

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

### **Inclusion du genre non binaire dans la stratégie de calage pour les poids d'enquête de l'échantillon du questionnaire détaillé canadien**

par Alexander Imbrogno

Date de diffusion : le 8 septembre 2025



Statistique  
Canada

Statistics  
Canada

Canada

## Inclusion du genre non binaire dans la stratégie de calage pour les poids d'enquête de l'échantillon du questionnaire détaillé canadien

Alexander Imbrogno<sup>1</sup>

### Résumé

Pour répondre aux besoins récents en matière de données désagrégées, en 2021, le Canada est devenu le premier pays au monde à recueillir et à diffuser des données sur la diversité de genre, tirées d'un recensement national offrant à la population canadienne la possibilité de déclarer être de genre masculin, féminin ou non binaire. En raison de leur petit nombre et du risque d'augmentation de la variance des estimations, les chiffres de population non binaire n'ont pas été utilisés dans la procédure de calage de l'échantillon du questionnaire détaillé du Recensement de 2021. La présente étude propose une nouvelle stratégie de calage pour le questionnaire détaillé permettant d'intégrer des populations de petite taille, comme la population non binaire, tout en traitant des préoccupations méthodologiques. La stratégie présentée peut intégrer plusieurs petites populations simultanément, tout en demeurant suffisamment souple pour s'adapter aux systèmes de calage d'autres organismes nationaux de statistique. Les résultats d'une simulation Monte Carlo sont présentés afin de démontrer l'amélioration de la qualité des données pour la population non binaire dans le cadre de cette stratégie de calage de rechange.

Mots clés : Calage linéaire; données désagrégées; Recensement de la population du Canada; optimisation par décomposition.

### 1. Introduction

Au cours des dernières années, les organismes nationaux de statistique (ONS) et les gouvernements ont exprimé le besoin de données désagrégées sur les groupes marginalisés et les minorités afin de mieux traiter des inégalités grâce à l'élaboration de politiques fondées sur des données (UNCEB, 2017). Pour répondre à ce besoin, en 2021, le Canada est devenu le premier pays à recueillir et à diffuser des données fondées sur la diversité de genre tirées d'un recensement national. Le Recensement de la population du Canada brosse un portrait statistique détaillé de la population canadienne et joue un rôle important dans l'élaboration de politiques. Réalisé tous les cinq ans (le cycle de 2021 étant le plus récent), le recensement comprend deux volets. Le premier est un recensement qui vise à dénombrer l'ensemble de la population et à recueillir des renseignements démographiques de base. Le deuxième est une enquête par sondage, appelée l'échantillon du questionnaire détaillé, qui recueille des données sociales et économiques plus détaillées auprès d'un échantillon d'environ 25 % de la population. En 2021, la volet du recensement comprenait une question sur le genre, fondée sur les catégories cisgenre, transgenre et non binaire.

À l'étape de l'estimation de l'échantillon du questionnaire détaillé, les poids d'échantillonnage font l'objet d'une série d'ajustements de la pondération, afin de corriger le biais de couverture et de non-réponse et de réduire la variance au moyen d'une procédure de calage linéaire final. Le processus de pondération utilise les chiffres de population obtenus du recensement pour améliorer les estimations pondérées produites à partir de l'échantillon du questionnaire détaillé. Même si l'état de genre non binaire est une caractéristique du recensement, il n'a pas été utilisé dans le processus de pondération en raison de la petite taille de la population non binaire. Le calage sur des caractéristiques rares de la population peut donner lieu à des poids de calage extrêmes (grands ou petits); ce qui peut accroître la variance des estimations pondérées.

---

<sup>1</sup> Alexander Imbrogno, Statistique Canada, 170, Pré Tunney, Canada, K1A 0T6 ([alexander.imbrogno@statcan.gc.ca](mailto:alexander.imbrogno@statcan.gc.ca)).

En réponse à ces préoccupations, une variable de genre à deux catégories (hommes+ et femmes+) a été utilisée en 2021.

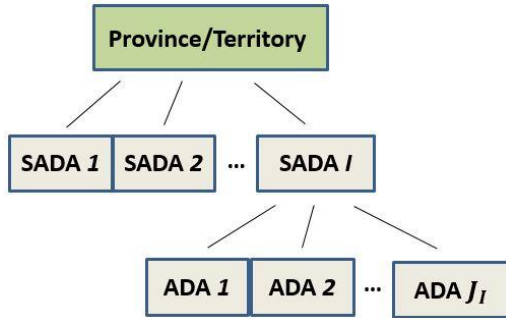
Le présent article propose une nouvelle stratégie de calage final pour intégrer les petites populations au processus de calage du questionnaire détaillé du recensement tout en traitant des préoccupations d'ordre méthodologique. L'application de la stratégie proposée à la population non binaire est démontrée. La solution peut être étendue pour intégrer plusieurs petites populations simultanément tout en étant adaptable aux systèmes de calage d'autres ONS. Le reste de l'article est organisé comme suit : dans la deuxième partie, le processus habituel de calage du questionnaire détaillé est décrit; dans la troisième, la nouvelle stratégie de calage proposée est présentée; dans la quatrième, les résultats d'une simulation MC examinant l'incidence du calage sur la qualité des données pour des chiffres de population non binaires est décrite; en dans la cinquième partie, les principales conclusions sont présentées.

## 2. Stratégie d'estimation du questionnaire détaillé

Lors du calage du questionnaire détaillé, les poids sont ajustés pour réduire la variance d'échantillonnage et assurer la cohérence numérique des estimations par rapport aux chiffres de recensement connus. Le calage est effectué de façon indépendante au sein des super aires de diffusion agrégées (SADA). Il s'agit de régions géographiques contiguës créées pour le processus de pondération du recensement. Les aires de diffusion agrégées (ADA), qui sont des subdivisions des SADA, sont également utilisées comme pondération géographique.

**Figure 2-1**

**Représentation hiérarchique des régions géographiques standard utilisées dans le processus de calage du recensement canadien**



Un processus de sélection est mené dans chaque SADA/ADA pour sélectionner l'ensemble des caractéristiques de la population, appelées contraintes de calage, en fonction desquelles effectuer le calage. Un ensemble de contraintes est sélectionné pour chaque SADA/ADA. Les ADA étant intégrées aux SADA, les contraintes des SADA sont utilisées pour désigner les contraintes des SADA et des ADA. Une fois les contraintes sélectionnées, les poids calés pour chaque SADA sont obtenus en résolvant le problème de calage linéaire énoncé par Deville et Särndal (1992) :

$$\text{minimiser } \sum_{k \in s_i} \frac{(w_k - d_k)^2}{2d_k} \quad (1)$$

$$\text{en fonction de } \sum_{k \in s_i} w_k x_k^{(i)} = t_{x^{(i)}} \text{ et } 1 \leq w_k \leq 20,$$

où  $s_i$  est l'ensemble de ménages faisant l'objet d'un calage dans la SADA  $i$ , pour les ménages  $k$ ;  $w_k$  est le poids calé,  $d_k$  est le poids avant calage,  $x_k^{(i)}$  est le vecteur des variables de calage et  $t_{x^{(i)}}$  est le vecteur des chiffres du recensement pour  $x_k^{(i)}$ . Pour ajouter une protection supplémentaire contre les poids de calage extrêmes, chaque poids calé est limité

entre 1 et 20. Le poids de calage obtenu peut être écrit comme étant le poids avant calage multiplié par un facteur d'ajustement reposant sur  $\mathbf{x}_k^{(i)}$  et les paramètres du modèle estimés  $0 < \hat{\Phi}_k \leq 1$  et  $\hat{\lambda}_k$  :

$$w_k = d_k (1 + \hat{\Phi}_k \hat{\lambda}_k^T \mathbf{x}_k^{(i)}). \quad (2)$$

L'algorithme décrit par Singh et Mohl (1996) peut être utilisé pour estimer les paramètres  $\Phi_k$  et  $\lambda_k$ . Le problème du calage du questionnaire détaillé est résolu à l'aide du Système d'estimation généralisé (SGE), outil interne mis au point par Statistique Canada.

### 3. Calage des poids du questionnaire détaillé en fonction des totaux de population non binaire

Au lieu de caler les poids en fonction des totaux des personnes non binaires des SADA et ADA, la stratégie proposée cale les poids en fonction des totaux des personnes non binaires au niveau provincial/territorial. Les provinces et territoires étant des agrégats de SADA, cela atténue le risque d'augmentation de la variance en raison de la taille accrue du groupe de calage. Cette stratégie maintient les contraintes au niveau de la SADA parallèlement à cette nouvelle contrainte au niveau provincial, déplaçant le problème de calage de la SADA au niveau provincial. Ce problème au niveau provincial peut être résolu à l'aide de méthodes d'optimisation standard, mais pour de nombreuses provinces, l'augmentation de la taille du calage exige une puissance et des ressources de calcul dépassant celles disponibles dans l'environnement de calage du questionnaire détaillé. Cette question est abordée à la section 3.2, qui présente un schéma de décomposition pour subdiviser le problème de grandeur provinciale en problèmes au niveau de SADA indépendantes de taille gérable.

#### 3.1 Problème de calage non binaire au niveau provincial

Soit  $s^{(p)}$ , l'ensemble des ménages faisant l'objet d'un calage dans une province donnée et  $I$ , le nombre de SADA dans la province. Les poids de calage sont la solution au problème d'optimisation suivant :

$$\begin{aligned} & \text{minimiser} \sum_{k \in s^{(p)}} \frac{(w_k - d_k)^2}{2d_k}, \\ & \text{en fonction de} \sum_{k \in s^{(p)}} w_k \tilde{\mathbf{x}}_k = \mathbf{t}_{\tilde{\mathbf{x}}} \text{ et } 1 \leq w_k \leq 20. \end{aligned} \quad (3)$$

Pour le ménage  $k$  dans la SADA  $i$ ,  $\tilde{\mathbf{x}}_k = [x_{k*}, 0, \dots, 0, \mathbf{x}_k^{(i)T}, 0, \dots, 0]^T$  est un vecteur contenant la variable d'état non binaire dans le premier terme et le vecteur des variables de calage au niveau de la SADA dans le terme  $i + 1^{\text{e}}$ . Le vecteur total  $\mathbf{t}_{\tilde{\mathbf{x}}} = [t_*, \mathbf{t}_{x(1)}^T, \mathbf{t}_{x(2)}^T, \dots, \mathbf{t}_{x(I)}^T]^T$  contient le total provincial des personnes non binaires dans le premier terme et les vecteurs totaux au niveau de la SADA pour chaque SADA de la province.

#### 3.2 Décomposition du problème de calage provincial

Le schéma de décomposition repose, en premier lieu, sur la reformulation de la contrainte provinciale relative aux personnes non binaires en une somme des contraintes au niveau de la SADA :

$$\sum_{k \in s^{(p)}} w_k x_{k*} - t_* = \sum_{i=1}^I \left\{ \sum_{k \in s_i} w_k x_{k*} - t_{i*} \right\} = 0. \quad (4)$$

Les quantités  $t_{i*}$ ,  $i = 1, \dots, I$  sont appelées totaux artificiels de SADA des personnes non binaires, ou totaux artificiels. Pour que (4) soit satisfait, les totaux artificiels doivent être égaux au total provincial,  $\sum_{i=1}^I t_{i*} = t_*$ .

Le terme « totaux artificiels » est utilisé parce que les  $t_{i*}$  ne visent pas à représenter les totaux réels au niveau de la SADA. Les totaux artificiels sont considérés comme un outil pour subdiviser la contrainte provinciale en contraintes indépendantes au niveau de la SADA; ce qui permet la décomposition du calage. Dans chaque SADA, la contrainte suivante est introduite :

$$\sum_{k \in S_i} w_k x_{k*} = t_{i*} \quad (5)$$

Le problème au niveau provincial se décompose en problèmes au niveau de la SADA en augmentant le calage de la SADA et les vecteurs totaux respectivement avec la variable indicatrice non binaire et le total artificiel. Le problème de calage en décomposition pour la SADA  $i$  est :

$$\text{minimiser } \sum_{k \in S_i} \frac{(w_k - d_k)^2}{2d_k} \quad (6)$$

$$\text{en fonction de } \sum_{k \in S_i} w_k x_{k*}^{(i)} = t_*^{(i)} \text{ et } 1 \leq w_k \leq 20,$$

où  $x_{k*}^{(i)} = [x_{k*}, x_k^{(i)T}]^T$  et  $t_*^{(i)} = [t_{i*}, t_x^{(i)T}]^T$ . Si  $\sum_{i=1}^I t_{i*} = t_*$ , le calage en fonction des totaux artificiels assure le calage en fonction du total provincial. Toutefois, pour tout choix arbitraire de  $\{t_{i*}\}_{i=1}^I$  résolvant  $\sum_{i=1}^I t_{i*} = t_*$ , il n'est pas garanti que les poids de calage obtenus à partir des problèmes décomposés soient égaux aux poids qui auraient été obtenus en résolvant le problème provincial. En fait, nos études ont montré que le choix des  $t_{i*}$ , qui sont loin des totaux estimés non calés au niveau de la SADA, peut entraîner des poids de calage extrêmes. Dans la section suivante, les totaux artificiels sont calculés pour veiller à ce que les poids de calage obtenus à partir des problèmes décomposés au niveau de la SADA soient égaux aux poids de calage qui auraient été obtenus en résolvant le problème provincial.

### 3.3 Calcul des totaux artificiels des personnes non binaires au niveau de la SADA

Cette dérivation est effectuée indépendamment dans chaque SADA en établissant la solution aux poids de calage obtenus à partir des problèmes au niveau de la SADA et de la province, puis en résolvant le problème de  $t_*^{(i)}$ . Il convient de souligner que trouver  $\{t_{i*}\}_{i=1}^I$  de cette façon résout implicitement  $\sum_{i=1}^I t_{i*} = t_*$ . Pour faciliter la dérivation,  $x_{k*}^{(i)}$  et  $t_*^{(i)}$  sont reformulés de sorte que leur dimension corresponde à leurs homologues du problème au niveau provincial. Les vecteurs de l'unité  $k$  dans la SADA  $i$  sont reformulés comme  $\tilde{x}_k = [x_{k*}, 0, \dots, 0, x_k^{(i)T}, 0, \dots, 0]^T$  et  $t_*^{(i)} = [t_{i*}, 0, \dots, 0, t_x^{(i)T}, 0, \dots, 0]^T$ , où le premier terme contient la contrainte artificielle pour les personnes non binaires et les contraintes au niveau de la SADA se trouvent dans le terme  $i + 1^e$ . Les autres éléments sont 0. À l'aide de l'équation (2), les formules des poids de calage sont définies comme étant égales entre elles. Les composantes appartenant au problème au niveau provincial sont désignées en utilisant  $p$  et les composantes du problème au niveau de la SADA, en utilisant  $s$ . Pour les ménages  $k$  dans la SADA  $i$ , l'égalité suivante est obtenue :

$$w_{i,k}^{(p)} = w_{i,k}^{(s)} \Leftrightarrow d_k (1 + \hat{\Phi}_{i,k}^{(p)} \hat{\lambda}_{i,k}^{(p)T} \tilde{x}_k) = d_k (1 + \hat{\Phi}_{i,k}^{(s)} \hat{\lambda}_{i,k}^{(s)T} \tilde{x}_k) \Leftrightarrow \hat{\Phi}_{i,k}^{(p)} \hat{\lambda}_{i,k}^{(p)} = \hat{\Phi}_{i,k}^{(s)} \hat{\lambda}_{i,k}^{(s)} \quad (7)$$

Il peut être démontré que les paramètres estimés  $\hat{\Phi}_{i,k}$  et  $\hat{\lambda}_{i,k}$  pour les problèmes au niveau de la SADA et de la province sont égaux, si le total artificiel garantit que les poids calés à partir des problèmes sans contraintes de plage sont égaux. Soient  $w_{i,k}^{-(p)}$  et  $w_{i,k}^{-(s)}$ , les poids calés pour le ménage  $k$  dans la SADA  $i$  obtenus à partir des problèmes au niveau provincial et de la SADA sans contraintes de plage. À l'aide de la solution de Deville et Särndal (1992), il s'ensuit que

$$w_{i,k}^{-(p)} = w_{i,k}^{-(s)} \Leftrightarrow \left[ \sum_{k \in S^{(p)}} d_k \tilde{x}_k \tilde{x}_k^T \right]^{-1} (t_{\tilde{x}} - \hat{t}_{\tilde{x}}) = \left[ \sum_{k \in S_i} d_k \tilde{x}_k \tilde{x}_k^T \right]^{-1} (t_*^{(i)} - \hat{t}_*^{(i)}), \quad (8)$$

où  $\hat{\mathbf{t}}_{\bar{x}}$  et  $\hat{\mathbf{t}}_*^{(i)}$  sont les vecteurs totaux estimés à l'aide des poids avant calage. Résoudre pour  $\mathbf{t}_*^{(i)}$  donne

$$\mathbf{t}_*^{(i)} = \sum_{k \in S_i} d_k \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \left[ \sum_{k \in S(p)} d_k \tilde{\mathbf{x}}_k \tilde{\mathbf{x}}_k^T \right]^{-1} (\mathbf{t}_{\bar{x}} - \hat{\mathbf{t}}_{\bar{x}}) + \hat{\mathbf{t}}_*^{(i)}, \quad (9)$$

qui contient le total artificiel dans le premier terme.

## 4. Simulation de Monte Carlo

### 4.1 Configuration

L'étude de simulation était fondée sur une pseudo-province/SADA créée à l'aide des réponses de l'échantillon du questionnaire détaillé de 2021 obtenues dans la province de la Colombie-Britannique. En tout, 500 échantillons MC ont été tirés de la pseudo-province par échantillonnage aléatoire simple stratifié des ménages sans remplacement en utilisant une fraction d'échantillonnage de 25%. Pour chaque échantillon MC, deux ensembles de poids calés ont été produits. L'un d'eux a été construit en suivant le processus habituel de calage du questionnaire détaillé. Un deuxième ensemble de poids a été produit en calant en fonction de l'ensemble de contraintes sélectionnées par la procédure habituelle plus une contrainte artificielle non binaire au niveau de la SADA dérivée pour chaque échantillon.

Chaque échantillon MC a également fait l'objet de la création de poids de rééchantillonnage correspondant à la méthodologie d'estimation de la variance de répliques répétées équilibrées partiellement  $\varepsilon$  utilisée pour estimer la variance des estimations de l'échantillon du questionnaire détaillé (Devin et Verret, 2016). Les poids de rééchantillonnage ont fait l'objet des mêmes calages que le principal poids d'enquête; ce qui a permis d'étudier l'incidence du calage non binaire à la fois sur les estimations ponctuelles et sur leur erreur-type estimée (ET).

### 4.2 Résultats

Les figures 4.2-1 et 4.2-2 présentent le biais relatif de Monte Carlo (BRMC) et le coefficient de variation de Monte Carlo (CVMC) pour les estimations ponctuelles et d'ET au niveau de la pseudo-SADA. L'axe des x des tracés correspond à la nouvelle stratégie de calage (*avec valeurs non binaires*) et l'axe des y correspond au calage habituel du questionnaire détaillé (*sans valeurs non binaires*). Une ligne de référence correspondant à  $y = x$  est superposée pour aider le lecteur. On résume les figures comme suit :

- Les deux procédures de calage produisent des estimations ponctuelles approximativement sans biais.
- La nouvelle procédure de calage produit des CVMC plus petits pour les dénombrements et leurs ET estimés. Les CVMC de la moyenne (et ses ET) étaient comparables entre les procédures.
- Les deux procédures de calage produisent des ET estimées ayant un petit BRMC (négatif). Les deux procédures sont comparables.

**Figure 4.2-1**

**BRMC et CVMC des estimations ponctuelles avec et sans calage non binaire au niveau de la pseudo-SADA**

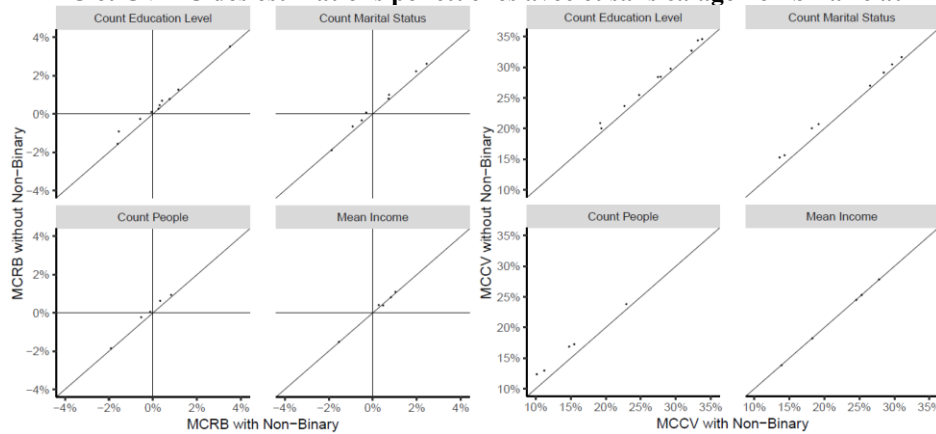
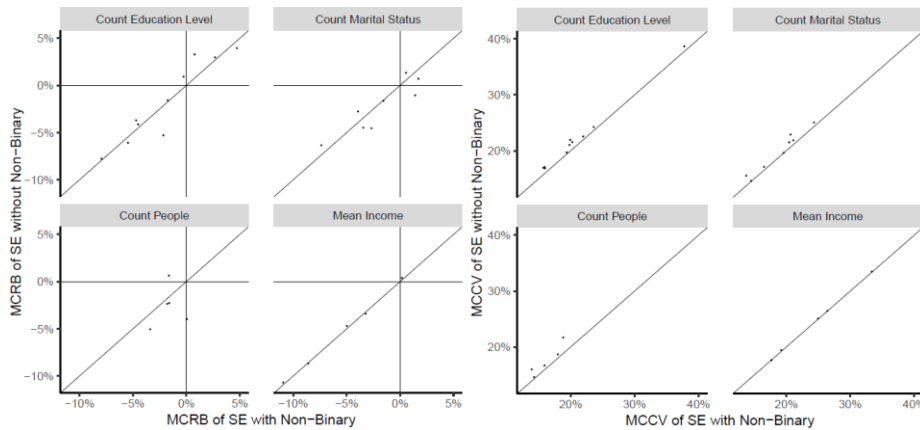


**Figure 4.2-2**

**BRMC et CVMC des ET estimés avec et sans calage non binaire au niveau de la pseudo-SADA**



## 5. Conclusion

Cette étude présente une nouvelle stratégie de calage pour le questionnaire détaillé permettant d'intégrer de petites populations, comme la population non binaire, tout en traitant des préoccupations méthodologiques. La stratégie proposée dérive des totaux artificiels de SADA pour faciliter le calage des poids du questionnaire détaillé sur les chiffres provinciaux des personnes non binaires, en ajoutant une protection contre l'inflation de la variance. Les totaux artificiels permettent la décomposition du grand problème au niveau provincial qui en résulte en problèmes réalisables sur le plan des calculs au niveau de la SADA, en ne nécessitant que des changements minimes au système de calage actuel. La mise en œuvre de cette nouvelle stratégie de calage continue d'être évaluée aux fins d'application à l'échantillon du questionnaire détaillé de 2026 et a suscité de l'intérêt dans le cadre du projet de microsimulation démographique de l'organisme. L'étude de simulation a démontré l'incidence positive du calage non binaire sur la qualité des données pour le domaine des personnes non binaires. Le calage non binaire a entraîné une plus grande précision des estimations ponctuelles et des ET pour les caractéristiques du domaine des personnes non binaires comparativement au calage sans cette contrainte.

## Bibliographie

- Deville, J.-C. et C.-E. Sarndal. (1992), « Calibration estimators in survey sampling », *Journal of the American Statistical Association*, 87 (418), p. 376-382.
- Devin, N. et F. Verret. (2016), « The development of a variance estimation methodology for large-scale dissemination of quality indicators for the 2016 Canadian census long form sample », *JSM Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association*, p. 1977-1991
- Singh, A.C. et C.A. Mohl (1996), « Comprendre les estimateurs de calage dans les enquêtes par échantillonnage », *Techniques d'enquête*, 22, p. 107-116.
- UNCEB (2017), « Leaving no one behind: Equality and non-discrimination at the heart of sustainable development », Organisation des Nations Unies.