspondant [4].
On adopte une représentation sous forme de graphe des données géographiques présentées précedemment : chaque zone du premier découpage (zonage A) est représentée par un sommet (ou *noeud*). De plus, deux sommets sont reliés par une arête s'il existe au moins une zone du deuxième découpage (zonage B) qui a des observations dans chacune des zones du zonage A correspondant aux deux sommets. Afin de tenir compte de ce nombre d'observations, on crée un graphe orienté avec des arêtes pondérées. Si z_A^i et z_A^j sont deux zones du zonage A et que $\{z_B^{k_1}, ..., z_B^{k_p}\}$ sont les zones du zonage B ayant des observations dans z_A^i et dans z_A^j, alors l'arête de z_A^i vers z_A^j porte la valeur $Card\{z_A^i \cap \left[\cup_1^p z_B^{k_l}\right]\}$ (et de même pour l'arête de z_A^j vers z_A^i). Ainsi la carte de la figure 2 peut être représentée sous la forme du graphe de la figure 3. On voit que les observations des carreaux z_B^3 et z_B^4 ne sont pas représentées sur le graphe, car elles appartiennent à des carreaux qui sont entièrement compris dans la commune z_A^2. Comme on le verra par la suite, seuls comptent les carreaux "aux frontières" des communes. Au contraire, certaines observations apparaissent plusieurs fois dans le graphe : c'est le cas des observations situées dans le carreau z_B^1 qui recouvre 3 communes. Comme on considère les liens entre communes deux à deux, ces observations sont toujours comptées deux fois. Ainsi, la somme des pondérations des arêtes du graphe est égale au nombre d'observations des carreaux z_B^2, z_B^5 et z_B^6, plus deux fois le nombre d'observations du carreau z_B^1.

La représentation qui comporte le plus d'information est celle de la carte, car elle donne des précisions sur la géométrie des contours et sur l'emplacement précis des observations au sein des zones. Les représentations "tableaux" que sont la table individuelle, la table croisée et la matrice de croisement sont équivalentes en termes d'information : on peut déduire l'une d'une autre. Enfin, la représentation sous forme de graphe est la moins riche en information : c'est justement

individu	**zonage A**	**zonage B**
1	z_A^1	z_B^1
2	z_A^1	z_B^1
3	z_A^1	z_B^2
4	z_A^2	z_B^1
5	z_A^2	z_B^2
6	z_A^2	z_B^3
7	z_A^2	z_B^4
8	z_A^2	z_B^4
9	z_A^2	z_B^5
10	z_A^2	z_B^6
11	z_A^3	z_B^5
12	z_A^3	z_B^6
13	z_A^4	z_B^1

TABLE 1 – Représentation de l'exemple de la figure 2 sous forme de table individuelle.

zonage A	**zonage B**	**nombre d'observations**
z_A^1	z_B^1	2
z_A^1	z_B^2	1
z_A^2	z_B^1	1
z_A^2	z_B^2	1
z_A^2	z_B^3	1
z_A^2	z_B^4	2
z_A^2	z_B^5	1
z_A^2	z_B^6	1
z_A^3	z_B^5	1
z_A^3	z_B^6	1
z_A^4	z_B^1	1

TABLE 2 – Représentation de l'exemple de la figure 2 sous forme de table croisée.

	z_B^1	z_B^2	z_B^3	z_B^4	z_B^5	z_B^6
z_A^1	2	1	0	0	0	0
z_A^2	1	1	1	2	1	1
z_A^3	0	0	0	0	1	1
z_A^4	1	0	0	0	0	0

TABLE 3 – Représentation de l'exemple de la figure 2 sous forme de matrice de croisement.

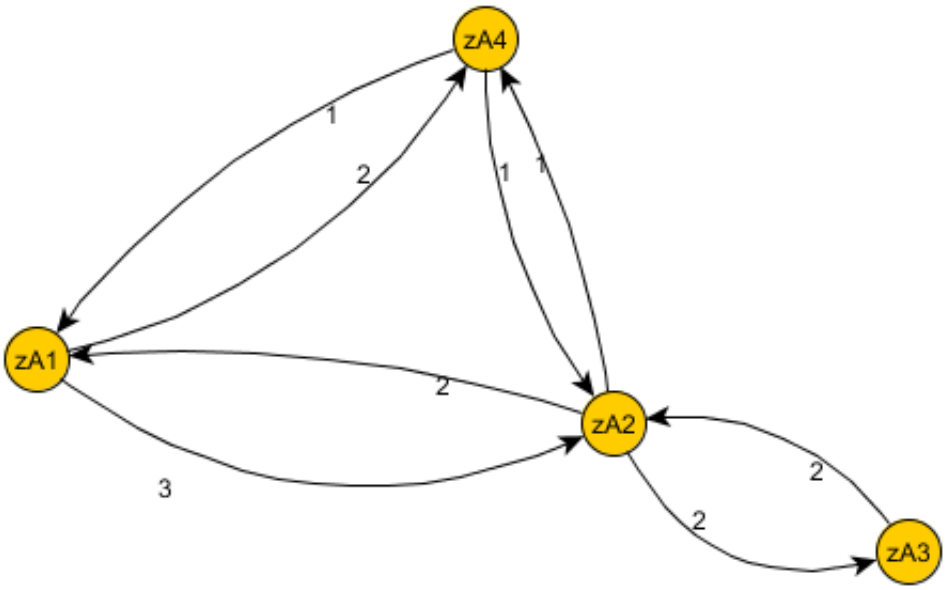


FIGURE 3 – Représentation de l'exemple de la figure 2 sous forme de graphe.

là son but, elle comporte seulement les informations essentielles à la résolution du problème. On y reviendra par la suite.

1.2 Un nombre de combinaisons très grand

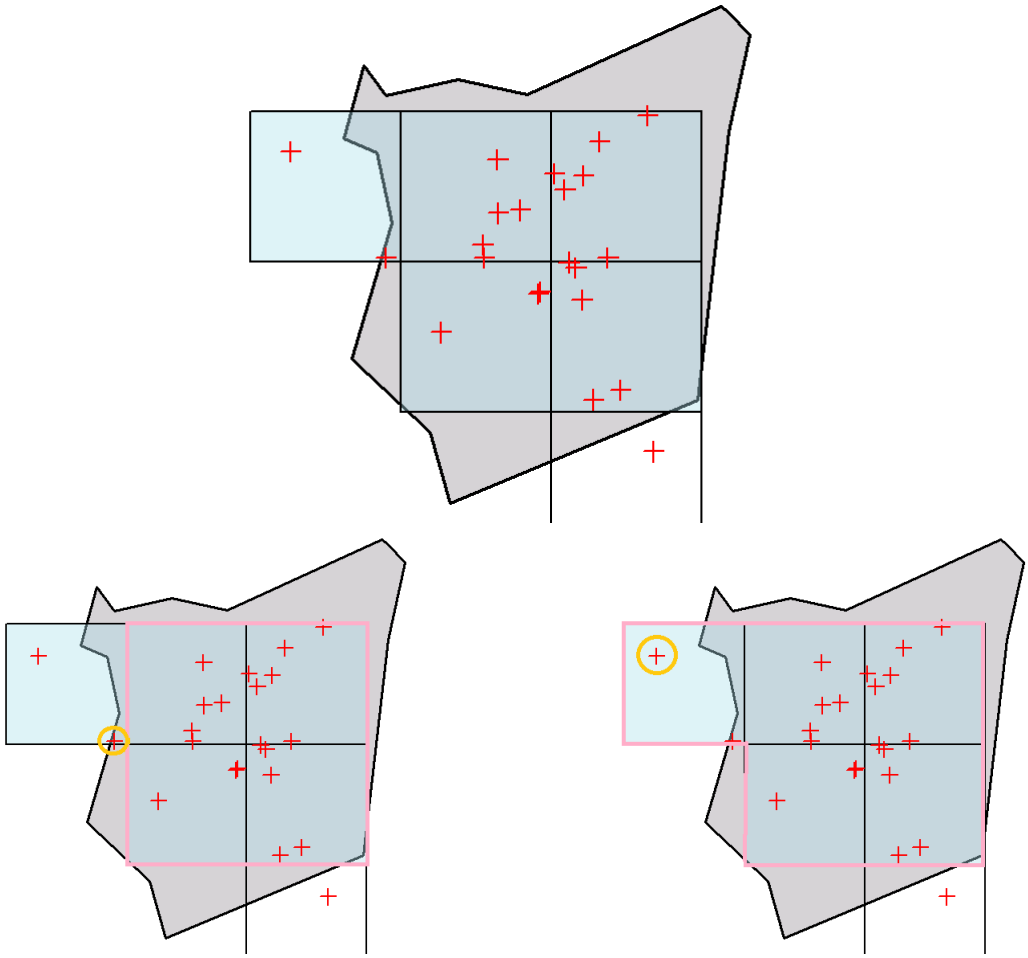
La figure 4 montre une configuration de points qui conduit à un problème de différenciation géographique entre une commune et des carreaux. Il y a deux manières de réaliser l'opération de différenciation.

La première est de soustraire au total de la variable additive considérée sur la commune, le total de la variable sur les carreaux "internes" à la commune (c'est-à-dire les carreaux dont toutes les observations sont situées dans la commune). On en déduit alors la valeur de la variable sur les observations situées sur des carreaux "à la frontière" et à *l'intérieur* de la commune. La figure 4a indique l'observation à risque si on effectue une telle différenciation. On parle alors de différenciation "interne".

La seconde manière est de soustraire au total de la variable additive sur les carreaux intérieurs et à la frontière, le total de la variable sur la commune. On en déduit alors le total de la variable sur l'ensemble des observations situées sur les carreaux "frontières" et à *l'extérieur* de la commune (voir figure 4b). On parle alors de différenciation "externe".

Il est possible de généraliser ces deux types de différenciation, *interne* et *externe*, en ne se restreignant plus à une seule commune, mais en combinant plusieurs communes. La figure 5 en est un exemple. De plus, cette figure illustre bien qu'il peut ne pas y avoir de problème de différenciation sur les deux communes considérées séparément, alors qu'un problème de différenciation apparaît lorsqu'on considère les deux communes ensemble, comme formant une même zone.

On peut ainsi former des zones complexes et potentiellement très grandes composées d'un groupe de communes conduisant à un problème de différenciation. Dans ces groupes, les communes doivent être contigües, c'est-à-dire que des carreaux frontières les relient les unes aux autres. Si ce n'est pas le cas, et que le groupe peut se décomposer en deux sous-groupes non contigüs, alors si un problème de différenciation apparaît sur le groupe, forcément le même problème doit apparaître sur au moins un des sous-groupes. Par conséquent, il est inutile de rechercher des problèmes de différenciation sur des groupes non contigüs car ces problèmes pouraient



(a) Différenciation "interne" : la zone englobante est la commune et la zone englobée est constituée des quatre carreaux intérieurs (en rose). L'observation entourée en jaune est l'observation à risque de différenciation.

(b) Différenciation "externe" : la zone englobante est composée des cinq carreaux en rose et la zone englobée est la commune. L'observation entourée en jaune est l'observation à risque de différenciation.

FIGURE 4 – Exemple de différenciation sur une commune conduisant à une rupture de la confidentialité.

déjà être identifiés au niveau des groupes plus petits qui le composent.

C'est justement pour cette raison (les groupes doivent être composés de zones contigües) que la représentation sous forme de graphe énoncée plus haut est utile. Cette représentation permet de visualiser les agrégats possibles et de séparer les composantes connexes.

De plus, comme on l'a vu sur les exemples, les observations à risque de différenciation se situent toujours sur des carreaux "aux frontières" et non pas sur des carreaux "internes". Les carreaux "internes" n'apportent donc aucune information utile à la résolution du problème, c'est pourquoi on peut les ignorer. La représentation sous forme de graphe concentre donc l'information utile à la résolution du problème.

La première stratégie qui vient à l'esprit est de tester, pour tous les agrégats possibles, s'il y a un problème de différenciation ou non, en effectuant la différence du nombre d'observations, "interne" et "externe". Il s'agit alors de repérer les agrégats dont la différence du nombre d'observations est plus petite que le seuil de confidentialité.

L'inconvénient de cette méthode est qu'elle n'est pas possible à mettre en oeuvre pratiquement

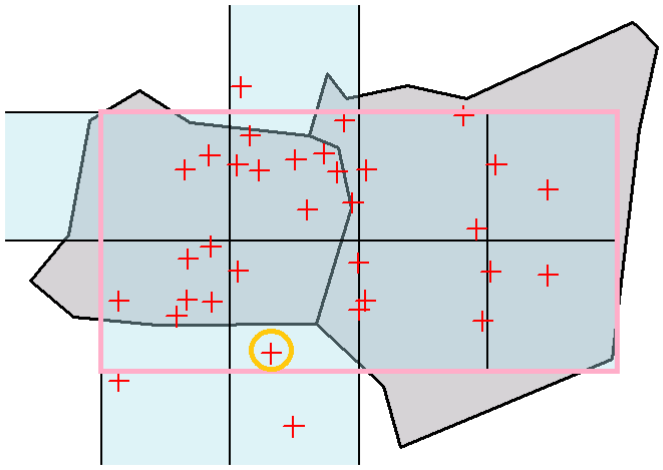


FIGURE 5 – Exemple de différenciation sur plusieurs communes conduisant à une rupture de la confidentialité.
Note : sur chacune des communes séparement il n'y a pas de problème de différenciation, mais lorsqu'on considère les deux communes ensemble alors apparaît un problème de différenciation.La zone englobante est constituée des 8 carreaux entourés en rose et la zone englobée des deux communes. L'observation entourée en jaune est l'observation à risque lorsqu'on effectue la différenciation.

dès lors que le nombre de zones N_A du zonage A devient grand, car alors le nombre d'agrégats à tester dépasse largement les capacités de calcul de n'importe quel ordinateur actuel.

1.3 Stratégies pour simplifier le problème

Dès lors que les zonages présentent un grand nombre de zones, il apparaît impossible de tester systématiquement toutes les combinaisons possibles pour savoir celles qui conduisent à une rupture du secret secondaire par différenciation. Dans la suite, on présente une méthode qui permet de simplifier les données en entrée, afin de réduire le nombre de cas à tester.

Ne considérer que les carreaux aux frontières. Les carreaux à l'intérieur d'une commune, c'est-à-dire dont toutes les observations sont situées dans la commune (même si une partie non habitée de la surface du carreau recoupe d'autres communes), peuvent être retirés des données. En effet, les observations qui se situent dans ces carreaux ne seront jamais à risque de différenciation, car par différenciation on ne peut retrouver que des croisements entre carreaux et communes.

On peut alors constituer le graphe qui représente les liens entre communes : deux communes sont liées s'il existe au moins un carreau dont une partie des observations sont réparties sur ces deux communes. La figure 6 montre cette étape sur un exemple très simple qui représente 7 communes et 16 carreaux habités.

Il est possible de simplifier le graphe ainsi constitué selon deux méthodes. Simplifier le graphe revient à fusionner des noeuds et à supprimer les arêtes qui relient ces noeuds fusionnés. Avant d'aller plus loin, il faut évoquer le critère sur lequel on peut savoir s'il y a un problème de différenciation ou non à partir de la représentation sous forme de graphe.

Choisir un groupe de communes contigües revient à choisir un sous-graphe connexe. Les arêtes reliant ce sous-graphe connexe au reste du graphe donnent le nombre d'observations situées à la frontière du groupe et donc le nombre d'observations dont on peut déduire des informations par différenciation entre le groupe de communes et les carreaux "internes" à ce groupe. Il y a

deux types d'arêtes du fait que le graphe est orienté : les arêtes "sortantes", du sous-graphe connexe vers un autre sommet du graphe et les arêtes "entrantes", d'un sommet du graphe vers le sous-graphe connexe. La somme des valeurs associées aux arrêtes "sortantes" donnent le nombre d'observations sur la frontière "interne" tandis que celles associées aux arêtes "entrantes" concernent la frontière "externe".

Une subtilité apparaît lorsque que certains carreaux sont à cheval sur trois communes ou plus. Dans ce cas, les observations aux frontières seront alors comptés plusieurs fois dans le graphe. Par la suite, on adopte la règle suivante. Pour un sous-graphe connexe donné, on note $\{\epsilon_{1,s}, \epsilon_{1,e}, ..., \epsilon_{m,s}, \epsilon_{m,e}\}$ la valeur des arêtes (e pour arête entrante pour s pour arête sortante) en considérant les carreaux recouvrant 3 communes ou plus et on note $\{\eta_{1,s}, \eta_{1,e}, ..., \eta_{n,s}, \eta_{n,e}\}$ la valeur des arêtes **sans considérer** les carreaux recouvrant 3 communes ou plus. Ainsi, $n \leq m$ et $\sum_k \eta_k \leq \sum_k \epsilon_k$.

- si $\sum_k \epsilon_{k,e} < seuil$ **ou** $\sum_k \epsilon_{k,s} < seuil$ alors il est certain qu'il y a un problème de différenciation ;
- si $\sum_k \eta_{k,e} \geq seuil$ **et** $\sum_k \eta_{k,s} \geq seuil$ alors il est certain qu'il n'y a **pas** de problème de différenciation.

Fusionner les communes - méthode 1 Soient deux noeuds A et B du graphe et soit ϵ_A et ϵ_B les valeurs des arêtes de A vers B et de B vers A. On décide de fusionner A et B s'il ne peut pas y avoir, de façon certaine, de problème de différenciation sur n'importe quelle zone contenant A mais ne contenant pas B et sur n'importe quelle zone contenant B mais ne contenant pas A. Si on considère donc une zone constituée d'une ou plusieurs communes, et qui contient la commune A mais qui ne contient pas la commune B, alors par différenciation entre cette zone et les carreaux internes à cette zone, on en déduit la zone frontière constituée des intersections des carreaux "aux frontières" avec cette même zone. Par construction, cette zone frontière contient au moins ϵ_A observations. Donc si ϵ_A est supérieur ou égal au seuil de confidentialité, on en déduit qu'il n'y a pas de problème de différenciation sur la zone en question. Si ϵ_A est supérieur ou égal au seuil de confidentialité, ceci reste vrai pour n'importe quelle zone contenant A mais ne contenant pas B. On obtient le résultat symétrique en testant si ϵ_B est supérieur ou égal au seuil de confidentialité.

Ainsi, si ϵ_A et ϵ_B sont tous les deux supérieurs ou égaux au seuil de confidentialité, on peut fusionner les noeuds A et B.

La fusion de deux noeuds conduit parfois à modifier la valeur de certaines arêtes : on additionne en effet les valeurs des arêtes qui relient les mêmes noeuds. Ainsi, au fur et à mesure de l'agrégation du graphe, les valeurs des arêtes tendent à augmenter, rendant de plus en plus probables de nouvelles agrégations. C'est un effet "boule de neige".

Fusionner les communes - méthode 2 La deuxième façon de tester si on peut fusionner deux noeuds est plus subtile. On reprend les mêmes notations que précedemment et on suppose cette fois-ci, sans perte de généralité, que ϵ_A est inférieur au seuil de confidentialité. On suppose de plus qu'il existe un *autre* chemin qui relie A à B par une succession d'arêtes $\mathcal{E} = \{\eta_1, ..., \eta_n\}$ reliant les sommets $A, S_1, ..., S_{n-1}, B$. Ce chemin doit faire partie du graphe **sans** les carreaux recouvrant 3 communes ou plus.

Considérons maintenant un sous-graphe connexe $\mathcal{G}_c$ contenant A mais ne contenant pas B : alors ce sous-graphe connexe est relié au reste du graphe par une des arêtes de $\mathcal{E}$. En effet, soit il existe $i \in \{1, ..., n-1\}$ tel que S_i n'appartienne pas au sous-graphe connexe $\mathcal{G}_c$, et dans ce cas η_i relie $\mathcal{G}_c$ au reste du graphe. Soit tous les S_i appartiennent à $\mathcal{G}_c$ et dans ce cas, η_n relie le sous-graphe connexe au reste du graphe. Soit donc k tel que η_k relie $\mathcal{G}_c$ au reste du graphe.

Si la somme des arêtes sortantes qui relient $\mathcal{G}_c$ au reste du graphe est supérieure ou égale au seuil de confidentialité, alors il n'y a pas de problème de différenciation sur $\mathcal{G}_c$. Or, cette somme est supérieure ou égale à $\epsilon_A + \eta_k$. Si donc $\eta_k \geq seuil - \epsilon_A$, alors il n'y a pas de problème de

différenciation sur $\mathcal{G}_c$.

On en déduit alors, que s'il existe un chemin reliant A à B, sans emprunter ϵ_A, tel que toutes les arêtes ont une valeur plus grande ou égale à $seuil - \epsilon_A$, et inversement de B vers A,alors on peut fusionner A et B.

Illustration des simplifications possibles sur un exemple simple La figure 7 montre comment est simplifié, étape par étape, le graphe de départ de l'exemple de la figure 6. Les étapes (a) et (c) résultent de la fusion selon la méthode 1 et l'étape (b) résulte de la fusion selon la méthode 2.

1.4 Une recherche exhaustive

Le graphe simplifié comporte beaucoup moins de sous-graphes connexes que le graphe de départ. Simplifier le graphe revient en réalité à réduire le nombre de sous-graphes connexes à tester. Si cette réduction est importante, il est envisageable de tester de manière exhaustive tous les sous-graphes connexes restant.

En réalité, il n'y a pas besoin de tester tous les sous-graphes connexes un par un, mais il suffit de tester les sous-graphes connexes dont la taille (i.e. le nombre de sommets) est inférieure à la moitié de la taille du graphe total. En effet, lorsqu'on teste s'il y a un problème de différenciation sur un sous-graphe connexe, on teste en même temps s'il y a un problème de différenciation sur le complémentaire de ce sous-graphe. Ceci est dû au fait que la différenciation "interne" sur un sous-graphe donné correspond à la différenciation "externe" sur le sous-graphe complémentaire, et inversement.

Si le graphe simplifié est composé de N sommets, on en teste alors chaque sous-graphe connexe dont la taille (i.e. le nombre de sommets) est inférieure à $N/2$. En pratique, si le graphe simplifié est toujours très grand, malgré la simplification, il sera impossible de tester tous les sous-graphes connexes, car il seront trop nombreux. On peut alors se contenter de tester les sous-graphes connexes de taille limitée, par exemple seulement les sous-graphes connexes de taille comprise entre 1 et 10.

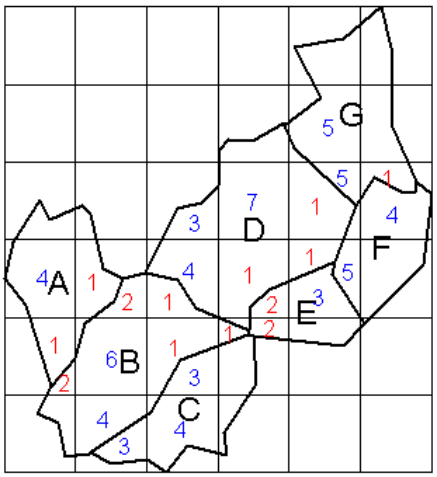
Si le graphe simplifié est composé de plusieurs composantes connexes, alors on applique la stratégie ci-dessus à chacune des composantes connexes.

Arlindo Dos Santos et François Sémécurbe de la Division des Statistiques et de l'Analyse Urbaine (DSAU) à l'Insee ont développé un programme en C++ permettant de réaliser cette recherche exhaustive de manière efficace et rapide. Leur programme prend en entrée un tableau de type "table croisée" (voir section 1.1) ainsi qu'une taille d'agrégat N visée, en ressort une liste d'agrégats de zones du zonage A de taille N conduisant à un problème de différenciation "interne" ou "externe". Pour cela, il effectue successivement sur tous les agrégats possibles de taille N une différenciation et garde en mémoire les agrégats dont la différenciation avec les zones du zonage B conduit à avoir des informations sur un nombre d'observations plus petit que le seuil considéré. La liste de tous les agrégats possibles de taille N est construite à l'aide d'une fonction récursive. La difficulté est que le nombre d'agrégats de taille N augmente de façon quasi-exponentielle avec N, il devient alors impossible, avec les moyens informatiques actuels à disposition, de tester tous les agrégats possibles dès lors que N devient trop grand.

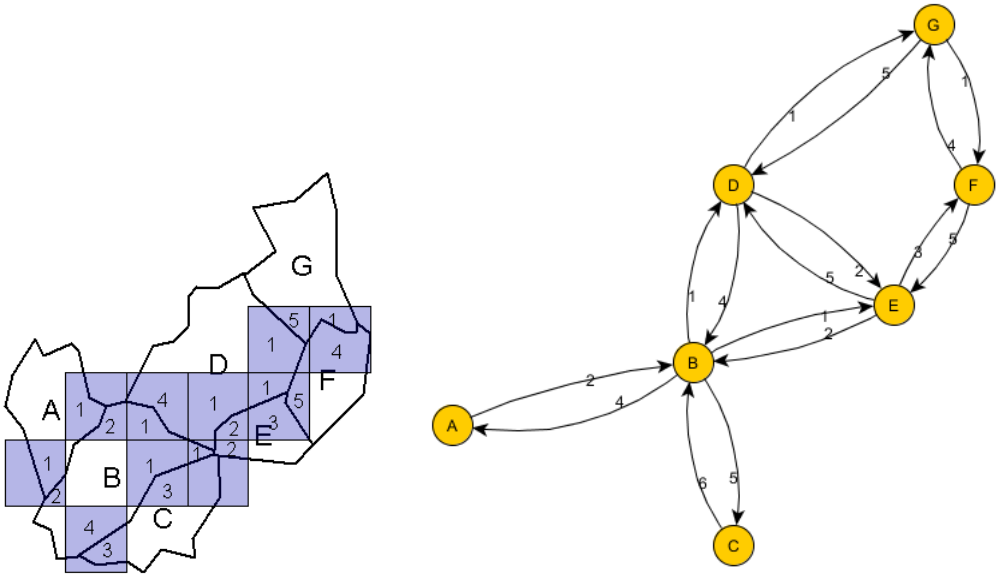
2 Application aux données carroyées

2.1 Les données carroyées de Filosofi

On cherche ici à détecter les problèmes de différenciation lors de la diffusion de la source Filosofi (milésime 2014) sur les communes d'une part et sur des carreaux d'autre part. Pour le carroyage, on retient ici la méthode des grilles superposées au niveau naturel explicitée dans

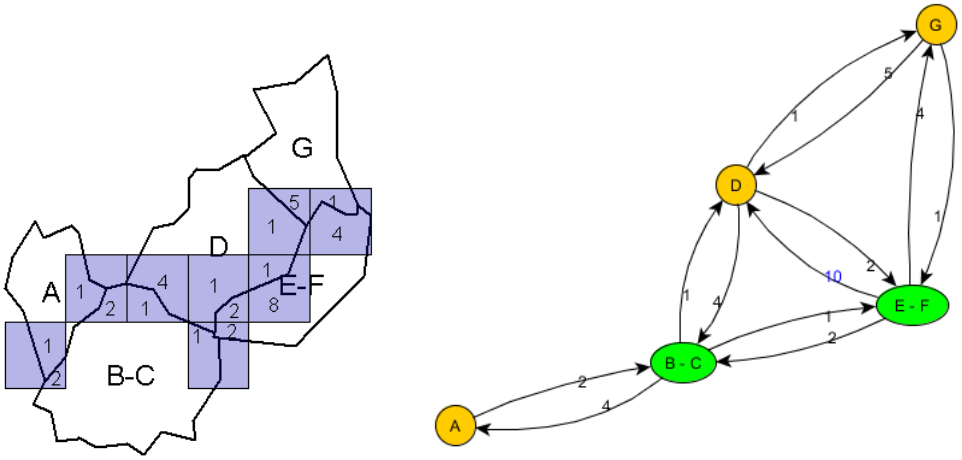


(a) Les chiffres indiquent le nombre de personnes présentes sur les intersections entre carreaux et communes. Les chiffres en rouge sont ceux qui sont en dessous du seuil de confidentialité (1 ou 2 personnes dans notre exemple).

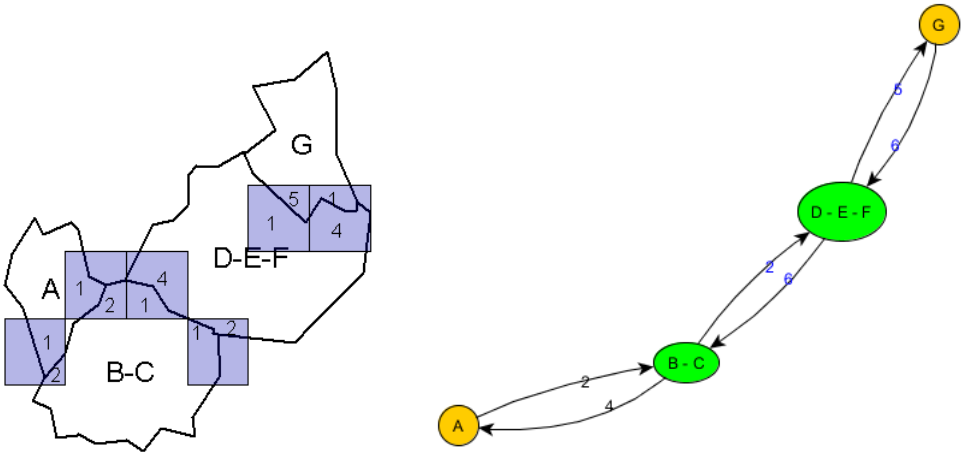


(b) À gauche on a retiré les carreaux "internes" et à droite on a représenté le graphe correspondant.

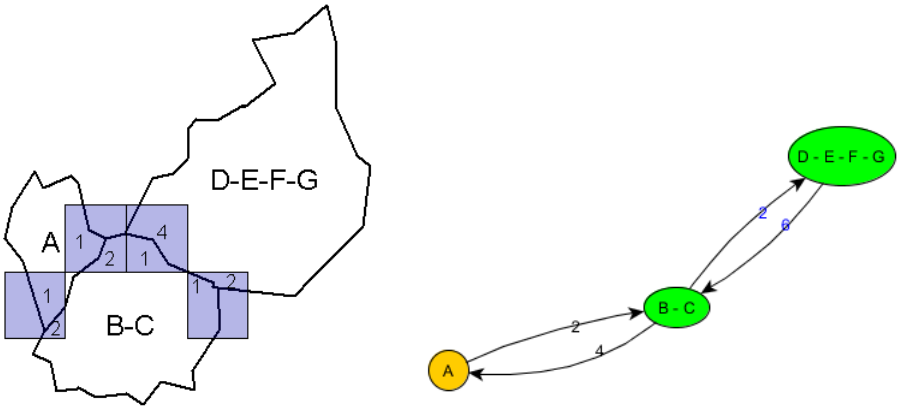
FIGURE 6 – Exemple de simplification des données sur 7 communes. On considère un seuil de confidentialité de 3.



(a) Fusion des communes B et C d'une part et E et F d'autre part.



(b) Fusion des commmunes D et E-F.



(c) Fusion des communes D-E-F et G

FIGURE 7 – Illustration de la simplification du graphe, issu de l'exemple de la figure 6, en trois étapes.

Note : noeuds du graphe en vert représentent les communes fusionnées.

l'article de Branchu et al., 2018, [1], qui fournit un pavage du territoire par des carrés de tailles différentes, allant de 200 mètres pour les plus petits, à 32 kilomères pour les plus grands. Cette méthode s'apparente à la méthode du *quadtree* consistant en un découpage itératif du territoire, en carrés de plus en plus petits, jusqu'à ne plus pouvoir découper sans rompre les règles du secret statistique. En l'occurence, le secret primaire consiste à respecter le seuil de confidentialité de 11 ménages fiscaux qui régit toute source fiscale. Au finale chaque carré contient 11 ménages fiscaux et par conséquent la taille des carrés est inversement proportionnelle à la densité de la population. On se restreint de plus au champ de la France métropolitaine.

La table individuelle comporte 27 625 783 ménages, répartis dans 36 671 communes et 144 706 carreaux. Parmi ces carreaux, 13 % sont de taille 200 m, 51 % de taille 1 km, 20 % de taille 2 km et 16 % de taille 4 km ou plus. Il y a 244 424 croisements carreaux-communes dont 28 % (68 418 intersections) sont sous le seuil de 11 ménages.

De plus, 78 714 carreaux sont strictement inclus dans une commune, soit 54 % des carreaux, 42 663 (30 %) sont à l'intersection de deux communes et 23 329 (16 %) sont à l'intersection de trois communes ou plus [1].

2.2 Premières simplifications et agrégations des communes

La première simplification consiste à retirer les carreaux entièrement compris dans une commune et à fusionner entre eux les carreaux qui recouvrent les mêmes communes.

Comme le montre le tableau 4, la simplification des données (i.e. premières simplifications sur les carreaux + agrégation méthode 1 + agrégation méthode 2) permet de réduire les nombres de zones des zonages A et B ainsi que le nombre d'intersections entre ces deux zonages présentant moins de 11 ménages. Par conséquent cela réduit aussi le nombre de ménages potentiellement à risque qui passe de 285 727 à 10 884 (soit - 96%) après ces trois étapes de simplification.

De plus, avec la première méthode d'agrégation, de nombreuses composantes connexes disparaissent : on passe de 1 147 composantes connexes à 327 composantes connexes. Cela est dû au fait que pour certaines petites composantes connexes, on peut les agréger jusqu'à ce qu'il n'y ait plus qu'une seule zone sur laquelle il n'y a pas de problème de différenciation. Par ailleurs, la taille de ces composantes connexes tend à diminuer avec l'agrégation du graphe. Initialement, chaque composante connexe compte en moyenne 32 sommets, tandis qu'après la première agrégation il n'y en a en moyenne plus que $14,3$ et seulement $8,4$ après la deuxième agrégation.

L'étape de simplification des données permet donc de réduire considérablement la complexité du problème : cela évite de tester des combinaisons dont on sait qu'elles ne mèneront pas à un problème de différenciation.

On peut visualiser cette simplification sur la carte de la figure 8 où chaque agrégat de communes résultant de l'agrégation du graphe selon les deux méthodes présentées auparavant, est représenté dans une même couleur. On voit immédiatement une différence marquée entre l'ouest et l'est de la France métropolitaine. À l'ouest, et particulièrement en Bretagne, la plupart des communes ont été agrégées à leurs voisines. Au contraire, les communes du quart nord-est de la France métropolitaine sont très peu agrégées : cela signifie qu'à leurs frontières se situent des ménages à risque de différenciation. Cette différence provient de la répartition locale des habitants. À l'est, les habitants sont plutôt concentrés dans des bourgs ou des centre-villes, avec une faible densité de population autour. C'est du fait de cette faible densité de population aux frontières des communes que dans la majorité des cas on n'a pas pu agréger les communes entre elles.

2.3 Les composantes connexes obtenues

Parmi les 327 composantes connexes obtenus au final, 317 composantes sont constituées de 10 sommets ou moins, seulement 9 composantes ont entre 11 et 30 sommets et il reste une

1. Dans la majorité des cas il s'agit de carreaux de grande taille

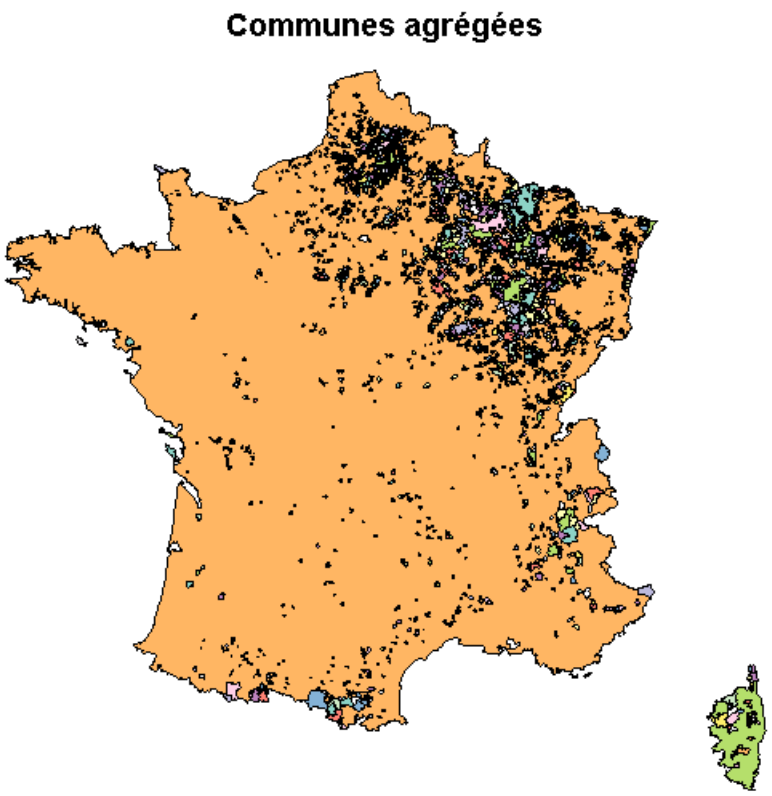


FIGURE 8 – Carte de la France métropolitaine des communes agrégées après avoir appliqué les deux méthodes d'agrégation.

	N_A	N_B	Nb. intersections sous le seuil	Nb. ménages sous le seuil	Nb. composantes connexes	Taille moyenne composantes connexes
État initial	36 671	144 706	68 418	285 727		
Premières simplifications	35 851	52 235	53 557	222 833	1 147	32, 0
Première agrégation	4 674	4 886	5 977	21 259	327	14, 3
Deuxième agrégation	2 740	2 822	2 937	10 884	327	8, 4

TABLE 4 – Nombre de communes, carreaux, intersections carreaux-communes sous le seuil de confidentialité et composantes connexes du graphe, selon l'état d'agrégation du graphe.

composante, la plus grande de toutes, qui comporte 1 895 sommets.

Cette dernière composante connexe a une forme particulière : elle est constituée d'un noeud central auquel sont reliés les autres sommets formant un réseau en étoile. Ce noeud central correspond à l'agrégation de 30 289 communes (il correspond donc à la grande zone en orange sur la carte de la figure 8). Cette composante connexe est représentée dans la figure 9.

L'idée avec une structure de graphe en étoile est de tester séparément chaque branche de l'étoile. Cela vient du fait que tester si un sous-graphe conduit à un problème de différenciation est équivalent à effectuer le test sur le sous-graphe complémentaire. Or, le complémentaire d'un sous-graphe contenant le noeud central, sera composé de différentes branches non connectées entre elles par des arêtes. Donc, tester séparément chaque branche suffit à repérer l'ensemble des problèmes de différenciations qui peuvent exister.

Ici, le noeud central a 2 516 branches. Ce qui fait qu'on dispose finalement de 2 842 composantes connexes (326 + 2 516) à tester, chacune de ces composantes connexes possédant 30 sommets ou moins.

2.4 Les ménages à risque pour la différenciation

Au final, après avoir testé tous les sous-graphes connexes des 2 842 composantes connexes, on a détecté 10 485 ménages à risque de différenciation, soit 0.04 % des ménages de France métropolitaine. C'est pratiquement l'ensemble des ménages présents dans des intersections carreaux-communes agrégées, après agrégation (voir tableau 4). Ceci prouve que l'agrégation du graphe permet de cibler précisément les ménages à risque.

Tous les ménages ainsi détectés ne font pas exactement face au même risque (voir tableau) : certains (13.7 %) sont situés sur des zones obtenues par différenciation comportant 10 ménages et d'autres (6.9 %) sur des zones ne comportant qu'un seul ménage. Ainsi, si on diffuse les données de Filosofi selon les communes et selon un carroyage en grilles superposées au niveau naturel avec des carreaux les plus fins de 200 m de côté, on diffuse aussi indirectement les variables additives de 719 ménages.

Comme on le voit sur la figure 10, la plupart des ménages à risque sont situés dans la partie nord-est de la France : c'est là en effet que beaucoup de communes n'avaient pu être agrégées. On retrouve néanmoins des ménages à risque dans l'ensemble des régions.

Au total, on a repéré 3 599 groupes de communes qui conduisent à une rupture du secret statistique secondaire par différenciation. Ces groupes sont de tailles différentes : 1 664 sont

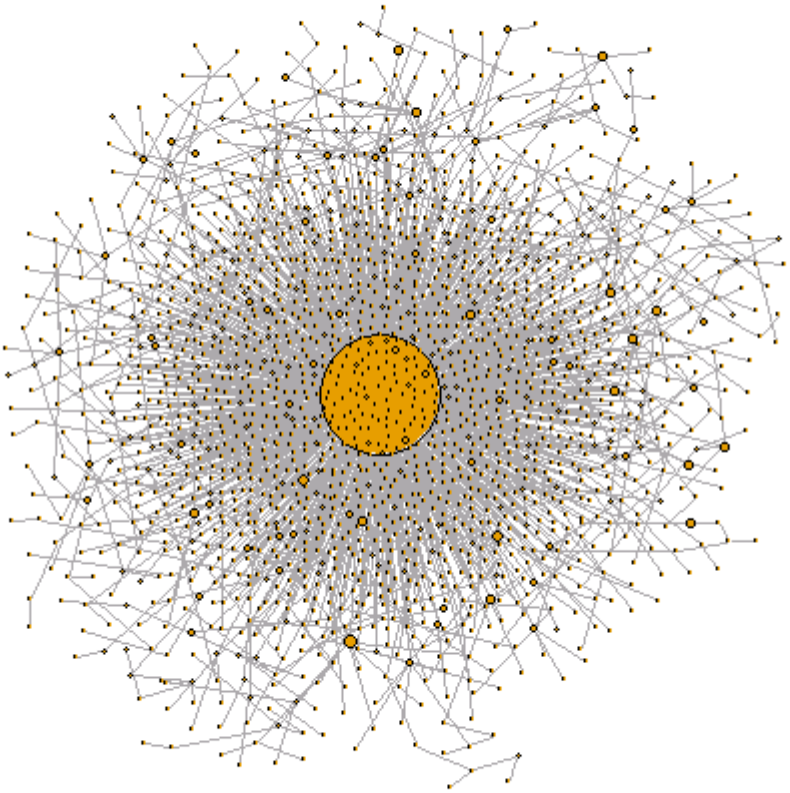


FIGURE 9 – Représentation sous forme de graphe de la composante connexe constituée d'un noeud central correspondant à l'agrégation de 30 289 communes. Ce graphe comporte 1 895 sommets.

Taille agrégat	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Nb. ménages	719	836	888	989	1074	1052	968	1232	1288	1439	10485
Prop. (en %)	6,9	8,0	8,5	9,4	10,2	10,0	9,2	11,8	12,3	13,7	100

TABLE 5 – Nombre de ménages à risque de différenciation selon la taille de l'agrégat qu'on peut déduire par différenciation.

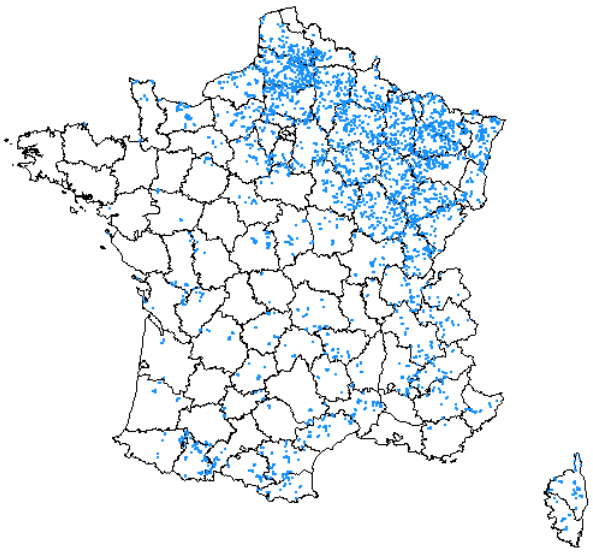


FIGURE 10 – Localisation des 10 485 ménages à risque de différenciation.

composées d'une seule commune, 1 147 entre deux et dix communes et 788 avec plus de onze communes. Le groupe le plus grand est composé de 93 communes.

Au départ, avant simplification, agrégation et recherche exhaustive, il y avait 285 727 ménages potentiellement à risque et au final on a détecté 10 485 ménages réellement à risque, soit une proportion 27 fois plus faible.

Quelques exemples de différenciation

2.5 Limites

Le processus proposé ici ne permet pas *a priori* de traiter les cas où des données sont diffusées selon trois zonages (ou nomenclatures de façon plus générale). Dans ce dernier cas, les possibilités de recouper l'information sont beaucoup plus importantes.

Références

[1] Branchub, M., Costemalle, V., and Fontaine, M. Données carroyées et confidentialité. *Actes des Journées de Méthodologies Statistiques, Insee* (2018).

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

[3] Duke-Williams, O., and Rees, P. Can census offices publish statistics for more than one small area geography ? an analysis of the differencing problem in statistical disclosure. 579–605.

[4] Eusebio, P., Floch, J.-M., and Levy, D. Manuel d'analyse spatiale. *Eurostat, EFGS, Insee* (2018).

[5] Organisation nationale des Nations Unies. Division de statistique. Manuel des systèmes d'information géographique et de cartographie numérique.

Annexe : Une autre piste de réflexion - le graphe dual

Les paragraphes qui suivent ne sont ni des méthodes ni des résultats. Ils se proposent d'exposer des réflexions qui ne sont pas encore vraiment abouties, mais tentent néanmoins d'apporter de nouvelles perspectives.

On se place dans le cas où il n'y a pas de carreaux recouvrant trois communes ou plus, mais seulement des carreaux entièrement compris dans une commune ou recouvrant deux communes. Dans ce cas particulier, la représentation sous forme de graphe orienté des données nous permet d'avoir toute l'information nécessaire pour détecter les zones sur lequelles des problèmes de différenciation apparaissent. Il s'agit des zones correspondant aux sous-graphes connexes dont la somme des arêtes "sortantes" est inférieure au seuil de confidentialité.

L'idée est ici de passer par le *graphe dual* : chaque face du graphe correspond à un sommet du graphe dual, et deux tels sommets sont reliés si les faces partagent une arête commune (voir figure 12). Ainsi, il y a autant d'arêtes dans le graphe que dans le graphe dual (on peut prendre en compte également le fait que le graphe de départ est orienté avec des arêtes pondérées, et reporter cette information sur le graphe dual).

Le problème qui consiste à trouver les sous-graphes connexes dont la somme des arêtes "sortantes" est inférieure au seuil revient à trouver, dans le graphe dual, des cycles dont la somme des arêtes est inférieure à ce même seuil. Le problème est-il alors plus simple à résoudre dans le graphe dual ?

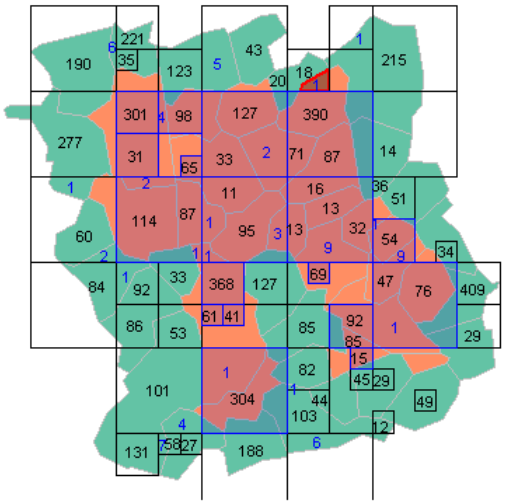
Il semble en tout cas *a priori* plus facile de simplifier le graphe dual : il n'y a plus besoin de "fusionner" des sommets entre eux, on peut dorénavant supprimer des arêtes ou des sommets du graphe dual. Le problème paraît plus élégant une fois posé dans le graphe dual.

La simplification du graphe dual correspond aux opérations suivantes :

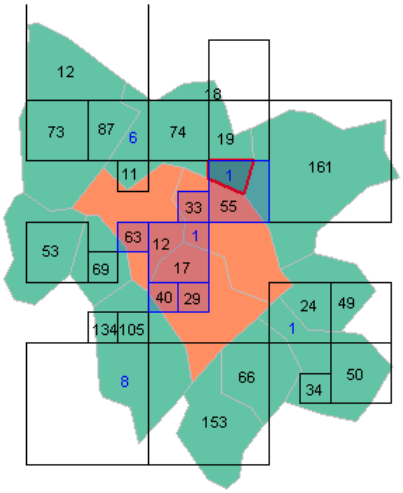
- **supprimer une arête :** si une arête a une valeur plus grande que le seuil de confidentialité, alors on peut la supprimer, car aucun des cycles recherchés n'utilisera cette arête. De même, si une arête a une valeur ϵ inférieur au seuil, mais que n'importe quelle arête qui suit cette arête a une valeur plus grande ou égale à $seuil - \epsilon$, alors on peut également supprimer l'arête ;
- **supprimer un sommet :** si quel que soit le couple de deux arêtes, une orientée vers le sommet et l'autre pointant vers l'extérieur, la somme de ces deux arêtes est supérieure ou égale au seuil, alors on peut supprimer le sommet du graphe dual (et les arêtes correspondantes), car aucun cycle recherché ne passera par ce sommet.

L'idée de passer au graphe dual semble séduisante, mais elle ne semble fonctionner que dans des cas particuliers et simples. Les obstacles sont les suivants :

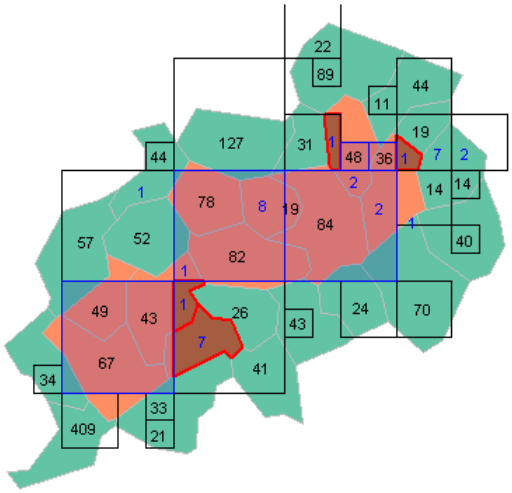
— il ne semble pas y avoir d'algorithme pour construire concrètement un graphe dual à partir d'un graphe donné, ceci étant peut-être dû au fait que la représentation duale n'est pas toujours unique ;
— dans le cas où le graphe initial n'est pas planaire, il semble plus difficile de déterminer ce qu'est un graphe dual car la notion de "face" est alors moins claire ;
— la recherche des sous-graphes connexes, et donc des cycles dans le graphe dual, n'apporte toutes les solutions au problème de la différenciation que s'il n'y a pas de carreaux recouvrant trois communes ou plus. Or ce n'est pas toujours le cas en réalité.



(a) Une observation à risque par différenciation "interne"



(b) Une observation à risque par différenciation "externe"



(c) Dix observations à risque par différenciation "interne"

FIGURE 11 – Trois exemples de problème de différenciation.
Note : Les communes en orange sont celles sur lesquelles on effectue la différenciation avec les carreaux en bleu. Les intersections ainsi déduites apparaissent en plus foncé.

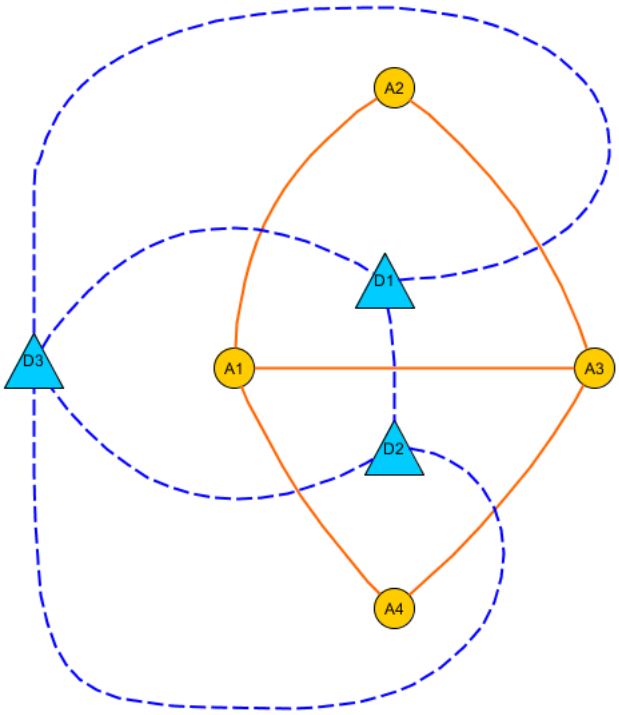


FIGURE 12 – Exemple d'un graphe dual.
Note : Le graphe en bleu avec les arêtes en pointillées est le dual du graphe en jaune.