Les « *big data* » sont-elles l'avenir de la géographie [théorique et quantitative] ?

Frédéric Audard[1], Samuel Carpentier[1], Sébastien Oliveau[1,2]

[1]Aix-Marseille Université, CNRS, ESPACE UMR 7300
Pôle Géographie, Aménagement, Environnement
29 avenue Robert Schuman F -13621 Aix-en-Provence
[2] Groupe Dupont
frederic.audard@univ-amu.fr, samuel.carpentier@univ-amu.fr, sebastien.oliveau@univ-amu.fr

---

"Raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care".

G.C. Bowker[1]

Les *big data* sont aujourd'hui désignées comme une des perspectives majeures des sciences (Marx, 2013), ainsi que des sciences sociales. La géographie, discipline largement impliquée dans la construction et l'analyse de données, n'échappe pas à ce mouvement, bien au contraire. La généralisation des outils de géolocalisation (par GPS, GSM, Wifi ou encore adresse IP) aboutit à une explosion de la disponibilité des données à référence spatiale. Envisagées généralement comme une opportunité, ces masses de données offriraient au chercheur un champ nouveau et jusqu'alors inaccessible, celui de l'exhaustivité et du temps réel (Batty, 2012).

La géographie théorique et quantitative, parfois critiquée pour son aspect « formel » qui la tiendrait trop loin de la « réalité du terrain », a-t-elle trouvé dans ces masses de données l'interface empirique idéale de ses modélisations ? De même, les géographies sociale et culturelle peuvent-elles trouver dans l'accès et la manipulation facilitée[2] des données numériques une « scientificité » qui leur ferait défaut ? La thèse de cette communication est que cette question de l'accès généralisé à ces *big data*, et surtout de ses conséquences sur la pratique de la recherche dans notre discipline, n'a pas encore fait l'objet d'une nécessaire controverse parmi la communauté des géographes (francophones).

Big data ?

Si, en première approche, les *big data* peuvent être définies comme de grandes bases de données, plusieurs caractéristiques complémentaires permettent de préciser ce concept. Outre le nombre d'individus statistiques (d'enregistrements), qui tend vers l'exhaustivité, ces masses de données sont caractérisées par leur variété, leur granularité fine, leur caractère relationnel ou encore leur flexibilité – permettant par exemple une certaine modularité ainsi que des facilités d'appariement (Boyd et Crawford, 2012). Elles se distinguent également des bases de données conventionnelles (enquêtes, recensement, inventaires…) par le caractère dynamique de leur production, qui s'effectue de manière continue (Kitchin, 2013). Les sources de ces grandes bases de données, souvent géoréférencées et longitudinales, sont principalement liées au secteur privé à travers les grandes sociétés de la téléphonie ou de l'internet ; là où, jusqu'à présent, la plupart des grandes bases de données émanaient du secteur public.

[1]Bowker, G. C. (2005). *Memory Practices in the Sciences*. MIT Press, Cambridge, Massachusetts
[2]À travers les logiciels en ligne produisant des cartes, graphiques et rapports automatisés, comme Géoclip, les outils du Géoportail ou encore ceux de l'INSEE…

« Garbage in »…

L'enthousiasme général autour de l'émergence de ces *big data* s'accompagne néanmoins de quelques mises en garde. Parmi les réserves formulées à l'encontre des masses de données, on peut retenir principalement trois registres, les conditions de production, les conditions d'accès et les limitations éthiques, lesquels se déclinent selon le type de producteur de la donnée.

Si on s'intéresse dans un premier temps aux données publiques, force est de constater que l'accès à ces dernières se trouve de plus en plus facilité par les nombreuses initiatives de type *open access*. En Europe c'est notamment la directive INSPIRE de la commission européenne qui donne un cadre à ce mouvement de diffusion massive des données géographiques. De par ce cadre règlementaire et l'expertise souvent ancienne des organismes publics en la matière, les métadonnées sont en principe bien renseignées. Le problème principal est alors celui de l'appariement, notamment pour des questions de différences de nomenclature d'une source à l'autre.

Les données issues de sociétés privées constituent quant à elles la source potentiellement la plus importante de diffusion de ces grandes bases de données. L'accès à ces sources de données privées reste plus problématique dans la mesure où ces dernières constituent pour les sociétés une potentielle source de revenus. Par ailleurs on déplore un déficit de description des métadonnées, voire des erreurs (Goodchild, 2007). Mais la réserve la plus fondamentale, qui doit nous amener à considérer ces données avec la plus grande prudence, tient plus à la finalité commerciale de ces sociétés et *in fine* de ces bases de données. Le point de vue particulier lié à l'activité commerciale n'est ainsi pas nécessairement compatible avec l'exigence d'objectivation d'une démarche scientifique. Par ailleurs, l'utilisation privilégiée de ces ressources comme matériau principal des travaux de recherche pourrait, à terme, induire une forme d'asservissement en créant une relation de dépendance.

Enfin, s'agissant des données libres issues de processus collaboratif, qui constituent potentiellement une alternative aux données privées, les limitations tiennent en deux points principaux. Tout d'abord, le caractère collectif de leur construction suppose des difficultés d'harmonisation ainsi que de multiples erreurs (Flanagin et Metzger, 2008). Ensuite, ces bases de données exploitent souvent, au moins en partie, d'autres sources de données (notamment issues du secteur privé) et les fusionnent avec plus ou moins de rigueur, conduisant à la création d'un corpus hétérogène.

Dans tous les cas de figure, mais sans doute de manière plus prononcée pour les données issues de sociétés privées, la question du respect de la vie privée représente le dernier écueil de taille de la question des *big data*.

La tentation du « *end user* »

Au-delà des limitations inhérentes aux données elles-mêmes, l'usage de ces données suppose également quelques restrictions. En effet, l'accès facilité à de grandes quantités de données aboutit, d'après nous, au paradoxe suivant : délesté (en apparence du moins) de la question de la collecte des données, indispensable à l'approche expérimentale, le géographe quantitativiste se trouve dans la posture inédite de l'utilisateur final. Or « *il n'est en général pas possible de redonner* a posteriori *une cohérence à des informations collectées sur des bases conceptuelles différentes* » (Terrier, 2011). Ce faisant, il donne ainsi à des données qu'il n'a pas produites une emprise plus forte sur ses résultats en réduisant sa capacité de contrôle de ses expérimentations. Autrement dit, en s'épargnant l'étape, souvent ingrate et fastidieuse, de constitution de son corpus, le chercheur géographe

renonce au contrôle de la validité interne de ses expérimentations, faute de maîtriser les conditions de productions des données.

Il est alors tentant d'objecter que ce que ces bases de données perdent en précision, elles le gagnent en représentativité (N=all selon Kitchin, 2013). Or, une base de données, ou un plus largement un corpus, ne s'oriente pas obligatoirement vers une objectivation accrue de par son exhaustivité apparente. S'il y a un intérêt à analyser ces immenses bases de données, c'est sans doute moins pour leur représentativité justement que pour y déceler des tendances nouvelles, marginales, émergentes, que d'autres analyses plus rigoureuses et approfondies, portant sur des données *ad hoc*, viendraient étayer par la suite.

Si les *big data* ne semblent pas être l'outil privilégié de l'expérimentation, de par le manque de contrôle des conditions expérimentales, faut-il pour autant rejeter les perspectives extrêmement stimulantes qui découlent de cette abondance d'information ? Assurément non. Les grandes masses de données n'en restent pas moins une immense opportunité d'exploration de pistes nouvelles ; à défaut d'inférence, elles permettent sans doute l'émergence. Si les données ne peuvent à elles seules constituer l'alpha et l'oméga d'une recherche scientifique, force est de constater que les traditionnelles démarches inductives et hypothético-déductives doivent désormais s'articuler avec une troisième voie, non exclusive, mais sans doute complémentaire : l'abduction (Banos, 2005).

Les *big data* catalyseur ou inhibiteur des clivages traditionnels de la géographie ?

Finalement, si la fouille de données n'est pas une perspective totalement nouvelle, depuis notamment les travaux de Tukey sur l'analyse exploratoire de données (1977) ou ceux portant sur l'analyse exploratoire de données spatiales (Monmonier, 1989 ; Haining, 1990 ; Anselin, 1994), la masse de données et la relative facilité d'accès sont aujourd'hui telles qu'elles bousculent les pratiques de recherche contemporaines. On peut alors s'interroger sur l'impact de ces masses de données sur la pratique même de la géographie.

L'accès généralisé à ces grandes bases de données s'accompagne notamment de la mise à disposition de nombreux outils permettant de les traiter, y compris en ce qui concerne les corpus « qualitatifs », voire de les visualiser (Antoni *et al*., 2004 ; Cheshire et Batty, 2012). C'est d'ailleurs sans doute là une des perspectives les plus stimulantes liées au développement, tant de ces grandes bases de données que des outils d'analyse : l'apparition de méthodes d'analyse numérique de données qualitatives (textuelles notamment). En démocratisant la manipulation des données – ce qui n'est pas sans risque du point de vue des compétences méthodologiques nécessaires au choix et à l'interprétation des analyses – ainsi qu'en articulant l'analyse numérique de variables de toutes natures (nominales, ordinales, discrètes, continues, textuelles…), la révolution du *big data* est-elle alors de nature à transcender le clivage entre géographie qualitative et quantitative (DeLyser et Sui, 2013) ?

Bibliographie

Anselin, L., (1994). Exploratory spatial data analysis and geographic informations systems, *in* Painho, M., (Ed.), *Proceedings of the Workshop on New Tools for Spatial Analysis*, EUROSTAT, pp. 45-54.

Antoni, JP., Klein, O.,Moisy, S., (2004). Cartographie interactive et multimédia : vers une aide à la réflexion géographique, *Cybergeo: European Journal of Geography*, en ligne.http://cybergeo.revues.org/2621

Banos, A., (2001). À propos de l'analyse spatiale exploratoire des données. *Cybergeo : European Journal of Geography*, en ligne. http://cybergeo.revues.org/4056

Banos, A., (2005). La voie de l'étonnement : favoriser l'abduction dans les Systèmes d'Information Géographique, *in* Fotsing, JM., (dir.), *Apport des SIG à la recherche*, Presses Universitaires d'Orléans, p. 237-254.

Batty, M., (2012). Smart cities, big data. *Environment and planning B: Planning & design*, 39(2): 191-193.

Boyd, D., Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society*, 15(5): 662-679.

Cheshire, J., Batty, M., (2012). Visualisation tools for understanding big data. *Environment and planning B: Planning & design*, 39(3): 413-415.

DeLyser, D.,Sui., D., (2013). Crossing the qualitative- quantitative divide II: Inventive approaches to big data, mobile methods, and rhythmanalysis. *Progress in human geography*, 37(2): 293-305.

Flanagin, A.J., Metzger, M.J., (2008). The credibility of volunteered geographic information, *GeoJournal*, 72(3-4): 137-148.

Goodchild, M.F., (2007). *Citizens as sensors: the world of volunteered geography*, *GeoJournal*, 69(4): 211-221.

Haining, R., (1990). *Spatial data analysis in the social and environmental sciences*, Cambridge University Press, Cambridge, 409 p.

Kitchin, R. (2013). Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography*, 3(3): 262–267.

Marx, V., (2013). The big challenges of big data. *Nature*, 498 (2013) : 255-260.

Monmonier, M., (1989). Geographic Brushing enhancing exploratory analysis of the scatterplot matrix, *Geographical analysis*, 21(1): 81-84.

Terrier, C., (2011)., La valeur des données géographiques, *L'espace Géographique*, 40(2): 103-108.

Tukey, J.W., (1977). *Exploratory data analysis*, Addison-Wesley, Reading, Massachusetts, 688 p.