DIFFUSION DE DONNÉES FINEMENT LOCALISÉES : NE LAISSER PERSONNE SUR LE CARREAU

Marc BRANCHU, Vianney COSTEMALLE, Maëlle FONTAINE

Insee, Division des Méthodes et Référentiels Géographiques (DMRG), Direction de la méthodologie et de la coordination statistique et internationale

marc.branchu@insee.fr

Mots-clés : données carroyées, confidentialité, swapping

Résumé

Aujourd'hui, les enquêtes de l'Insee et les sources administratives tendent à un géoréférencement plus systématique. Il devient possible de diffuser de l'information statistique sur une grille de carreaux : on parle alors de données « carroyées ». Le carroyage permet de s'affranchir des découpages administratifs et ainsi d'étudier les phénomènes sur des zones précises, infracommunales notamment.

Cependant, plus la taille de ces carreaux est petite, plus se posent des problèmes de confidentialité. On risque en effet de divulguer des informations sensibles sur des individus, des ménages ou des entreprises facilement identifiables par leur position géographique.

Dans l'optique de la diffusion carroyée de la source Filosofi en 2018, la DMRG a implémenté différentes méthodes possibles pour gérer la confidentialité tout en prenant en compte les réalités géographiques, afin de s'éloigner le moins possible de l'objectif initial d'analyses statistiques à un niveau fin. L'objet de la présentation est d'exposer quatre de ces méthodes et d'évaluer dans quelle mesure elles préservent ou altèrent l'utilité des données en vue des analyses futures. Les méthodes présentées sont les suivantes :

- **Diffusion sur une partition moins fine en agrégeant des carreaux**

Ces méthodes ne sont pas perturbatives, dans le sens où les valeurs diffusées sont les « vraies valeurs » observées dans la source à carroyer. Les deux méthodes présentées consistent à agréger les carreaux dans des polygones de telle sorte que chacun de ces polygones contienne un nombre minimum d'observations.

  - *méthode dite « des rectangles »*

On présente un algorithme itératif de désagrégation qui minimise à chaque itération la dispersion des carreaux habités au sein des sous-rectangles. Il aboutit à une partition du territoire en rectangles de dimensions variables et de surfaces décroissantes en fonction de la densité de population.

  - *méthode dite « des grilles superposées »*

On présente un autre algorithme itératif de désagrégation qui divise les carreaux en sous-carreaux, et qui s'arrête dès lors qu'un des sous-carreaux contient moins d'observations que le seuil. On obtient alors un premier niveau de diffusion possible dit « naturel », pour lequel l'ensemble de l'information est diffusé. En allant plus loin, on peut autoriser la diffusion sur des sous-carreaux diffusables à condition de ne pas diffuser l'information sur les sous-carreaux « non-diffusables » : on obtient un second niveau possible, dit « creusé », qui permet de diffuser plus d'information finement localisée que le niveau « naturel ».

- **Diffusion sur la grille la plus fine de valeurs modifiées**

Ces méthodes sont dites perturbatives, dans le sens où les valeurs diffusées sont modifiées. Dans cette optique, on vise cependant à minimiser la perte d'utilité des données, sous contrainte de respecter le secret statistique.

  - *imputation locale par clés de répartition*

Après l'une des méthodes d'agrégation présentées plus haut, on mobilise une méthode d'imputation de valeurs pour petits carreaux non diffusables, par une méthode de clés de répartition, c'est-à-dire une ventilation de la valeur observée dans le dernier carreau diffusable au *pro-rata* du nombre d'observations dans les sous-carreaux

  - *swapping prétabulé*

Une autre méthode consiste à modifier directement les valeurs dans la table individuelle, en procédant à des échanges de ménages situés dans des carreaux différents (*swapping*). Cette méthode consiste dans un premier temps à repérer les ménages à risque, puis à constituer des paires sous contrainte d'un niveau de ressemblance minimum et en essayant de minimiser le total des distances géographiques entre ménages échangés. Pour ce faire, une utilisation originale des techniques d'*optimal matching*, répandus en évaluation de politiques publiques, est proposée.

Ces quatre méthodes seront ensuite comparées entre elles à l'aune d'indicateurs jugés pertinents : part d'information diffusée, différentes mesures du niveau de perturbation introduit, et conservation de l'utilité des données. Il s'agit en particulier de tester dans quelle mesure les indices d'autocorrélation spatiale sont déformés avant et après gestion de la confidentialité.

Bibliographie

[1] *Opportunities and challenges of grid-based statistics*. Tammilehto-Luode, Marja, World Statistics Congress of the International Statistical Institute, 2011

[2] *La gestion de la confidentialité pour les données individuelles. Maxime Bergeat, Document de travail Insee M2016/07, 2016*

[3] *Diffusion de données carroyées. Documentation complète sur les données carroyées à 200 mètres*. Insee, 2013, lien ici

[4] *Data Swapping for Protecting Census Tables*. Shlomo, Natalie and Tudor, Caroline and Groom, Paul, Privacy in Statistical databases, pp. 41–51, 2010

[5] *Geographically intelligent disclosure control for flexible aggregation of census data*. Young, Caroline and Martin, David and Skinner, Chris, International Journal of Geographical Information Science, N°4, Vol. 23, pp. 457-482, 2009

[6] *Targeted record swapping on grid-based statistics in Hungary*. Beata Nagy, Hungarian Central Statistical Office, Submission for the 2015 IAOS Prize for Young Statisticians, 2015

[7] *Optimal full matching and related designs via network flows*. Hansen, B.B. and Klopfer, S.O., Journal of Computational and Graphical Statistics, 15, 609–627. 2006