

Recueil du Symposium de 2024 de Statistique Canada : Le futur des statistiques officielles

Élaboration de poids ajustés en fonction du couplage pour rendre compte du genre dans la Cohorte santé et environnement du recensement canadien de 2021

par Yubin Sung et Eric Hortop

Date de diffusion : le 8 septembre 2025



Statistique
Canada

Statistics
Canada

Canada

Élaboration de poids ajustés en fonction du couplage pour rendre compte du genre dans la Cohorte santé et environnement du Recensement canadien de 2021

Yubin Sung et Eric Hortop¹

Résumé

La plus récente Cohorte santé et environnement du recensement canadien (CSERCan) s'inscrit dans une série de couplages de microdonnées fondés sur la population qui appuient la recherche en santé de la population selon certaines caractéristiques démographiques, sociales et économiques. La CSERCan de 2021 contient 95,5 % des enregistrements de l'échantillon du questionnaire détaillé du Recensement de 2021. Les enregistrements ne pouvant pas être couplés aux données du Dépôt d'enregistrements dérivés (DED) et ceux présumés être des doublons représentent les 4,5 % restants. Les poids principaux et les poids de répliques rajustés en fonction du couplage permettent aux chercheurs d'estimer et d'évaluer la variance des mesures agrégées de la santé de la population en présence d'appariements manquants, afin de mieux comprendre les expériences des différents groupes de population.

Mots clés : Couplage d'enregistrements; poids de couplage ajustés; calage; poststratification; genre.

1. Introduction

La Cohorte santé et environnement du recensement canadien (CSERCan) de 2021 est le plus récent d'une série de couplages de microdonnées fondés sur la population qui appuient la recherche en santé de la population tenant compte de caractéristiques démographiques, sociales et économiques. La CSERCan de 2021 contient 95,5 % des enregistrements de l'échantillon du questionnaire détaillé du Recensement de 2021. Les enregistrements des répondants au recensement qui n'ont pu être appariés aux données du Dépôt d'enregistrements dérivés (DED) dans l'Environnement de couplage de données sociales (ECDS) et les enregistrements présumés être des doublons, identifiés au moyen d'un couplage interne, représentent les 4,5 % restants. La version initiale de la CSERCan de 2021, publiée le 27 novembre 2024, était fondée sur un couplage des données de l'échantillon du questionnaire détaillé du Recensement de 2021 aux données de la Base canadienne de données de l'état civil – Décès (du 11 mai 2021 au 31 décembre 2022), ainsi qu'aux codes postaux annuels pour les adresses postales (1981 à 2022). La CSERCan de 2021 était très attendue, car le Recensement de 2021 comprend du nouveau contenu qui permet de traiter de tendances et d'enjeux émergents (sexe à la naissance et genre, expérience dans les forces militaires canadiennes, etc.) (Statistique Canada, 2022).

Pour permettre aux chercheurs d'estimer et d'évaluer la variance des mesures agrégées de la santé de la population lorsqu'il manque des appariements entre l'échantillon du questionnaire détaillé du Recensement de 2021 et les données du DED, les poids principaux et de rééchantillonnage produits et associés aux membres des ménages ayant rempli le questionnaire détaillé du recensement ont été ajustés et calés pour produire des poids principaux et de rééchantillonnage rajustés en fonction du couplage. Les décisions prises au moment de la pondération peuvent avoir des effets sur l'exactitude des estimations pour les petits groupes de population, comme les hommes transgenres, les femmes transgenres et les personnes non binaires. Lors du Recensement de 2021, pour la toute première fois, des données ont été recueillies spécifiquement à propos du genre du répondant, un concept différent de celui du sexe à la naissance, rendant possible l'examen des résultats en matière de santé chez les personnes de la diversité de genre. Les nouveaux points de données appuient les efforts du gouvernement et de la société en général visant à combler les lacunes connues en matière de données et de connaissances, à remédier aux inégalités et à appuyer la prise de décisions

¹ Yubin Sung, Statistique Canada, 100 promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, yubin.sung@statcan.gc.ca; Eric Hortop, Statistique Canada, 100 promenade Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6, eric.hortop@statcan.gc.ca

justes et inclusives. Nous comparons les comptes de population estimés au moyen des données de la CSERCan de 2021 avec les comptes de population fondés sur les données d'échantillon du questionnaire détaillé du Recensement de 2021, selon le statut de diversité de genre et certaines caractéristiques démographiques, sociales et économiques, dans le but d'évaluer les méthodes de pondération que nous utilisons.

2. Méthodologie

2.1 Processus de pondération

Une façon d'examiner les réalités changeantes et les inégalités vécues par divers groupes de personnes est au moyen de données désagrégées détaillées, et la diversité de genre est une dimension essentielle de cette désagrégation. Cependant, il est important d'être conscient des biais et des erreurs statistiques lors de l'analyse de petites sous-populations. Les biais (et erreurs) de couplage sont préoccupants, car des résultats biaisés peuvent amener les chercheurs à tirer des conclusions non fondées à l'égard de la population. Notre analyse révèle que 88,4 % des hommes transgenres et 89,7 % des femmes transgenres énumérés au moyen des données d'échantillon du questionnaire détaillé du Recensement 2021 ont pu être appariés aux données du DED. Ces taux sont inférieurs à ceux observés pour les hommes cisgenres (95,8 %), les femmes cisgenres (95,8 %) et les personnes non binaires (92,4 %). Pour améliorer les estimations produites au moyen de la CSERCan de 2021, nous décrivons brièvement la façon dont les méthodes de pondération utilisées pour ajuster les données au niveau de la population ont été modifiées pour que les estimations soient représentatives de la population cible des ménages privés.

Pour le Recensement de 2021, dans l'ensemble, les poids principaux (finaux) ont produit des écarts minimes entre les comptes de population fondés sur les données du questionnaire abrégé du recensement (c.-à-d. comptes du recensement) et les comptes fondés sur les données d'échantillon du questionnaire détaillé du recensement (c.-à-d. estimations du questionnaire détaillé du recensement) (Statistique Canada, 2023). Pour obtenir les estimations des comptes de population fondées sur les données de la CSERCan de 2021 (c.-à-d. estimations de la CSERCan), les poids principaux produits pour les estimations du questionnaire détaillé ont été ajustés et calés.

L'estimateur de rééchantillonnage choisi pour l'échantillon du questionnaire détaillé du Recensement de 2021 est dérivé de la méthode des demi-échantillons équilibrés de Fay, qui détermine la création de répliques, le calcul des poids de répliques ainsi que le facteur multiplicatif utilisé pour estimer la variance (Statistique Canada, 2023). Afin de produire les estimations de variance pour les estimations de la CSERCan, l'ensemble des 100 poids de répliques produits pour estimer la variance des estimations de l'échantillon du questionnaire détaillé du recensement a fait l'objet d'un ajustement et d'un calage.

Hortop (2024) décrit en détail la méthode de pondération utilisée pour ajuster et caler les poids principaux et les poids fondés sur le rééchantillonnage équilibré de Fay pour la CSERCan de 2021. Il s'agit d'une modification de la méthode de pondération élaborée par Hortop et Aubin (2020) et appliquée à la CSERCan de 2016. Les processus de pondération utilisés pour créer les poids principaux et les poids de répliques de la CSERCan de 2021 se sont déroulés en deux étapes :

1. Tout d'abord, des groupes de réponse homogènes ont été créés à l'aide du score de propension du couplage obtenu d'un modèle logistique ajusté à l'échantillon du questionnaire détaillé pour ajuster les poids principaux et les poids de répliques.
2. Ensuite, un processus de calage des cellules a été utilisé pour ajuster davantage les poids principaux et de rééchantillonnage de sorte que les comptes pondérés tirés d'une totalisation multivariée de la CSERCan de 2021 correspondent aux totaux pondérés tirés de l'échantillon du questionnaire détaillé du Recensement de 2021. Les variables de la totalisation comprennent, entre autres, la variable du genre à deux catégories (hommes+ et femmes+). Hommes+ comprend les hommes (et/ou les garçons), de même que certaines personnes non binaires. Femmes+ comprend les femmes (et/ou les filles), de même que certaines personnes non binaires).

Dans un premier temps, nous avons supposé une relation linéaire entre les probabilités logarithmiques de l'évènement selon lequel les enregistrements de l'échantillon du questionnaire détaillé du Recensement de 2021 sur le formulaire

long du recensement ont été inclus dans la CSERCan de 2021 (c'est-à-dire couplé avec le DED et les doublons ont été supprimés) et les variables indépendantes suivantes :

- Âge,
- Province ou territoire de résidence le jour du recensement,
- Genre (deux catégories ou cinq catégories),
- Fréquentation scolaire,
- Genre de questionnaire (classification) (questionnaire détaillé 2A-L ou questionnaire détaillé 2A-R),
- Année d'immigration,
- Identité autochtone,
- Mobilité : Lieu de résidence 1 an auparavant (2020),
- Peuples autochtones, groupes racialisés et population non racialisée, non autochtone
- Travail : situation d'activité

Après avoir effectué l'ajustement des modèles, les scores de propension prédits pour la relation ont été calculés pour chaque enregistrement de l'échantillon du questionnaire détaillé du recensement. Deux ensembles de groupes de réponse homogènes ont été créés au moyen des scores de propension du couplage de deux modèles logistiques adaptés pour l'échantillon du questionnaire détaillé, afin d'ajuster les poids principaux et les poids de répliques. Le premier modèle logistique (modèle G-2) a été utilisé pour estimer les probabilités de couplage selon un ensemble donné de variables indépendantes qui incluent la variable du genre à deux catégories et d'autres caractéristiques. Pour le deuxième modèle logistique (modèle G-5), la catégorie de genre à cinq catégories ([hommes cisgenres](#), [femmes cisgenres](#), [hommes transgenres](#), [femmes transgenres](#) et [personnes non binaires](#)) a été utilisée au lieu de la variable du genre à deux catégories dans les variables indépendantes. Les groupes de réponses homogènes ont été constitués pour des plages de probabilité prédites telles que :

- Chaque groupe contenait au moins 1 000: enregistrements au total,
- Le facteur d'ajustement du poids doit être inférieur à 2,
- La différence entre les extrémités de l'intervalle était supérieure à la valeur suivante :

$$\begin{cases} 0.05 & \text{if } (1 - p_\ell)/2 > 0.05, \\ 0.01 & \text{if } (1 - p_\ell)/2 < 0.01, \\ (1 - p_\ell)/2 & \text{autrement,} \end{cases}$$

où p_ℓ est la probabilité de couplage la plus faible observée dans la plage. Ce processus a permis d'obtenir 15 groupes de réponses homogènes pour les deux ensembles.

Dans la deuxième étape, les cellules d'étalonnage étaient le produit cartésien des variables d'étalonnage. Le même ensemble de variables d'étalonnage a été utilisé pour la procédure d'étalonnage des cellules et elles étaient :

- Groupe d'âge (12 catégories),
- Genre (deux catégories),
- Province ou territoire de résidence le jour du recensement (11 catégories, les territoires étant regroupés),
- Indicateur d'identité autochtone,
- Indicateur d'immigration.

Cette simple modification apportée à la méthode de modélisation fait en sorte que notre processus de pondération est plus pratique, car un processus normalisé de poststratification par regroupement de cellules en fonction des seuils de taille de l'échantillon peut être sensible à une hausse ou à une baisse de la variance estimée, entraînant ainsi un accroissement des biais (Kim et coll., 2007).

Dans le présent manuscrit, nous comparons deux ensembles d'estimations de la CSERCan (produites au moyen des poids principaux et des poids de répliques ajustés à l'aide des deux modèles G-2 et G-5, puis calés à l'aide du processus de calage des cellules) avec les estimations du questionnaire détaillé du recensement. Les intervalles de confiance ont été construits comme indicateur de la qualité fondée sur la variance pour accompagner les estimations du questionnaire

détaillé du recensement, ainsi que les deux ensembles d'estimations de la CSERCan. La méthode d'intervalle de confiance utilisée pour les comptes est décrite à la section 2.2.

2.2 Intervalle de confiance modifié de Wilson pour les comptes

Pour les estimations tirées du questionnaire détaillé du recensement, Statistique Canada (2023) recommande d'utiliser la méthode d'intervalle de confiance modifié de Wilson, en raison de sa couverture généralement supérieure (Neusy et coll., 2021) et des aspects pratiques de sa mise en œuvre. Pour les comptes selon le statut de diversité de genre et selon une variable d'intérêt secondaire, en raison de la petite taille des échantillons, les intervalles de confiance de Wald et de Student produiraient de piètres résultats pour les hommes transgenres, les femmes transgenres et les personnes non binaires. Pour la CSERCan de 2021, la méthode d'intervalle de confiance modifié de Wilson a été utilisée pour les comptes estimés. La borne inférieure et la borne supérieure de l'intervalle de confiance modifié de Wilson de 95 % pour un compte Y sont données par :

$$LB = \hat{Y} + t^2 \frac{1}{2} \frac{\widehat{Var}(\hat{Y})}{\hat{Y}} - \sqrt{t^2 \hat{Y}(\hat{Y}) + \left[t^2 \frac{1}{2} \frac{\widehat{Var}(\hat{Y})}{\hat{Y}} \right]^2},$$

$$UB = \hat{Y} + t^2 \frac{1}{2} \frac{\widehat{Var}(\hat{Y})}{\hat{Y}} + \sqrt{t^2 \hat{Y}(\hat{Y}) + \left[t^2 \frac{1}{2} \frac{\widehat{Var}(\hat{Y})}{\hat{Y}} \right]^2},$$

où

- $\hat{Y} = \sum_{k \in S} w_k$ est une estimation de Y ,
- t est le 97,5^e centile de la distribution t de Student à R degrés de liberté,
- $\widehat{Var}(\hat{Y})$ est la variance estimée de $\hat{Y}$. L'équation pour l'estimateur de la variance $\hat{Y}$ d'un compte Y pour un ensemble de R répliques est illustrée ci-dessous :

$$\widehat{Var}(\hat{Y}) = \frac{1}{R/2} \sum_{r=1}^R (\hat{Y}^{(r)} - \hat{Y})^2,$$

où

- $\hat{Y} = \sum_{k \in S} w_k$,
- $\hat{Y}^{(r)} = \sum_{k \in S} w_k^{(r)}$,
- $R = 100$ poids de répliques.

Le poids principal de la CSERCan est représenté par w_k , $w_k^{(r)}$ est le poids de répliques de la CSERCan $r = 1, \dots, R$, k est un membre de la cohorte ayant des caractéristiques d'intérêt (p. ex. intersection de la race noire et d'une identité non binaire), et s est la CSERCan (c.-à-d. un sous-ensemble de l'échantillon du questionnaire détaillé du recensement).

Il est à noter que les règles de confidentialité pour la CSERCan de 2021 ont été élaborées en fonction des normes relatives à la confidentialité ainsi que des lignes directrices sur la diffusion des données du Recensement de la population de 2021 et des données administratives. Diverses règles de confidentialité ont été appliquées aux estimations présentées dans ce manuscrit afin d'éviter la publication ou la divulgation de toute information jugée confidentielle.

3. Résultats

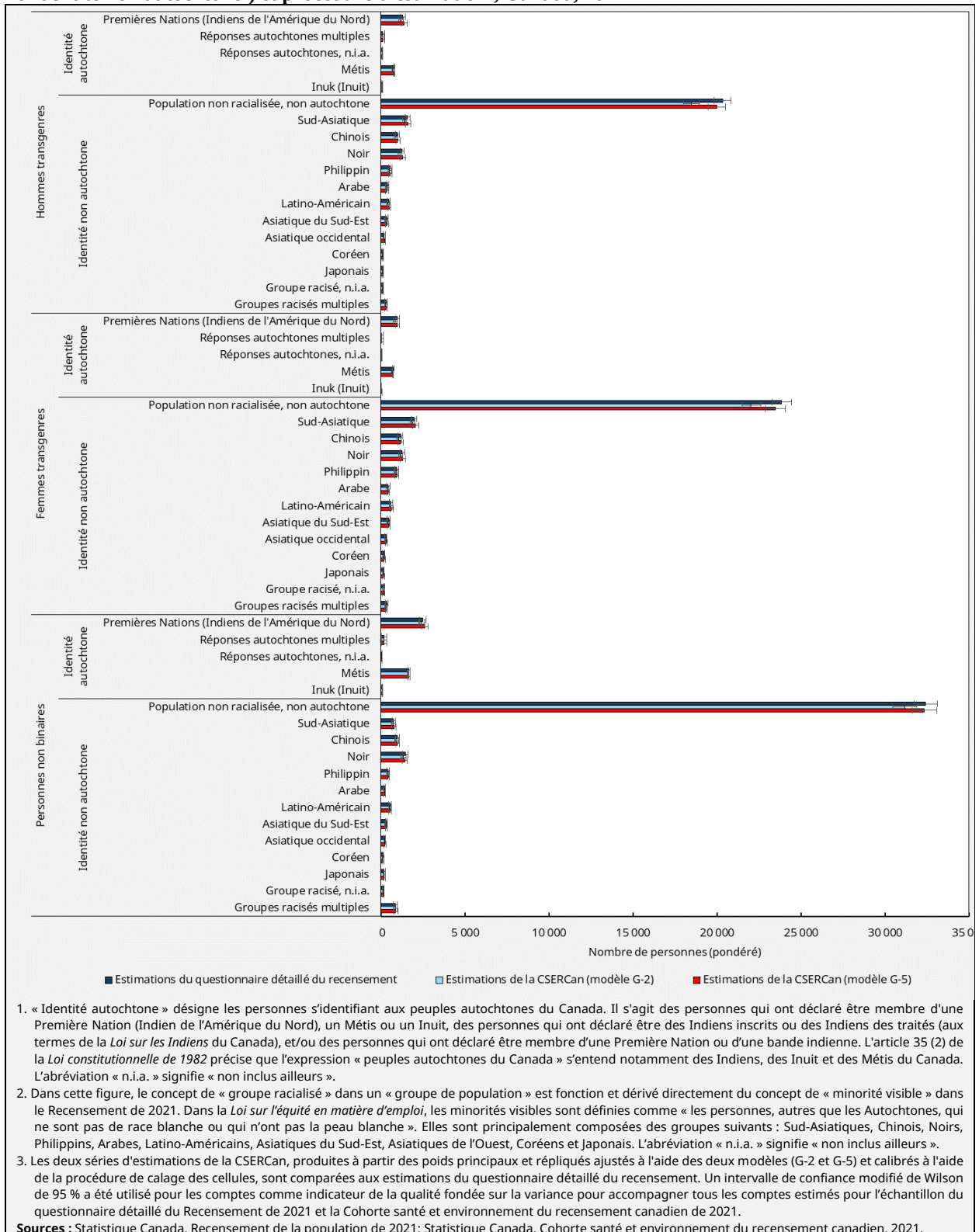
Le compte estimé d'hommes transgenres, de femmes transgenres et de personnes non binaires de 15 ans ou plus au Canada dans les ménages privés, par groupe de population (population d'identité autochtone et population d'identité non autochtone) et par procédure d'estimation, est présenté à la figure 3-1. Les intervalles de confiance modifiés de Wilson de 95 % pour les comptes sont aussi présentés avec des moustaches au-dessus des barres. Étant donné que les différences entre les estimations du compte d'hommes cisgenres et de femmes cisgenres âgés de 15 ans et plus au Canada, dans les ménages privés, par groupe de population et par procédure d'estimation, sont négligeables et ne sont donc pas indiquées.

Pour chaque intersection des variables présentées dans la figure 3-1, les largeurs des intervalles de confiance autour des trois estimations sont similaires. De plus, en nous basant sur la longueur des intervalles de confiance qui se chevauchent entre l'estimation du questionnaire détaillé du recensement et chacune des estimations de la CSERCan (modèles G-2 et G-5), nous avons constaté que la prise en compte de la diversité de genre (c'est-à-dire la variable de genre à cinq catégories) lors de l'ajustement des poids principaux et répliqués dans la procédure de modélisation permettait généralement une meilleure concordance entre les estimations du questionnaire détaillé du recensement et les estimations de la CSERCan (modèle G-5). Les différences entre les estimations du questionnaire détaillé du recensement et les estimations de la CSERCan (modèle G-2) sont particulièrement notables chez les hommes transgenres non racialisés et non autochtones, les femmes transgenres et les personnes non binaires, où les intervalles de confiance pour tous, à l'exception des personnes non binaires non racialisées et non autochtones, ne se chevauchent pas.

Notre analyse a également comparé les estimations du compte d'hommes et de femmes cisgenres, d'hommes et de femmes transgenres et de personnes non binaires âgées de 15 ans et plus au Canada, dans les ménages privés, pour chacune des variables du questionnaire détaillé du recensement et de la procédure d'estimation suivantes : statut d'immigrant (trois catégories), plus haut certificat, diplôme ou grade (six catégories) et travail (situation d'activité) (trois catégories, dérivées). Là encore, la prise en compte de la diversité des sexes lors de l'ajustement des poids principaux et répliqués dans la procédure de modélisation a permis d'obtenir un meilleur accord général entre les estimations du questionnaire détaillé du recensement et les estimations de la cohorte (modèle G-5).

Figure 3-1

Comptes estimés d'hommes transgenres, de femmes transgenres et de personnes non binaires de 15 ans ou plus dans les ménages privés, par groupe de population (population d'identité autochtone¹ et population d'identité non autochtone²) et procédure d'estimation³, Canada, 2021



4. Discussion

Le présent manuscrit souligne l'importance d'évaluer régulièrement la qualité des couplages d'enregistrements et d'examiner les processus de pondération subséquents. Nous avons constaté que des décisions concernant la pondération peuvent avoir une incidence sur l'exactitude des estimations pour les petits groupes de population, dont les personnes de la diversité de genre. Nous avons comparé les estimations de la CSERCan de 2021 avec les estimations du questionnaire détaillé du Recensement de la population, à l'aide des données recueillies au moyen du questionnaire détaillé du Recensement de 2021 (questionnaires 2A-L et 2A-R). Selon nos observations, lorsque la diversité de genre (c.-à-d. la variable de genre à cinq catégories) est prise en compte dans l'ajustement des poids principaux et de rééchantillonnage dans le cadre de la modélisation, cela permet d'obtenir une concordance généralement meilleure entre les estimations du questionnaire détaillé du recensement et les estimations de la CSERCan (modèle G-5).

Une limite connue de notre processus de modélisation de la propension et de calibration est que les poids principaux et les poids de répliques ne prennent en compte que le couplage des données d'échantillon du questionnaire détaillé du recensement et des données du DED, alors que de nombreuses analyses de la santé de la population devraient tenir compte de couplages additionnels avec les données administratives sur la santé.

Bibliographie

- Hortop E. (2024), « Adjusting and calibrating Fay's Balanced Repeated Replication (BRR) weights for the 2021 Canadian Census Health and Environment Cohort (CanCHEC) », rapport interne, Statistique Canada, Ottawa, Ontario.
- Hortop E. et Aubin P. (2020), « Adjusting and calibrating Fay's Balanced Repeated Replication (BRR) weights for the 2016 Canadian Census Health and Environment Cohort (CanCHEC) », rapport interne, Statistique Canada, Ottawa, Ontario.
- Kim, J. J., J. Li. et R. Valliant. (2007), « Regroupement de cellules dans la poststratification », *Techniques d'enquête*, 32(2), p. 139-150, produit n° 12-001-X au catalogue de Statistique Canada, Ottawa, Ontario.
- Neusy E., S.-A. Savard, M. Hidioglou et V. Martin. (2021), « Modified Wilson Intervals for Estimated Counts with Application to Census 2021 Long Form Estimation », présentation donnée au Comité consultatif des méthodes statistiques, mai 2021, rapport interne, Statistique Canada, Ottawa, Ontario.
- Statistique Canada. (2022), *Guide du Recensement de la population, 2021*, produit n° 98-304-X2021001 au catalogue de Statistique Canada, Ottawa, Ontario. Version mise à jour le 30 novembre 2022.
- Statistique Canada. (2023), *Rapport technique sur l'échantillonnage et la pondération, Recensement de la population, 2021*, produit n° 98-306-X2016001 au catalogue de Statistique Canada, Ottawa, Ontario.