La géographie et les données massives

Denise Pumain

Chapitre in Bouzeghoub M. et Mosseri R. (dir.) - *Les Big Data à découvert*. Paris, CNRS Editions, 2017, 30-31.

La géographie est depuis longtemps familière des grandes masses de données, sous au moins deux formes qui ont précédé de quelques décennies le « déluge » récent des « big data » (Cukier et Mayer-Schönberger, 2013) : des millions de pixels des images satellite qui ont dû être traités pour réaliser par exemple des cartes d'utilisation du sol, et par ailleurs des « interactions spatiales », échanges de toutes sortes (migrations de personnes, commerce de biens, circulation d'information) dont le nombre élève au carré celui des unités géographiques considérées (soit par exemple près d'un milliard et demi de relations entre les quelque 36 000 communes françaises), qui ont posé maints défis face aux capacités jusqu'à récemment limitées des ordinateurs.

La géographie s'est aussi très vite jetée dans la déferlante des *big data* car beaucoup de ces données apportent des nouveautés intéressantes, du moins dans leur forme : elles sont individuelles, ce qui intéresse toutes les sciences sociales, mais surtout elles sont très souvent *géolocalisées* (le pléonasme de l'expression en dit assez la nouveauté), qu'elles soient issues de capteurs disséminés dans l'environnement, des multiples usages de l'Internet, ou d'objets connectés mobiles. Là où les géographes se heurtaient bien souvent autrefois à la frilosité des opérateurs et des statisticiens, réticents pour des raisons éthiques ou commerciales à communiquer les précieuses coordonnées nécessaires à la cartographie et porteuses du risque d'identification[1], le nombre a permis de balayer ces craintes en les noyant sous des masses plus facilement anonymisables. Désormais, grâce à leurs coordonnées, les informations individuelles peuvent être intégrées dans des systèmes d'information géographique pour être mises en relation avec des structures et des dynamiques spatiales d'entités géographiques à différentes échelles.

Comme l'indique Michael Goodchild (2015) parmi les fameux trois ou quatre « V » qui caractérisent les big data, les nouveautés qui restent un défi pour les géographes ne sont pas tant le volume de ces données, non plus que leur variété, mais leur vélocité et sans doute aussi la véracité ou validité qu'il est possible de leur attribuer. S'agissant de vélocité, des applications « en temps réel » ont été conçues, par exemple pour localiser les foules rassemblées à la faveur de grands évènements en utilisant les traces de téléphonie mobile (Lucchini et Elissalde, 2013), pour analyser les pulsations de la vie urbaine (Fen-Chong, 2012) ou encore pour restituer rapidement des cartographies d'accidents catastrophiques – le séisme de 2010 en Haïti en fut l'un des premiers exemples.

[1] C'est ainsi que les plus petites unités géographiques dans lesquelles les données du recensement français peuvent être communiquées sont des IRIS qui doivent rassembler environ 2000 habitants.

Les quantités d'information sont telles que des méthodes adaptées sont nécessaires et désormais disponibles pour extraire (« moissonner ») ces données, pour les traiter (analyses statistiques multivariées, analyses des réseaux avec la théorie des graphes, analyses des contenus sémantiques etc.), pour en visualiser les grandes structures ou pour faire comprendre les résultats de leurs traitements. Non seulement ces données exigent de nouvelles méthodes informatisées (algorithmes adaptés), mais pour certains auteurs elles conduiraient aussi à transformer les bases épistémologiques du travail scientifique, en apportant un substitut à la démarche classique des enquêtes et des expériences : la masse des informations compenserait en quelque sorte leurs imperfections, et leur finesse de résolution (en général il s'agit d'informations à un échelon « individuel ») permettrait de réviser les fondements de certaines théories construites pour d'autres niveaux d'observation.

Encore faut-il pouvoir se fier à la qualité des données, et se donner la capacité d'intégrer des informations variées, multi-sources. Cette question a fait l'objet de nombreux débats et expériences, par exemple autour de l'*Openstreetmap* (cartographie des routes réalisée par des bénévoles à partir de sources GPS) considérée comme un succès mondial. Cependant, plutôt que de considérer qu'il s'agit d'une « néogéographie », les spécialistes s'accordent désormais pour parler plus modestement de « géographie volontaire » au sujet de l'apport et du partage d'informations géographiques par de multiples et diverses personnes remplaçant leur distribution centralisée par des institutions dûment estampillées. Les outils de visualisation employés par exemple pour figurer l'expansion des réseaux sociaux à la surface du globe ont aussi fait l'objet de débats (Joliveau, 2011 et 2013). En effet, la plupart des documents cartographiques « bruts » représentant les flux échangés sur les réseaux sociaux, ou les émissions de messages sur Twitter, ne montrent pas tellement des images de la mondialisation des échanges, mais illustrent d'abord la localisation des principales masses de la population connectée, celle des pays riches, tout autant que l'expression sur les échanges de la « première loi de la géographie » selon laquelle deux individus proches ont plus de chances d'interagir que deux individus éloignés, si bien que ces cartes n'apportent guère de nouveauté en matière de théorie géographique, ou bien à la marge seulement. Ainsi Carlo Ratti et ses collègues (2010) ont pu montrer que la carte du réseau construit par les échanges téléphoniques en Grande Bretagne identifiait des noyaux de relations les plus fréquentes selon des configurations très proches de la délimitation des régions administratives.

Une exploitation bien plus prometteuse des big data en géographie semble devoir être celle des données de synthèse engendrées par des modèles de simulation, qui pour la première fois permettent de valider des modèles multi-agents employés en sciences sociales comme substitut à l'expérimentation (Schmitt et al., 2015). Au lieu des quelques centaines de simulations effectuées pour calibrer un modèle « à la main », les algorithmes évolutionnaires et le calcul distribué ont permis d'en calculer des centaines de millions et d'explorer dans sa quasi-totalité l'espace de variation des paramètres du modèle. Cela produit un saut qualitatif dans la validation des hypothèses théoriques introduites dans le modèle, dont il est désormais possible d'affirmer le caractère « nécessaire » et « suffisant ». Big data d'observation et de synthèse deviennent ainsi des instruments complémentaires au service de l'avancement de la science géographique.

Références

Cukier K. Mayer-Schönberger V. 2013, *Big Data: A Revolution That Will Transform How We Live, Work, and Think.* Boston, Houghton Mifflin Harcourt.

Fen-Chong J. 2012, *Organisation spatio-temporelle des mobilités révélées par la téléphonie mobile*. Université Paris I, thèse de doctorat.

Goodchild M.F. 2016, GIS in the Era of Big Data, *Cybergeo : European Journal of Geography.* http://cybergeo.revues.org/27647 ; DOI : 10.4000/cybergeo.27647

Joliveau T. 2011 et 2013, 500 millions d'amis, la carte de Facebook. *https://mondegeonumerique.wordpress.com/*

Lucchini F., Elissalde B., 2013, L'Armada des vieux gréements à Rouen. Une ville en scène et des pulsations urbaines observées par la téléphonie mobile. *Etudes Normandes*, 1, 2013, 17-30.

Ratti C, Sobolevsky S, Calabrese F, Andris C, Reades J, Martino M, et al. (2010) Redrawing the Map of Great Britain from a Network of Human Interactions. *PLoS ONE* 5(12): e14248. doi:10.1371/journal.pone.0014248

Schmitt C., Rey-Coyrehourcq S., Reuillon R., Pumain D., 2015, Half a billion simulations, Evolutionary algorithms and distributed computing for calibrating the SimpopLocal geographical model, *Environment and Planning B*, 42, 2,300-315.