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

Projet pilote de mégadonnées pour les données des compteurs intelligents (version abrégée)

Lily Ma[1]

Résumé

Qu'entend-on par mégadonnées? Peuvent-elles remplacer ou compléter les enquêtes officielles? Quels sont certains des défis liés à l'utilisation des mégadonnées pour les statistiques officielles? Quelles sont certaines des solutions possibles? L'automne dernier, Statistique Canada a investi dans un projet pilote sur les mégadonnées afin de répondre à certaines de ces questions. Il s'agissait du premier projet d'enquête auprès des entreprises de cette sorte. Le présent document abordera certaines des leçons apprises dans le cadre du projet pilote de mégadonnées pour les données des compteurs intelligents.

Mots-clés : Mégadonnées; compteurs intelligents; statistiques officielles.

1. Introduction

L'automne dernier, Statistique Canada a manifesté son enthousiasme à l'égard des mégadonnées en finançant un projet pilote de mégadonnées qui émanait de la Conférence des grandes idées et de La voie de l'innovation. Les deux principaux objectifs du projet pilote de mégadonnées étaient les suivants : 1) utiliser les données de compteurs intelligents comme exemples de mégadonnées afin d'explorer ce qui est faisable ou non, ainsi que les outils et les compétences requis, de même que les avantages et les lacunes possibles de l'utilisation de données de cette ampleur à Statistique Canada, et 2) évaluer la faisabilité de remplacer et/ou de compléter les données d'enquête de Statistique Canada sur la consommation résidentielle d'électricité par les données des compteurs intelligents.

Les objectifs du présent document sont les suivants : tout d'abord, décrire le projet en fournissant un aperçu des données, de l'approche et des méthodes utilisées; puis, présenter les résultats du point de vue des six dimensions de la qualité des données définies par Statistique Canada; et enfin, formuler un certain nombre de recommandations sur la base des constatations du projet, en vue d'aller de l'avant avec les données de compteurs intelligents (de façon particulière) et les mégadonnées (en général).

2. Sommaire du projet

2.1 Compteurs intelligents au Canada

Les compteurs intelligents sont des compteurs électroniques qui permettent la collecte automatisée de données sur la consommation d'électricité des ménages et des petites entreprises (IESO, 2014). Selon l'Agence internationale de l'énergie, entre 2008 et 2012, le déploiement cumulatif des compteurs intelligents dans le monde a augmenté de 500 %, et l'installation cumulative mondiale de compteurs intelligents augmentera, pour passer de moins de 300 millions de compteurs en 2012 à 1 milliard en 2018 (AIE, 2013). Au Canada, des compteurs intelligents sont installés partout en Ontario et en Colombie-Britannique. Les provinces qui restent en sont à diverses étapes de l'étude des compteurs intelligents ou de leur installation.

---

[1] Il s'agit d'une version condensée d'un document plus long. Pour la version complète, veuillez communiquer avec l'auteure : Lily Ma, Statistique Canada, Ottawa (Ont.) Canada, K1A 0T6 (lily.ma@statcan.gc.ca). Cette recherche a été financée par la Conférence des grandes idées de Statistique Canada. L'auteure aimerait remercier Michael Scrim et Yves DeGuire pour leur soutien continu, ainsi qu'André Bernard, Larry Mckeown, Russell Kowaluk, Karla Fox, Hélène Bérard, Jean Pignal, Jean-Pierre Simard et André Loranger pour leurs commentaires concernant les versions antérieures du document.

2.2 Données de compteurs intelligents comme exemples de mégadonnées

Les données de compteurs intelligents sont un bon exemple de mégadonnées, parce que, non seulement elles répondent aux critères de volume élevé et de rapidité élevée des mégadonnées, mais aussi parce que leur source est fiable et que leur présentation est semi-structurée, ce qui les rend utilisables avec les outils existants de Statistique Canada. Pour donner un aperçu de l'échelle, en 2014, environ 4,8 millions de compteurs intelligents étaient installés en Ontario et recueillaient des données auprès de presque tous les ménages et petites entreprises, à un taux de 3,456 milliards de points de données par mois (IESO, 2013).

2.3 Données de compteurs intelligents et enquêtes de Statistique Canada

À l'heure actuelle, plusieurs enquêtes et programmes statistiques à Statistique Canada recueillent ou utilisent des données sur la consommation résidentielle d'électricité ou des données liées à cette consommation, y compris l'Écoulement de l'électricité – trimestriel – secteur résidentiel, la Disponibilité et écoulement de l'électricité – Enquête annuelle, l'Enquête sur les ménages et l'environnement, l'Enquête sur les dépenses des ménages, les Dépenses de consommation finales des ménages, trimestrielles, les Dépenses de consommation finales des ménages, détaillées, l'Indice des prix à la consommation, les Parités de pouvoir d'achat, les Indices comparatifs des prix de détail entre les villes, les programmes du Recensement et du Système de comptabilité nationale. Chacun de ces programmes et enquêtes pourrait profiter de données de compteurs intelligents, et chacun représente une occasion à explorer.

Les compteurs intelligents offrent la possibilité, à l'avenir, de recueillir des données directement auprès de l'Entité responsable des compteurs intelligents (une fois l'autorisation obtenue), plutôt que de faire enquête auprès des services publics et des ménages individuels. Cela réduirait le fardeau de réponse, accroîtrait l'efficacité de la collecte des données et pourrait améliorer l'exactitude, l'actualité, la cohérence et la pertinence des données.

2.4 Obtention d'ensembles de données

À l'heure actuelle, seuls les sociétés de distribution locale (SDL) et leurs agents autorisés peuvent transmettre ou demander de l'information du système de gestion et d'entreposage des données des compteurs, qui traite et entrepose les données de compteurs intelligents d'un emplacement géographique. Ainsi, pour pouvoir obtenir des données de compteurs intelligents, nous devions d'abord conclure des ententes de partage des données avec les SDL.

Une fois un premier contact établi et après une rencontre avec l'Entité responsable des compteurs intelligents et deux SDL, deux ententes de partage de données adaptées ont été conclues après plusieurs mois de négociations. Afin de préserver l'anonymat des deux compagnies, on les appelle SDL A et SDL B.

Au début de 2014, nous avons reçu plus de 200 Go de données de compteurs intelligents des deux SDL. Nous avons été informés que les ensembles de données étaient des échantillons représentatifs de leur clientèle résidentielle et de petites entreprises[2]. Chaque ensemble de données représente un panel non équilibré comprenant des données anonymes, sélectionnées de façon aléatoire[3] et horodatées sur la consommation horaire en kWh, au niveau du ménage, entre 2007 et 2013.

Pour mieux tirer parti du potentiel analytique des données de compteurs intelligents, nous avons aussi obtenu des données sur la température horaire de stations météorologiques d'Environnement Canada (Gouvernement du Canada, 2014), ainsi que des données sur la tarification horaire de la Commission de l'énergie de la province (Commission de l'énergie de l'Ontario, 2014). Toutes ces données ont été appariées aux emplacements géographiques et aux périodes pendant lesquelles les données de compteurs intelligents ont été obtenues.

[2] Clients dont la demande est inférieure à 50 kW par mois.
[3] Conformément à la définition de chaque SDL.

2.5 Transfert des données

Comme la taille des fichiers de données obtenus dépassait les limites du système actuel de transfert électronique de fichiers (TEF) de Statistique Canada, les fichiers ont été transférés au moyen de disques durs chiffrés, selon les procédures de transfert de données de nature délicate approuvées par la Gendarmerie royale du Canada (GRC).

2.6 Entreposage des données

Une fois à Statistique Canada, les fichiers de données ont été téléchargés dans le réseau d'essai SAS au moyen de Secure Shell et du protocole de transfert de fichiers sécuritaires (SSH/SFTP). Au total, 1,2 To d'espace de serveur a été obtenu de Services partagés Canada et utilisé, une fois les autres ensembles de données fusionnés et ayant fait l'objet d'une copie de sauvegarde.

2.7 Variables de comparaison

Compte tenu des contraintes de budget et de temps de l'étude, nous avons mis l'accent sur la comparaison de l'échantillon de données de compteurs intelligents et des données de l'enquête Écoulement de l'électricité - trimestriel – secteur résidentiel, étant donné que cela permettait la comparaison la plus directe des concepts (c.-à-d. la consommation) et de la fréquence (c.-à-d. trimestrielle). En outre, afin d'explorer davantage les applications analytiques possibles des données de compteurs intelligents, nous avons couplé les données sur la consommation horaire d'électricité et les données sur la tarification horaire, ainsi que sur la température horaire. Nos données et variables de comparaison figurent dans le tableau 2.7-1.

Tableau 2.7-1
Ensembles de données et variables de comparaison

Données	Variables	Source
Écoulement de l'électricité – trimestriel – secteur résidentiel	Électricité livrée aux clients du secteur résidentiel (mWh)	Services publics A et B
Compteur intelligent	Numéro d'identification de client anonyme, consommation horaire d'électricité (kWh), horodateur (JJMMAA : HH : MM : SS)	Services publics A et B
Météo	Température (C), horodateur (JJ/MM/AAAA/HH:MM)	Environnement Canada
Données sur la tarification horaire	Prix des périodes de consommation faible, moyenne et élevée (¢)	Commission de l'énergie de la province
Variables créées	Année (AAAA), Mois (MM), Jour (JJ), Heure (HH), Éteint (0,1), Moyen (0,1), Allumé (0,1), Fin de semaine (0, 1), Congé (0, 1)	Créé selon le calendrier local

2.8 Période de comparaison

Une fois les données converties et téléchargées, des fréquences ont été exécutées concernant le nombre de lectures de compteurs intelligents selon le trimestre. Pour la SDL A, les fréquences se sont stabilisées, avec une différence trimestrielle moyenne de moins de 1 % après le quatrième trimestre (T4) de 2011. Pour la SDL B, les fréquences se sont stabilisées, avec une différence trimestrielle moyenne d'environ 1 % entre le premier (T1) et le deuxième trimestre (T2) de 2009. Ainsi, les périodes de référence utilisées pour les comparaisons dans la présente étude commencent avec le T4 de 2011 pour la SDL A et le T1 de 2009 pour la SDL B.

2.9 Traitement des données

À partir de ces périodes, nous avons élaboré des statistiques descriptives au niveau du ménage pour déterminer les problèmes de données, comme les données en double, les données manquantes et les valeurs aberrantes importantes. Souvent, les valeurs aberrantes importantes ne représentaient que des données mal catégorisées. Après avoir

examiné la ou les causes, les valeurs aberrantes similaires, ainsi que les données en double et les données manquantes, ont été automatiquement détectées et supprimées au moyen d'algorithmes à partir de programmes SAS.

Une fois les données épurées, les données sur la température horaire et les données sur la tarification horaire ont été appariées aux données horaires des compteurs intelligents, selon le moment de la consommation. Des statistiques descriptives et des graphiques ont alors été exécutés à l'égard de divers ensembles de données, afin de déterminer leurs rapports.

Enfin, nous avons agrégé la consommation horaire d'électricité sur une base trimestrielle, afin qu'elle corresponde aux données sur l'électricité livrée sur une base trimestrielle aux consommateurs résidentiels tirées de l'enquête sur l'écoulement d'électricité, et nous avons effectué une analyse des tendances et des comparaisons de la variation d'une période à l'autre entre les deux ensembles de données.

L'ensemble du traitement des données décrit précédemment a été effectué grâce à un traitement en parallèle au moyen du réseau d'essai SAS à Statistique Canada. Le calcul en réseau peut être considéré comme une architecture de traitement en parallèle, dans laquelle les ressources informatiques sont partagées dans l'ensemble d'un réseau, ce qui permet à un ou plusieurs utilisateurs d'utiliser pleinement les processeurs distribués entre plusieurs machines, ainsi que les processeurs multiples installés sur une machine[4]. Le réseau d'essai SAS est composé de six nœuds, chacun comprenant 16 cœurs, soit 96 cœurs au total[5]. Pour donner une idée de l'échelle, un ordinateur de bureau type comporte 2 cœurs. Pour notre projet, un programme, pour lequel il faudrait généralement plusieurs heures d'exécution séquentielle sur un poste de travail, a nécessité moins de 15 minutes, grâce au traitement parallèle au moyen du réseau d'essai SAS.

2.10 Dimensions de la comparaison

On a évalué la qualité des données des compteurs intelligents en comparaison avec celle de l'enquête sur l'écoulement d'électricité, à partir des six dimensions de la qualité de Statistique Canada : exactitude, pertinence, actualité, cohérence, intelligibilité et accessibilité (Statistique Canada, 2014).

3. Résultats du projet

3.1 Dimension de la qualité 1 : Exactitude

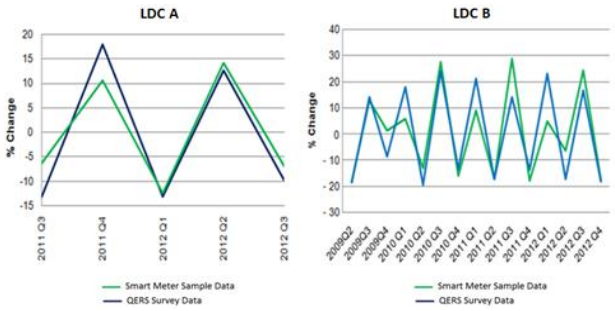
L'*exactitude* des données statistiques est le degré auquel elles décrivent correctement les phénomènes qu'elles visent à mesurer.

Comme les répondants à l'enquête dotés de compteurs intelligents fonderaient leurs réponses sur les données agrégées de ces compteurs, on peut conclure a priori que les données des compteurs intelligents sont au moins aussi exactes, sinon plus, que les données d'enquête. Les données des compteurs intelligents, si on pouvait toutes les obtenir, ne comporteraient pas d'erreurs de réponse (p. ex. déclaration de chiffres inexacts), de biais de non-réponse, d'erreurs de traitement (p. ex. erreurs d'entrée des données) et d'erreurs d'imputation. Même s'il y a des possibilités d'erreurs mécaniques, peu importe les erreurs qui se produiraient par suite de l'utilisation des compteurs intelligents, ces erreurs se manifesteraient aussi dans les données d'enquête, parce que les répondants à l'enquête se fient aux compteurs intelligents pour obtenir les données pour répondre à l'enquête. En outre, il est dans l'intérêt primordial de chaque service public de disposer de lectures de compteurs précises, et tous ont mis en place des contrôles réglementés pour réduire ces erreurs au minimum. Par ailleurs, si on ne pouvait obtenir que des échantillons des données des compteurs intelligents, la non-représentativité des échantillons représenterait une source possible d'erreurs. Une solution possible consisterait à obtenir la collaboration des services publics pour s'assurer que les échantillons sont représentatifs (p. ex. établir des règles pour la sélection des données).

[4] Consultation avec Yves Deguire, 1er novembre 2013.
[5] Consultation avec l'équipe de la Division de l'ingénierie des systèmes, 1er avril 2014.

A posteriori, nous avons effectué une analyse des tendances et comparé les variations d'une période à l'autre entre nos données de l'échantillon de compteurs intelligents et les données de l'enquête sur l'écoulement d'électricité. Nous avons choisi cette enquête parce qu'elle permettait la comparaison la plus directe du point de vue des concepts et de la fréquence. Même si les totaux des échantillons ne sont pas directement comparables aux données d'enquête, comme l'enquête couvre un plus grand nombre de ménages que l'échantillon, et comme l'échantillon est représentatif[6], nous étions en mesure d'agréger la consommation horaire d'électricité pour chaque trimestre, afin qu'elle corresponde aux données sur la quantité totale d'électricité livrée aux consommateurs résidentiels tirées de l'enquête. Une autre approche consisterait à effectuer des comparaisons par ménage, mais l'enquête ne comprend pas de données au niveau du ménage, ni le nombre de ménages. Une autre possibilité consisterait à accroître l'agrégat fondé sur les compteurs intelligents selon la fraction des ménages échantillonnés, mais jusqu'à maintenant, nous disposons uniquement de données brutes à ce sujet et nous ne voulons pas émettre d'hypothèses. Par conséquent, nous présentons les comparaisons entre les deux ensembles de données du point de vue des taux de croissance trimestriels (voir la figure 3.1-1).

Figure 3.1-1
Total de la consommation résidentielle, variation trimestrielle en pourcentage



Lorsque nous avons comparé les deux ensembles de données, nous avons déterminé que les tendances de la variation d'une période à l'autre correspondaient généralement bien, tant du point de vue de la direction que de l'amplitude. Les points pour lesquels l'appariement n'était pas aussi bon étaient ceux où les données d'enquête étaient imputées. Une fois que nous avons exclu les données d'enquête imputées, la corrélation entre les deux sources de données s'est améliorée, ce qui indique que certaines inexactitudes sont causées par la méthode actuelle d'imputation.

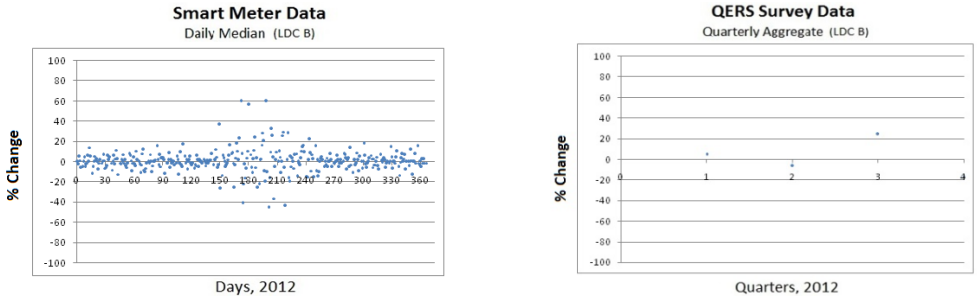
3.2 Dimension de la qualité 2 : Pertinence

La *pertinence* rend compte du degré auquel une mesure répond aux besoins des clients en éclairant les questions qui sont importantes pour eux.

Étant donné que les données de compteurs intelligents sont plus détaillées et fréquentes que les données de l'enquête sur l'écoulement d'électricité, elles peuvent reproduire l'ensemble de l'information que l'enquête peut fournir et, ainsi, tous les besoins des clients qui peuvent être comblés au moyen de l'enquête peuvent aussi l'être au moyen des données des compteurs intelligents. En outre, il est possible de détecter une volatilité à fréquence élevée dans la consommation d'électricité à partir des données détaillées, mais pas à partir des agrégats trimestriels (voir la figure 3.2-1).

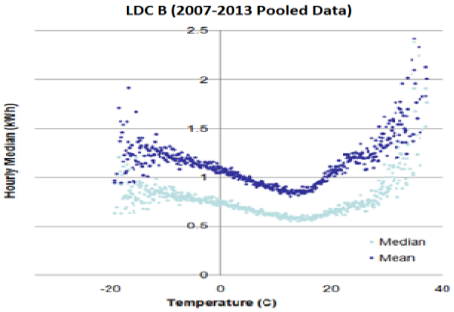
[6] Conformément à la définition des SDL.

Figure 3.2-1
Fréquence des données de compteurs intelligents et des données d'enquête



En outre, les données de compteurs intelligents plus désagrégées peuvent élargir les utilisations par les clients. Par exemple, les données de compteurs intelligents peuvent nous aider à déterminer le rapport entre la température et la consommation d'électricité par les ménages (voir la figure 3.2-2).

Figure 3.2-2
Effets de la température



Le tracé en V de la figure 3.2-2 a été obtenu en couplant les données sur la température horaire et les données sur la consommation horaire, et il montre la mesure dans laquelle la majeure partie des augmentations de la consommation d'énergie sont le résultat du chauffage et de la climatisation. Ce tracé est conforme aux ouvrages publiés. Des données détaillées décrivant les rapports entre la température et la consommation d'électricité pourraient avoir des répercussions sur les politiques, par exemple, sur le comportement de consommation résidentielle et sur l'étalonnage social.

Les données de compteurs intelligents peuvent aussi nous aider à déterminer l'élasticité des prix de la demande d'électricité. Par exemple, selon la figure 3.2-3, au fur et à mesure que le prix augmente, la consommation horaire d'électricité semble diminuer, ce qui indique que la structure de tarification horaire, qui vise à réduire la consommation pendant les heures de pointe, fonctionne peut-être. Il s'agit d'un sujet qui pourrait être exploré davantage et qui pourrait intéresser les décideurs, ainsi que les consommateurs.

Figure 3.2-3
Effets du prix

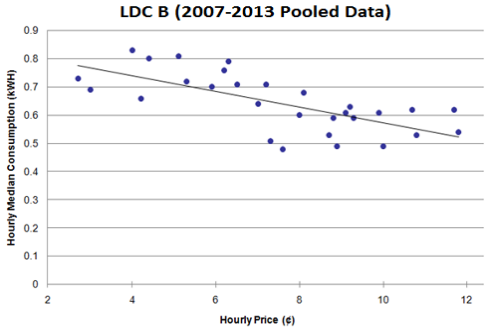
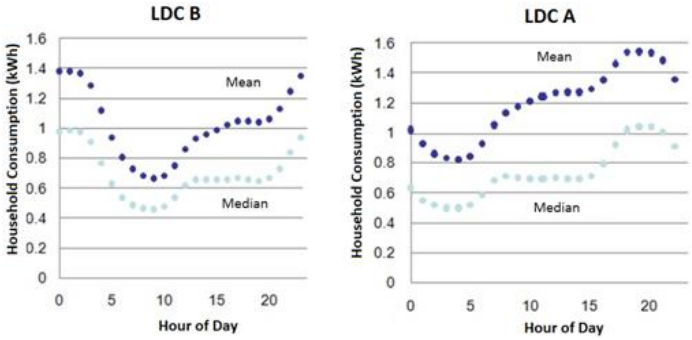


Figure 3.2-4
Écarts géographiques



Dans cet autre exemple (voir la figure 3.2-4), les données de compteurs intelligents pourraient nous aider à déterminer les écarts géographiques entre les modèles de consommation. À partir de seulement deux services publics de la même province, nous pouvons établir les différences dans les modèles de consommation entre les emplacements géographiques. Cela montre que les résultats agrégés pourraient potentiellement cacher des différences régionales importantes, que les données de compteurs intelligents pourraient faire ressortir, ce qui pourrait intéresser les responsables régionaux de la réglementation et d'autres utilisateurs des données.

Enfin, les données de compteurs intelligents plus détaillées nous offrent aussi la possibilité, par exemple, d'exécuter une analyse en grappes ou de réseaux neuronaux. Les possibilités sont vastes.

3.3 Dimension de la qualité 3 : Actualité

L'*actualité* est définie comme le temps écoulé entre la période de référence et la date à laquelle les données sont mises à la disposition des utilisateurs.

Comme les données de compteurs intelligents sont produites sur une base horaire, elles offrent une fourchette plus grande de périodes de référence possibles que les données d'enquête existantes. Toutefois, tout comme les données d'enquête, les données de compteurs intelligents ont le même niveau d'actualité que le taux auquel les données peuvent être obtenues, traitées et diffusées. Comme on l'a indiqué précédemment, l'obtention de données de compteurs intelligents dépend de facteurs comme l'obtention d'autorisations de SDL et l'accès aux outils appropriés. Compte tenu de la priorité des activités au quotidien de l'Entité responsable des compteurs intelligents et des SDL, les données des compteurs intelligents ne peuvent être extraites sur demande pour l'utilisation par un tiers. En outre, même si on nous accorde l'autorisation d'obtenir les données dont nous avons besoin, lorsque nous en avons besoin, nous devons disposer des outils nécessaires pour transférer et traiter les données.

3.4 Dimension de la qualité 4 : Intelligibilité

L'*intelligibilité* indique la disponibilité des métadonnées et des renseignements supplémentaires nécessaires pour interpréter et utiliser les données de façon appropriée.

Bien que les métadonnées que nous avons reçues (p. ex. les étiquettes de données et les descriptions) étaient exactes et suffisantes, la disponibilité de renseignements supplémentaires (p. ex. taille de l'échantillon, méthode d'échantillonnage) dépend de la collaboration des SDL.

3.5 Dimension de la qualité 5 : Cohérence

Par *cohérence*, on entend la mesure dans laquelle les données peuvent être jumelées à d'autres sources dans un cadre analytique global à travers le temps.

Comme nous l'avons démontré précédemment, la nature détaillée, malléable, fréquente et transparente des données des compteurs intelligents permet de les apparier avec d'autres sources de données et de les intégrer à d'autres bases plus facilement. Si, à l'avenir, nous devions obtenir des données additionnelles, comme les codes postaux, avec les données de compteurs intelligents, celles-ci pourraient être utilisées pour mettre à jour les bases de sondage. Par exemple, s'il y a un nombre *X* de compteurs intelligents dans l'emplacement *Y*, on pourrait présumer avec assez de certitude qu'il y a au moins un nombre *X* de ménages dans l'emplacement *Y*. Ces renseignements pourraient notamment profiter à notre recensement.

3.6 Dimension de la qualité 6 : Accessibilité

L'*accessibilité* est définie comme la facilité avec laquelle les clients peuvent obtenir et utiliser des données pertinentes d'un organisme statistique.

Nous ne nous attendons pas à des différences significatives entre l'accessibilité des données de compteurs intelligents et d'autres sources de microdonnées confidentielles ici à Statistique Canada, étant donné qu'elles sont toutes protégées par la *Loi sur la statistique*.

4. Recommandations

4.1 Données de compteurs intelligents

Comme le Canada en est toujours aux premières étapes de la mise en œuvre des compteurs intelligents, il faudra cultiver les rapports avec les services publics, les entités de compteurs intelligents et les commissions de l'énergie compétentes et Services partagés Canada (SPC) pour que cette initiative aille de l'avant, afin de déterminer les options possibles et durables à long terme. Par exemple, les équipes de la technologie de l'information des deux côtés pourraient examiner des façons d'accélérer les solutions de transfert des données. Les préoccupations du public concernant la protection des renseignements personnels devraient être prises en compte. En outre, des études portant sur l'utilité des données pour d'autres enquêtes et programmes, devraient être aussi explorées pour maximiser les avantages possibles.

4.2 Acquisition de mégadonnées

Sauf pour les concurrents analytiques[7] (p. ex. Google, Opower), de nombreuses sources de mégadonnées sont des sous-produits de processus générés par des machines d'organisations (p. ex. transactions commerciales, lectures de capteur, registres d'événements), ce qui fait qu'elles dépendent de la capacité analytique et des priorités opérationnelles de ces organisations. Leur capacité et leur volonté d'extraire des données pour un usage extérieur peuvent varier. Par exemple, alors que certaines organisations sont capables de recueillir, d'entreposer, d'épurer, d'analyser et d'extraire des données facilement, d'autres peuvent être limitées dans leur capacité de manipuler les données, en raison de calendriers opérationnels serrés et de la disponibilité des ressources nécessaires pour extraire les données pour un usage externe. Certaines organisations peuvent être réticentes à fournir leurs données exclusives ou services d'analyse de données à Statistique Canada sans compensation, tandis que d'autres peuvent avoir des raisons stratégiques de souhaiter garder privées leurs données exclusives. Ainsi, le temps et les efforts nécessaires pour que les organisations fournissent des données aux parties externes devraient être pris en compte.

En dernier ressort, il revient à l'organisation de fournir les données, et il doit y avoir des incitatifs pour que les organisations fournissent leurs données à Statistique Canada. Une approche possible pour faciliter l'acquisition des données est de convaincre nos organisations partenaires que la fourniture de leurs données peut profiter aux Canadiens. Une autre approche consisterait à mettre l'accent sur le fait que la *Loi sur la statistique* interdit strictement la divulgation publique de données confidentielles et applique cette interdiction au moyen de sanctions pouvant aller jusqu'à l'emprisonnement. Ces mesures protégeraient la confidentialité des données de compteurs intelligents, tout comme elles protègent déjà la confidentialité des autres données de nature délicate, par exemple, les données du recensement. En outre, de la rétroaction analytique pourrait être offerte aux fournisseurs des données en guise de compensation. Dans tous les cas, Statistique Canada devra constituer des partenariats avec les intervenants clés et créer des ententes de partage des données adaptées aux fournisseurs de données particuliers.

4.3 Transfert des données

Même si le transport des données au moyen d'un disque dur chiffré et de procédures approuvées par la GRC a satisfait à nos besoins pour ce projet, il ne s'agit pas de la méthode la plus efficace pour transférer des données de nature délicate à long terme. Étant donné que SPC, une nouvelle organisation créée pour centraliser l'infrastructure des technologies de l'information fédérale, doit appuyer tout changement dans la capacité d'entreposage et de traitement des données de Statistique Canada pour tenir compte des mégadonnées, les solutions possibles pourraient inclure des partenariats avec SPC pour trouver des façons d'augmenter la capacité du service actuel de TEF et/ou pour rendre les procédures approuvées par la GRC plus accessibles et rapides.

4.4 Entreposage et traitement des données

[7] L'analyse est définie comme l'utilisation exhaustive de données, d'analyses statistiques et quantitatives, ainsi que de modèles explicatifs et prédictifs pour prendre des décisions et des mesures, et un concurrent analytique est défini comme une organisation dont l'avantage concurrentiel repose sur ses capacités analytiques. Davenport,T. H. et J. G. Harris, (2007). p. 7.

Nous recommandons que Statistique Canada établisse un partenariat avec SPC pour explorer la possibilité d'accroître et d'élargir la capacité du réseau (ce qui donnerait à un plus grand nombre d'utilisateurs l'accès à une puissance de traitement supérieure), ainsi que de tirer parti de l'architecture d'entreposage et de traitement décentralisé Hadoop, ainsi que de certains autres outils (p. ex. RHadoop), que l'on pourrait utiliser pour des techniques plus complexes[8].

4.5 Visualisation des données

L'ensemble de la visualisation des données a porté sur des statistiques sommaires. Nous recommandons que Statistique Canada envisage la possibilité de mettre en œuvre des outils analytiques visuels des mégadonnées.

5. Autres considérations

Il sera essentiel d'obtenir la coopération et le partenariat de parties externes pour que les organismes de statistiques officielles puissent aller de l'avant avec les mégadonnées. Statistique Canada devra déterminer comment il obtiendra la collaboration des organismes externes, ainsi que la façon de procéder en l'absence de collaboration (c.-à-d. la conclusion d'une entente de partage de données). En outre, certaines compagnies dépensent des millions pour leur infrastructure analytique, et il peut être plus efficace et rentable pour les organismes statistiques officiels de procéder à des analyses de données au moyen de leur système, plutôt que de reproduire des ensembles de données et des systèmes.

Si Statistique Canada veut poursuivre l'exploitation des mégadonnées, il devra tenir compte des perceptions publiques de la protection des renseignements personnels. Par exemple, Statistique Canada peut souligner au public qu'il a une longue tradition d'utilisation de sources de données externes et qu'il dispose d'un système robuste pour protéger les données confidentielles. Dans les cas où les sources de mégadonnées servent à remplacer les enquêtes existantes, Statistique Canada peut souligner que l'information est déjà recueillie et que les méthodes axées sur les mégadonnées ne font qu'accroître l'exactitude et réduire le fardeau de réponse.

Bibliographie

BC Hydro (2014). *Meter Choices*. Disponible à : https://www.bchydro.com/energy-in-bc/projects/smart_metering_infrastructure_program/smart_meter_installation/installation_preparation/meter-choice.html (consulté le 1er juin 2014).

Davenport, T. H. et J. G. Harris (2007). *Competing on analytics: The new science of winning.* Boston, MA: Havard Business School.

EMC^2 (2014). *Digital Universe*. Disponible à : http://www.emc.com/leadership/programs/digital-universe.htm (consulté le 1er juin 2014).

Fédération canadienne des municipalités (2013). *Parc éolien et programme pilote de réseau d'efficacité énergétique*. Disponible à : http://www.fcm.ca/home/awards/fcm-sustainable-communities-awards/2013-winners/2013-energy-projects- co-winner-2.htm (consulté le 1er juin 2014).

Fortis Alberta (2012). *Our Meters*. Disponible à : http://www.fortisalberta.com/residential/customerservice/meters/Pages/FortisAlberta-Meters.aspx (consulté le 1er juin 2014).

Gouvernement du Canada (2014). *Données climatiques historiques*. Disponible à : http://climate.weather.gc.ca/ (consulté le 1er février 2014).

Hydro Québec (2014). *Projet*. Disponible à : http://meters.hydroquebec.com/questions-answers/project/meter-replacement-hydro-quebec (consulté le 1er juin 2014).

IDC (2011). *Top 10 Predictions*. Disponible à : http://cdn.idc.com/research/Predictions12/Main/downloads/IDCTOP10Predictions2012.pdf (consulté le 1er juin 2014).

[8] Hadoop est un cadre ouvert pour l'entreposage et le traitement décentralisé de grands ensembles de données au moyen de matériel standard.

IESO (2014). *How Your Smart Meter Works*. Disponible à : https://www.ieso.ca/imoweb/siteshared/smart_meter_information.asp?sid=ic (consulté le 1er juin 2014).

IESO (2013). *Ontario Smart Grid Progress Assessment: A Vignette*. Disponible à : http://www.ieso.ca/documents/smart_grid/Smart_Grid_Progress_Assessment_Vignette.pdf (consulté le 1er juin 2014).

Agence internationale de l'énergie (2013). Tracking Clean Energy Progress 2013. Paris : AIE.

Manitoba Hydro (2010). Disponible à : http://www.hydro.mb.ca/regulatory_affairs/electric/gra_2010_2012/Appendix_24.pdf (consulté le 1er juin 2014).

Newfoundland and Labrador Hydro (juillet 2012). *A Report To The Board Of Commissioners Of Public Utilities*. Disponible à : http://www.pub.nf.ca/applications/NLH2013Capital/files/application/NLH2013Application-VolumeII-Report23.pdf (consulté le 1er juin 2014).

Nova Scotia Power (2014). *Nova Scotia Power Answered*. Disponible à http://tomorrowspower.ca/answer/106 (consulté le 1er juin 2014).

Commission de l'énergie de l'Ontario (29 avril 2014). *Time-of-use (TOU) Prices*. Disponible à : http://www.ontarioenergyboard.ca/ OEB/Consumers/Electricity/Electricity+Prices (consulté le 1er juin 2014).

SaskEnergy (n.d.). *Advanced Metering Infrastructure*. Disponible à : http://www.saskenergy.com/residential/AMI.asp (consulté le 1er juin 2014).

Smartmeters (13 février 2013). *New Brunswick Opens Smart Grid Center*. Disponible à : http://www.smartmeters.com/the-news/smart-grid-news/3844-new-brunswick-opens-smart-grid-center.html (consulté le 1er juillet 2013).

Statistique Canada (10 avril 2014). *La définition de la qualité*. Disponible à : http://www.statcan.gc.ca/pub/12-539-x/4147797-eng.htm#elements (consulté le 1er juin 2014).