The Development of a Variance Estimation Methodology for Large-Scale Dissemination of Quality Indicators for the 2016 Canadian Census Long Form Sample

Nancy Devin[1], François Verret[1]
[1]Statistics Canada, 100 Tunney's Pasture Driveway, Ottawa, Ontario K1A 0T6

Abstract

The Canadian Census long form is a quinquennial large-scale sample survey for which millions of estimates on the Canadian population are published at various levels of geography. In 2016, to improve the analytical potential and the intelligibility of the published point estimates, Statistics Canada wants to be able to calculate a variance-based quality-indicator (QI) for each estimate. In addition, for the first time, analysts having access to microdata will be provided replicate weights enabling them to produce variance estimates on their own. This paper summarizes the development of a replication variance estimator that uses few replicates to be integrated into the existing dissemination systems. Emphasis will be put on the challenges of developing the variance estimator and the results of a Monte Carlo simulation supporting the choice of the method to be used. These challenges include the very large sample size along with the large sampling fraction, the need to calibrate the replicate weights and the numerous variance estimates being calculated for both smooth and non-smooth statistics in a limited timeframe while respecting confidentiality of the data provided.

Key Words: Large-scale survey, variance estimation, dissemination, balanced half-samples, Jackknife, finite population correction

1. Introduction

The Canadian Census program has a five year cycle and consists of both a census of the population and, since 1971, a sample survey of households receiving a more detailed census questionnaire called the Census long form. The Census long form was mandatory until 2006. In 2011, it was made voluntary and the data was collected through the National Household Survey (NHS). The Canadian government reinstated the Census mandatory long form for 2016. Every census cycle, millions of estimates on the Canadian population are published at various levels of geography using the Census long form data. In 2011, a sample design based on the subsampling scheme of Hansen and Hurwitz (1946) was used to follow up nonrespondents in an effort to minimize the impact of nonresponse caused by the voluntary nature of the NHS. Because total nonresponse was not expected to be as much of an issue in 2016 due to the mandatory nature of the Census long form, no follow-up subsample of nonrespondents was taken. The 2016 sample design consists of taking a stratified systematic sample with a sampling fraction of one-fourth. This is the same as the sample design in 2006 except that the sampling fraction was one-fifth.

Before 2011, neither variance estimates nor quality indicators (QI) based on variance estimates were released. Instead, the formula for calculating the standard error of an estimator of a total under simple random sampling without replacement, adjustment factors for the design and weighting effect and instructions on how to use them for variance estimation were provided to analysts in Census technical reports. To derive the adjustment factors provided in the reports, variance estimates were produced using classic Taylor linearization under the assumption that systematic samples can be approximated by simple random samples. Multiplicative adjustment factors were calculated using a Monte Carlo simulation and applied to the variance estimates to compensate for a downward bias. In 2011, the same general variance estimation methodology was used. However a different dissemination strategy was adopted since the more complex design and the more pronounced nonresponse made the design effects a lot less homogeneous. Variance estimates were disseminated for key characteristics and geography. With both dissemination approaches, the variance estimates were published many months after the publication of the point estimates. The goal of Statistics Canada for 2016 is to be able to calculate a variance-based QI for all published long form estimates.

This paper summarizes the development of a replication variance estimator that uses few replicates to be integrated to the existing dissemination systems. The selection of the replication method to be used with the 2016 Census long form data is based on the results of a Monte Carlo simulation study based on a pseudo population constructed from the 2006 Census long form responses. Section 2 describes the methodological challenges faced in the development of a variance estimator. Section 3 gives an overview of a few replication methods. Section 4 describes the various components of the analysis including the Monte Carlo simulation setup, the two contending replication methods, the weight calibration strategy and various statistics used in the Monte Carlo study. Section 5 gives the results of the study and a conclusion is given in Section 6.

2. Challenges in developing a variance estimation methodology

Developing a variance estimator meeting all of Statistics Canada's needs is a challenge. Firstly, millions of estimates are produced by Statistics Canada from the Canadian Census long form data using the dissemination system of the census. They cover various topics and geographies. In addition, analysts having access to the Census long form microdata files can also calculate their own estimates using data analysis software. The variance estimation methodology needs to cover both situations. Secondly, in order to inform users of data quality in a timely manner, the Census long form variance estimates should be calculated promptly. To achieve this, and because of the great volume of variance estimates that needs to be calculated and delivered, the variance estimation methodology needs to be integrated into the existing dissemination systems. Thirdly, point estimates released for the Canadian Census long form consist of totals, means, ratios and percentiles. Other types of point estimates may also be calculated from the various microdata files. The chosen method will thus have to produce good quality indicators for linear and non-linear as well as smooth and non-smooth point estimates. In other words it should be versatile. Finally and most importantly, the design weights go through a series of adjustments before they are used for estimation. These adjustments, which consist of a total nonresponse adjustment and a calibration to known totals, need to be taken into account when calculating the variance estimates.

Classic Taylor linearization, which was used in previous cycles, requires that the estimator be linearized. This is not possible with non-smooth estimators such as medians.

Furthermore, the linearization method is too complex to be implemented within the dissemination systems. To address the various challenges, replication methods that use a small number of replicates were considered to estimate the variance of the Census long form estimates. Each method considered has its own weaknesses and strengths in terms of convergence, bias and versatility for instance. The number of replicates was set to 16 in the first steps of development, but was later increased to 32 to insure a minimum stability of the variance estimator.

3. Overview of various replication methods

The first method studied in the development process was the variance estimation method of the long form public-use microdata files, namely the Dependent Random Group (DRG) method, see for instance Wolter (1985, or the 2007 reedition). The DRG method consists of partitioning the observed sample, often referred to as the parent sample, into *R* disjoint random groups or replicates using the same sample design as the one used for the parent sample. The main advantage of the method is its implementation simplicity. Variance estimation of both smooth and non-smooth estimators can also be undertaken. A disadvantage is that one has to make a compromise between stability of the variance estimator, which is a function of the number of groups, and desirable asymptotic properties of replicate estimators (Wolter, 1985, page 23, or the 2007 reedition, page 25), which is a function of the size of each group. Favouring stability (i.e. numerous small size groups) might generate problems with some types of estimators, for example when, for a given replicate, a ratio estimate has no unit contributing to the denominator. Calibration of the replicates may also be affected by small replicate sizes when the calibration method is intricate or tight, meaning that it uses many calibration constraints on a limited number of responding units. This might result in either an artificial increase of the replicate weight variances or the inability to replicate the calibration performed on the parent sample. To circumvent this problem, Särndal et al. (1992, page 430) propose to repeat *T* times the DRG method using a small number of replicates and then to average the *T* variance estimates to obtain the final variance estimate. The resulting estimate would be more stable and would have a smaller bias under tight calibration. The approach was studied with 16 replicates. To form the replicates with the largest group size possible, the DRG method was repeated eight times using two groups. However, even with this variant, it was not always feasible to perform estimation and calibration on the replicates. Furthermore, the method does not achieve the precision of other replication methods using the same number of replicates.

Repeating the DRG method many times with two large replicates led us to consider the Balanced Repeated Replication (BRR) method, which also uses half-samples to form the replicates. The balancing property ensures that the variance estimator corresponds to the variance estimator one would obtain if all possible half-samples were used for variance estimation. As with the DRG method, the BRR method can estimate the variance of both smooth and non-smooth estimators. Estimation and calibration on the replicates can still be problematic however since the replicates do not contain all of the sampled units. Another weakness of the method is that the number of replicates needed is close to the number of first-phase strata. This makes the method impractical when the number of first-phase strata is very large, such as in the Census long form. To reduce the number of replicates, a Partially Balanced Repeated Replication method (PBRR) was considered. The PBRR balances the replicates within groups of first-phase strata rather than across the sample.

The problem with the replication of ratio estimates and calibration of the replicate weights was solved by using the modified BRR method also known as the BRR-epsilon or Fay's

BRR as described in Rao and Shao (1999). With the BRR-epsilon method, a factor perturbing value –epsilon– is introduced in the definition of the replicate weights in order to have replicate weights closer to the design weights. The replicate weights of the BRR-epsilon take the values $(1-\varepsilon)d_k$ or $(1+\varepsilon)d_k$, $0<\varepsilon<1$, instead of 0 or $2d_k$ as with the classic BRR. With the introduction of an epsilon in the replicate weights, the entire sample contributes to each replicate. This may facilitate replicate estimation and calibration. It also allows the inclusion of a finite population correction factor inside the epsilon to reduce the biases caused by the large sampling fraction of the Census long form survey, as will be described in Section 4.3.1. The BRR-epsilon method and the PBRR method were combined to produce the first contender, the PBRR-epsilon.

A variant of the Jackknife method was considered as a second contender for variance estimation. The customary delete-one Jackknife method as described in Wolter (1985, or the 2007 reedition) was not considered since it does not properly estimate the variance of non-smooth estimators due to the lack of variation between the replicate estimates. Instead the delete-a-group variation of the Jackknife, DAGJK, defined by Kott (2001), was examined. With the DAGJK method, a random group of units is removed from the parent sample to form each replicate. The method can estimate the variance of non-smooth estimators provided that the size, *d*, of the group removed is large enough and that the set of replicates corresponds to a random sample taken from all possible groups of size $(n-d)$ taken from the parent sample of size *n*. More specifically, the size of the group, *d*, should satisfy the relation $\sqrt{n}<d<n$ to ensure convergence of the variance estimator for non-smooth estimators, see Efron and Tibshirani (1993). Problems may again occur with the replication of some ratio estimators and tight calibration since the replicates do not contain all the sampled units. As with the PBRR-epsilon, an epsilon was introduced in the creation of the replicate weights in order to include all sampled units in each replicate and to correct for the large sampling fraction, as will be described in Section 4.3.2.

4. Analysis setup

4.1 Monte Carlo simulation

The PBRR-epsilon approach and the DAGJK-epsilon approach were compared through a Monte Carlo simulation.

4.1.1 Monte Carlo population

The 2016 Monte Carlo study on variance estimation was initiated before the 2016 survey was made mandatory. The study was based on the Monte Carlo simulation done in 2006 to estimate the bias of the variance estimators, see Benjamin (2008) for more details. For simplicity the same simulation was used. It is based on a pseudo population created from the 2006 Census long form responses. The set of responses represents close to 20% of the population because of the 20% sampling rate and the high response rate to the Census long form. To create the pseudo population, 12 Census Divisions, CDs, have been selected across the country based on size and operational considerations. Canada is comprised of approximately 300 CDs and the median size of each CD is about 38,000 people. Within each selected CD, pseudo weighting areas, WAs, were formed by combining the WAs in homogeneous groups of five to create pseudo WAs. The pseudo WA is thus representative of the Canadian population in a WA. The WA is the level of geography at which calibration was performed in 2006. It contains between 1,000 and 3,000 households and is formed of approximately eight dissemination areas, DAs. Similarly, DAs were also combined in groups of five to create pseudo DAs. Then, one pseudo WA was randomly selected by CD. Hence the pseudo population consists of the 12 selected pseudo WAs for a total of 20,259 households. In the rest of the paper, the pseudo population, the pseudo WA and the pseudo DA will be referred to as the population, the WA and the DA.

4.1.2 Monte Carlo samples and design weights

The Monte Carlo simulation uses 500 samples. Each sample is created by selecting a stratified simple random sample of households by DA without replacement from the population using a sampling fraction of one-fifth (Benjamin, 2008). Each sample has 4,089 households which represents approximatively one-fifth of the population since the size of a stratum is not always a multiple of five. Even though the survey's design is stratified systematic, generating simple random samples in the simulation is justified because each stratum is relatively small and homogeneous. Sampling for the Monte Carlo simulation is illustrated in Figure 1.

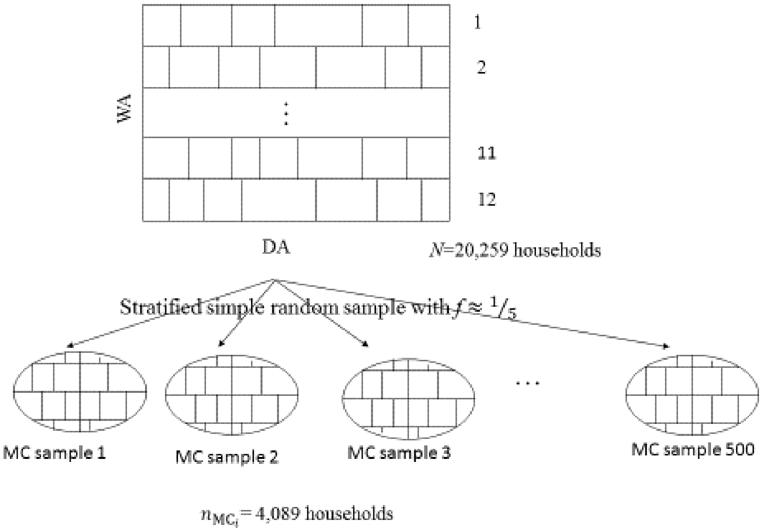
1
2
WA
⋮
11
12
DA
N=20,259 households
Stratified simple random sample with $f \approx {}^1/_5$
...
MC sample 1
MC sample 2
MC sample 3
MC sample 500
n_{MC_i}= 4,089 households

Figure 1: Monte Carlo simulation setup

4.2 Calibration

Initial design weights are calculated for each sample which are then calibrated to match population totals. The current calibration strategy has two steps. The first step consists of selecting the calibration constraints to be used out of a pre-defined set of potential constraints. Constraint selection is performed independently once for each WA and once for all WAs combined. Selection also does not depend on the selected sample. One of the reasons for the constraint selection process is to avoid direct or indirect calibration on constraints targeting too few households, called small constraints. The small constraints are first identified by comparing the number of households in the population having the characteristic of the constraint to a pre-set threshold. Once a constraint has been identified as small, it cannot be used as a calibration constraint. A forward selection process is then used to add the remaining calibration constraints one at a time, starting with these two which are mandatory: total number of households and total number of persons in the geography. At each step of the process, potential constraints are evaluated one by one and they get discarded when: 1) they are judged to be too collinear with the selected set of constraints; or 2) they would cause implicit calibration to any small constraints if they were added to the selected set. Among the evaluated constraints, the one that splits best the population of the geography in two equal parts is chosen. The second step is calibration per se to known totals based on the chosen constraints. Calibration is performed on all selected constraints at once. The constraint totals are derived from the Canadian census variables and from administrative data linked to the census records. In the long form survey these variables are known for the entire population, whereas in the simulation even the variables of interest of the survey are known for the entire population. Calibration variables are referred to as “2A” variables since the long form was called the 2A form in 2006. Conversely variables coming from the Census long form sample only are referred to as “2B” variables. Characteristics under study in the simulation are derived from both sets of variables.

4.3 Contending Replication methods

This section presents in more details the final two replication methods studied to estimate the variances of the Census long form estimates.

4.3.1 Partially Balanced Repeated Replication – epsilon (PBRR-epsilon)

The first contender for variance estimation is the Partially Balanced Repeated Replication method described in Section 3, in which the replicates are balanced within groups of two strata. The goal was to create 15 substrata of two or more sampled households for each stratum and hence obtain 30 substrata by pair of strata. This is rather straightforward for strata with 30 or more sampled households: households are sorted randomly and are distributed consecutively to each substratum. For strata with fewer than 30 sampled households, only $\lfloor n_h/2 \rfloor$ substrata having at least two sampled households were generated. If n_h is odd, the last sampled household is randomly assigned to one of the existing $\lfloor n_h/2 \rfloor$ substrata. Finally, households within each substratum were divided into two clusters. The SAS procedure PROC SURVEYMEANS was used to generate the 32 BRR replicates by pair of strata. The procedure uses a 32 by 32 Hadamard matrix to determine the clusters that will compose each replicate. The BRR replicates of each pair of strata form the PBRR replicates. This is illustrated in Figure 2.

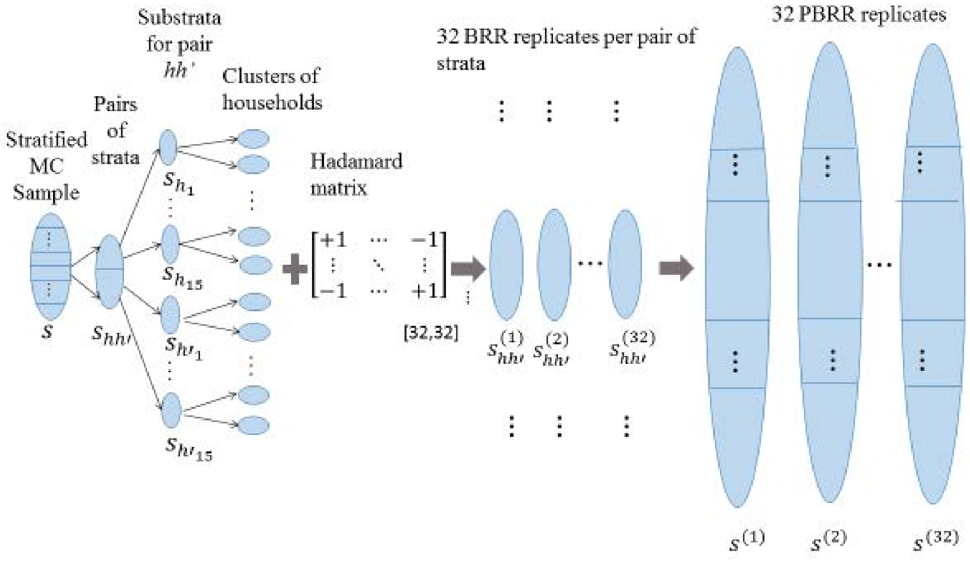

Figure 2: Creation process of the PBRR replicates

Once the replicates are formed, the replicate weights are calculated. The general PBRR-epsilon weights are given by:

$$d_k^{(r)} = \begin{cases} (1+\varepsilon)d_k, \; k \in s^{(r)} \\ (1-\varepsilon)d_k, \; k \notin s^{(r)} \end{cases},$$

where $s^{(r)}$ is the rth replicate subsample and r = 1, …,32.

Because the sizes of the two clusters within a substratum are not always equal, the replicate weights were first adjusted as described in Rao and Shao (1999) for the case where the number of PSUs is larger than two in some strata. If n_{h_i} represents the number of households in the ith substratum of stratum h, then each cluster contains either $\lfloor n_{h_i}/2 \rfloor$ or $\lceil n_{h_i}/2 \rceil$ households when n_{h_i} is odd. When n_{h_i} is even, both clusters contain $\left[n_{h_i}/2 \right]$ households. The adjustment factor, $\mathrm{RA}_{h_i}^{(r)}$, takes into account the size of each cluster. The replicate weights thus become:

$$d_k^{(r)} = \begin{cases} \left(1+\varepsilon\sqrt{\mathrm{RA}_{h_i}^{(r)}}\right)d_k, \quad k \in s^{(r)} \\ \left(1-\varepsilon\sqrt{1/\mathrm{RA}_{h_i}^{(r)}}\right)d_k, \; k \notin s^{(r)} \end{cases},$$

where $\mathrm{RA}_{h_i}^{(r)} = \left(n_{h_i} - n_{h_i}^{(r)}\right)/n_{h_i}^{(r)}$, $s^{(r)}$ is the rth replicate subsample and r = 1, …,32. The large sampling fraction was corrected by adding a finite population correction factor, $1-f$, to the definition of the epsilon value. One has to be careful in the selection of the epsilon. On the one hand, a value close to 1 gives results similar to those of the original PBRR method and hence introduces a lot of fluctuation in the replicate weights. On the other hand, a value close to 0 gives results similar to the Taylor linearization or the delete-one-Jackknife, which are superior to the BRR for smooth estimators but are not appropriate for non-smooth estimators (Rao & Shao, 1999). Based on simulations, Judkins (1996) suggests that an epsilon value close to one-half is a good compromise if one needs to estimate the variance of various estimators. Different values of epsilon were studied and this paper presents only the results obtained with the most promising value which is: $\varepsilon_k = \sqrt{(1-f_k)/2}$, where $f_k = 1/d_k$. Finally, the replicates were calibrated using the same calibration strategy as the one used for the full Monte Carlo sample. Only variance estimates of totals were studied in the Monte Carlo simulation. The PBRR-epsilon variance estimator is given by:

$$\hat{\mathrm{Var}}_{\text{PBRR-epsilon}}\left(\hat{T}\right) = \frac{1}{32\left(\sqrt{1/2}\right)^2}\sum_{r=1}^{32}\left(\hat{T}^{(r)} - \overline{\hat{T}^{(r)}}\right)^2,$$

where $\hat{T}=\sum_{k\in s}w_k y_k$, $\hat{T}^{(r)}=\sum_{k\in s}w_k^{(r)}y_k$, $\overline{\hat{T}^{(r)}}=\sum_{r=1}^{32}\hat{T}^{(r)}\big/32$, $w_k=g_k d_k$, $w_k^{(r)}=g_k^{(r)}d_k^{(r)}$ and $g_k^{(r)}$ is the calibration factor of the rth replicate. Note that the squared term in the denominator of the variance estimator would normally be equal to epsilon squared. A modified version of the epsilon was used in the denominator in order to make the finite population correction effective in the variance estimator.

4.3.2 Delete-a-group-Jackknife-epsilon (DAGJK-epsilon)

To obtain a DAGJK-epsilon variance estimator with 32 replicates, the average of two DAGJK-epsilon variance estimators with 16 replicates was used. Firstly, the two independent sets of 16 replicates were created by dividing the households of the parent sample into two independent sets of 16 DRGs. Secondly, the replicate subsamples, $s^{(r)}$, were constructed by removing each DRG from the parent sample. This is illustrated in Figure 3.

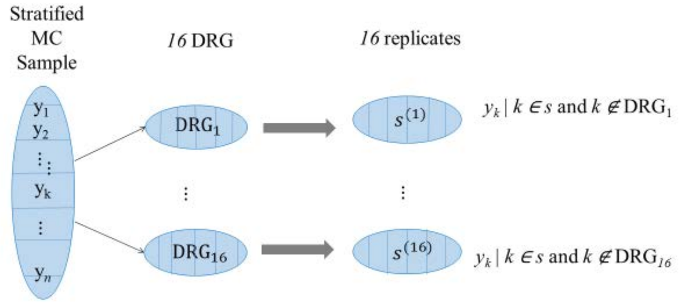

Figure 3: Creation process of the DAGJK replicates

For negligible sampling fractions, Kott (2001) defines the DAGJK weights of the rth replicate as:

$$d_k^{(r)}=\mathrm{RA}_h^{(r)}\times d_k\times \mathrm{I}\{k\notin \mathrm{DRG}_r\},$$

where $\mathrm{RA}_h^{(r)}=n_h/n_h^{(r)}$, n_h is the number of units in stratum h, $n_h^{(r)}$ is the number of units in stratum h but not in DRG r and $k\in s$. To correct for the large sampling fraction and to ensure that all households contribute to every replicate, an epsilon was added to the definition of the replicate weights. The final weights are defined as follows:

$$d_k^{(r)}=d_k\left[1+\varepsilon_k\left(\mathrm{RA}_h^{(r)}\times \mathrm{I}\{k\notin \mathrm{DRG}_r\}-1\right)\right],$$

where $\varepsilon_k=\sqrt{1-f_k}$, $f_k=1/d_k$, $r=1,\ldots,16$ and $k\in s$. Finally, each replicate was calibrated to the totals of the population. Since only totals where studied in the Monte Carlo simulation, the variance estimator of the DAGJK-epsilon method is:

$$\hat{\mathrm{Var}}_{\mathrm{DAGJK}}\left(\hat{T}\right)=\sum_{t=1}^{2}\hat{\mathrm{Var}}_{\mathrm{DAGJK}t}\left(\hat{T}\right),$$

where $\overset{\wedge}{\text{Var}}_{\text{DAGJK}t}\left(\hat{T}\right)=15\sum_{r=1}^{16}\left(\hat{T}^{(r)}-\overline{\hat{T}^{(r)}}\right)^2\Big/16$, $\hat{T}=\sum_{k\in s}w_k y_k$, $\hat{T}^{(r)}=\sum_{k\in s}w_k^{(r)}y_k$, $\overline{\hat{T}^{(r)}}=\sum_{r=1}^{16}\hat{T}^{(r)}\Big/16$, $w_k=g_k d_k$, $w_k^{(r)}=g_k^{(r)}d_k^{(r)}$ and $g_k^{(r)}$ is the calibration factor of the rth replicate.

4.4 Statistics under study

Many statistics are studied in the Monte Carlo simulation. They are used to compare the two contending replication methods.

4.4.1 Targeted statistics based on the Monte Carlo samples estimates

The calibrated estimated totals, $\hat{T}_Y^{\text{MC}_i}$, and their expectations, are given by:

$$\hat{T}_Y^{\text{MC}_i}=\sum_{k\in s_{\text{MC}_i}}w_k y_k \quad ,i=1,\ldots,500$$

and

$$\hat{\text{E}}_{\text{MC}}\left(\hat{T}_Y\right)=\frac{1}{500}\sum_{i=1}^{500}\hat{T}_Y^{\text{MC}_i}\ .$$

The targeted variance and coefficient of variation are approximated by:

$$\overset{\wedge}{\text{Var}}_{\text{MC}}\left(\hat{T}_Y\right)=\frac{1}{500}\sum_{i=1}^{500}\left[\hat{T}_Y^{\text{MC}_i}-\hat{\text{E}}_{\text{MC}}\left(\hat{T}_Y\right)\right]^2$$

and

$$\overset{\wedge}{\text{CV}}_{\text{MC}}\left(\hat{T}_Y\right)=\frac{\sqrt{\text{Var}\left(\hat{T}_Y\right)}}{T_Y}\ .$$

In the rest of the paper, the Monte Carlo estimates obtained will be considered the real values. Hence $\hat{\text{E}}_{\text{MC}}\left(\hat{T}_Y\right),\overset{\wedge}{\text{Var}}_{\text{MC}}\left(\hat{T}_Y\right)\text{and}\,\overset{\wedge}{\text{CV}}_{\text{MC}}\left(\hat{T}_Y\right)$ will be referred to as $\text{E}\left(\hat{T}_Y\right),\text{Var}\left(\hat{T}_Y\right)\text{and}\,\text{CV}\left(\hat{T}_Y\right)$.

4.4.2. Estimated statistics based on the replication methods

For each Monte Carlo sample, the replicate estimates and the corresponding estimated variances are calculated as defined in Section 4.3. From these statistics, the expected estimated variance and coefficient of variation of the point estimate are approximated by the following formulas:

$$\hat{\text{E}}_{\text{MC}}\left[\overset{\wedge}{\text{Var}}\left(\hat{T}_Y\right)\right]=\frac{1}{500}\sum_{i=1}^{500}\overset{\wedge}{\text{Var}}\left(\hat{T}_Y^{\text{MC}_i}\right)$$

and

$$\widehat{\mathrm{CV}}\left(\hat{T}_Y\right) = \frac{\sqrt{\hat{\mathrm{E}}_{\mathrm{MC}}\left[\widehat{\mathrm{Var}}\left(\hat{T}_Y\right)\right]}}{T_Y}.$$

Then, the variance of the variance estimator and the coefficient of variation of the variance estimator are calculated as follows:

$$\widehat{\mathrm{Var}}_{\mathrm{MC}}\left[\widehat{\mathrm{Var}}\left(\hat{T}_Y\right)\right] = \frac{1}{500}\sum_{i=1}^{500}\left\{\widehat{\mathrm{Var}}\left(\hat{T}_Y^{\mathrm{MC}_i}\right) - \hat{\mathrm{E}}_{\mathrm{MC}}\left[\widehat{\mathrm{Var}}\left(\hat{T}_Y\right)\right]\right\}^2$$

and

$$\widehat{\mathrm{CV}}\left[\widehat{\mathrm{Var}}\left(\hat{T}_Y\right)\right] = \frac{\sqrt{\widehat{\mathrm{Var}}_{\mathrm{MC}}\left[\widehat{\mathrm{Var}}\left(\hat{T}_Y\right)\right]}}{\hat{\mathrm{E}}_{\mathrm{MC}}\left[\widehat{\mathrm{Var}}\left(\hat{T}_Y\right)\right]}.$$

The above statistics are used to estimate the biases of the variance estimator and the coefficient of variation through:

$$\widehat{\mathrm{B}}\left[\widehat{\mathrm{Var}}\left(\hat{T}_Y\right)\right] = \hat{\mathrm{E}}_{\mathrm{MC}}\left[\widehat{\mathrm{Var}}\left(\hat{T}_Y\right)\right] - \widehat{\mathrm{Var}}_{\mathrm{MC}}\left(\hat{T}_Y\right) \text{ and } \widehat{\mathrm{B}}\left[\widehat{\mathrm{CV}}\left(\hat{T}_Y\right)\right] = \widehat{\mathrm{CV}}\left(\hat{T}_Y\right) - \widehat{\mathrm{CV}}_{\mathrm{MC}}\left(\hat{T}_Y\right). \quad (1)$$

In addition, the bias estimates are approximately normally distributed because of the central limit theorem. Let $\overline{z} = \sum_{i=1}^{500} z^{\mathrm{MC}_i} \big/ 500,$ $z^{\mathrm{MC}_i} = \widehat{\mathrm{Var}}\left(\hat{T}_Y^{\mathrm{MC}_i}\right) - \left[\hat{T}_Y^{\mathrm{MC}_i} - \mathrm{E}\left(\hat{T}_Y\right)\right]^2$ and $s = \sqrt{\sum_{i=1}^{500}\left(z^{\mathrm{MC}_i} - \overline{z}\right)^2 \big/ 499}$, a 95% confidence interval for the bias is given by:

$$\left(\overline{z} - 1.96\frac{s}{\sqrt{500}}, \overline{z} + 1.96\frac{s}{\sqrt{500}}\right). \quad (2)$$

5. Results

To evaluate the performance of both the PBRR-epsilon and the DAGJK-epsilon methods, the bias and the stability of the variance estimates were compared for the two types of characteristic. Section 5.1 looks at the bias while Section 5.2 looks at the stability of the variance estimates.

5.1 Comparison of the PBRR-epsilon and the DAGJK-epsilon variance estimators

Figure 4 compares the expected estimated CV, $\widehat{\mathrm{CV}}\left(\hat{T}_Y\right)$, to the target CV, $\mathrm{CV}\left(\hat{T}_Y\right)$, by method and by type of variable, 2A and 2B. The figure shows the CVs for population totals of 100 or more. The thick blue lines correspond to the regression lines through the origin and the thin red lines correspond to *Y*=*X*. The figure shows a consistent overestimation of $\hat{\text{CV}}\left(\hat{T}_Y\right)$ for the DAGJK method. The overestimation is not seen with the PBRR-epsilon method. All variants of the DAGJK-epsilon studied also overestimated the variances. Table 1 gives the regression slopes by method and by type of variable. It puts in number the overestimation of the DAGJK-epsilon method seen on Figure 4. The regression slopes are four hundredth point lower for the PBRR-epsilon than the DAGJK-epsilon for both types of characteristics. The overestimation is also slightly larger for 2A variables.

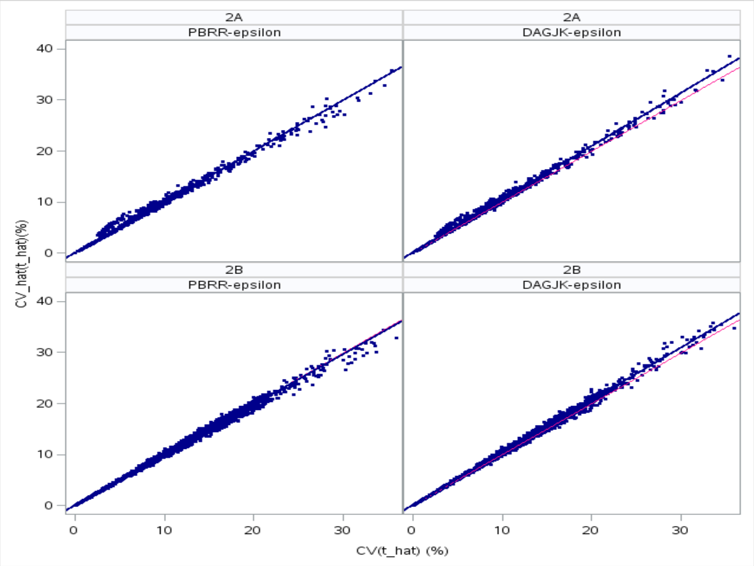

Figure 4: Graph of $\hat{\text{CV}}\left(\hat{T}_Y\right)$ (%) against $\text{CV}\left(\hat{T}_Y\right)$ (%) and the regression line by method and by type of characteristic where population totals are greater or equal to 100

Table 1: Estimated regression slopes for $\hat{\text{CV}}\left(\hat{T}_Y\right) = \hat{\beta} \times \text{CV}\left(\hat{T}_Y\right)$, where the population totals are greater or equal to 100

Variable	Method	Number of observations	Regression slope
2A	PBRR-epsilon	1243	1.00
	DAGJK-epsilon	1243	1.04
2B	PBRR-epsilon	2564	0.99
	DAGJK-epsilon	2564	1.03

Figure 5 again shows the slight overestimation of the expected estimated CVs for the DAGJK-epsilon method. Confidence intervals for the bias of the CV, as defined by Equation (2), can indicate the quality of the expected estimated CV, $\hat{\text{CV}}\left(\hat{T}_Y\right)$. Because of the central limit theorem, one would expect 95% of the intervals to contain the value 0. Table 2 shows that the PBRR-epsilon method surpasses the DAGJK-epsilon method in terms of bias of the variance estimator because the coverage rate is closer to 95% for each type of characteristic and more so for the 2B characteristics.

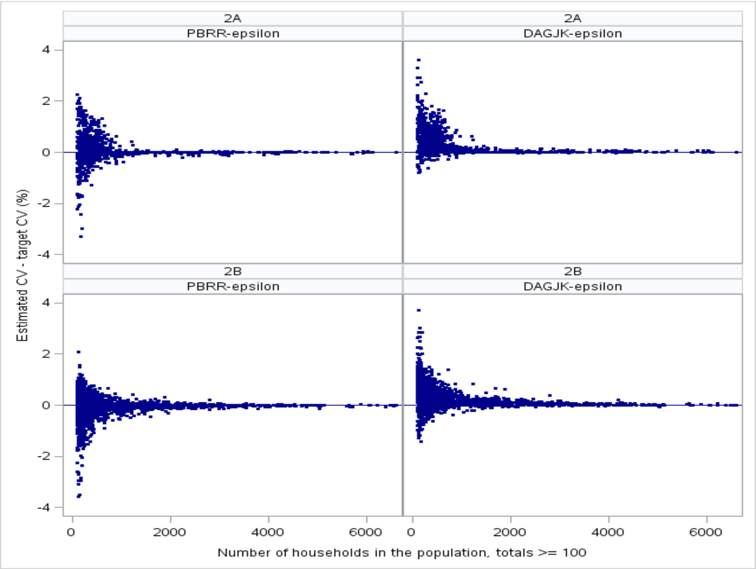

Figure 5: Graph of $\hat{\mathrm{B}}\left[\hat{\mathrm{CV}}\left(\hat{T}_Y\right)\right]$ against the number of households in the population having the characteristic by method and by type of characteristic where population totals are greater or equal to 100

Table 2: Percentage of confidence intervals for the bias of the CVs that contain 0, where population totals are greater or equal to 100

Variables	Method	Number of observations	% of the C.I. that includes 0
2A	PBRR-epsilon	1243	83.75
	DAGJK-epsilon	1243	70.47
2B	PBRR-epsilon	2564	92.75
	DAGJK-epsilon	2564	83.31

5.2 Comparison of the stability of both variance estimators

Another way of studying the quality of the variance estimator is to look its variance. Figure 6 shows the expected estimated CV of the variance estimates against the expected estimated CV of the point estimates. The horizontal line is set at 33.3%.

The figure shows that the variances of the variance estimates inflate when the expected variance estimates are small. The increase in the variances of the variance estimates is not as strong for large expected variance estimates. For small $\hat{\mathrm{CV}}\left(\hat{T}_Y\right)$, the large $\hat{\mathrm{Var}}_{\mathrm{MC}}\left[\hat{\mathrm{Var}}\left(\hat{T}_Y\right)\right]$ are not too much of a concern since the confidence intervals for $\hat{\mathrm{Var}}\left(\hat{T}_Y\right)$ are short. The figure also shows that the expected variances of the estimated variances are mostly under the reference line when the expected variances of the point estimates are within a certain interval.

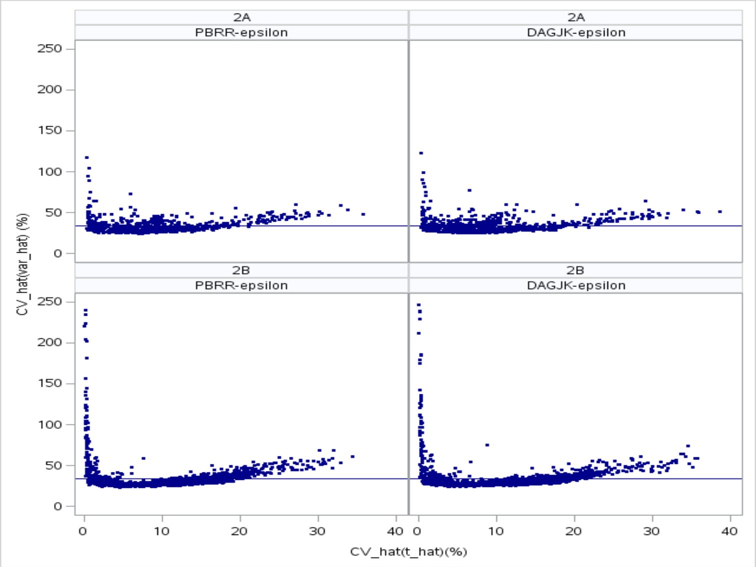

Figure 6: $\hat{\text{CV}}\left[\hat{\text{Var}}\left(\hat{T}_Y\right)\right]$ (%) against $\hat{\text{CV}}\left(\hat{T}_Y\right)$ (%) by method and by type of characteristic, where population totals are greater or equal to 100

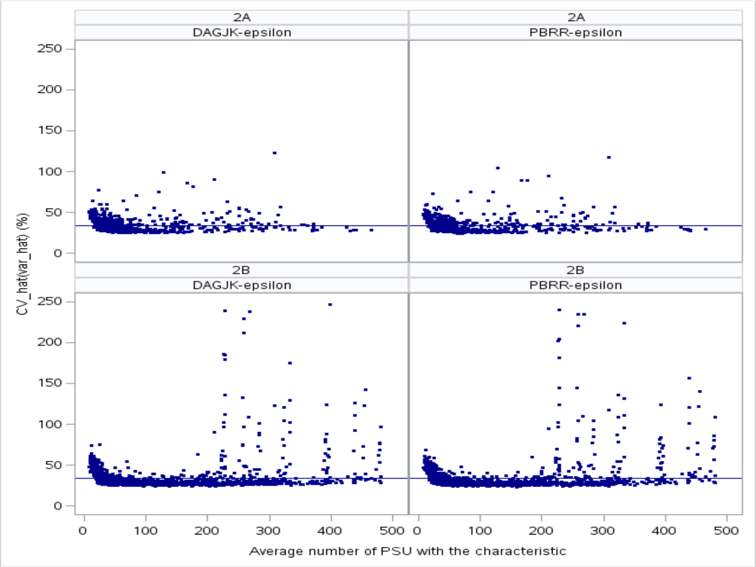

Figure 7: $\hat{\text{CV}}\left[\hat{\text{Var}}\left(\hat{T}_Y\right)\right]$ (%) against the Monte Carlo average number of households with the characteristic by method and type of characteristic

Figure 7 presents the Monte Carlo CV estimates of the variance estimates with respect to the average number of households having a given characteristic in the parent sample. The graphs were produced in order to identify the minimal number of households needed for a given characteristic to ensure that the variance estimator had a CV in percent smaller or equal to 33.3%. Asymptotes can be seen on the figure, especially with the 2B characteristics. These occur when the number of households having the characteristic corresponds to almost all the households of a given WA in the sample. Each WA thus has its own asymptote. These cases correspond to the largest CVs of the variance estimates of

Figure 6 and, as mentioned above, they are not viewed as problematic. From these results, the minimum number of households necessary to obtain a coefficient of variation of the variance estimates smaller than 33.3% was derived. More specifically, local regressions were done for each of the 4 combinations of variance estimation methods and types of characteristic, excluding the data points of the asymptotes. The minimum number of households necessary to produce a CV below 33.3% with the PBRR-epsilon method are 36 and 30 for the 2A variables and the 2B variables respectively. The numbers obtained with the DAGJK-epsilon method are 37 and 29 for the 2A variables and the 2B variables respectively.

6. Conclusion

The regression slopes of $\hat{\mathrm{CV}}\left(\hat{T}_Y\right)$ against $\mathrm{CV}\left(\hat{T}_Y\right)$ by method and type of characteristic show that the PBRR-epsilon produces variance estimates very close to the real variances while the DAGJK-epsilon tends to overestimate the variances. Also, a higher percentage of the 95% confidence intervals for the bias of the estimated CVs includes the value 0 with the PBRR-epsilon method than with the DAGJK-epsilon method. Both the PBRR-epsilon method and the DAGJK-epsilon method require a similar number of households having the characteristic to ensure that the estimated CV of the variance estimator is below 33.3%. For 2A variables, the minimum number of households obtained with the PBRR-epsilon method is 36 while it is 37 with the DAGJK-epsilon method. For 2B variables the minimum number of households obtained with the PBRR-epsilon method is 30 and 29 with the DAGJK-epsilon method. Based on the results of the study, the PBRR-epsilon method has been chosen to estimate the variances of the Census long form estimates. The feasibility of using 100 replicates to further increase the stability of the PBRR-epsilon variance estimator is now being assessed.

Throughout our study, the impact of different calibration strategies on variance estimation has in fact been evaluated. This paper does not show the results obtained with the different strategies but they can be summarized by the following statements:

- The calibration strategy affects the quality of the variance estimator;
- In order to reduce the bias of the variance estimator it is important to keep the number of calibration constraints under a certain threshold. Statistics Canada’s Advisory Committee on Statistical Methods suggested using as a rule of thumb for a given calibration geography no more calibration constraints than the square root of the number of responses; and
- It is important to have a fixed set of calibration constraints for the sample and all replicates to stabilize the variance estimator. Standard variance estimation methods are not designed to take into account the variability introduced by randomly selecting the calibration constraints based on the sample.

Acknowledgements

The authors would like to thank Arthur Goussanou, Sarah-Anne Savard, Ivan Carrillo-Garcia and Yannick Bridé for their contributions to the project.

References

Benjamin, W. 2008. 2006 Variance Study Processing Guide. *Statistics Canada internal document*.

Efron, B., and R. J. Tibshirani. 1993. An Introduction to the Bootstrap. Chapman & Hall.

Hansen, M. H., and W.N. Hurwitz. 1946. The Problem of Non-Response in Sample Surveys. In *Journal of the American Statistical Association*, Vol. 41, 517-429.

Judkins, D. R. 1996. Fay's method for Variance Estimation. In *Journal of official Statistics*, Vol.6, No. 3. 223-239.

Kott, P.S. 2001. The Delete-a-Group Jackknife. In *Journal of Official Statistics*, Vol. 17, No. 4. 521-526.

Rao, J. N. K., and J. Shao. 1999. Modified balanced repeated replication for complex survey data. In *Biometrika*, Vol. 86, No. 2. 403-415.

Särndal, C.-E., B. Swensson, and J. Wretman. 1992. Model Assisted Survey Sampling. New York: Springer-Verlag, Inc.

Wolter. K. M. 1985. Introduction to Variance Estimation. New York: Springer-Verlag, Inc.