Fiabilité des Big Data du point de vue de la statistique publique

Bertrand Loison, vice-directeur de l'Office fédéral de la statistique
Diego Kuonen, fondateur et directeur général de Statoo Consulting

L'avènement des Big Data modifie le contexte dans lequel les organisations produisant des statistiques officielles opèrent. Bien que les Big Data offrent des opportunités, il n'en demeure pas moins vrai qu'un certain nombre de défis importants doivent encore être relevés afin d'en optimiser leur utilisation dans le contexte de la statistique publique.

L'ère des Big Data devrait avoir un impact important sur les organisations pour lesquelles la production et l'analyse de données et d'informations constitue le cœur de métier. Les instituts nationaux de statistique (INS) n'y font pas exception. Ils sont responsables de la production de la statistique publique qui est largement utilisée par les décideurs politiques et d'autres acteurs importants de la société. On peut raisonnablement poser comme postulat que la façon dont les INS adopteront ou pas les Big Data aura des implications pour l'ensemble de la société.

Les statistiques officielles sont souvent considérées comme allant de soi. Cependant, là où la confiance fait défaut, la société manque d'un pilier important pour une discussion pragmatique et l'élaboration de politiques publiques fondées sur des données probantes. Les normes et standards professionnels jouent un rôle vital pour assurer la confiance envers les statistiques officielles. La statistique publique dispose de ses propres codes de déontologie[1], [2], [3]. La prise en compte des Big Data dans la production de la statistique publique devra se faire dans le respect de la déontologie scientifique.

La confiance engendrée par le respect de ces codes de déontologie offre une position privilégiée aux INS en matière d'acquisition de données. De nombreux INS à travers le monde, dont l'Office fédéral de la statistique (OFS) en Suisse, ont déjà accès, conformément à la loi, aux sources de données gouvernementales. Certains pays ont légiféré ou sont en train de le faire pour permettre aux producteurs de statistiques officielles de pouvoir accéder gratuitement aux données de tierces parties (entreprises, ...). De plus, à des fins statistiques, de nombreux INS sont autorisés à appareiller des données provenant des différentes sources. La Suisse n'y fait pas exception[4].

Un nouvel écosystème

L'émergence de nouvelles sources de données crée pour les INS un bénéfice potentiel, mais cela rend aussi leurs produits moins uniques, puisque d'autres acteurs du marché de l'information ont commencé à produire des statistiques.

Le potentiel pour de nouvelles statistiques officielles est cependant bien réel. Par exemple, les données de localisation des téléphones mobiles pourraient être utilisées pour des statistiques quasi instantanées sur la population diurne et le tourisme. Les messages issus des médias sociaux pourraient être utilisés pour plusieurs types d'indicateurs, comme par exemple un indicateur précoce de la consommation. L'inflation pourrait être estimée à partir de l'information sur les prix disponible sur le web, et ainsi de suite. Toutefois, pour saisir ces opportunités, un certain nombre de défis doivent être surmontés.

Défis

Le principal défi auxquels les statisticiens officiels sont confrontés dans leur utilisation des Big Data est celui de la véracité des données qui représente le fondement de la confiance dans les données. Elle comprend la fiabilité, la solidité et la validité des données, leur qualité, ainsi que la transparence des processus de production des données. Un autre défi de taille concerne la méthodologie. De nombreuses sources de type Big Data, comme par exemple les messages issus des médias sociaux, sont composés de données d'observation et ne sont pas délibérément conçus pour l'analyse des données, et n'ont donc pas de population cible, ni de structure et ni de qualité. C'est pourquoi il est difficile d'appliquer les méthodes statistiques traditionnelles basées sur la théorie de l'échantillonnage.

Pour les INS, la question est donc de savoir comment la qualité des statistiques officielles peut être garantie si elles sont tout ou partiellement produites à partir de Big Data. L'utilisation des Big Data va induire un changement de paradigme et une utilisation accrue des méthodes d'analyse complémentaires (p. ex. l'analyse prédictive par des techniques statistiques avancées, la science des données et/ou l'apprentissage automatique).

La protection de la vie privée et les questions juridiques constituent d'autres défis, de même que les droits d'auteur et de propriété des données.

Pour les INS, il est essentiel de répondre à ces préoccupations par le biais de pratiques telles que la transparence quant à l'utilisation des sources de données et à la manière dont elles sont utilisées. Il en va de la crédibilité de la statistique publique.

L'avenir de la statistique publique

En cette période d'abondance croissante de données, la production d'informations statistiques potentiellement pertinentes pour la société n'est plus une activité intrinsèquement limitée aux INS.

Etant donné la concurrence croissante que les données générées par d'autres sources représentent vis-à-vis des INS en tant que porteurs des statistiques officielles, une réévaluation du positionnement stratégique de ceux-ci est nécessaire.

L'avenir des statistiques officielles à l'ère des Big Data fait encore l'objet de discussions. Le fait que la communauté internationale de la statistique publique doive s'adapter à une nouvelle réalité et répondre aux opportunités et aux défis auxquels elle est confrontée ne fait, lui, aucun doute.

L'OFS qui est membre du Global Working Group on Big Data for Official Statistics depuis 2017 a identifié ces défis et y a apporté une première réponse en publiant, en novembre 2017, sa stratégie sur l'innovation des données[5].

Les auteurs

Bertrand Loison

Le Prof. Dr Bertrand Loison, MPA IDHEAP, est vice-directeur de l'Office fédéral de la statistique (OFS) et chef de la division des registres, membre nommé du Comité de planification de la Cyberadministration Suisse et représentant de la Suisse au sein du «UN Global Working Group on Big Data for Official Statistics (ONU)». Il est également responsable du groupe de travail «New Data Sources» en charge de l'implémentation au sein de l'OFS de la stratégie d'innovation sur les données. Ses travaux se focalisent sur les changements induits par les nouvelles sources de données sur les offices nationaux de statistiques. Il est également Professeur en systèmes d'information au sein de la Haute école de gestion Arc (HES-SO).

Diego Kuonen

Le Prof. Dr Diego Kuonen, PhD en statistiques, statisticien accrédité (CStat et PStat) et scientifique accrédité (CSci), est fondateur et directeur général de Statoo Consulting (www.statoo.ch). Le Prof. Dr Diego Kuonen, CStat PStat CSci, intervient depuis de nombreuses années auprès de grands groupes industriels et de services en Europe. Depuis 2016, il est également Professeur en Data Science au sein de la Faculté d'économie et de management (GSEM) de l'Université de Genève, et fondateur et directeur de son nouveau programme de Master en Business Analytics. Actuellement, il est également le principal conseiller stratégique et scientifique externe de la Direction et du conseil de Direction de l'OFS dans le domaine d'expertise Big Data Analytics.

Notes

[1] United Nations (A/RES/68/261 from 29 January 2014) Fundamental principles of official statistics. Disponible à l'adresse: https://unstats.un.org/unsd/dnss/gp/FP-Rev2013-F.pdf (consulté le 25 mai 2018).

[2] La Suisse est membre du SSE depuis la signature le 26.10.2004 de l'accord bilatéral Suisse – Union européenne sur la statistique. Disponible à l'adresse: https://www.eda.admin.ch/dea/en/home/bilaterale-abkommen/ueberblick/bilaterale-abkommen-2/statistik.html (consulté le 25 mai 2018).

[3] Le code de bonnes pratiques des statistiques européennes est également valable en Suisse. Disponible à l'adresse: https://www.bfs.admin.ch/bfs/fr/home/ofs/engagement-qualite.html (consulté le 12 mai 2018).

[4] L'appariement de données à des fins statistiques est réglé à l'art. 14a de la loi sur la statistique fédérale (LSF; RS 431.01). Disponible à l'adresse: https://www.bfs.admin.ch/bfs/fr/home/services/appariement-donnees/generalites.html (consulté le 26 mai 2018).

[5] Stratégie d'innovation sur les données. Disponible à l'adresse: https://www.bfs.admin.ch/bfs/fr/home/actualites/quoi-de-neuf.gnpdetail.2017-0673.html