POURQUOI DES DONNÉES PLUS NOMBREUSES ET DE MEILLEURE QUALITÉ SONT LA CLÉ D'UN CANADA FONDÉ SUR DES DONNÉES PROBANTES

JAN KESTLE

Jan Kestle est présidente et fondatrice d'Environics Analytics, un fournisseur de premier plan de données et de connaissances aux entreprises, aux gouvernements et aux organismes sans but lucratif du Canada. Elle travaille dans le milieu de la statistique au Canada depuis 50 ans, est membre du Conseil national de la statistique et membre du conseil d'administration de l'Association canadienne du marketing.

Lorsque le recensement détaillé obligatoire a été annulé et remplacé par l'enquête à participation volontaire en 2011, il a été surprenant d'entendre des chercheurs expérimentés déclarer que nous avions tellement de données volumineuses que nous n'avions pas besoin d'un recensement détaillé. La vérité est que, même si bon nombre des nouvelles sources peuvent augmenter notre production de données, des statistiques officielles fiables sont nécessaires pour créer une économie fondée sur des preuves ainsi qu'un cadre efficace de politique sociale.

Avec le rétablissement du questionnaire détaillé du recensement, la poussière est retombée — du moins pour le moment. Tous sont d'accord que les résultats de l'Enquête nationale auprès des ménages (ENM), qui a temporairement remplacé le questionnaire détaillé du recensement, étaient discutables au mieux. En tant que compagnie privée qui crée des données régionales pour les entreprises, les gouvernements et les organisations à but non lucratif, nous examinons les prévisions de nos modèles pour voir comment elles se comparent à celles mesurées par la suite. La plupart des résultats provenant de l'ENM n'étaient pas utilisables ; en ce qui concerne les données de qualité suffisante, il existait toujours des « ruptures » évidentes dans la série, révélées lors d'une revue historique, malgré l'application des meilleures pratiques.

Espérons que nos politiciens ont tiré les leçons de cette décision mal conçue, mais pour que cela ne se reproduise plus, la communauté statistique doit convaincre les électeurs du fait que le recensement est la pierre angulaire des données canadiennes. Alors que Statistique Canada entre dans son deuxième siècle d'existence, deux choses sont essentielles. Premièrement, nous devons protéger et développer le recensement, les comptes écono-

miques, le programme de données administratives sur les petites régions et les programmes d'enquêtes clés. Deuxièmement, Statistique Canada doit développer de nouvelles méthodes de recherche et de production de données. Les efforts de modernisation en cours, notamment la coopération entre le secteur public et le secteur privé, l'extraction de données volumineuses et l'utilisation accrue des données administratives, sont cruciaux pour le futur de Statistique Canada. Mais ces initiatives ne réussiront que si de nouvelles approches sont combinées avec un fort soutien pour les programmes fondamentaux.

Un bon exemple des défis et des opportunités que Statistique Canada peut s'attendre à rencontrer alors qu'ils testaient et utilisaient de nouvelles approches, a récemment fait l'objet d'une autre chronique. Statistique Canada allait mener un projet pilote en consultant des données détaillées sur les dépenses à partir d'enregistrements bancaires personnels pour tester l'utilisation de ces enregistrements administratifs en tant que source possible de remplacement d'Enquête sur les dépenses des ménages. Pourquoi ? Parce qu'il est de plus en plus difficile d'amener les Canadiens à remplir des sondages approfondis, comme celui-ci, qui est si crucial pour l'économie.

Le plan consistait à obtenir, en vertu de la Loi sur la statistique, de l'information sur les dépenses des gens. A partir d'un grand fichier, un échantillon serait mis au point, puis les données seraient agrégées et utilisées pour produire des statistiques officielles. Un test clé dans la conception du projet pilote consistait à déterminer si l'échantillon aléatoire de dossiers bancaires serait suffisamment représentatif pour reproduire les résultats du sondage ou si les données seraient de meilleure qualité si elles étaient pondérées en fonction des données démographiques qui sont habituellement utilisées dans la conception du sondage.

Statistique Canada prévoyait acquérir les données sur les dépenses ainsi que les identificateurs personnels afin d'ajouter certains attributs du recensement aux registres bancaires. Ces identifiants seraient supprimés après avoir été utilisés pour faire correspondre les enregistrements. Les enregistrements anonymes seraient traités pour produire les statistiques. Cette approche, utilisant les données de recensement annexées, serait comparée à l'utilisation des mêmes données de dépenses bancaires sans ajouter les attributs du recensement pour le plan de sondage et la pondération. Statistique Canada avait discuté de l'approche avec l'Association des banquiers canadiens (ABC) et le Commissariat à la protection de la vie privée du Canada (CPVP). Le fait que les données seraient transférées sous la protection de la Loi sur la statistique aurait assuré qu'aucune information personnelle ne serait divulguée ou stockée.

Cela semblait être un plan raisonnable pour obtenir des données de bonne qualité sans un fardeau important pour les répondants tout en assurant la protection totale de la confidentialité des données utilisées. Mais cela ne s'est pas bien passé. Les médias et le partis d'opposition ont vivement désapprouvé cet « espionnage » des Canadiens sans leur permission et le projet pilote a finalement été mis en veilleuse.

Il convient d'ajouter que Statistique Canada est bien équipé pour traiter de gros volumes de données avec toutes les dispositions de confidentialité et de sécurité qui s'y rapportent. Toutefois, que vous travailliez avec un seul enregistrement ou des millions

de fichiers, l'important est de savoir comment ces données ont été collectées. Nous pensons que les analystes peuvent tirer des conclusions géographiquement pertinentes sans collecter d'informations personnelles.

QU'AVONS-NOUS APPRIS ?

Pour que les citoyens partagent ce type de données sensibles avec leur gouvernement, ils doivent lui faire confiance, ce qui ne semble pas être le cas. Un rapport critique a évoqué la possibilité qu'autoriser le gouvernement à consulter les registres bancaires permettrait d'identifier les personnes qui fraudent leurs impôts. Bien sûr, cela ne serait pas possible, car aucune donnée ne serait autorisée à quitter Statistique Canada (c'est-à-dire qu'elle ne pourrait pas être légalement partagée avec un autre ministère), mais la préoccupation semble en outre un peu ironique.

Il est clair que la communauté des données ne fait pas assez bien son travail d'aider les Canadiens à comprendre pourquoi de bonnes données sont si importantes dans leur vie — pour la sécurité communautaire, la compensation et les politiques reliés à l'emploi, ainsi que pour la recherche, l'innovation, la santé, l'éducation et la planification de services sociaux.

De plus, le fait que Statistique Canada soit depuis longtemps considéré comme l'un des plus importants organismes statistiques au monde et que l'on puisse faire confiance à des informations confidentielles ne semble pas être bien connu ni accepté. Lorsque j'ai indiqué aux gens que les informations personnelles ne seraient utilisées que pour annexer d'autres données, puis retirées, de nombreuses personnes ont clairement indiqué leur incrédulité : « Une fois que Statistique Canada aura obtenu les données, le gouvernement (Big Brother) nous surveillera. »

Le fait que de nombreux pays utilisent ce type de sources de données administratives associé à des données de recensement et autres pour remplacer des enquêtes a été perdu dans le débat. S'il est interdit à une organisation indépendante de classe mondiale telle que Statistique Canada d'accéder à des sources de données administratives et alternatives, nos décideurs et dirigeants d'entreprise devront compter sur différentes organisations disposant de ce type de données pour nous indiquer leur version de ce qui se passe au Canada. Il est préférable que ces données importantes soient améliorées et intégrées dans le cadre statistique national sous la supervision de Statistique Canada.

ALORS OÙ ALLONS-NOUS MAINTENANT ?

Un grand nombre de données, des données anonymes de localisation de téléphone mobile aux rapports sur l'Internet des objets (IdO) sur des sujets tels que la consommation d'énergie, peuvent être utilisées pour concevoir, développer et mesurer des programmes gouvernementaux. Mais ils doivent être normalisés, nettoyés, pondérés et agrégés à l'aide de données plus traditionnelles pour assurer leur exactitude et leur fiabilité. Les données officielles peuvent rendre les mégadonnées utilisables et fiables.

Les expériences doivent continuer à garantir que non seulement nous préservons ce que nous avons, mais que nous comblons les nombreuses lacunes. Les silos entre les ministères fédéraux et entre les gouvernements fédéral et provincial doivent être surmontés. La production du volume et de la variété

des données nécessaires dans notre monde en mutation exigera d'importants changements dans nos approches.

Les citoyens ont certainement le droit de voir leurs données personnelles protégées en vertu des lois qui consacrent le droit à la vie privée. Mais si les producteurs de données peuvent montrer que non seulement les données seront protégées, mais qu'elles seront réellement utilisées pour améliorer la vie des gens, nous pourrons nous rapprocher du processus décisionnel fondé sur des preuves au Canada.

La politique mise à part, ces nouvelles approches doivent être explorées. Nous nous efforçons de déchiffrer les faits de la fiction. Statistique Canada est la bonne organisation pour diriger ces efforts continus de modernisation. Les Canadiens et leur gouvernement devraient leur donner l'occasion de le faire et de faire en sorte que les 100 prochaines années soient aussi fructueuses que les années précédentes.