

OPEN DATA, BIG DATA : QUEL RENOUVEAU DU RAISONNEMENT CARTOGRAPHIQUE ?

par Emilie Lerond

ThéMA, UMR 6049 CNRS, Université Bourgogne Franche-Comté
4 Boulevard Gabriel, 21000 Dijon
emilie.lerond@univ-fcomte.fr

Olivier Klein

Luxembourg Institute of Socio-Economic Research (LISER)
Maison des Sciences Humaines
11, Porte des Sciences, L-4366 Esch-sur-Alzette/Belval
olivier.klein@liser.lu

et Jean-Philippe Antoni

ThéMA, UMR 6049 CNRS, Université Bourgogne Franche-Comté
4 Boulevard Gabriel, 21000 Dijon
jean-philippe.antoni@u-bourgogne.fr

Le mouvement d'open data, qui permet l'accès gratuit à un grand nombre de données spatiales ou démographiques, associé au développement d'outils de cartographie ou de visualisation libres (SIG, programme), a permis l'augmentation de la production de cartes. Seulement, l'automatisation des traitements permise par ces outils tend à gommer le raisonnement cartographique et peut conduire à des erreurs de construction cartographique, d'autant plus que la phase d'analyse des données peut s'avérer de plus en plus complexe dans un contexte de big data (en tant que données massives, peu structurées et désagrégées). Le raisonnement cartographique est le processus menant d'une donnée brute à une donnée transformée, interprétée et représentée. Il consiste en l'analyse statistique du jeu de données, en l'émission d'hypothèse(s) sur le jeu de données, aux choix de la variable du jeu de données à représenter, de l'unité spatiale de base de la cartographie, et de la discrétisation retenue pour représenter au mieux le jeu de données. L'effet de cette représentation est important sur le résultat obtenu et ainsi sur la manière dont sera ensuite interprétée la variable. L'objectif de cette communication est de présenter un exemple de raisonnement cartographique à partir d'un jeu de données massives (plus de deux millions d'individus statistiques dans une zone d'étude d'environ 300 km²), en insistant sur l'influence des choix réalisés au cours du raisonnement cartographique : comment, à partir d'une même variable, la seule manière de cartographier peut nous conduire à des résultats graphiques très différents ?

Contexte

La diversification des sources de données en géographie et leur plus grande facilité d'accès, liée aux données ouvertes (*open data*), offrent plus de possibilités pour les analyses statistiques et géographiques, ainsi que pour leurs représentations cartographiques. Cette diversification conduit à l'introduction, dans le domaine de la géographie, des données massives (*big data*), en tant que grand volume de données, peu structurées, désagrégées et multidimensionnelles. Leurs attributs sont spatiaux (localisation, géométrie), thématiques et parfois temporels. Leur volume et leur structure les rendent

difficiles à manipuler, y compris dans des systèmes d'information géographique (SIG).

Ainsi les choix effectués au cours du raisonnement cartographique doivent porter à la fois sur les attributs thématiques (quel attribut représente-t-on ? Doit-on le transformer pour lui appliquer une discrétisation cohérente ? Quel type de discrétisation applique-t-on ?) et sur les attributs spatiaux (conserve-t-on la forme géométrique de départ de l'attribut ? Propose-t-on une discrétisation de l'espace ? Ces différents choix peuvent être réalisés de manière consciente au cours du raisonnement cartographique (processus guidant la création d'une carte), transformant les

possibilités offertes par les données massives et les données ouvertes en réelles opportunités pour leur représentation graphique. Cependant ces choix peuvent également être réalisés de manière automatique, suggérés notamment dans les interfaces de sémiologie des SIG, et les possibilités peuvent alors se transformer en véritables écueils et conduire à « faire mentir les cartes » (Monmonier 1993). A partir d'un exemple de données de mobilité, nous proposons de montrer comment certains choix cartographiques conduisent à des résultats graphiques et analytiques très différents.

Terrain d'étude et données

Dans le cadre de ce travail, nous disposons d'un jeu de données issu du modèle de simulation spatiale MobiSim, plateforme de simulation LUTI individu-centrée (Tannier *et al.* 2015 ; Antoni *et al.* 2016). MobiSim fournit des données de mobilités individuelles, désagrégées et structurées, qui correspondent aux positions successives des individus en déplacement dans l'espace urbain. Le jeu de données compte environ 180 000 individus, dont les déplacements sont connus à la seconde sur une journée type de simulation, à l'échelle de la Communauté d'Agglomération du Grand Besançon (CAGB). Ces déplacements sont constitués par un point d'origine et un point de destination connus à une heure de départ et à une heure d'arrivée (au format heure:minute:seconde) et dont le motif (travail, étude, domicile, loisirs, réseau social, accompagnement, commerce ou service) et le mode de déplacement (voiture, transports en commun, vélo, marche à pied) sont également connus. Ces informations sont stockées dans une table attributaire correspondant à ces points.

La plupart des individus effectuant plusieurs déplacements par jour, le nombre total de déplacements augmente très rapidement. On compte jusqu'à plus de deux millions de déplacements, stockés sous forme d'une matrice origine-destination (points de départ et de destination connus et reliés par une ligne directe) et sous forme de points (points de départ et de destination). Ainsi les déplacements dans notre jeu de données possèdent à la fois des attributs spatiaux linéaires et ponctuels, que l'on pourra choisir de conserver sous cette forme ou de modifier au besoin pour la représentation graphique.

Méthodologie

Nous choisissons de nous focaliser sur les points de départ en voiture des déplacements, sans distinction

d'horaires ou de motifs de déplacements, soit environ 550 000 déplacements. L'ensemble de ces points de départ sont localisés en coordonnées géographiques x,y exactes et l'espace d'étude n'est donc pas discrétisé *a priori* lors de la simulation des déplacements. Ainsi deux traitements particuliers doivent être réalisés pour obtenir une représentation cartographique des déplacements (fig.1) : en premier lieu, agréger les attributs spatiaux, et en second lieu, transformer l'attribut thématique et le discrétiser.

Agrégation des attributs spatiaux

Pour analyser et visualiser ces données, nous cherchons à en réduire le volume en essayant de conserver autant que possible leur caractère désagrégé. La solution apportée par l'agrégation, en tant que réalisation d'un système de partition (Openshaw 1981 ; Openshaw 1983) permet de gagner en maniabilité et en visualisation.

Les choix possibles de méthode d'agrégation de l'espace sont multiples et s'inscrivent, à ce titre, à part entière dans le raisonnement cartographique (Cauvin *et al.* 2007a ; Cauvin *et al.* 2007b). Plusieurs paramètres sont à prendre en compte :

- Le point d'origine de la grille,
- L'orientation des cellules de la grille,
- La régularité ou l'irrégularité de ces cellules,
- La forme des cellules (carrés, hexagones, etc. pour des grilles régulières, ou polygones de Thiessen pour des grilles irrégulières par exemple ; Antoni *et al.* 2017),
- La résolution des cellules de la grille (Lerond *et al.* 2017).

Transformation des attributs thématiques

Une fois l'espace d'étude discrétisé et l'attribut sommé dans des unités spatiales d'agrégation (cellules), nous nous intéressons aux attributs thématiques en tant que tels. Dans le cas retenu ici, chaque unité spatiale compte n départs de déplacements (0 exclu). Afin d'analyser cet attribut, la discrétisation de ces n valeurs en k classes s'impose pour comprendre plus clairement la distribution statistique des valeurs. Cette discrétisation nécessite de choisir le nombre k de classes qui sera réalisé et de choisir la méthode par laquelle les valeurs vont être « découpées » en classe, en s'appuyant sur la forme de la distribution statistique de l'attribut. Plusieurs méthodes de discrétisation des valeurs existent : intervalles égaux, effectifs égaux, analyse de la variance (dite méthode de Jenks) ou écart-type par exemple. Ces méthodes peuvent être employées de manière automatisée dans les SIG, tout en restant paramétrables (choix de la méthode, choix du nombre de classes, choix de la sémiologie).

La forme de la distribution statistique est également déterminante dans le choix de la méthode de discrétisation (Cauvin *et al.* 2007b). Dans l'idéal, afin de minimiser le risque d'erreur statistique, c'est à partir d'une distribution se rapprochant le plus possible de la loi normale qu'il faut discrétiser un attribut. Si la variable à cartographier ne suit pas une loi normale, il est nécessaire de la transformer mathématiquement de manière à obtenir une distribution plus proche d'une distribution normale, lorsque celle d'origine paraît trop disparate. Ces transformations mathématiques sont diverses, mais l'on utilise assez souvent des fonctions logarithmiques (\ln , \log_{10} , \log^2) ou puissances.

Paramétrage

Les effets du traitement des attributs spatiaux et de celui des attributs thématiques sont testés sur le jeu de données, à partir de l'attribut « départs de déplacements ». Seuls un petit nombre de paramètres des traitements sera testé, aussi bien pour les attributs spatiaux que thématiques (fig. 2). Le fait de modifier simultanément à la fois la résolution et la méthode de discrétisation permet de s'assurer que les effets supposés par ces deux paramètres existent dans différentes conditions, et ne sont donc pas des artefacts liés aux traitements effectués.

L'ensemble des paramètres du traitement des attributs spatiaux présentés en page 110 peut être testé séparément afin de constater leur effet. Pour l'instant, seule la résolution spatiale est testée, tous les autres paramètres étant fixés. Nous utilisons une grille régulière de cellules carrées et testons trois résolutions : 200 mètres (soit 11.795 cellules), 500 mètres (1.968 cellules) et 1000 mètres (520 cellules). Ces résolutions sont choisies afin que la différence entre le nombre de cellules par résolution soit bien visible et son effet appréciable à l'œil nu.

Concernant les attributs thématiques, nous choisissons de les discrétiser en cinq classes par la méthode de Jenks sur chacune des résolutions proposées, à partir de la distribution originelle de l'attribut et à partir de sa distribution transformée. Cette méthode est fondée sur la variance et vise à maximiser la variance inter-classes et à minimiser la variance intra-classe. Un dernier exemple est proposé à partir d'une discrétisation par la méthode des écart-types en six classes (cette méthode étant fondée sur la moyenne et l'écart-type, il est préférable d'employer un nombre pair de classes) sur la résolution de 200 mètres. Cette méthode permet de mettre en valeur la dispersion des valeurs de l'attribut autour de la moyenne de la distribution.

Résultats

Les résultats obtenus permettent de visualiser les effets de la transformation de l'attribut « nombre de départs de déplacements », dont la comparaison conduit à revoir l'interprétation thématique de cet attribut. Nous proposons également une rapide comparaison de la méthode de discrétisation sur la résolution de 200 mètres à partir de l'attribut transformé.

Comparaison des effets de la transformation de l'attribut

Dans les trois cas de résolution, la comparaison des distributions statistiques originelles et transformées révèle des formes de distributions très différentes (fig. 3) : si les distributions originelles présentent une dissymétrie à gauche vers les très petites valeurs et quelques valeurs extrêmes éloignées de l'ensemble des valeurs, les distributions transformées par logarithme naturel se rapprochent de distributions normales, à l'exception de celle réalisée à 1000 mètres de résolution. La difficulté à la rapprocher de la loi normale est peut-être liée au petit nombre d'unités spatiales (520 individus statistiques) couvrant une très grande étendue de valeurs (de 1 à 19 500 départs). Ce constat laisse supposer qu'une résolution trop petite par rapport à la zone d'étude ne permet pas de proposer de représentation graphique statistiquement valable. Ainsi, bien que les cartes obtenues à partir des distributions transformées aux résolutions de 200 et 500 mètres (fig. 5 et 9) semblent sur-représenter les classes 4 et 5, elles sont statistiquement plus correctes pour l'analyse de l'attribut « nombre de départs de déplacements ».

Les cartes obtenues par discrétisation sans transformation de l'attribut paraissent montrer une grande différenciation entre le nombre de départs de déplacements en voiture dans le centre-ville de la CAGB et dans les communes périphériques (fig. 4, 6 et 8). A l'inverse, les cartes obtenues par discrétisation avec transformation révèlent finalement que, dans les communes périphériques, le nombre de départs en voiture est élevé (fig. 5, 7 et 9) ; moindre certes que dans la ville de Besançon (déjà plus présente que son seul centre-ville), mais la périphérie reste très visible dans les classes 4 à 5 pour cet attribut. Ainsi, sans transformation, on pourrait conclure que les ménages vivant en périphérie, soit se déplacent moins que les ménages du centre, soit ont moins recours à l'automobile. Or, après transformation de l'attribut (préférable pour appliquer la discrétisation), on observe que les ménages vivant en périphérie se déplacent également beaucoup en voiture : la périphérie compte un grand nombre de cellules présentes dans les classes 3 à 5. De plus, on

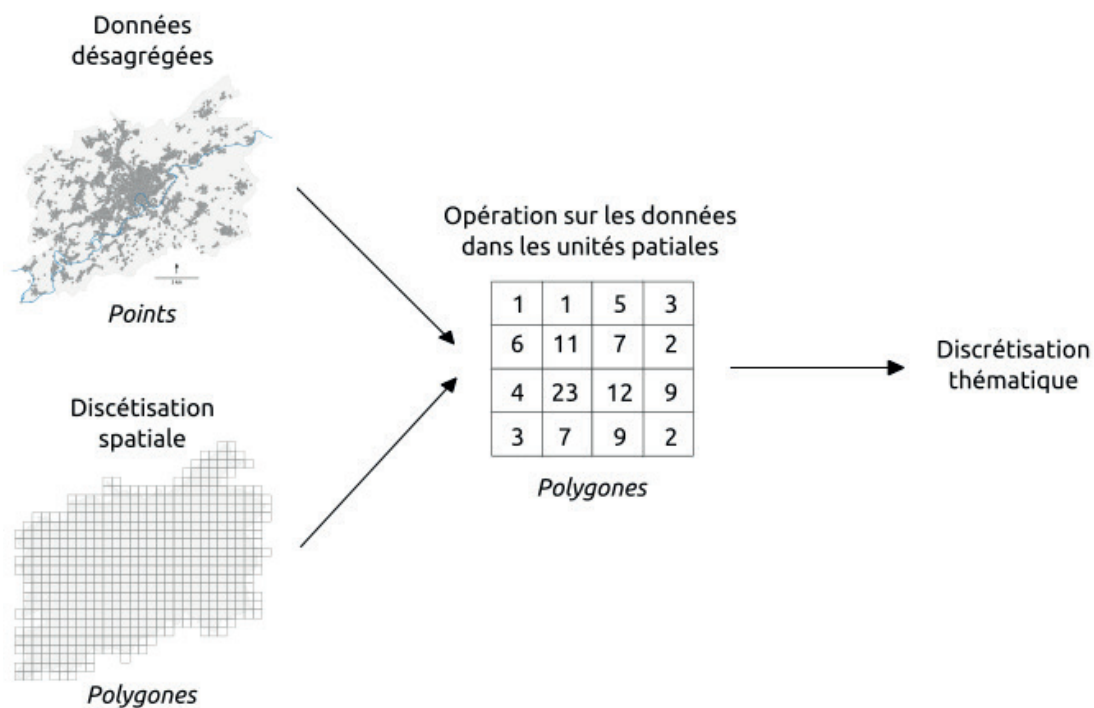


Figure 1 : Traitements des données pour la représentation cartographique

Attributs spatiaux						
Origine	fixe		fixe		fixe	
Orientation	fixe		fixe		fixe	
Régularité	fixe		fixe		fixe	
Forme	fixe		fixe		fixe	
Résolution	200		500		1000	
Nombre de cellules	11.795		1.968		520	
Attributs thématiques						
Méthode de discrétisation	Jenks	Ecart-type	fixe		fixe	
Transformation de l'attribut	non	$\ln(n) \times 10$	non	$\ln(n) \times 10$	non	$\ln(n) \times 10$
Nombre de classes	5	6	fixe		fixe	

Figure 2 : Résumé des paramètres de traitement

peut également supposer que, si ce nombre de départs était rapporté au nombre de ménages ou d'individus par cellule, les communes périphériques seraient encore plus représentées dans les classes élevées, en partant de l'hypothèse d'une plus grande dépendance à l'automobile dans ces parties de la communauté d'agglomération. Ainsi la transformation de l'attribut avant la cartographie et l'analyse des données donnent à voir une structure de déplacement automobile assez différente des hypothèses *a priori* que l'on peut avoir sur les comportements de mobilité.

Comparaison des effets de la méthode de discrétisation

Avec une distribution statistique proche de la loi normale pour l'attribut transformé, nous testons maintenant l'effet de la méthode de discrétisation en elle-même sur la représentation graphique. Nous comparons la méthode de Jenks à celle par écarts-types pour la résolution de 200 mètres (fig. 5 et 10).

Le recours à chacune de ces deux méthodes permet d'obtenir deux discrétisations différentes. La méthode par écarts-types (fig. 10) met en avant les valeurs centrales (classe 4) et crée plus de classes pour les valeurs faibles (classes 1 et 2), qui témoignent d'un comportement assez rare : la CAGB compte peu de lieux avec un très petit nombre de départs de déplacements. Avec la méthode de Jenks, on observe que les classes élevées ont de grandes étendues (57-248 pour la classe 4, 249-2936 pour la classe 5). Si ces classes sont homogènes malgré leurs étendues, cela témoigne que les lieux de la CAGB comptant un grand nombre de départs de déplacements sont plus fréquents. On obtient ainsi, malgré deux méthodes de discrétisation différentes, des conclusions thématiques similaires. Cela laisse à supposer que lorsqu'on dispose d'un attribut, dont la distribution statistique suit la loi normale, il n'y a pas qu'une seule méthode de discrétisation appropriée pour sa représentation graphique. Si le principe de la méthode utilisée est connu (de manière à ce que la discrétisation soit correctement interprétée), alors l'attribut sera correctement analysé, même s'il est issu d'une source de données massive, peu structurée et désagrégée. Cependant, même si ces deux méthodes conduisent à des analyses semblables, on peut supposer que le fait d'obtenir plus de classes de petites valeurs ou au contraire de grandes valeurs (c'est-à-dire plus de classes de couleurs claires ou de couleurs foncées) peut amener le lecteur de la carte à un *a priori* différent sur le phénomène qu'il observe. Ce point reste encore difficile à explorer et à évaluer, et rejoint les problématiques traitées par le domaine de la cognition (Garlandini et Fabrikant 2009).

Conclusion et perspectives

Les résultats obtenus à partir d'un jeu de données massif et désagrégé (*big data*) rappellent l'importance de l'analyse statistique préalable des variables à cartographier. Cette analyse guide les choix à réaliser au cours du raisonnement cartographique et permet d'éviter de commettre des erreurs statistiques susceptibles de conduire à une mauvaise représentation et, de fait, à une mauvaise interprétation des données. Il est par exemple possible de choisir d'utiliser les cartes réalisées sans transformation de la variable (fig. 4, 6 et 8) pour amener à croire que les ménages vivant en périphérie ont peu recours à l'automobile, se reportent très probablement sur les transports en commun, qui n'ont donc pas besoin d'être plus développés vers ces communes : il est ainsi possible d'utiliser ces cartes pour faire « mentir » (délibérément ou non) les données de mobilité. De manière générale, l'importance de l'analyse statistique concerne n'importe quel jeu de données. Cela est d'autant plus le cas lorsqu'il est massif, désagrégé et peu structuré, car il nécessite, en fin de compte, de bien l'appréhender afin de le représenter au mieux.

Cependant ces résultats sont partiels, car seuls quelques paramètres ont été testés et comparés. Du point de vue des attributs spatiaux, de plus nombreux tests sont nécessaires pour évaluer l'effet de la création des unités spatiales d'agrégation sur l'attribut thématique à cartographier. Non seulement leur forme et leur régularité, mais aussi leur origine et leur orientation sont susceptibles de produire des résultats différents. Cette problématique, qui relève du *Multiple area unit problem (MAUP)*, reste encore un obstacle conséquent en géographie, obstacle dont les effets sont de surcroît difficiles à évaluer. Du point de vue des attributs thématiques, une seule variable a été testée (nombre de départs de déplacements). Cet attribut pourrait être thématiquement enrichi en le rapportant au nombre de ménages, d'habitants, ou à la superficie de bâti dans les cellules, afin de proposer une analyse plus fine du rapport entre déplacements et démographie ou occupation du sol. Néanmoins, en démontrant l'effet de la transformation de l'attribut sur la représentation graphique et l'interprétation thématique qui en résulte, il est d'ores et déjà possible de visualiser toute l'importance et l'intérêt des choix effectués au cours du raisonnement cartographique pour l'analyse des données.

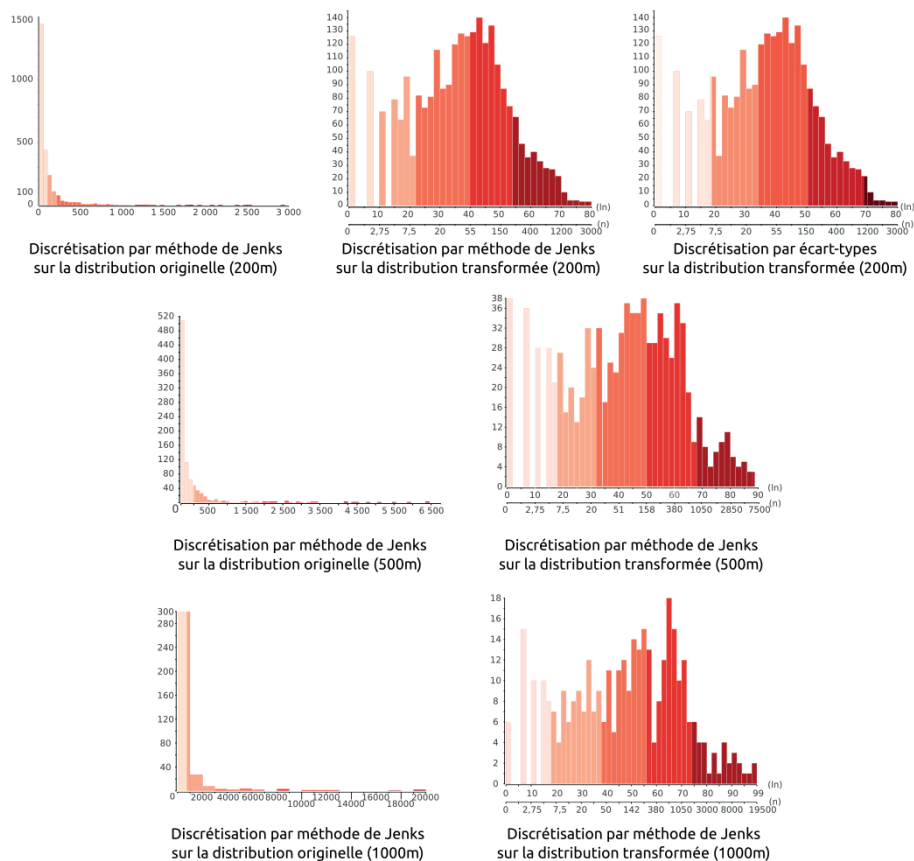


Figure 3 : Distributions du nombre de départs de déplacements (nombre de cellules en abscisses, nombre n de déplacements en ordonnées)

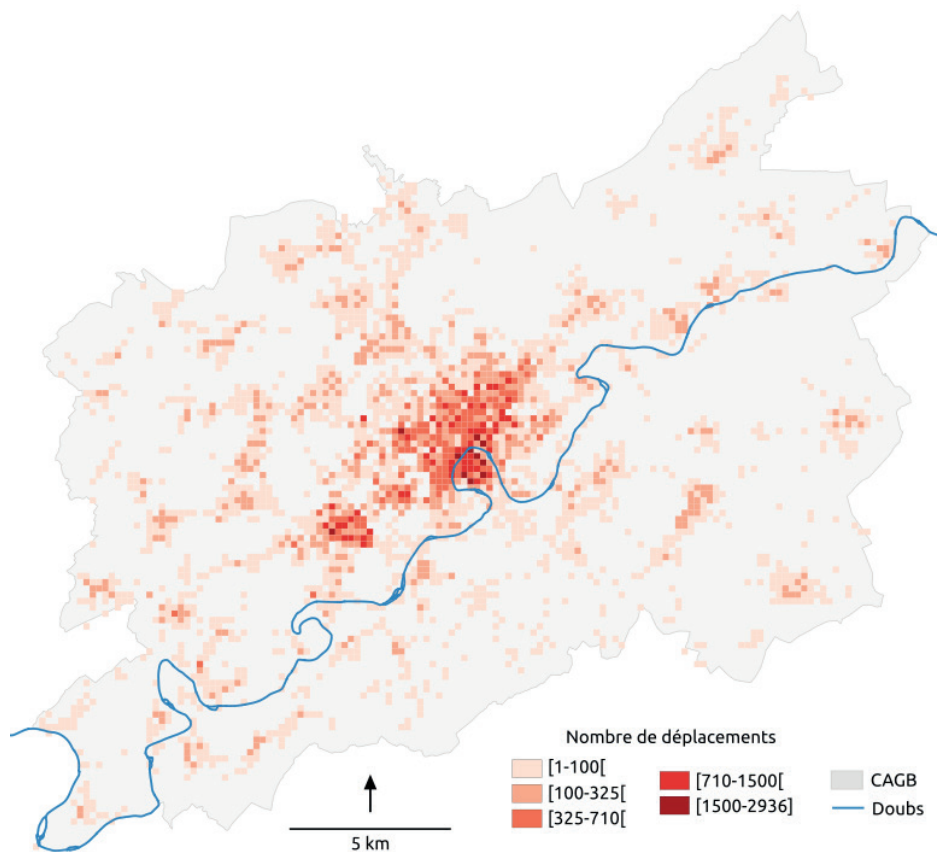


Figure 4 : Nombre de départs de déplacements en voiture sans transformation à 200 mètres de résolution (méthode de Jenks)

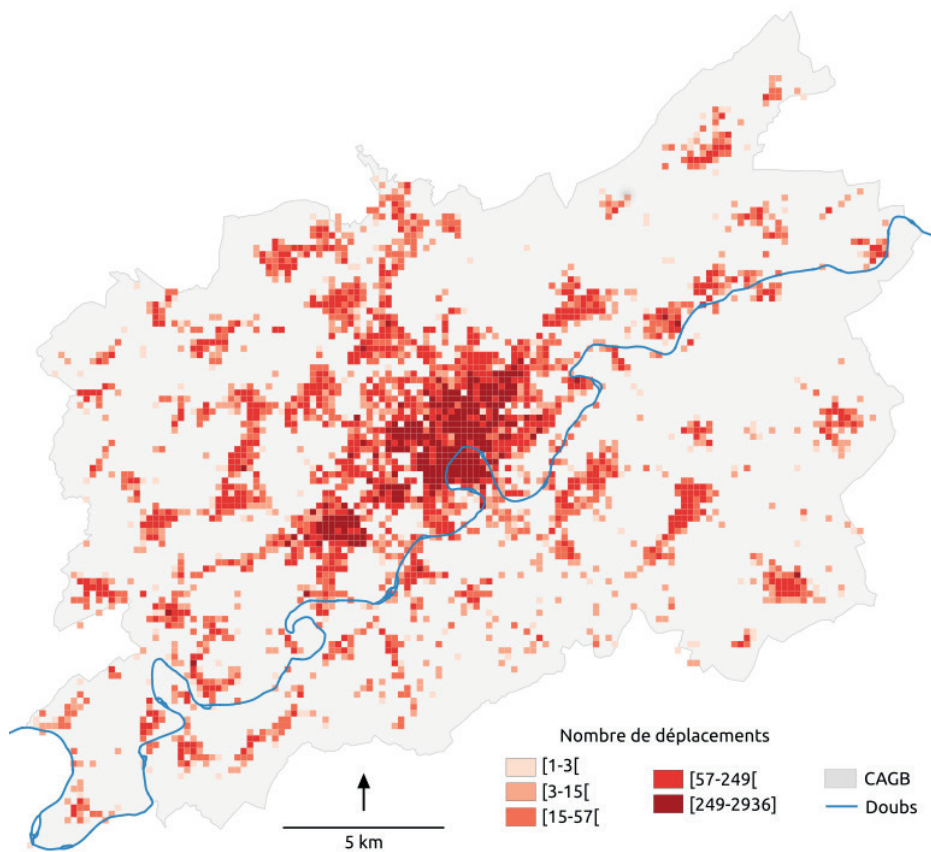


Figure 5 : Nombre de départs de déplacements en voiture avec transformation à 200 mètres de résolution (méthode de Jenks)

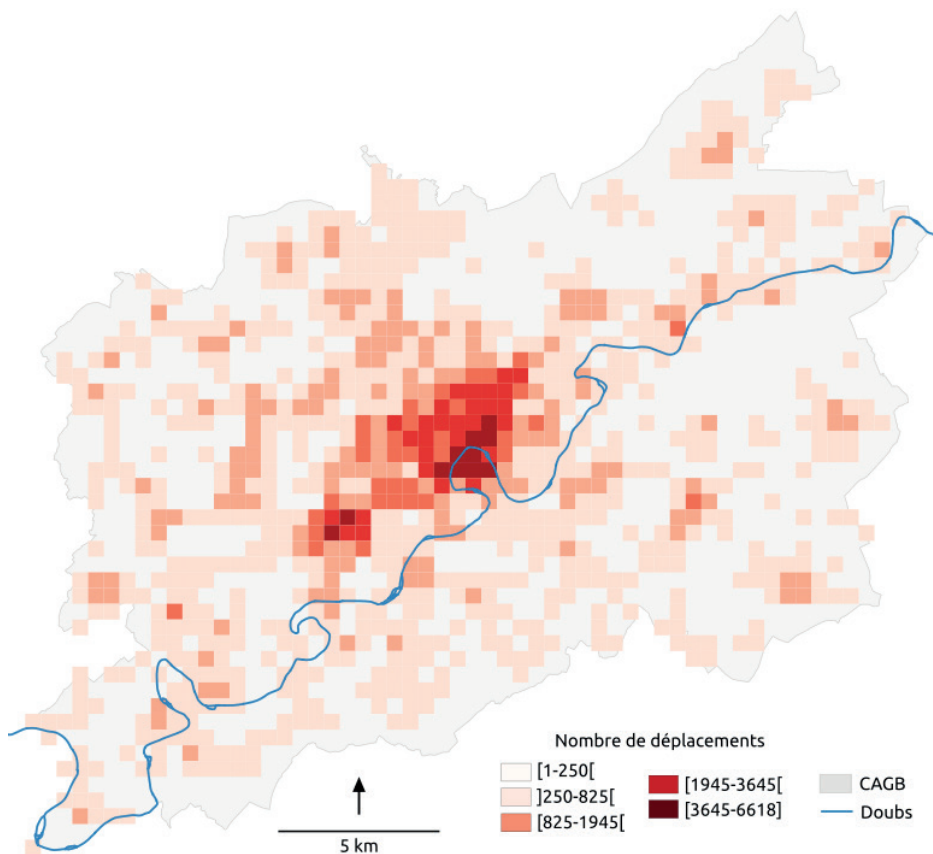


Figure 6 : Nombre de départs de déplacements en voiture sans transformation à 500 mètres de résolution

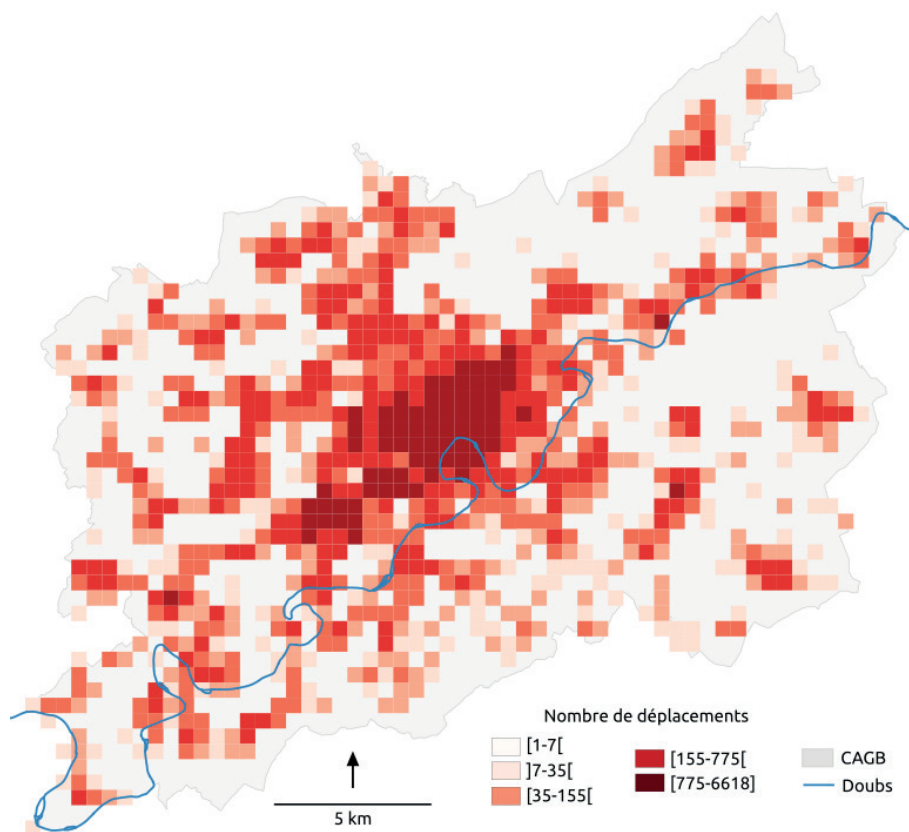


Figure 7 : Nombre de départs de déplacements en voiture avec transformation à 500 mètres de résolution

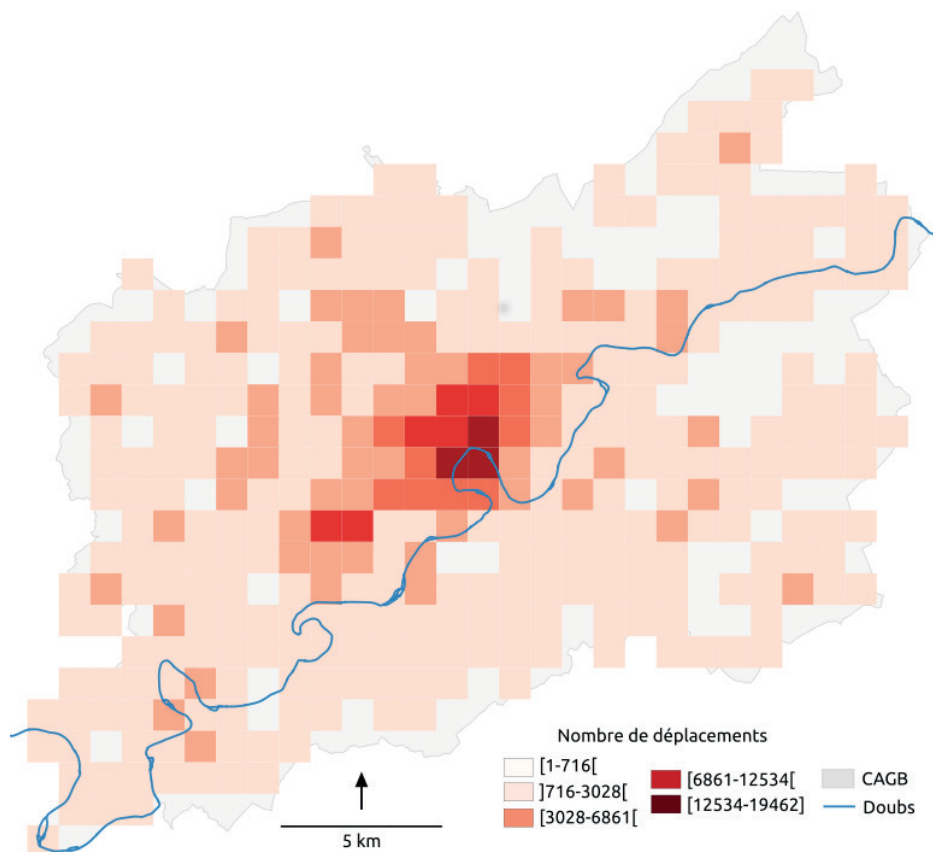


Figure 8 : Nombre de départs de déplacements en voiture sans transformation à 1000 mètres de résolution

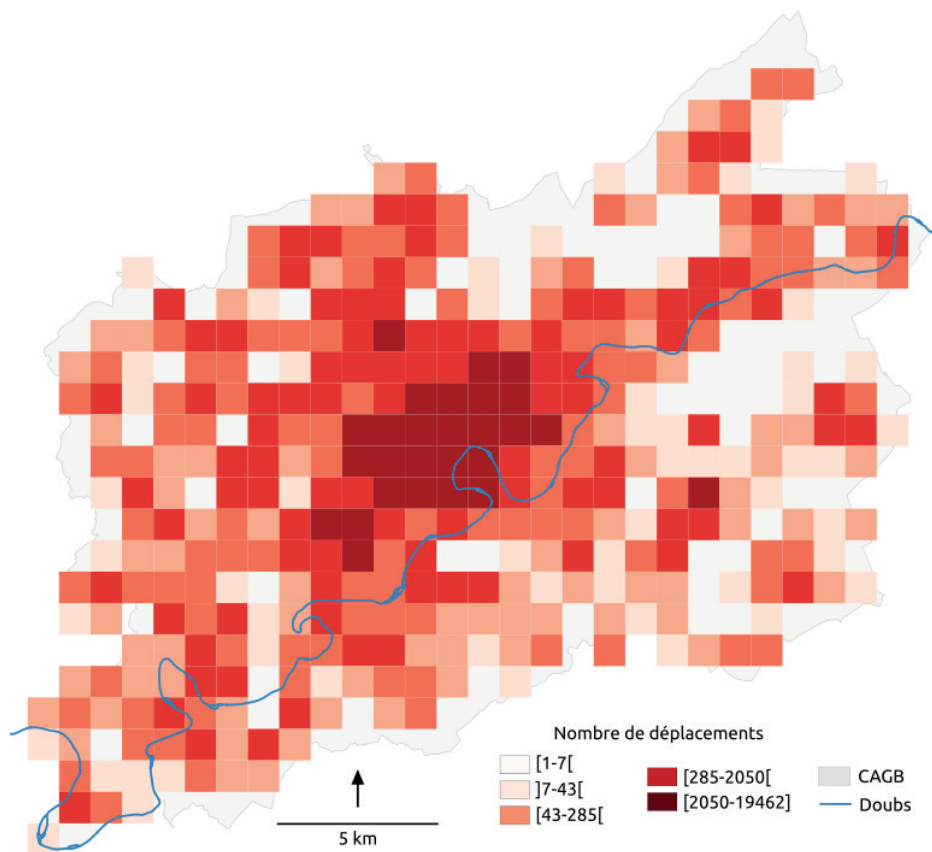


Figure 9 : Nombre de départs de déplacements en voiture avec transformation à 1000 mètres de résolution

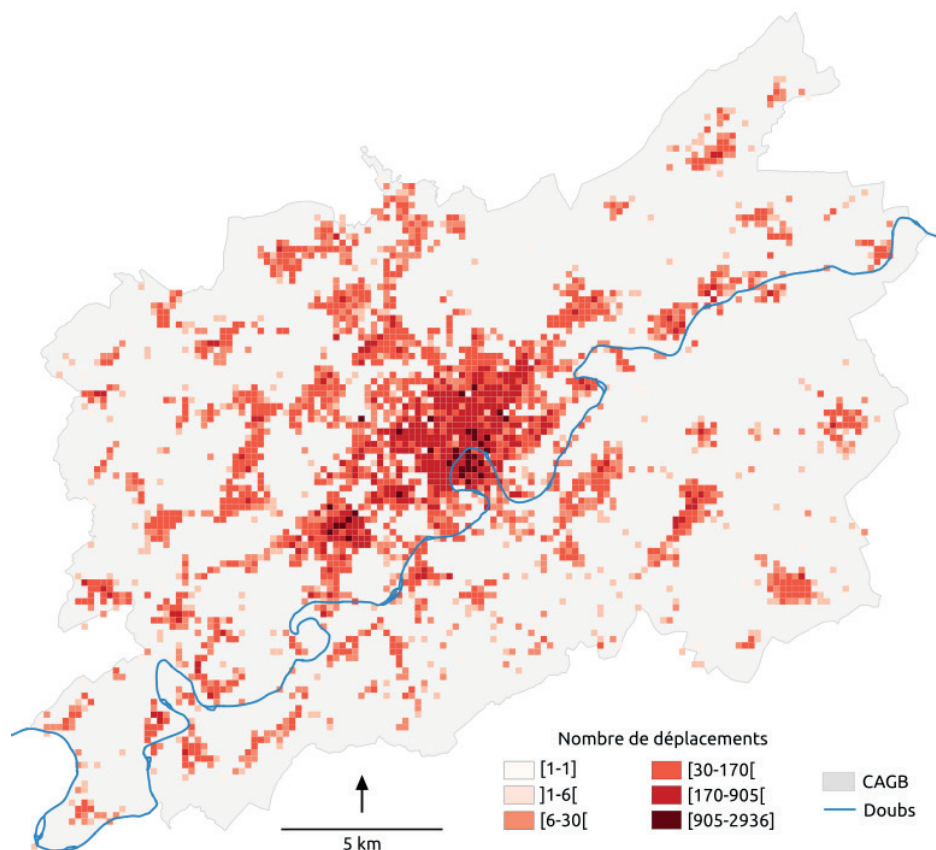


Figure 10 : Nombre de départs de déplacements en voiture avec transformation à 200 mètres de résolution (discrétisation par écart-type)

Bibliographie

Antoni J.-Ph., Lerond E., Klein O., Moissy S., 2017, « Six nuances d'a-grey-gation. L'influence du choix des unités spatiales d'agrégation sur la lecture et l'interprétation des résultats cartographiés », *ThéoQuant*, Besançon, mai 2017.

Antoni J.-Ph, Lunardi N., Vuidel G., 2016, « Simuler les mobilités individuelles. Les enjeux de l'information géographique », *Revue internationale de géomatique*, vol. 2, p. 237–262.

Cauvin C., Escobar F. et Serradj A., 2007a, *Cartographie thématique. Tome 1* : Paris, Ed. Hermès-Lavoisier.

Cauvin C., Escobar F. et Serradj A., 2007b, *Cartographie thématique. Tome 3 : méthodes quantitatives et transformations attributaires*, Paris, Ed. Hermès-Lavoisier.

Garlandini S., Fabrikant S. I., 2009, « Evaluating the effectiveness and efficiency of visual variables for geographic information visualization », *International Conference on Spatial Information Theory*, Aber Wrac'h (Landéda, France), septembre 2009.

Lerond E., Klein O., Antoni J.-Ph., 2017, « An aggregation method for mobility data analysis and visualization », *International Cartographic Conference*, Washington, juillet 2017.

Monmonier M.S., 1993, *Comment faire mentir les cartes ou Du mauvais usage de la géographie*, Paris, Ed. Flammarion.

Openshaw S., 1981, « Le problème de l'agrégation spatiale en géographie », *L'espace géographique*, n°1, p. 15-24

Openshaw S., 1983, *The Modifiable Areal Unit Problem*, Norwick, Géo Books, collection CATMOG.

Tannier C., Hirtzel J., Stephenson R., Couillet A., *et al.*, 2015, « Conception and use of an individual-based model of residential choice in a planning decision process. Feedback from an experimental trial in the city of Besançon, France », *Progress in planning*, 108, p. 1-38.