

TRANSITION VERS UN NOUVEAU SYSTÈME DE TABULATION POUR LE RECENSEMENT DE 2021 – ASPECTS MÉTHODOLOGIQUES LIÉS À L'ESTIMATION ET À LA PROTECTION DE LA CONFIDENTIALITÉ

Sarah-Anne Savard¹

RÉSUMÉ

À chaque cycle du Recensement de la population canadienne, un très grand nombre d'estimations sont produites et publiées par Statistique Canada sous forme de profils et de tableaux croisés. Pour le Recensement de 2021, le système actuellement utilisé pour produire ces estimations, le Système de spécification de produits assisté par ordinateur (SSPAO), devrait être remplacé par l'Outil généralisé de totalisation de Statistique Canada (GTAB). Du point de vue méthodologique, cette transition présente plusieurs défis, mais aussi de nombreuses opportunités en ce qui concerne les statistiques produites, la qualité des données publiées et la protection de la confidentialité des répondants. Ces défis et opportunités font l'objet de cet article.

MOTS CLÉS : Recensement; Production de tableaux; Confidentialité; Qualité des données.

ABSTRACT

Each Census cycle, a great number of statistics are produced and disseminated by Statistics Canada in Census Profiles and various data tables. For the 2021 Census, the system currently used to produce these estimations, the Computer-assisted Product Specification System (CAPSS), is expected to be replaced by Statistics Canada's Generalized Tabulation Tool (GTAB). From a methodological viewpoint, this transition brings its share of challenges, as well as opportunities regarding availability of statistics, data quality and confidentiality. These challenges and opportunities are further explored throughout this paper.

KEY WORDS: Census; Tabulation; Confidentiality; Data quality.

1. INTRODUCTION

Le Recensement de la population est un programme majeur de Statistique Canada. Tous les Canadiens sont tenus de remplir leur questionnaire en vertu de la *Loi sur la statistique*. La production de tableaux de données pour le recensement est une opération d'envergure en raison du grand volume de données à traiter et de l'imposante quantité de tableaux diffusés. Comme plusieurs de ces tableaux sont produits pour de petites régions géographiques, il est fréquent qu'un très faible nombre de répondants contribuent au chiffre ou à l'estimation pour certaines cellules. Dans un tel contexte, il est indispensable d'appliquer des procédures automatiques afin de prévenir la divulgation d'information au sujet des répondants. Il est également souhaitable d'informer les utilisateurs lorsque la qualité d'une estimation est compromise par le petit nombre de répondants qui y contribuent. Enfin, des méthodes d'estimation et d'inférence qui tiennent compte des particularités du recensement doivent être proposées aux utilisateurs des données.

Le Système de spécification de produits assisté par ordinateur (SSPAO) utilisé jusqu'à maintenant pour la production des tableaux n'étant plus en mesure de répondre aux besoins croissants du Programme du recensement, il est prévu qu'il soit remplacé par l'Outil généralisé de totalisation de Statistique Canada (GTAB) pour le cycle de 2021.

Cet article présente de manière générale les enjeux et les opportunités liés à cette transition. Dans la section 2, le Recensement de la population canadienne est introduit afin de mettre le lecteur en contexte. La section 3 traite des fonctionnalités principales du SSPAO et de GTAB. Enfin, la section 4 donne un aperçu des évaluations méthodologiques qui ont été menées à ce jour pour assurer le succès de la transition. Les enjeux traités sont liés aux statistiques disponibles dans les deux systèmes, à la qualité des données et à la protection de la confidentialité des répondants.

¹ Sarah-Anne Savard, 100 promenade Tunney's Pasture, Ottawa, ON, Canada, K1A 0T6, sarah-anne.savard@canada.ca.

2. RECENSEMENT DE LA POPULATION

2.1 Aperçu

Le Recensement de la population canadienne a lieu tous les 5 ans et le dernier recensement s'est tenu en mai 2016. Le recensement fournit un portrait statistique détaillé du Canada et de sa population en fonction de ses caractéristiques démographiques, sociales et économiques. Les données recueillies sont importantes pour les collectivités, car elles sont essentielles pour planifier des programmes et offrir des services à la population. De plus, le recensement est la principale source de données sociodémographiques pour certains groupes de population.

Le Recensement de la population est formé de deux volets : un recensement et une enquête-échantillon. Le recensement permet de recueillir des renseignements tels que l'âge, le sexe, les relations entre les membres du ménage et la langue. Pour sa part, l'enquête-échantillon vise à fournir des renseignements supplémentaires comme la scolarité, le travail, la mobilité, le lieu de naissance, l'immigration et le logement. Son objectif est de produire des estimations fiables à différents niveaux géographiques ainsi que pour de petits domaines de la population.

Les données sont recueillies au moyen des questionnaires abrégé et détaillé et, en 2016, les données sur le revenu et certaines données sur l'immigration ont été obtenues au moyen d'un couplage avec des fichiers administratifs.

2.2 Stratégie d'estimation pour l'enquête-échantillon

La sélection des ménages pour l'enquête-échantillon est effectuée selon un plan de sondage systématique stratifié au niveau de petites régions géographiques appelées unités de collecte. La fraction de sondage est d'environ 25 % et les poids initiaux sont donc environ égaux à 4. Les poids sont ajustés pour la couverture et la non-réponse, puis sont calés à certains totaux du recensement calculés avec les données du questionnaire abrégé ou de sources administratives. Les poids finaux ont une valeur entre 1 et 20. Il y a une exception au plan de sondage. Les ménages dans les territoires et les réserves indiennes sont tous sélectionnés dans l'échantillon (et reçoivent tous le questionnaire détaillé) et leur poids de sondage est donc égal à 1. Dans ces régions, la non-réponse est compensée par imputation et les ménages gardent leur poids de 1 comme poids final.

Pour l'estimation de la variance, une version modifiée de la méthode des demi-échantillons partiellement équilibrés avec 32 répliques est utilisée (Devin et Verret, 2016). Cette méthode tient compte de la procédure de pondération et de l'imputation dans les régions où les poids sont égaux à 1. Elle a été développée pour le Recensement de 2016 afin d'être en mesure d'estimer des erreurs-types pour une variété d'estimateurs tout en utilisant un petit nombre de répliques.

2.3 Programme de diffusion

Le Programme de diffusion du Recensement de la population a pour principal objectif de rendre l'information accessible, tout en veillant à ce que les produits diffusés répondent aux principaux besoins de la majorité des utilisateurs de données et que la confidentialité des répondants soit préservée. Statistique Canada tente de fournir gratuitement de plus en plus de données au public et s'efforce de trouver des moyens de publier les résultats dans des délais raisonnables et de manière accessible sur Internet.

Les produits offerts par Statistique Canada sur son site web sont variés. On y trouve entre autres des tableaux de données et des graphiques interactifs. Des menus permettent aux utilisateurs de choisir les catégories à inclure dans les totalisations, pour des variables comme la géographie, le groupe d'âge ou le sexe par exemple. Les Profils du recensement et les tableaux croisés sont tous produits par le système de tabulation. Des produits personnalisés sont aussi disponibles sur demande pour répondre aux besoins spécifiques de certains utilisateurs. Des fichiers de micro-données sont également disponibles. En effet, les chercheurs canadiens ayant prêté serment de discrétion ont accès aux données dans les centres de données de recherche. De plus, des fichiers de micro-données à grande diffusion sont créés pour le grand public (au niveau de la personne, avec ou sans structure hiérarchique au niveau du ménage).

La diffusion du Programme du recensement se démarque des autres enquêtes sociales de Statistique Canada de par son volume : des centaines de millions d'estimations sont produites à partir des données du Programme du recensement et sont publiées sur le site web de Statistique Canada. De plus, il n'y a pas de pondération pour le volet recensement et certains ménages ont un poids de 1 à l'enquête-échantillon. De nombreuses vérifications doivent être effectuées automatiquement afin de préserver la confidentialité et la qualité des données.

3. TRANSITION VERS UN NOUVEAU SYSTÈME DE TABULATION

3.1 L'ancien système : le Système de spécification de produits assisté par ordinateur (SSPAO)

Le SSPAO est un système de production basé sur TPL Tables. Il est utilisé depuis une trentaine d'années à Statistique Canada et est propre au Programme du Recensement de la population. Le SSPAO est muni d'une interface utilisateur qui permet de spécifier toutes les caractéristiques du tableau à produire : l'année de référence, la région géographique, l'unité d'analyse, l'univers à considérer et les variables à inclure. Les statistiques qu'il peut produire sont majoritairement des statistiques descriptives simples. Le SSPAO permet aussi d'appliquer certaines règles de confidentialité directement et d'intégrer un indicateur de qualité global pour chaque niveau géographique d'un tableau. Cependant, plusieurs processus doivent être appliqués après le passage des données dans le SSPAO.

Bien qu'il ait longtemps répondu aux besoins de la diffusion du Programme du recensement, le SSPAO atteint aujourd'hui ses limites. En effet, il est difficile, voire impossible dans certains cas, d'y ajouter de nouvelles fonctionnalités. Par exemple, il est impossible d'y intégrer une méthode d'estimation de la variance utilisant un grand nombre de répliques.

3.2 Le nouveau système : l'Outil généralisé de totalisations de Statistique Canada (GTAB)

Une transition est actuellement en cours vers GTAB, le système généralisé de production de tableaux de Statistique Canada, un système basé sur SAS qui a été développé pour les enquêtes auprès des ménages. Le mandat de GTAB est d'élaborer un système de tabulation commun, destiné à normaliser le calcul de statistiques, la production de tableaux et l'application des règles de confidentialité aux sources de données d'enquête et administratives en vue de la diffusion de données. Éventuellement, GTAB pourrait être utilisé pour produire toutes les estimations diffusées, que ce soit au bureau central de Statistique Canada, dans les bureaux régionaux ou dans les centres de données de recherche.

Les fonctionnalités de GTAB sont nombreuses. Évidemment, l'outil permet le calcul de statistiques descriptives simples, mais aussi de statistiques plus complexes, notamment de statistiques définies par un algorithme et de statistiques ayant une dimension temporelle (par exemple, le coefficient de Gini et la moyenne mobile). Il permet également de produire davantage d'indicateurs de qualité, comme des coefficients de variation et des intervalles de confiance. Un module de confidentialité propre aux données est appliqué aux estimations produites en vue de la diffusion.

Une équipe multidisciplinaire formée de spécialistes du recensement, de gestionnaires, d'informaticiens et de méthodologistes s'occupe du projet de transition vers GTAB pour la diffusion du Recensement de 2021. Les défis sont nombreux car le système n'a pas été pensé pour le recensement. Le rôle des méthodologistes impliqués est de s'assurer que les statistiques nécessaires sont disponibles dans le nouvel outil, que les produits diffusés répondent aux normes de qualité de Statistique Canada et que la confidentialité des répondants soit préservée. La transition vers GTAB offre l'opportunité de réviser la définition des statistiques utilisées et les règles appliquées pour protéger la confidentialité des répondants et assurer la qualité des données. Dans plusieurs cas, ces règles peuvent être améliorées car les limitations imposées par l'ancien système ne sont plus présentes. Cela dit, certaines fonctionnalités essentielles du SSPAO n'existent pas dans GTAB et doivent y être implémentées afin de répondre aux besoins du recensement.

4. ENJEUX MÉTHODOLOGIQUES LIÉS À LA TRANSITION

4.1 Disponibilité des statistiques

L'une des premières évaluations méthodologiques menées avait pour objectif de vérifier que toutes les statistiques produites par le SSPAO puissent l'être aussi par GTAB. Les statistiques disponibles dans le SSPAO sont des statistiques descriptives simples, comme des comptes, des sommes, des moyennes, des ratios et des pourcentages, de même que certaines statistiques d'ordre, notamment la médiane, les quartiles, les quintiles et les déciles. Depuis le cycle de 2016, le SSPAO permet aussi d'estimer l'erreur-type pour ces statistiques dans le cadre de l'enquête-échantillon à partir de 32 répliques.

Pour sa part, GTAB permet de calculer toutes les statistiques disponibles dans le SSPAO, et plus encore. Entre autres, GTAB permet l'estimation de statistiques selon les quantiles d'une variable donnée dans la population et de différences relatives entre les niveaux d'une variable, ainsi que les comparaisons entre les cycles. Les erreurs-types sont calculées avec la méthode du Bootstrap de Rao-Wu et il est possible de produire des intervalles de confiance et des indicateurs de qualité basés sur le coefficient de variation pour la majorité des statistiques.

L'un des défis rencontrés concernant les statistiques disponibles est que l'approche utilisée pour l'estimation des quantiles n'est pas la même dans les deux systèmes. Avec le SSPAO, les quantiles sont estimés différemment selon le type de valeurs que prend la variable. Pour les variables dont les valeurs sont des entiers (nombre de personnes dans un logement, âge, etc.), une méthode propre au recensement est utilisée. Elle est différente de celle utilisée par GTAB. L'exemple de la médiane est utilisé ci-dessous pour illustrer cette différence.

Soit une variable d'enquête discrète Y dont la fonction de répartition est donnée par $F(t) = Prob(Y \leq t)$. La valeur définie par $Y_p = \inf\{y : F(y) \geq p\}$ est le quantile d'ordre p de Y . Par exemple, la médiane de Y est donnée par $Y_{0,5} = \inf\{y : F(y) \geq 0,5\}$. Soient $\{y_i, w_i\}$ les valeurs observées de Y et leur poids respectif, et $y_{(1)}, y_{(2)}, \dots, y_{(n)}$ les statistiques d'ordre de l'échantillon. La distribution $F(t)$ est inconnue, mais elle peut être estimée par la fonction de répartition empirique $\hat{F}(t)$ en utilisant les valeurs observées et les poids. Soit k le rang tel que $\hat{F}(y_{(k)}) \leq 0,5 < \hat{F}(y_{(k+1)})$.

Dans GTAB, une méthode très simple est utilisée pour l'estimation de la médiane de Y dans le cas où les données sont pondérées. Dans le cas où la probabilité $p = 0,5$ est atteinte exactement par la fonction de répartition empirique, la médiane est estimée par la moyenne entre les deux valeurs $y_{(k)}$ et $y_{(k+1)}$. Sinon, la médiane est estimée par $y_{(k+1)}$, la valeur de y pour laquelle la fonction de répartition dépasse p pour la première fois.

Dans le SSPAO, une approche différente est utilisée. Dans le cas d'égalité, la médiane est estimée par la moyenne entre $y_{(k)} + 1$ et $y_{(k+1)}$. Sinon, la médiane est estimée par la valeur $y_{(k+1)}$ augmentée d'une quantité entre 0 et 1 qui reflète la position de 0,5 dans l'intervalle $(\hat{F}(y_{(k)}), \hat{F}(y_{(k+1)}))$. Autrement dit, si $\hat{F}(y_{(k)}) = 0,5$, la médiane est estimée par

$$\hat{Y}_{0,5} = \frac{y_{(k)} + 1 + y_{(k+1)}}{2}$$

Si $\hat{F}(y_{(k)}) < 0,5$, la médiane est estimée par

$$\hat{Y}_{0,5} = y_{(k+1)} + \frac{0,5 - \hat{F}(y_{(k)})}{\hat{F}(y_{(k+1)}) - \hat{F}(y_{(k)})}$$

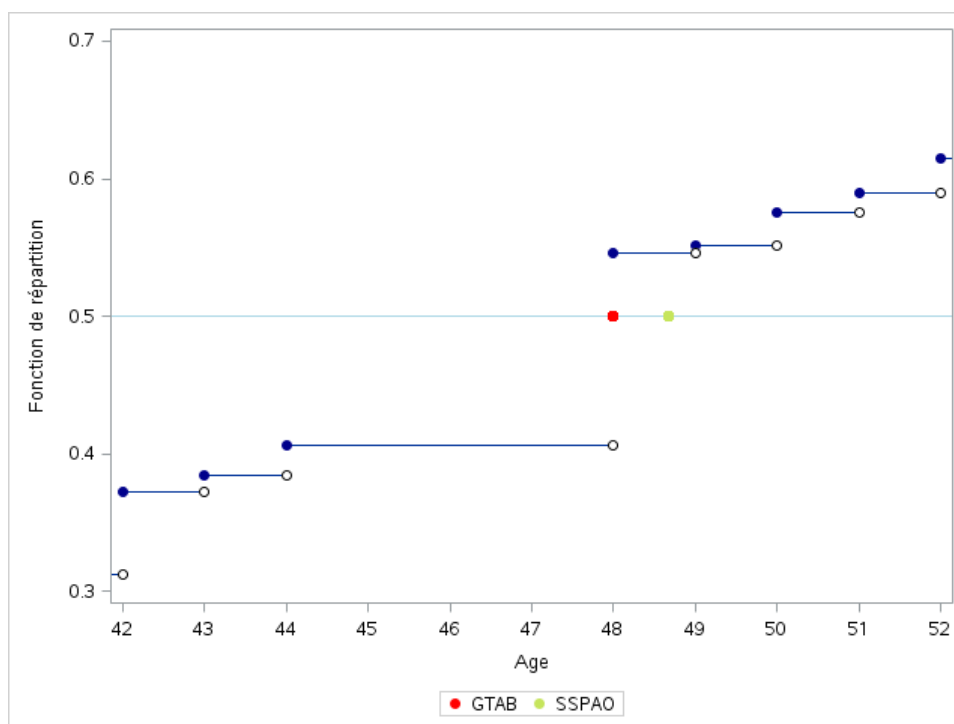
Par exemple, nous considérons la variable de l'âge en valeurs discrètes. Dans une région géographique donnée, la fonction de répartition empirique prend les valeurs présentée au tableau 1. Dans ce cas, on a $\hat{F}(y_{(k)}) = 0,4060$ et $\hat{F}(y_{(k+1)}) = 0,5463$. Comme la probabilité $p = 0,5$ n'est pas atteinte exactement, la médiane obtenue avec la méthode de GTAB est 48, soit la valeur pour laquelle la fonction de répartition empirique dépasse p pour la première fois. La méthode du SSPAO donne plutôt 48,67 comme estimation de la médiane.

La figure 1 est une représentation graphique de la fonction de répartition empirique et des valeurs estimées de la médiane avec les deux méthodes pour la variable de l'âge. La faiblesse de la méthode GTAB est qu'elle ne permet pas beaucoup de valeurs possibles. Dans le cas de certaines variables, comme la taille du ménage, cela limite les comparaisons. En effet, la majorité des régions ont alors une taille de ménage médiane égale à 2. La méthode utilisée par le SSPAO a l'avantage d'apporter plus d'information. Il a été décidé d'implémenter la méthode d'estimation de la médiane du SSPAO dans GTAB pour les variables discrètes.

Tableau 1 – Fonction de répartition empirique de la variable de l'âge

Âge	Fonction de répartition empirique
...	...
42	0,3728
43	0,3843
44	0,4060
48	0,5463
49	0,5513
...	...

Figure 2 – Fonction de répartition empirique et estimation de la médiane avec GTAB et le SSPAO



4.2 Qualité des données

Un deuxième aspect méthodologique important pour le projet de transition vers GTAB est de continuer de garantir la qualité des données. Les estimations produites, que ce soit à partir des données du recensement ou de l'enquête-échantillon, sont sujettes à diverses erreurs. La non-réponse fait partie des erreurs non dues à l'échantillonnage. En effet, certaines personnes ne répondent pas à toutes les questions ou ne répondent pas du tout. Cette non-réponse, bien que faible dans le contexte du recensement, peut induire un biais dans les chiffres calculés et dans les estimations, et ce biais est très difficile à mesurer.

Le taux global de non-réponse (TGN) est un indicateur de la qualité des données qui combine les non-réponses complètes et partielles à l'enquête. Un TGN plus faible indique que le risque d'un biais de non-réponse est plus faible. Le SSPAO permet d'intégrer le TGN pour chacune des régions géographiques pour lesquelles des tableaux sont produits. Les régions qui présentent un TGN supérieur à un certain seuil sont supprimées des produits de données standards mais sont disponibles sur demande. Une fonctionnalité semblable n'existait pas dans GTAB au début du projet de transition et sera ajoutée pour le recensement, mais la manière dont elle sera intégrée est encore incertaine.

Évidemment, l'échantillonnage entraîne aussi une erreur dans les estimations. Dans le cadre du Programme du recensement, seules les estimations produites à partir des données de l'enquête-échantillon sont assujetties à une telle erreur. Différentes mesures de qualité dérivées de la variance estimée peuvent être utilisées pour rapporter l'erreur due à l'échantillonnage. En 2011, des coefficients de variation ont été produits pour certaines estimations et certains niveaux géographiques. En 2016, l'erreur-type a été publiée pour toutes les estimations des Profils du recensement à partir de l'aire de diffusion agrégée, une nouvelle région géographique de diffusion créée pour le Recensement de 2016 et qui compte de 5 000 à 15 000 habitants. Les erreurs-types sont aussi disponibles sur demande pour toutes les régions. À partir du cycle de 2021, il serait souhaitable de publier un indicateur de qualité basé sur la variance pour toutes les estimations diffusées. Un avantage est que GTAB permet le calcul d'intervalles de confiance appropriés pour la plupart des statistiques produites, ce qui facilite l'inférence pour les utilisateurs de données. La méthodologie doit par contre être adaptée pour tenir compte de la méthode d'estimation de la variance utilisée pour l'enquête-échantillon du Programme du recensement. Il reste aussi à évaluer si le volume des tableaux produits serait raisonnable dans ce contexte. De plus, il sera possible d'utiliser davantage de répliques pour l'estimation de la variance, ce qui apportera plus de stabilité.

4.3 Protection de la confidentialité des répondants

Finalement, un troisième aspect méthodologique touché par le changement de système est la façon de protéger la confidentialité des répondants. Plusieurs règles visant à préserver la confidentialité sont appliquées automatiquement dans le SSPAO, comme la suppression de régions, la suppression de statistiques, l'arrondissement aléatoire et les calculs statistiques spéciaux. Ces règles doivent pour la plupart être ajoutées à GTAB car elles sont différentes de celles déjà implémentées.

Premièrement, la suppression de régions est appliquée pour les régions géographiques dont la taille de la population, ou son estimation, est inférieure à un seuil donné. Pour ces régions, aucune donnée n'est diffusée. Des seuils plus stricts sont appliqués pour les caractéristiques de revenu.

Deuxièmement, la suppression de statistiques est appliquée aux variables dont les valeurs sont exprimées en dollars. La suppression est effectuée chaque fois que l'une ou l'autre des trois conditions suivantes est satisfaite :

- il y a dominance ;
- le nombre d'enregistrements à partir desquels la statistique a été calculée est inférieur à un seuil donné ;
- l'étendue des valeurs prises par la variable est trop faible.

Les règles appliquées pour traiter ces cas particuliers sont présentement examinées et les solutions aux problèmes ne sont pas simples. Lorsqu'une stratégie sera élaborée, de nouvelles règles seront ajoutées à GTAB.

Troisièmement, l'arrondissement aléatoire est appliqué à tous les comptes et totaux publiés (sauf les chiffres de logements et des personnes), qu'ils proviennent des données du recensement ou du questionnaire détaillé. L'arrondissement aléatoire transforme toutes les estimations brutes en estimations arrondies aléatoirement, ce qui réduit la possibilité de révéler l'identité de personnes dans les totalisations. L'algorithme d'arrondissement aléatoire utilise une valeur de départ aléatoire pour déclencher le processus d'arrondissement pour les tableaux. Avec GTAB, il sera possible d'utiliser la même valeur de départ aléatoire à chaque fois qu'un même tableau est produit. Cet arrondissement plus cohérent devrait être utilisé en 2021.

Enfin, un autre moyen utilisé dans le SSPAO pour protéger la confidentialité des répondants est d'utiliser des spécifications particulières pour certaines statistiques. En effet, lors du calcul de certaines statistiques, l'arrondissement est fait lors d'une étape intermédiaire. Ces spécifications particulières font partie intégrante de la stratégie de confidentialité appliquée au recensement. Cependant, dans GTAB, il est seulement possible d'appliquer l'arrondissement après l'estimation. Des pistes sont présentement explorées pour élaborer une nouvelle stratégie qui permettrait de mieux protéger la confidentialité des répondants tout en s'insérant dans GTAB. Une approche envisagée pourrait être d'arrondir les estimations produites en contrôlant la différence entre l'estimation et sa valeur arrondie. De plus, un bruit aléatoire sera ajouté à certaines estimations (totaux, moyennes, ratios, etc.).

5. CONCLUSION

En somme, la transition vers GTAB devrait apporter davantage d'autonomie à Statistique Canada et améliorer la qualité des produits diffusés et la protection de la confidentialité. Les travaux futurs impliquent la mise à l'essai du système lorsque les statistiques et les règles auront été implémentées ainsi que la formation du personnel en vue de la production de tableaux pour le Recensement de 2021. De plus, des alternatives pour l'accès aux données seront explorées, notamment l'accès aux données à distance, c'est-à-dire de manière à calculer des statistiques à partir des micro-données sans avoir un accès réel à ces dernières, par l'entremise de GTAB.

REMERCIEMENTS

Je tiens à remercier Camille Charbonneau, Guylaine Dubreuil, Johane Dufour, Sri Kanagarajah, Tyler Kirkland, Marie-Pier Lemieux, Vincent Martin, David Price et Julie Trépanier pour la révision de cet article.

RÉFÉRENCES

Devin, N. et Verret, F. (2016). « The Development of a Variance Estimation Methodology for Large-Scale Dissemination of Quality Indicators for the 2016 Canadian Census Long Form Sample ». *JSM 2016 Proceedings*.

Statistique Canada. *Guide du Recensement de la population, 2016*. N° 98-304-X au catalogue.