IDENTIFICATION DES PROBLÈMES DE DIFFÉRENCIATION GÉOGRAPHIQUE À L'AIDE DE LA THÉORIE DES GRAPHES

Marc BRANCHU, Vianney COSTEMALLE, Maëlle FONTAINE

Insee, Division des Méthodes et Référentiels géographiques, Direction de la méthodologie et de la coordination statistique et internationale

vianney.costemalle@insee.fr

Mots-clés : secret statistique secondaire, différenciation, géographie, carroyage

Résumé

La rupture du secret par différenciation

Le secret statistique secondaire concerne les informations qu'un utilisateur peut déduire indirectement à partir de l'ensemble des informations diffusées. On se place dans le cadre de la diffusion d'information sur deux zonages géographiques distincts : d'un côté certaines données sont diffusées au niveau de la commune et d'un autre côté les mêmes données sont diffusées sur une grille de carreaux.

La différenciation géographique consiste à faire la différence entre une zone englobante (composée de communes) et une zone englobée (composée de carreaux) : si les variables diffusées sur ces deux zones sont des variables additives (somme des revenus par exemple), on peut alors en déduite la valeur des variables sur la zone frontière, c'est-à-dire la zone géographique à l'intérieur de la zone englobante, mais ne recoupant pas la zone englobée. Il y a rupture de la confidentialité s'il est possible, par cette technique de la différenciation, de déduire des informations concernant un nombre d'observations inférieur à un certain seuil, fixé par avance et dépendant de la source de données.

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, afin de détecter les ensembles qui conduisent à une rupture de la confidentialité par différenciation. En effet, ce nombre de combinaisons est bien trop grand.

Comment identifier les cas à problème ?

Les travaux présentés visent à modéliser le problème de la différenciation géographique sous forme de graphe orienté : chaque sommet du graphe correspond à une commune, et une arête relie deux tels sommets s'il existe des carreaux à cheval entre les deux communes correspondantes. Cette modélisation prend en compte le fait qu'un carreau peut recouvrir plus de trois communes.

Une fois cette modélisation effectuée, tout l'enjeu est de repérer les sous-graphes connexes conduisant à un problème de différenciation. Pour cela on effectue une simplification du graphe de départ : on fusionne des sommets si la valeur des arêtes les reliant est supérieur au seuil de confidentialité. Par cette méthode, on rétrécit la taille du graphe, jusqu'à converger vers les zones les plus problématiques.

Une dernière étape est alors nécessaire pour repérer tous les ensembles de communes menant à une rupture de la confidentialité par différenciation. 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 repérer des *clusters* et ainsi identifier plus rapidement (mais non exhaustivement) les problèmes de différenciation les plus évidents.

Application

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. L'avantage de pouvoir repérer précisément les endroits où le problème de différenciation se pose est de limiter la perte d'information. En effet, il est toujours possible de régler le problème de la différenciation en perturbant beaucoup les données de base où en ne diffusant l'information que de façon parcellaire.

Bibliographie

[1] Nations Unies, 2004. « Manuel des systèmes d'information géographique et de cartographie numérique ».