Recueil du Symposium 2014 de Statistique Canada
Au-delà des méthodes traditionnelles d'enquêtes : l'adaptation à un monde en évolution

Ce que les mégadonnées peuvent signifier pour les enquêtes

Mick P. Couper[1]

Résumé

Deux tendances convergentes soulèvent des questions concernant l'avenir des enquêtes probabilistes à grande échelle menées par ou pour des instituts nationaux de statistique (INS). Tout d'abord, l'augmentation des coûts et des taux de non-réponse menace potentiellement la rentabilité et la valeur inférentielle des enquêtes. En deuxième lieu, l'intérêt est de plus en plus grand à l'égard des mégadonnées en remplacement des enquêtes. Il existe de nombreux types différents de mégadonnées, mais je mettrai l'accent particulièrement sur les données produites par l'entremise des médias sociaux. Le présent document complète et met à jour un document antérieur sur le sujet (Couper, 2013). Je passerai en revue certaines des préoccupations concernant les mégadonnées, particulièrement dans la perspective des enquêtes. Je soutiens qu'il y a place pour des enquêtes de grande qualité et des analyses de mégadonnées dans les travaux des INS. Même s'il est peu probable que les mégadonnées remplacent les enquêtes de grande qualité, je crois que les deux méthodes peuvent remplir des fonctions complémentaires. Je tente de déterminer certains des critères qui doivent être satisfaits, ainsi que les questions auxquelles on doit répondre, avant que les mégadonnées puissent être utilisées pour des inférences fiables au niveau de la population.

Mots clés : mégadonnées, enquêtes, médias sociaux.

1. Introduction

1.1 Définition des mégadonnées et types de mégadonnées

Le présent document constitue une brève mise à jour et un prolongement d'un document publié précédemment sur le sujet (voir Couper, 2013). Dans ce document, j'ai soulevé un certain nombre de préoccupations concernant les mégadonnées du point de vue d'un chercheur d'enquête, mais j'ai démontré que nous ne pouvons pas laisser de côté le phénomène des mégadonnées et que nous devons tenter de trouver des façons de collaborer avec les experts de ce domaine pour évaluer la qualité et comprendre les forces et les faiblesses des divers types de mégadonnées, comme nous le faisons depuis des décennies pour les enquêtes.

Un des premiers problèmes auxquels nous devons faire face en ce qui a trait aux « Mégadonnées » est qu'il existe de nombreux types différents de mégadonnées (comme il existe en fait de nombreux types différents d'enquêtes). La tendance à utiliser la majuscule pour Mégadonnées (comme je l'ai fait jusqu'à maintenant) laisse supposer dans les faits qu'il s'agit d'un phénomène unitaire. Nous devrions prendre garde aux affirmations générales (positives ou négatives) concernant les mégadonnées, tout comme nous devrions éviter de prétendre que toutes les enquêtes sont aussi bonnes ou aussi mauvaises. Le regroupement d'un nombre aussi important d'approches différentes en une seule catégorie mène de façon inévitable à une généralisation à outrance.

Tout d'abord, qu'entend-on par mégadonnées? Les trois critères les plus courants utilisés pour définir les mégadonnées sont le volume, la rapidité et la variabilité (voir Daas et coll., 2002; TechAmerica Foundation, 2012). Même si un élément qui caractérise les mégadonnées est la taille (ce qui nécessite des ressources différentes pour le calcul), les autres caractéristiques exigent aussi des approches de traitement et d'analyse statistiques différentes.

Il existe de nombreux types différents de mégadonnées, tout comme il existe de nombreux types différents d'enquêtes. Ces types varient du point de vue de la qualité et du contenu de l'information. Les défis inférentiels qui

[1]Survey Research Center, University of Michigan, C.P. 1248, Ann Arbor, Michigan, 48106, États-Unis (mcouper@umich.edu).

se posent sont différents pour les différents types de mégadonnées. Parmi les variétés de mégadonnées figurent les suivantes :

- Données administratives – fournies par des personnes ou des organisations pour l'administration d'un programme (p. ex., dossiers médicaux électroniques, dossiers d'assurance, dossiers bancaires, dossiers fiscaux, registres).
- Données transactionnelles – générées comme sous-produit automatique de transactions et d'activités (p. ex., transactions par carte de crédit, transactions en ligne).
- Données de capteurs (p. ex., imagerie par satellite, capteurs routiers, capteurs climatiques).
- Données de dispositif de repérage (p. ex., GPS, téléphones mobiles).
- Données comportementales (p. ex., recherches en ligne, consultations de pages, témoins).
- Données de médias sociaux – créées par des personnes expressément aux fins du partage avec d'autres (quelques-uns à tout le moins) (p. ex., Facebook, Twitter).

Groves (2011) a utilisé le terme « données organiques » pour faire une distinction par rapport aux « données conçues » qui découlent des enquêtes. Toutefois, cette appellation s'applique davantage à certains types de mégadonnées qu'à d'autres. Par exemple, les mégadonnées administratives sont pour une large part des données conçues, mais souvent à des fins différentes que les données d'enquête. Toutefois, les données administratives sont souvent très structurées et, dans le cas des données de registres (par exemple), peuvent ne pas être affectées par les mêmes biais de sélection que d'autres types de mégadonnées. De même, les données de capteurs, comme les capteurs routiers, sont conçues pour recueillir certains types de renseignements, mais peuvent être utilisées à des fins différentes de celles auxquelles elles étaient destinées au départ. Par ailleurs, les données comportementales générées par les activités en ligne et les données des médias sociaux comportent les caractéristiques de rapidité et de variabilité, mais ne sont pas particulièrement volumineuses par rapport aux mégadonnées des sciences naturelles ou physiques. Par exemple, Twitter génère environ 10 gigaoctets ($1\,000^3$ octets) de données par jour et le Web comprend environ 4 zettaoctets ($1\,000^7$ octets) de données, tandis qu'une petite fraction (0,001 %) des capteurs du Grand collisionneur de hadrons génère environ 25 pétaoctets ($1\,000^5$ octets) de données chaque année. Selon cette norme, les données des médias sociaux pourraient être considérées comme des données très volumineuses, mais pas comme des données exceptionnellement volumineuses, et des échantillons de données de Twitter (p. ex., le tuyau d'arrosage ou l'échantillon de 10 %) pourraient être analysés au moyen d'ordinateurs de bureau puissants.

Le présent document met principalement l'accent sur les deux derniers types de mégadonnées, celles qui peuvent être décrites comme étant générées de façon organique par les activités en ligne des êtres humains, directement et intentionnellement (dans le cas de Twitter) ou indirectement (dans le cas des recherches sur Internet et de l'exploration), et celles que Lampe et ses collègues (2014) appellent des « mégadonnées sociales ». Ce genre de « données à l'état sauvage » présente les plus grands défis inférentiels.

1.2 La « menace » des mégadonnées

De nombreux partisans des mégadonnées prétendent qu'elles remplaceront les enquêtes. Par exemple, Mayer-Schönberger et Cukier (2013, p. 31) écrivent : « L'utilisation d'un échantillon aléatoire à l'ère des mégadonnées correspond à s'agripper à un fouet de cheval à l'ère de l'automobile ». Cette affirmation rappelle des affirmations similaires les premières années de la recherche d'enquête. Par exemple, en 1999, Gordon Black (alors PDG de Harris Interactive) disait : « La recherche sur Internet est une « technologie de remplacement », c'est-à-dire une invention révolutionnaire dont les avantages sont tellement marqués qu'elle finit pratiquement par éliminer les technologies traditionnelles qu'elle remplace, comme l'automobile pour le cheval et la calèche » (cité dans Couper, 2000). Je crois que, tout comme les enquêtes en ligne sont devenues une partie importante de la recherche d'enquête, sans remplacer complètement les autres modes, les mégadonnées trouveront une place similaire dans la production statistique, sans que les enquêtes deviennent redondantes. Même si la recherche d'enquête fait face à de nombreux défis, pour paraphraser Monty Python, « elle n'est pas encore morte ».

Une partie du problème en ce qui a trait à ces affirmations est qu'elles mettent habituellement l'accent sur la taille en excluant d'autres critères comme la qualité ou la représentation. Même si les mégadonnées ont une valeur incontestable, elles diffèrent des données d'enquête. Chaque type de données comporte ses forces et ses faiblesses, et une compréhension approfondie des limites des mégadonnées est essentielle pour pouvoir exploiter pleinement leur potentiel.

2. Limites des mégadonnées

Les défenseurs des mégadonnées, tout comme les défenseurs de toute autre nouvelle méthode, ont tendance à mettre l'accent sur les aspects prometteurs des nouvelles sources de données. Il revient aux chercheurs qui les suivent de procéder à une évaluation soigneuse des avantages et des inconvénients des mégadonnées. Au fur et à mesure que les enquêtes probabilistes ont pris de la maturité, le secteur a développé un intérêt légitime à comprendre les faiblesses de la méthode (pour un document classique, voir Deming, 1944). De même, les aspects innovateurs initiaux des enquêtes en ligne étaient axés sur l'acceptation de la nouvelle méthode, plutôt que sur ses faiblesses. Des recherches subséquentes ont mis l'accent sur les défis inférentiels des panels en ligne et d'autres types d'enquêtes par Internet (p. ex., American Association for Public Opinion Research, 2013; Callegaro et coll., 2014). Il est maintenant temps (selon moi) d'avoir un même regard critique sur les mégadonnées.

Dans ma communication précédente, j'ai déterminé un certain nombre de limites dans les mégadonnées. Il s'agissait notamment des suivantes :

- **Biais de sélection**
- Biais de mesure (autoprésentation)
- Covariables limitées
- **Volatilité ou absence de stabilité**
- Problèmes de protection des renseignements personnels
- Problèmes d'accès
- **Possibilité de méfaits**
- La taille n'est pas tout (plus volumineux ne signifie pas nécessairement mieux)
- **Problème du tiroir de classeur**
- **Corrélation n'est pas synonyme de causalité**

Dans le présent document, je mets l'accent sur un sous-ensemble déterminé de ces limites, c'est-à-dire celles en caractères gras. Je les aborde dans les sections ci-après. Les lecteurs intéressés par les autres pourront consulter Couper (2013).

2.1 Biais de sélection

Le problème de biais de sélection – qu'il soit le résultat de la non-couverture ou de la non-réponse – est bien connu des chercheurs d'enquête. Il s'agit d'une préoccupation majeure pour les mégadonnées sociales, mais ceux qui analysent les données des médias sociaux ne s'en occupent pas beaucoup. Il existe deux éléments importants liés à l'inférence à partir des données des médias sociaux. Tout d'abord, ce ne sont pas tous les gens qui utilisent les médias sociaux. Même si le nombre d'utilisateurs de Facebook à l'échelle mondiale est impressionnant, du point de vue de la proportion de la population d'un pays, l'utilisation de Facebook est loin d'être universelle. De même, on estime que seulement 15 % environ des adultes aux États-Unis utilisaient Twitter en 2013 (Pew Research Center), tandis qu'au Canada, la proportion était estimée à environ 21 % à la fin de 2014 (http://www.itbusiness.ca/news/canada-becoming-one-of-worlds-top-twitter-strongholds-emarketer-finds/49190).

Toutefois, ce n'est pas seulement le nombre (ou la proportion) d'utilisateurs des médias sociaux qui est important. C'est plutôt la façon dont les différents utilisateurs (et particulièrement les « grands » utilisateurs) diffèrent de ceux qui n'utilisent pas les médias sociaux. Dans un certain nombre de documents, on s'est penché sur la façon dont les utilisateurs des médias sociaux diffèrent des non-utilisateurs (voir, p. ex., Blank, 2014; Couper, 2014). Toutefois, la définition même d'« utilisateurs » est épineuse. Par exemple, selon http://twopcharts.com/, environ 44 % des utilisateurs de Twitter créent un compte, mais ne l'utilisent jamais, et seulement 13 % environ des comptes enregistrés ont servi à l'envoi d'un gazouillis au cours des 30 derniers jours. De même, Arnaboldi et coll. (2013) estiment qu'environ 68 % des comptes Twitter sont des comptes de personnes, le reste étant des comptes d'entreprises ou des systèmes automatisés. Un certain nombre d'études ont démontré la nature très asymétrique des activités dans les médias sociaux (p. ex., Bakshy et coll., 2011; Cha et coll., 2010; Lorince et coll., 2014). Cela rend compte de la règle souvent citée de 90-9-1 de Jakob Nielsen en ce qui a trait à l'inégalité de la participation, selon

laquelle : « Dans la plupart des communautés en ligne, 90 % des utilisateurs sont des observateurs qui ne contribuent jamais, 9 %, des utilisateurs qui participent occasionnellement, et 1 %, des utilisateurs qui sont à l'origine de la presque totalité de l'action. » (http://www.nngroup.com/articles/participation-inequality/). Peu de documents (voire pas du tout) qui utilisent les données des médias sociaux ont tenu compte de la nature hautement asymétrique des données dans leurs analyses, traitant habituellement chaque gazouillis ou affichage comme une observation indépendante avec un poids égal aux autres.

2.2 Volatilité ou absence de stabilité

Une caractéristique clé des statistiques nationales produites par les INS est la cohérence de la mesure au fil du temps, ce qui permet l'analyse de tendances temporelles concernant des indicateurs clés de la société. Par contre, une des caractéristiques qui définit les mégadonnées est la volatilité, et cela est particulièrement vrai dans le cas des données des médias sociaux.

Même si cela peut rendre les mégadonnées sociales utiles pour les tendances à court terme ou la compréhension d'un événement particulier (comme des élections), cela peut ne pas convenir pour la mesure des tendances temporelles qui durent des années ou même des décennies. Les médias sociaux qui sont populaires aujourd'hui pourraient ne plus exister (à tout le moins dans la même forme) demain. Google a été fondé en 1998, ce qui en fait un adolescent. Le premier gazouillis a été envoyé en 2006, et Twitter a connu une expansion de 5 000 % en l'espace de cinq ans. Par ailleurs, d'autres sites de réseautage social ont perdu du terrain ou ont cessé leurs activités (p. ex., MySpace, Second Life, Hyves aux Pays-Bas), et d'autres font tout ce qu'ils peuvent pour obtenir de l'attention (p. ex., Google+). Compte tenu du succès des premiers sites, nous assistons à une recrudescence des options de rechange (Instagram, Pinterest, Reddit, Snapchat, Tumblr, Yammer, etc.). La popularité de ces sites, et le genre de personnes qui les utilisent, sont susceptibles de varier au fil du temps, ce qui présente des défis pour les tendances temporelles à plus long terme. Même Google (utilisé pour Google Trends) peut servir un auditoire différent que (disons) Bing ou Yahoo, ou d'autres moteurs de recherche. Il serait utile de savoir si les résultats de Google Trends (par exemple) se répètent lorsque l'on utilise un moteur de recherche différent.

Même si le caractère immédiat des mégadonnées représente une caractéristique attrayante pour comprendre la situation présente, les médias sociaux ne sont peut-être pas une plateforme appropriée pour élaborer des indicateurs statistiques à long terme.

2.3 Possibilité de méfaits

Parmi les défis des utilisations secondaires des données des médias sociaux figure le fait que l'analyste (contrairement au concepteur d'enquête) a très peu de contrôle sur la façon dont le support est utilisé. Les tendances temporelles concernant les utilisateurs des sites et les fins de l'utilisation (examinées ci-dessous) sont une source de volatilité ou d'incertitude. Un autre défi toutefois a trait à la manipulation possible des données des médias sociaux.

Les médias sociaux sont utilisés à des fins particulières. Une part importante d'entre eux tourne autour de la culture populaire, et la popularité ou l'influence représente parfois un objectif en soi. Cela fait augmenter le risque de manipulation. Par exemple, en février 2014, on a estimé qu'entre 5,5 % et 11,2 % de tous les comptes sur Facebook étaient des doubles, des comptes malveillants ou d'autres types de « faux » comptes (voir https://nakedsecurity.sophos.com/2014/02/10/). Même si cela s'apparente aux problèmes de bourrage des urnes des sondages téléphoniques dans lesquels ce sont les sondés qui appellent, il existe plusieurs facteurs qui augmentent la probabilité de méfaits dans les médias sociaux par rapport aux autres médias :

- le relatif anonymat d'Internet facilite ce comportement;
- il est pour ainsi dire gratuit de s'enregistrer et d'afficher dans les médias sociaux;
- des systèmes automatisés (agents numériques) peuvent être élaborés pour générer du contenu (p. ex., désinformation populaire planifiée);
- des gains personnels (notoriété, attention, gratification financière, etc.) sont possibles par suite de l'adoption d'un tel comportement.

Les sondages en ligne partagent les trois premières caractéristiques, tandis que la dernière est propre aux médias sociaux.

Dans leur examen de Google Flu Trends, Lazer et coll. (2014) font une distinction entre la dynamique de l'équipe des bleus et de l'équipe des rouges, empruntant à la terminologie militaire. La dynamique de l'équipe des bleus se manifeste lorsque l'algorithme produisant les données (et, par conséquent, l'utilisation par les utilisateurs) a été modifié par le fournisseur de services en conformité avec son modèle d'affaires. Les fournisseurs de services changent constamment leurs modèles d'affaires et révisent ou remanient leurs produits, et cela peut avoir des répercussions pour ceux qui tentent d'utiliser les données à d'autres fins.

La dynamique de l'équipe des rouges se produit lorsque les sujets de recherche (p. ex., les utilisateurs des médias sociaux) tentent de manipuler le processus de production de données pour atteindre leurs propres objectifs, comme des gains économiques ou politiques. Comme le notent Lazer et coll. (2014) : « Ironiquement, plus nous réussissons à contrôler le comportement des personnes qui utilisent ces sources ouvertes d'information, plus il devient tentant de manipuler ces signaux ». Même s'il existe des exemples d'influence sur les sondages en ligne en libre accès (p. ex., Ubermotive Guide to Media Influence, à http://www.ubermotive.com/?p=68), il est plus difficile de trouver des preuves de la manipulation des médias sociaux. Toutefois, en 2011, *The Guardian* révélait que le gouvernement américain manipulait les médias sociaux (voir http://www.guardian.co.uk/technology/2011/mar/17/us-spy-operation-social-networks). Dans la culture populaire, des faux comptes sont utilisés pour promouvoir des célébrités et des causes. Par exemple, on estime qu'environ la moitié des 37 millions des abonnées au compte Twitter de Justin Bieber sont fictifs ou inactifs (voir http://www.digitalspy.com/music/news/a471915/justin-bieber-twitter-followers-50-percent-are-fake-says-report.html). Cela est pertinent parce que nombre des principaux comptes Twitter sont ceux de célébrités ou de compagnies médiatiques, ce qui laisse supposer que la culture populaire est au centre des médias sociaux. Parmi les 10 principaux comptes Twitter en juin 2014, sept (Katy Perry, Justin Bieber, Taylor Swift, Lady Gaga, Britney Spears, Rihanna et Justin Timberlake) étaient des comptes de musiciens, tandis que deux (YouTube et Instagram) étaient des comptes de compagnies; le compte Twitter du président Obama s'est classé au troisième rang du point de vue du nombre d'abonnés (voir http://twopcharts.com/twoplist?source=atlan). Le propre compte de Twitter se classe au 12e rang en ce qui a trait au nombre d'abonnés. Un article du *New York Times* (voir http://bits.blogs.nytimes.com/2014/04/20/friends-and-influence-for-sale-online/?_r=0) décrit comment on peut acheter des amis et de l'influence en ligne.

La facilité de manipulation des médias sociaux soulève une question concernant l'utilisation des données qu'ils comprennent pour des estimations gouvernementales officielles, particulièrement lorsqu'il devient connu que ces données sont utilisées de cette façon. Encore une fois, il ne s'agit pas d'invalider ces efforts, mais plutôt de suggérer un peu de prudence pour séparer le bon grain de l'ivraie, ou le signal du bruit, lorsque l'on analyse de telles sources de données organiques.

2.4 Problème du tiroir classeur

Une part importante des premières analyses de mégadonnées, particulièrement celles axées sur les médias sociaux, a mis l'accent sur les « prédictions » réussies. Les échecs de telles approches sont susceptibles d'obtenir moins d'attention, ce qui nous donne un aperçu potentiellement déformé de la valeur des nouvelles méthodes et sources de données. Cela est souvent vrai dans le cas des nouveaux progrès de la recherche, les défenseurs de la nouvelle approche tentant de susciter un intérêt à l'égard de leurs méthodes. Deux exemples du problème dans le contexte des mégadonnées sociales suffiront.

Le premier a trait à Google Flu Trends. Le premier document montrant que les recherches sur Google concernant les symptômes de l'influenza suivaient de près les données des Centers for Disease Control and Prevention (Ginsberg et coll., 2008) a reçu beaucoup d'attention, tant dans les ouvrages scientifiques que dans la presse populaire. Toutefois, des documents ultérieurs (voir, p. ex., Cook et coll., 2011; Butler, 2013; Dugas et coll., 2013; Lazer et coll., 2014) ont laissé supposer que les prédictions n'étaient pas aussi réussies les années subséquentes. En outre, les premiers résultats ont été obtenus grâce à l'adaptation des données aux tendances existantes, l'algorithme initial trouvant les meilleurs appariements parmi 50 millions de termes de recherche pour correspondre à 1 152 points de données. L'algorithme a été modifié plusieurs fois par la suite pour correspondre aux données plus récentes. Des résultats similaires ressortent pour les tendances de la grippe sur Twitter, les premiers résultats montrant une corrélation de 0,96 avec les données des Centers for Disease Control and Prevention (Paul et Drezde, 2011), mais ne se confirmant pas dans les analyses subséquentes (p. ex., Murphy, 2013).

Le deuxième exemple vient des élections de 2009 en Allemagne. À partir des données de Twitter, Tumasjan et coll. (2011) ont fait état d'une erreur absolue moyenne de 1,65 quant à la répartition des votes entre les partis politiques et ont prétendu que « le nombre même de gazouillis mentionnant un parti politique peut être considéré comme un reflet plausible de la part de vote, et que son pouvoir de prédiction se rapproche même de celui des sondages électoraux traditionnels » (Tumasjan et coll., 2011, p. 15). Toutefois, une réplique des analyses par Jungherr et ses collègues (2012) incluait aussi le Parti Pirate (un parti ayant une grande présence en ligne, mais obtenant une petite part du vote) et concluait qu'il aurait été prédit comme gagnant des élections. Par ailleurs, le fait de faire varier la période choisie par Tumasjan et coll. (2011) pour les prédictions a donné lieu à des conclusions différentes.

Les types d'analyses secondaires et de répliques décrits ci-dessus sont essentiels pour une évaluation soigneuse de la valeur des données des médias sociaux. Un prérequis de ces analyses est que les données originales et les algorithmes soient mis à la disposition des autres. S'ils demeurent secrets pour des motifs liés aux droits de propriété exclusifs ou d'autres raisons, la capacité de reproduire est gravement compromise. Les domaines de la méthodologie d'enquête et de l'analyse des mégadonnées doivent être ouverts à la publication de résultats nuls ou d'échecs de répliques. Si seuls les résultats « réussis » sont rendus publics, nous obtiendrons un aperçu déformé de la valeur des mégadonnées sociales. Le problème du tiroir classeur (qui a été mentionné pour la première fois par Rosenthal, 1979) n'est pas propre à l'analyse des mégadonnées. Pour un exemple récent, voir Franco, Malhotra et Simonovits (2014).

2.5 Corrélation n'est pas synonyme de causalité

Un dernier défi des mégadonnées dont il est question ici est lié au problème du tiroir classeur. Du fait de la taille des ensembles de données analysés, presque toute relation peut atteindre la signification statistique. Un exemple amusant se trouve dans Leinweber (2007) qui « a prédit » l'indice du marché boursier Standard and Poor 500 aux États-Unis avec un R^2 de 0,99, en utilisant seulement cinq variables : la production de beurre au Bangladesh et aux États-Unis, la production de fromage aux États-Unis, et la population de moutons au Bangladesh et aux États-Unis. Des exemples similaires de fausses corrélations se trouvent dans http://www.tylervigen.com/. Même si elles n'utilisent pas les mégadonnées, du fait du nombre important d'observations servant à l'analyse des mégadonnées et des techniques d'exploration des données utilisées, des corrélations importantes sont faciles à établir. Nous devons trouver des façons de tenir compte de la probabilité de faux positifs, et nous devons nous rappeler de séparer la signification statistique de la signification de fond. Dans presque toutes les analyses, les mégadonnées sont surreprésentées par rapport aux enquêtes. Les analyses de réplique et de sensibilité peuvent contribuer à nous protéger de ce que Lazer et ses collègues (2014) appellent la présomption des mégadonnées.

3. Discussion et conclusion

Cette brève revue de certains des problèmes que pose l'analyse des mégadonnées sociales ne se veut pas une critique globale des mégadonnées. Avant de trop se plaindre concernant les mégadonnées, nous devons nous rappeler que la qualité des enquêtes varie aussi, et qu'il existe de nombreuses mauvaises enquêtes. Mon intention est plutôt de rappeler que l'analyse et l'utilisation des mégadonnées sociales, particulièrement en remplacement ou en complément des données d'enquête, n'en sont encore qu'à leurs débuts, et que nous avons encore beaucoup à apprendre. Le domaine de la recherche d'enquête a une longue tradition d'autocritique et met l'accent sur les diverses sources d'erreur qui peuvent nous faire nous questionner sur les inférences tirées d'une enquête. La notion d'incertitude statistique, et donc une partie importante du paradigme de l'échantillonnage probabiliste, demeure un concept important, même dans le monde de N = Tout, pour utiliser le terme de Mayer-Schönberger et Cukier (2013). Même si le paradigme fréquentiste ne s'applique pas aux mégadonnées sociales, ni à une part importante de la recherche d'enquête, nous devons élaborer et appliquer des indicateurs de l'incertitude à partir d'approches bayésiennes. Comme Hal Varian de Google (2014, p. 24) le souligne : « Dans cette ère de « mégadonnées », il semble étrange de mettre l'accent sur l'incertitude de l'échantillonnage, qui a tendance à être faible dans les grands ensembles de données, tout en laissant de côté entièrement l'incertitude de modèle, qui peut être assez importante ».

Les mégadonnées, et toutes leurs formes, offrent un potentiel énorme. Même si les partisans des mégadonnées prétendent souvent qu'elles évoluent et qu'elles rendront les méthodes existantes (particulièrement les enquêtes) obsolètes, je crois que cela est davantage vrai dans certains domaines que dans d'autres. Même si les mégadonnées sont peu susceptibles de rendre obsolète les méthodes d'enquête, cela ne signifie pas que nous pouvons faire preuve de suffisance. L'exploration des mégadonnées offre des possibilités excitantes, et je suis ravi de l'attention que leur accordent les chercheurs. Il se peut que certains des cadres et des méthodes qui sont bien connus des chercheurs d'enquête (par exemple, le cadre de l'erreur totale d'enquête) s'appliquent bien aux mégadonnées, mais que d'autres méthodes doivent être élaborées davantage. Je crois que la recherche d'enquête a beaucoup à offrir au monde de l'analyse des mégadonnées, et plutôt que de percevoir ces progrès comme une source de concurrence, nous devrions trouver des façons de collaborer, afin d'améliorer notre compréhension des avantages et des limites possibles des mégadonnées et de la recherche d'enquête. Nous pouvons aussi en apprendre beaucoup des mégadonnées, particulièrement en ce qui a trait aux défis de l'inférence à partir de données qui ne sont peut-être pas pleinement représentatives, et sur lesquelles nous avons peu de contrôle.

Bibliographie

American Association for Public Opinion Research. 2013. *Report of the AAPOR Task Force on Non-Probability Sampling*, Deerfield, IL: American Association for Public Opinion Research.

ARNABOLDI, Valerio, Marco CONTI, Andrea PASSARELLA et Robin DUNBAR. 2013. « Dynamics of Personal Social Relationships in Online Social Networks: a Study on Twitter », dans *COSN '13 Proceedings of the First ACM Conference on Online Social Networks*, p. 15 à 26, doi: 10.1145/2512938.2512949.

BAKSHY, Eytan, Jake M. HOFMAN, Winter A. MASON et Duncan J. WATTS. 2010. « Everyone's an Influencer: Quantifying Influence on Twitter », *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, p. 65 à 74.

BLANK, G. 2014. « Who Uses Twitter? Representativeness of Twitter Users », document présenté à la General Online Research Conference, Cologne, mars.

BUTLER, Declan. 2013. « When Google Got Flu Wrong: US Outbreak Foxes a Leading Web-based Method for Tracking Seasonal Flu », *Nature*, 13 février 2013, http://www.nature.com/news/when-google-got-flu-wrong-1.12413.

CALLEGARO, Mario, Reginald P. BAKER, Jelke BETHLEHEM, Anja S. GÖRITZ, Jon A. KROSNICK et Paul J. LAVRAKAS (sous la dir.). 2014. *Online Panel Research: A Data Quality Perspective*, New York: Wiley.

CHA, Meeyoung, Hamed HADDIDI, Fabricio BENEVENUTO et Krishna P. GUMMADI. 2010. « Measuring User Influence in Twitter: The Million Follower Fallacy », *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*.

COOK, S., C. CONRAD, A.L. FOWLKES et M.H. MOHEBBI. 2011. « Assessing Google Flu Trends Performance in the United States during the 2009 Influenza Virus A (H1N1) Pandemic », *PLoS ONE*, vol. 6, nº 8 : e23610, doi:10.1371/journal.pone.0023610.

COUPER, Mick P. 2000. « Web Surveys: A Review of Issues and Approaches », *Public Opinion Quarterly*, vol. 64, nº 4, p. 464 à 494.

COUPER, Mick P. 2013. « Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys », *Survey Research Methods*, vol. 7, nº 3, p. 145 à 156.

COUPER, Mick P. 2014. « Social Media and Surveys: Collaboration, Not Competition », document présenté à la General Online Research Conference, Cologne, mars.

DAAS, Piet J.H., Marko ROOS, Mark VAN DE VEN et Joyce NERONI. 2012. « Twitter as a Potential Data Source for Statistics », La Haye/Heerlen, Pays-Bas : Statistics Netherlands, document de travail 201221.

DEMING, W. Edward. 1944. « On Errors in Surveys », *American Sociological Review*, vol. 9, nº 4, p. 359 à 369.

DUGAS, Andrea F., Mehdi JALALPOUR, Yulia GEL, Scott LEVIN, Fred TORCASO et coll. 2013. « Influenza Forecasting with Google Flu Trends », *PLoS ONE*, vol. 8, nº 2 : e56176, doi:10.1371/journal.pone.0056176.

FRANCO, Annie, Neil MALHOTRA et Gabor SIMONOVITS. 2014. « Publication Bias in the Social Sciences: Unlocking the File Drawer », *Science*, publié en ligne le 28 août 2014, DOI:10.1126/science.1255484.

GINSBERG, Jeremy, Matthew H. MOHEBBI, Rajan S. PATEL, Lynnette BRAMMER, Mark S. SMOLINSKI et Larry BRILLIANT. 2009. « Detecting Influenza Epidemics Using Search Engine Query Data », *Nature*, vol. 457, nº 7232, p. 1012 à 1014.

GROVES, Robert M. 2011. « Three Eras of Survey Research », *Public Opinion Quarterly*, vol. 75, nº 5, p. 861 à 871.

JUNGHERR, Andreas, Pascal JÜRGENS et Harald SCHOEN. 2012. « Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T.O., Sander, P. G., & Welpe, I.M. 'Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment », *Social Science Computer Review*, vol. 30, nº 2, p. 229 à 234.

LAMPE, Cliff, Josh PASEK, Lauren GUGGENHEIM, Fred CONRAD et Michael SCHOBER. 2014. « When Are Big Data Methods Trustworthy for Social Measurement? », document présenté à la réunion annuelle de l'American Association for Public Opinion Research, Anaheim, Californie, mai.

LAZER, David, Ryan KENNEDY, Gary KING et Alessando VESPIGNANI. 2014. « The Parable of Google Flu: Traps in Big Data Analysis », *Science*, vol. 343, nº 6176 (mars 2014), p. 1203 à 1205.

LEINWEBER, David J. 2007. « Stupid Data Miner Tricks: Overfitting the S&P 500 », *The Journal of Investing*, vol. 16, nº 1, p. 15 à 22.

LORINCE, Jared, Sam ZOROWITZ, Jaimie MURDOCK et Peter M. TODD. 2014. « 'Supertagger' Behavior in Building Folksonomies », *Proceedings of the 2014 ACM conference on Web science (WebSci '14)*, ACM, New York, New York, États-Unis, p. 129 à 138, DOI: 10.1145/2615569.2615686.

MAYER-SCHÖNBERGER, Viktor et Kenneth CUKIER. 2013. *Big Data; a Revolution That Will Transform How We Live, Work, and Think*, New York: Houghton Mifflin.

MURPHY, Joe. 2013. « 10 Things Every Survey Researcher Should Know about Twitter », document présenté au FedCASIC workshop, Washington, DC, mars.

PAUL, Michael J. et Mark DREDZE. 2011. « You Are What You Tweet: Analyzing Twitter for Public Health », *Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, July 17-21*, Palo Alto, Californie, AAAI Publications, p. 265 à 272.

ROSENTHAL, Robert. 1979. « The File Drawer Problem and Tolerance for Null Results », *Psychological Bulletin*, vol. 86, nº 3, p. 638 à 641.

TechAmerica Foundation. 2012. *Demystifying Big Data: A Practical Guide to Transforming the Business of Government*, Washington, DC: TechAmerica Foundation.

TUMASJAN, Andranik, Timm O. SPRENGER, Philipp G. SANDNER et Isabell M. WELPE. 2011. « Election Forecasts With Twitter: How 140 Characters Reflect the Political Landscape », *Social Science Computer Review*, vol. 29, nº 4, p. 402 à 418.

VARIAN, Hal R. 2014. « Big Data: New Tricks for Econometrics », *Journal of Economic Perspectives*, vol. 28, nº 2, p. 3 à 28.