

## Proceedings of Statistics Canada Symposium 2024: The Future of Official Statistics

### Including Non-binary Gender in the Calibration Strategy for the Canadian Long-Form Sample Survey Weights

by Alexander Imbrogno

Release date: September 8, 2025



# Including Non-binary Gender in the Calibration Strategy for the Canadian Long-form Sample Survey Weights

Alexander Imbrogno<sup>1</sup>

## Abstract

Aligning with recent needs for increased disaggregated data, in 2021 Canada became the first country to collect and disseminate data on gender diversity in a national census giving Canadians the option to select male, female, or non-binary. Due to their small size, non-binary population counts were not used in the 2021 Census long-form sample calibration procedure due to the risk of increasing the variance of estimates. This paper presents an alternative long-form calibration strategy which allows for small populations, such as the non-binary group, to be incorporated while mitigating methodological concerns. The strategy put forward can incorporate multiple small populations simultaneously while also being flexible enough to fit the calibration systems of other National Statistical Offices (NSOs). The results of a Monte Carlo (MC) simulation are presented showing improved data quality for the non-binary population under the alternative calibration strategy.

Key Words: Linear calibration; Disaggregated data; Canadian Census of Population; Decomposed optimization.

## 1. Introduction

In recent years, National Statistical Offices (NSOs) and governments have expressed the need for disaggregated data on marginalized and minority groups to better address inequalities through data driven policy making (UNCEB, 2017). Aligning with these needs, in 2021 Canada became the first country to collect and disseminate data based on gender diversity in a national census. The Canadian census of population provides a detailed statistical portrait of the Canadian population and plays an important role in policymaking. Conducted every 5 years, 2021 being the most recent, this program consists of 2 parts. The first is a census which enumerates the entire population and collects basic demographic information. The second is a sample survey, called the long-form sample, which collects more detailed social and economic data on approximately 25% of the population. In 2021 the census portion included a question measuring gender using the cisgender, transgender and non-binary categories.

During the estimation stage of the long-form sample, the sampling weights undergo a series of weighting adjustments to correct for coverage and non-response bias as well as to reduce variance through a final linear calibration procedure. The weighting process uses population counts obtained from the Census to improve weighted estimates produced from the long-form sample. Although non-binary gender status is a census characteristic, it was not used in the weighing process due to the small size of the non-binary population. Calibrating on rare population characteristics may result in extreme (large or small) calibration weights which can increase the variance of weighted estimates.

To alleviate these concerns, a 2-category gender variable with categories men+ and women+ was used in 2021.

This paper proposes an alternative final calibration strategy to incorporate small populations into the census long-form calibration process while mitigating the methodological concerns. The application of the proposed strategy to the non-binary population is demonstrated. The solution can be extended to incorporate multiple small populations simultaneously while being adaptable to the calibration systems of other NSOs. The remainder of the paper is organized as follows. Section 2 describes the usual long-form calibration procedure. Section 3 presents the proposed alternative long-form calibration strategy. Section 4 describes the results of a Monte Carlo (MC) simulation which

---

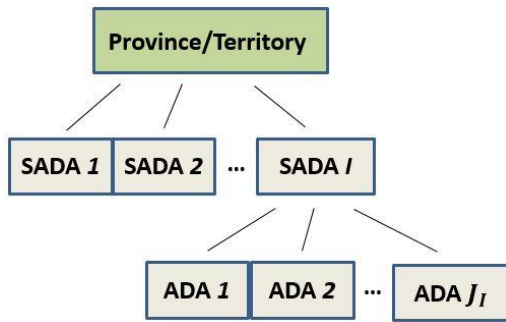
<sup>1</sup> Alexander Imbrogno, Statistics Canada, 170 Tunney's Pasture, Canada, K1A 0T6 ([alexander.imbrogno@statcan.gc.ca](mailto:alexander.imbrogno@statcan.gc.ca)).

studied the impact on data quality of calibration to non-binary population counts. Section 5 summarizes the main conclusions.

## 2. Long-Form Estimation Strategy

During long-form calibration, weights are adjusted to reduce sampling variance and ensure numerical consistency of estimates with known census counts. Calibration is carried out independently within Super Aggregate Dissemination Areas (SADAs). These are contiguous geographic areas created for the census weighting process. Aggregate Dissemination Areas (ADAs), which are partitions of SADAs, are also used as weighting geographies.

**Figure 2-1**  
**Hierarchical representation of the standard geographies used in the Canadian Census calibration process**



A selection process is carried out within each SADA/ADA to select the set of population characteristics, called calibration constraints, to calibrate on. One set of constraints is selected for each SADA/ADA. With ADAs being nested in SADAs, SADA constraints will be used to refer to both SADA and ADA constraints. Once constraints are selected, the calibrated weights for each SADA are obtained by solving the linear calibration problem stated by Deville and Särndal (1992):

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

$$\text{subject to } \sum_{k \in s_i} w_k \mathbf{x}_k^{(i)} = \mathbf{t}_{x^{(i)}} \text{ and } 1 \leq w_k \leq 20,$$

where  $s_i$  is the set of households undergoing calibration in SADA  $i$ , for household  $k$   $w_k$  is the calibrated weight,  $d_k$  is the pre-calibrated weight,  $\mathbf{x}_k^{(i)}$  is the vector of calibration variables and  $\mathbf{t}_{x^{(i)}}$  is the vector of census counts for  $\mathbf{x}_k^{(i)}$ . To add further protection against extreme calibration weights, each calibrated weight is range constrained between 1 and 20. The resulting calibration weight can be written as the pre-calibrated weight times an adjustment factor relying on  $\mathbf{x}_k^{(i)}$  and estimated model parameters  $0 < \hat{\Phi}_k \leq 1$  and  $\hat{\lambda}_k$ :

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

The algorithm described by Singh and Mohl (1996) can be used to estimate the parameters  $\Phi_k$  and  $\lambda_k$ . The long-form calibration problem is solved using the Generalized Estimation System (G-EST), an internal tool developed by Statistics Canada.

## 3. Calibrating Long-Form Weights to Non-Binary Population Totals

Instead of calibrating to SADA and ADA non-binary person totals, the proposed strategy calibrates weights to provincial/territorial non-binary person totals. Since provinces/territories are aggregates of SADAs, this mitigates the risk of increasing the variance due to the increased size of the calibration group. This strategy maintains the SADA-level constraints alongside this new provincial-level constraint, shifting the calibration problem from the SADA to provincial level. This provincial-level problem is solvable using standard optimization methods however, for many provinces, the increased size of the calibration requires computing power and resources which exceeds those available in the long-form calibration environment. This is addressed in section 3.2 which presents a decomposition scheme to partition the large provincial problem back into independent SADA problems of a manageable size.

### 3.1 The Provincial Non-Binary Calibration Problem

Let  $s^{(p)}$  be the set of households undergoing calibration within a given province and  $I$  be the number of SADA in the province. The calibration weights are the solution to the following optimization problem:

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

For household  $k$  in SADA  $i$ ,  $\tilde{\mathbf{x}}_k = [x_{k*}, 0, \dots, 0, \mathbf{x}_k^{(i)T}, 0, \dots, 0]^T$  is a vector containing the non-binary status variable in the first entry and the vector of SADA-level calibration variables in the  $i + 1^{th}$  entry. The total vector  $\mathbf{t}_{\tilde{\mathbf{x}}} = [t_*, \mathbf{t}_{x(1)}^T, \mathbf{t}_{x(2)}^T, \dots, \mathbf{t}_{x(I)}^T]^T$  contains the provincial non-binary person total in the first entry and the SADA-level total vectors for each SADA in the province.

### 3.2 Decomposing the Provincial Calibration Problem

The decomposition scheme relies on first re-writing the provincial non-binary person constraint as a sum of SADA-level constraints:

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

The quantities  $t_{i*}, i = 1, \dots, I$  are referred to as artificial SADA totals of non-binary people or *artificial totals* for short. For (4) to be satisfied, the artificial totals must sum to the provincial total,  $\sum_{i=1}^I t_{i*} = t_*$ .

The term ‘‘artificial totals’’ are used because the  $t_{i*}$  are not meant to represent the true SADA-level totals. The artificial totals are seen as a tool to partition the provincial constraint into independent SADA-level constraints allowing for the decomposition of the calibration. In each SADA, the following constraint is introduced:

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

The provincial problem is decomposed into SADA-level problems by augmenting the SADA calibration and total vectors with the non-binary indicator variable and artificial total respectively. The decomposed calibration problem for SADA  $i$  is:

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

$$\text{subject to } \sum_{k \in S_i} w_k \mathbf{x}_{k*}^{(i)} = \mathbf{t}_*^{(i)} \text{ and } 1 \leq w_k \leq 20,$$

where  $\mathbf{x}_{k*}^{(i)} = [x_{k*}, \mathbf{x}_k^{(i)T}]^T$  and  $\mathbf{t}_*^{(i)} = [t_{i*}, \mathbf{t}_{x^{(i)T}}]^T$ . If  $\sum_{i=1}^I t_{i*} = t_*$ , calibration to the artificial totals will ensure calibration to the provincial total. However, for any arbitrary choice of  $\{t_{i*}\}_{i=1}^I$  satisfying  $\sum_{i=1}^I t_{i*} = t_*$ , it is not guaranteed that the calibration weights obtained from the decomposed problems will be equal to the weights that would have been obtained from solving the provincial problem. In fact, our studies have shown that choosing  $t_{i*}$  which are far from the un-calibrated SADA-level estimated totals can lead to extreme calibration weights. In the following section, the artificial totals are derived to ensure the calibration weights obtained from the decomposed SADA problems are equal to the calibration weights that would have been obtained from solving the provincial problem.

### 3.3 Deriving the Artificial SADA-level Totals of Non-Binary Persons

The derivation is carried out independently in each SADA by setting the solution to the calibration weights obtained from the SADA and provincial problems equal and then solving for  $\mathbf{t}_*^{(i)}$ . It's noted that finding  $\{t_{i*}\}_{i=1}^I$  in this manner implicitly satisfies  $\sum_{i=1}^I t_{i*} = t_*$ . To facilitate the derivation,  $\mathbf{x}_{k*}^{(i)}$  and  $\mathbf{t}_*^{(i)}$  are re-written such that their dimension matches their provincial problem counterparts. The vectors for unit  $k$  in SADA  $i$  are re-written as  $\tilde{\mathbf{x}}_k = [x_{k*}, 0, \dots, 0, \mathbf{x}_k^{(i)T}, 0, \dots, 0]^T$  and  $\mathbf{t}_*^{(i)} = [t_{i*}, 0, \dots, 0, \mathbf{t}_{x^{(i)T}}, 0, \dots, 0]^T$ , where the first entry contains the artificial non-binary persons constraint and the SADA-level constraints are in the  $i + 1^{th}$  entry. The remaining elements are 0. Using equation (2), the formulas for the calibration weights are set equal to one another. Components belonging to the provincial problem are denoted using  $p$  and the SADA problem components using  $s$ . For household  $k$  in SADA  $i$  the following equation is obtained:

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

It can be shown that the estimated parameters  $\hat{\Phi}_{i,k}$  and  $\hat{\lambda}_{i,k}$  for the SADA and provincial problems are equal if the artificial total ensures the calibrated weights from the problems without range constraints are equal. Let  $w_{i,k}^{-(p)}$  and  $w_{i,k}^{-(s)}$  be the calibrated weights for household  $k$  in SADA  $i$  obtained from the provincial and SADA-level problems without range constraints. Using the solution from Deville and Särndal (1992), it follows that

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

where  $\hat{\mathbf{t}}_{\tilde{\mathbf{x}}}$  and  $\hat{\mathbf{t}}_*^{(i)}$  are total vectors estimated using the pre-calibrated weights. Solving for  $\mathbf{t}_*^{(i)}$  yields

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

which contains the artificial total in the first entry.

## 4. Monte Carlo Simulation

### 4.1 Setup

The simulation study was based on a pseudo-province/SADAs created using the responses to the 2021 long-form sample obtained in the province of British Columbia. 500 MC samples were drawn from the pseudo-province using a stratified simple random sample of households without replacement using a sampling fraction of 25%. In each MC sample, two sets of calibrated weights were produced. One set was constructed by carrying out the usual long-form calibration process. A second set of weights was produced by calibrating to the set of constraints selected by the usual procedure plus an artificial non-binary SADA constraint derived for each sample.

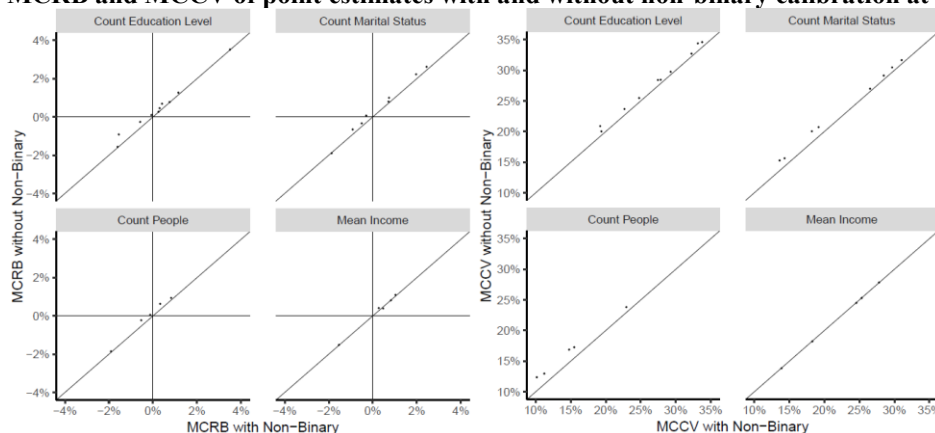
Each MC sample also underwent the creation of replicate weights corresponding to the Partially Balanced Repeated Replication- $\epsilon$  variance estimation methodology used to estimate the variance of long-form sample estimates (Devin and Verret, 2016). The replicate weights underwent the same calibrations as the main survey weight which allowed for the impact of non-binary calibration on both the point estimates and their estimated standard error (SE) to be studied.

### 4.2 Results

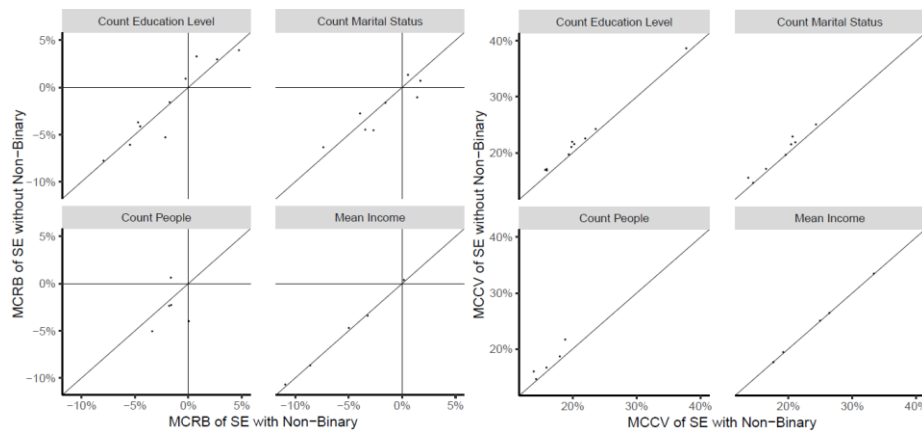
Figures 4.2-1 and 4.2-2 display the Monte Carlo relative bias (MCRB) and Monte Carlo coefficient of variation (MCCV) for the point and SE estimates at the pseudo-SADA level. The  $x$ -axis of the plots corresponds to the alternative calibration strategy (*with non-binary*) and the  $y$ -axis corresponds to the usual long-form calibration (*without non-binary*). A reference line corresponding to  $y = x$  is overlaid to assist the reader. The figures are summarized as follows:

- Both calibration procedures produce approximately un-biased point estimates.
- The alternative calibration procedure produces smaller MCCVs for counts and their estimated SEs. MCCVs of the mean (and its SE) were comparable between procedures.
- Both calibration procedures produce estimated SEs which have a small (negative) MCRB. The two procedures are comparable.

**Figure 4.2-1**  
**MCRB and MCCV of point estimates with and without non-binary calibration at the pseudo-SADA level**



**Figure 4.2-2**  
**MCRB and MCCV of estimated SEs with and without non-binary calibration at the pseudo-SADA level**



## 5. Conclusion

This paper presented an alternative long-form calibration strategy which allows for small populations, such as the non-binary group, to be incorporated while mitigating the methodological concerns. The proposed strategy derives artificial SADA totals to facilitate the calibration of long-form weights to provincial counts of non-binary persons – adding protection against variance inflation. The artificial totals allow for the decomposition of the resulting large provincial-level problem back into computationally feasible SADA-level problems requiring minimal changes to the current calibration system. The alternative calibration strategy is continuing to be evaluated for implementation in the 2026 long-form sample and has received interest from the Agency’s Demographic Microsimulation project. The simulation study demonstrated the positive impact that non-binary calibration had on data quality for the non-binary person domain. Non-binary calibration resulted in increased precision of point and SE estimates for characteristics of the non-binary person domain compared to calibration without the constraint.

## References

- Deville, J.-C., and Sarndal, C.-E. (1992), “Calibration estimators in survey sampling”, *Journal of the American Statistical Association*, 87 (418), pp. 376–382.
- Devin, N., and 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 Proceedings, Survey Research Methods Section. Alexandria, VA: American Statistical Association*, pp. 1977-1991
- Singh, A. C., and Mohl, C. A. (1996), “Understanding calibration estimators in survey sampling”, *Survey Methodology*, 22, pp. 107–116.
- UNCEB (2017), “Leaving no one behind: Equality and non-discrimination at the heart of sustainable development”, United Nations.