Weighting and variance estimation plans for the 2016 Census long form

François Verret, Arthur Goussanou & Nancy Devin

Statistics Canada
100 Tunney's Pasture Driveway, Ottawa, Ontario, Canada, K1A 0T6

1. Introduction

The Canadian census is conducted every five years. Until the 2006 Census, one in five households received the mandatory census long form, while the rest of the households received the mandatory census short form. In 2011, every household received a mandatory short form and the voluntary National Household Survey (NHS) was created to collect the long form data on a sample of households. In November 2015, it was decided by the newly elected Canadian government that the mandatory census long form would be reinstated in 2016 on a sample basis.

This paper presents the plans for 2016 Census long form sample weighting, variance estimation and dissemination of quality indicators (QIs) based on the estimated variance. Since much of the research on 2016 estimation has been done based on the 2011 NHS, this survey will also be discussed. In Section 2, the sampling design of the 2011 NHS and of the 2016 Census long form sample is described. In Section 3, 2011 weighting is summarised and the starting point for 2016 weighting is presented. Variance estimation is covered in Section 4. Variance estimation methods and associated variance-based QI dissemination strategies for the 2011 NHS and for most recent census long form samples are first summarised in this section. Two new dissemination objectives related to variance estimation for 2016 are then given, along with the plan for meeting these objectives. The section ends with the most recent developments on 2016 variance estimation. A discussion and next steps of development are presented in Section 5. Comments made by Dr. Phillip S. Kott during his discussion of the presentation of this paper at the 2015 Federal Committee on Statistical Methodology (FCSM) conference are also presented in this last section.

2. 2011 NHS and 2016 Census long form sampling design

The 2011 NHS was conducted at the same time as the 2011 Census. Both surveys shared the same general infrastructure and had the short form questions in common (10 questions). The sampling design of the 2011 NHS was based on the follow-up sub-sampling methodology of Hansen & Hurwitz (1946) for the vast majority of the Canadian population (98%). The first-phase sampling fraction of the survey was large (28.4%), as well as the sampling fraction of the non-response follow-up subsample (1/3 on average). Stratified systematic sampling was used for both phases of the sampling design. Nevertheless some aspects of the NHS do not correspond to the setting of Hansen & Hurwitz: responses were obtained from initial non-respondents not selected in the sub-sample (called "surprise respondents") and non-response was observed both in the sub-sample and amongst households that were identified as respondents at the time of sub-sampling. In areas that are either remote or Indian reserves (2% of the Canadian population), the sampling design consisted of a census without follow-up sub-sampling. The final unweighted response rate to the first phase sample of the 2011 NHS is 68.6%. When taking sampling and sub-sampling into account, the design-weighted response rate raises to 77.2%.

The 2016 Census long form sampling design is simpler than the 2011 NHS design and is similar to the 2006 Census long form design. In areas where sampling was performed in the 2011 NHS, a one in four stratified systematic sample of households will be asked to answer the census long form. The rest of the households in these areas will be asked to answer the census short form. No follow-up sub-sampling is planned since the expected response rate is of the same magnitude as the response rate to the 2006 Census long form (94%). In remote areas and in Indian reserves, census data will be collected by canvassers using only long form questionnaires.

3. 2011 NHS weighting and starting point for 2016 Census long form weighting

In this section, both 2011 NHS weighting and the starting point for 2016 Census long form weighting are discussed since the latter is based on the former and in order to provide basic weighting information for the research on variance estimation presented in Section 4.3.2. More details on 2011 NHS estimation are presented in Verret (2013). The households sampled in the 2011 NHS were first weighted to take into account the two-phase sampling design (sampling and sub-sampling). Weights were then adjusted to deal with surprise respondents and with unit non-response observed within the two-phase sample. To perform the non-response adjustment, good auxiliary information available for both respondents and non-respondents is key. The auxiliary information used either came from responses to the census questions or from administrative data linked to the census records. Information coming from the census consisted of geography, dwelling type, household size, demography and language. Information coming from linked administrative data consisted of 2010 Income Tax data, Immigration data and Indian register data. Imputed census data was used as reported auxiliary information for NHS non-respondents that did not respond to the 2011 Census (approximately 3% of Canadian households). Both adjustments consisted of weight transfers. In the case of surprise respondents, a total weight of 1 was transferred to each surprise respondent from their 20 nearest responding neighbours (i.e. nearest in terms of the auxiliary variables) within the sub-sample. In the case of non-respondents to the two-phase sample, their design weight was transferred to their 20 nearest responding neighbours. Nearest neighbour search was performed using the Canadian Census Edit and Imputation System (CANCEIS). To account for the fact that not every record should be linked to the administrative data sources, "non-link" was considered to be a valid value in the comparisons done to calculate the CANCEIS distance.

The resulting household weights were then calibrated to known census totals. This was done independently by geographical units called Weighting Areas (WA), which are contiguous and compact geographies representing between 1,000 and 3,000 households in the population (2,300 on average). Many constraints were considered for calibration. They were based on dwelling type, household size, demography and language. For the 2011 NHS, as well as for most recent census long form samples, the first step of the calibration process consisted of defining for each WA a set of calibration constraints by discarding some constraints from the set of potential constraints based on the observed sample. For the 2011 NHS, constraints with too few responding units (less than 30) contributing to the estimate, which are defined as "small" constraints, were first discarded. The rest of the constraints are referred to as "potential" constraints. A forward selection method was used to choose amongst the potential constraints. The set of selected constraints is first defined to include the two mandatory constraints "total number of households" and "total number of persons" in the WA. Potential constraints are then added one by one to this set starting with the constraint that is the least correlated with the constraints already in the set. This correlation is evaluated with the R-squared of a weighted linear regression. The potential constraint is regressed on the selected constraints using the sample of responding households and the non-response adjusted weights. During the selection process, potential constraints that had a high R-squared (i.e. redundant or nearly redundant constraints) were discarded. Those that were equivalent to calibrating on a small constraint were also discarded, again by doing collinearity checks using the sample-based and weighted R-squared of linear regressions. The second step of calibration is calibration per se on the final set of selected constraints. This was done using bounds on the calibrated weights.

In 2016, the one-phase design weights will first be calculated. Surprise respondents adjustment should not be required as no sub-sampling is expected. Nonetheless unit non-response is expected and weights will be adjusted accordingly. Furthermore, it is planned to divide the non-response adjustment in two based on the auxiliary information available for the non-respondents. For long form non-respondents who do not answer the questions in common with the short form, the known auxiliary data consists of geography, dwelling type and in some cases household size. It is planned to adjust the weight for this first type of non-response by forming non-response weighting classes using the available auxiliary data, mimicking the whole household imputation that will be done to adjust for non-response to the short form questions (i.e. to the census). For sampled households that provide answers only to the short form portion of their long form questionnaire, demography and language will be known. Linkage of census records to administrative data sources will be performed again in 2016. Income, Immigration and Indian register data will thus be known for census records that will be successfully linked. Values should be imputed when linkage is not successful. It is planned to deal with the second type of non-response to the long form with non-response reweighting adjustments that are more classical and based on statistical tests than the 2011 adjustments (e.g. the scores approach or regression trees such as the CART algorithm). The last step of weighting will still be calibration. Although 2016 calibration will be based on 2011 NHS calibration, several changes to the calibration strategy were made or are planned. For a start, calibration constraints based on language were revised with the

collaboration of subject matter experts. Constraints based on linked administrative data will also be considered. Bounds on the calibration adjustment will be put in addition to the bounds on the calibrated weights. The latter bounds will also be revised since the weight before calibration should vary a lot less in 2016 than in 2011.

The weighting methodology described in the last paragraph is the starting point for 2016 weighting. Other changes to 2016 weighting as well as planned research on weighting will be discussed in Section 5. In particular, the weighting methodology has a direct impact on variance estimation as will be seen in the next section. It is thus important to adapt the weighting methodology to limit this impact.

4. Variance estimation

4.1 Variance estimation methodology and variance-based QI dissemination strategy of the 2011 NHS and of most recent census long form samples

Variance estimation in the 2011 NHS and for the previous census long form samples consisted of two main steps. A first estimate of the variance was obtained using the Horvitz-Thompson (HT) variance estimator of the Taylor-linearized calibration/GREG estimator. This estimate was then inflated by a multiplicative adjustment based on Monte Carlo simulations to compensate for a downward bias observed in these simulations. For most recent census long form samples, the HT variance estimator took into account the single phase of sampling and the three successive waves of calibration that were applied to the design weights. In 2011, it took into account both phases of sampling, non-response to the two-phase sample and the single wave of calibration. For the second step, Monte Carlo simulations were conducted using the data of the latest census long form sample available. For variables that are long-form-specific, the median Monte Carlo adjustment to the variance was 1.64 in 2006, while it was 1.16 in 2011. The reduction in the adjustment from 2006 to 2011 can be explained by differences in the calibration methods: in 2006 three waves of calibration were performed compared to one in 2011, but most importantly the number and level of detail of the constraints used in calibration was far greater in 2006.

Monte Carlo simulations using the most recent long form data imposed a significant delay between the publication of point estimates and the availability of QIs based on the estimated variance. For most recent census long form samples, no such QIs were directly published, but analysts could estimate variances approximately by using the information contained in the censuses' Sampling and Weighting Technical Reports, which were published several months after the release of the point estimates. In these reports, the formula for estimating the variance of the estimator of a qualitative variable total under a simple random sampling without replacement design with a 20% sampling fraction is given along with various design effects or design effect distributions for several variables and geographies. A method is provided to the user for finding the most appropriate design effect to use depending on the variable(s) and geography of interest. These design effects took into account the two steps of variance estimation described in the previous paragraph. This strategy for informing users about sampling errors was judged unsatisfactory for the 2011 NHS given the great variation of overall sampling and response rates from domain to domain. Instead, coefficients of variation (CVs) for key qualitative characteristics and for large geographies were either published or made available on demand about a year after the release of the point estimates.

4.2 New objectives for the dissemination of variance-based QIs in the 2016 Census long form sample

For the 2016 Census long form sample, in order to improve timeliness and accessibility to QIs based on the estimated variance, two new dissemination objectives were set. Firstly, census management asked that a variance-based QI be published or made available for every published point estimate in standard dissemination products at the time of release and that suppression of point estimates based on this indicator be considered. Secondly, analysts should be able to either produce their own variance estimates or be able to obtain them through custom requests.

With respect to the first objective, the features of the dissemination of census long form standard products must be taken into account when designing a dissemination strategy for variance-based QIs within these products. Firstly, a very large number of tables and table cells is disseminated: it is estimated that at least 1 billion estimates are published for a given cycle. Secondly, a large proportion of these cells have no respondent contributing to them or almost none. Additionally, it is important to take into account some key facts with respect to data suppression: users will want as much information as possible; there exists no standard for suppression of estimates based on the estimated CV at Statistics Canada, however there are commonly accepted practices for such suppression; and at a

minimum, dissemination needs to be restricted in order to preserve confidentiality. With these considerations in mind, the following decisions were made conditional on operational feasibility, which is under assessment at the moment of writing this paper.

Variance estimation will be done with a replication method. This type of method has been selected for operational reasons. The only feasible option for publishing a variance-based QI with each point estimate is to have the census dissemination system calculate it automatically. Furthermore, programming a HT-type variance estimator within the dissemination system is not feasible, whereas replicating point estimation and programming a simple variance formula based on the replicate estimates is.

The replication method will use a small number of replicates to ensure timeliness of the dissemination of the huge number of point estimates. The number of replicates will be increased from an originally planned value of 16 to 32 to ensure minimum stability of variance estimators. The reasoning behind doubling this value follows from the recommendation given in Section 2.6 of Wolter (1985) for without replacement designs and dependent random groups (DRG) variance estimators. This recommendation is to use as a guideline the CV of the independent random groups (IRG) variance estimator under a simple random sampling with replacement design. The CV of the IRG variance estimator is 36.5% with 16 replicates when the kurtosis of the group/replicate estimators is that of a normal variable. The CV would thus be over the limit of 33.3%, which is commonly used at Statistics Canada to perform point estimate suppression. Conversely, with 32 replicates the CV of the variance estimator drops to 25.4%.

Weighting and variance estimation methods will be chosen so that Monte Carlo adjustments to variance estimates can be avoided. One reason is that it is not possible to perform the Monte Carlo simulations in a timely manner. Furthermore, it would not be feasible to apply domain-dependent and variable-dependent Monte Carlo adjustments within the census dissemination system.

As was done for the 2011 NHS and for past census long form samples, no point estimate suppression based on the estimated variance will occur. Instead, a variance-based QI will be published along with the point estimates for every estimate, namely the standard error, except in cases where confidentiality would be compromised. This will be done to make sure that user needs are met, to preserve consistency of dissemination with previous cycles' dissemination, to avoid suppression/non-suppression that could be inconsistent with one another and to prevent the potential increase of custom requests that suppression could generate. Furthermore, it is planned to provide a flag indicating standard errors calculated with few responding households. This would be done to give information to the users about the possible low quality of the QI and to help them in their statistical inference. In fact, given the great number of tables released and the large number of 0 and near 0 estimates published, the census long form release of estimates has a lot in common with the dissemination of a large microdata file. The flag would thus enable users to apply the informal rule often given to survey microdata users on the minimum number of units contributing to an estimate for this estimate to be considered of good quality for release.

Finally, to meet the second dissemination objective, the same general strategy and the same system would be used for dealing with custom requests made by external users. In addition, the replicate weights would be provided to analysts having access to microdata.

4.3 Recent developments on 2016 variance estimation

Recent research on the 2016 variance estimation methodology had two main goals. The first was to search for a variance estimation method that justifies not using Monte Carlo adjustments in variance estimation. The second was to adopt or design a replication variance estimation method with few replicates for the 2016 Census long form sample that takes into account the sampling design (in particular the large sampling fraction), unit non-response and the point estimation/weighting strategy. Section 4.3.1 describes a Monte Carlo analysis done to attempt to reach the first goal, while Section 4.3.2 presents an analysis done with 2011 NHS data that focuses on the second goal. Note that at the time the latter analysis was performed, it was expected that a voluntary NHS would be conducted in 2016.

4.3.1 Monte Carlo studies using 2006 Census long form data

The Monte Carlo simulations were done using the setting of the 2006 Census long form to better understand the causes of the downward bias in the Taylor-linearization variance estimators of census long form samples and of the

2011 NHS. They were also performed to search for an unbiased replication variance method given the point estimation methodologies. Responses to the 2006 Census long form were first aggregated to form a pseudo-population. In total 12 pseudo-weighting areas were created for the simulation. Since the sampling rate of the survey was 20%, each pseudo-WA was formed by combining the data of 5 WAs from a given province. Similarly, since WAs are made of smaller geographies called dissemination areas (DA), the DAs within pseudo-WAs were combined by groups of 5 to form pseudo-DAs.

Census long form sampling was replicated 500 times to create the Monte Carlo replicates using a stratified simple random sampling without replacement design with a sampling fraction of approximately 20% in each stratum. Note that the sampling fraction in a stratum is not always exactly equal to 20% when the stratum size is not a multiple of 5. Furthermore, although in practice DAs do not perfectly correspond to survey strata, pseudo-DAs were used as strata in the simulation for simplicity and since DAs and the survey strata have a similar size. Unit non-response was not replicated for simplicity and because the response rate of the 2006 long form was high (94% in 2006). Weighting procedures (calculation of design weights and both steps of calibration) and variance estimation of total estimates procedures were applied to each of the 500 Monte Carlo samples. The average of the 500 variance estimates was compared to the Monte Carlo estimate of the variance of the 500 point estimators of totals for various characteristics. The simulations are thus essentially designed to measure the effect of both steps of the calibration procedure on variance estimation under a simple design. This design is also similar to the 2016 design apart from the sampling fraction.

One would not expect to observe biases in the variance estimators when the theoretical conditions for using these estimators are met. However, these conditions do not seem to be met as significant and consistent downward biases were observed for the Taylor-linearization variance estimators in past Monte Carlo studies. Furthermore, these biases can be large, especially for estimators of the variance of characteristics that correspond to potential calibration constraints. Two factors that might explain the observed biases of variance estimators might be the following. First, variance estimation is usually done under the assumption that the set of calibration constraints is fixed, but in the NHS and census long form calibration, the set of constraints used is sample-dependent (as described in Section 3). This factor is discussed in Nascimento Silva and Skinner (1997), where the 1991 Canadian Census long form calibration is evaluated among other calibration approaches. Second, the estimator resulting from the calibration on several tens of constraints by WA might be too complex to justify linearization. For example, in the 2011 NHS, on average for a given WA, 453 responding households were obtained and 41.5 calibration constraints were selected by the Forward procedure.

Four variance estimation methodologies were analysed in the Monte Carlo simulation studies presented in this section. The first is traditional Taylor-linearization variance estimation. The other three are replication methods inspired by Dippo, Fay and Morganstein's (1984) modified balanced repeated replication (BRR), which is sometimes referred to as $\mathrm{BRR}(\varepsilon)$. To keep the number of replicates low, they were partially balanced. The three methods will be referred to as $\mathrm{PBRR}(\varepsilon)-1$, $\mathrm{PBRR}(\varepsilon)-2$ and $\mathrm{PBRR}(\varepsilon)-3$, PBRR standing for partially balanced repeated replication. In all three cases, to form the half-samples, first-phase strata were divided randomly into 15 sub-strata and 2 clusters of households of the same size (plus or minus one unit) were randomly formed within each sub-stratum. Sixteen BRR replicates were created in each stratum. Replicates were thus balanced within strata but not across strata. Formation of the sub-strata, clusters and half-samples was repeated independently for each Monte Carlo sample. This method for forming the replicates, in the fully balanced case, is closely related to the method described in Wolter (1985) on pages 132 and 133, where the strata are divided into substrata of 2 units instead. Rao and Shao (1996) refer to the latter method as alternative balanced half-samples (ABHS) and demonstrate the good asymptotic properties of the corresponding BRR variance estimator when the number of strata is fixed and as strata sizes tend to infinity. The two methods would actually correspond if in practice there had been no limit on the number of replicates.

For all three $\mathrm{PBRR}(\varepsilon)$ methods, the following variance formula was used to estimate the variance of estimator $\hat{\theta}$:

$$\hat{V}\left(\hat{\theta}\right)=\frac{1}{\varepsilon^2 16}\sum_{a=1}^{16}\left(\hat{\theta}_a-\bar{\hat{\theta}}\right)^2, \tag{1}$$

where $\hat{\theta}_a$, $a = 1,...,16,$ are the 16 replicate estimators and $\bar{\hat{\theta}}$ is the average of these estimators. Moreover, both steps of the calibration process were applied to the replicate weights before calculating estimator $\hat{\theta}_a$. For the first step, the definition of a small constraint needed to be redefined at the replicate level. To do this, the full sample definition of a small constraint was first expressed in terms of an HT estimator, namely that a constraint is small when the design-weighted sum of $1/d_k$ over the set of households from the sample contributing to the constraint is less than 30, where d_k is the main sample design weight of household *k*. Consequently, a constraint at the replicate level was considered small if the sum over the set of households from the sample contributing to the constraint of the ratio of the replicate weight d_{ak} and of the main sample weight d_k was less than 30. The standard BRR(ε) replicate weight for half-sample s_a is

$$d_{ak} = d_k\left[1+\varepsilon\left(2I\left\{k \in s_a\right\}-1\right)\right], \tag{2}$$

where *I* is the indicator function. Before describing further the three PBRR methods that were compared, parameter ε will be discussed. Its values are in the interval (0, 1]. Rao & Shao (1999) have shown that at the limit when ε tends to 0, the BRR(ε) variance converges to the Taylor-linearization variance. At the other extreme, an ε of 1 corresponds to standard BRR. Small epsilons favor variance estimation of smooth functions of totals. It is also easier to replicate calibration with small epsilons because the replicate weight is then more similar to the main weight. However, such epsilons can be detrimental to variance estimation of estimates of quantiles. The American Community Survey uses an ε of one half. Finally, in the context of the NHS and of the census long form, the ε value also has an impact on the variability of the constraints selected from one replicate to the next. Although a small value of ε should facilitate the second step of calibration (calibration per se on the selected constraints), if this value is too small then replication might not take into account the variability due to the first step of calibration (the sample-dependent constraint selection).

PBRR$(\varepsilon)-1$ consists of a) using $\varepsilon = \sqrt{1/2}$ for generating d_{ak}, b) applying calibration to the resulting replicate weights and c) using the following weights when calculating $\hat{\theta}_a$ to take into account the large sampling fraction:

$$w_{ak}^{*} = \sqrt{1-f_h}\,g_{ak}d_{ak} + \left(1-\sqrt{1-f_h}\right)g_k d_k, \tag{3}$$

where f_h is the sampling fraction in stratum *h*, g_{ak} is the calibration factor applied to d_{ak} and g_k is the calibration factor applied to d_k. The correction given in (3) is similar to the one applied in the American Community Survey for multiyear estimates (U.S. Census Bureau, 2009).

In PBRR$(\varepsilon)-2,$ instead of applying the post-calibration correction factor given in (3), the value of ε is allowed to vary from unit to unit when generating the replicate weights (2). More precisely, ε is replaced with $\varepsilon_k = \sqrt{1-f_h}$ in (2). Furthermore, ε is set to the constant 1 when calculating variance (1). This approach for taking the large sampling fraction into account is similar to the method described in Wolter (1985) on pages 120 and 121. It also bears some similarities with the generalized bootstrap of Beaumont & Patak (2012) in that the distribution of the Monte Carlo sub-sampling errors given the sample mimics the distribution of the sampling errors. This approach also has the advantage over PBRR$(\varepsilon)-1$ of having calibration as the last step of replicate weighting, which means calibration is preserved when calculating $\hat{\theta}_a$ and in turn guaranties that the estimated variance is null for a calibration constraint that is used in every replicate.

Finally, PBRR$(\varepsilon)-3$ is a mix of the other two approaches. It uses $\varepsilon_k = \sqrt{(1-f_h)/2}$ when generating replicate weights (2) and $\varepsilon = \sqrt{1/2}$ when calculating variance (1).

Table 1 gives summary statistics of the difference between the root relative expected variance and the Monte Carlo CV in percentage for the four variance estimation methods compared and for estimated totals of various variables of interest. More precisely, the root relative expected variance is the square root of the Monte Carlo average of the estimated variance, divided by the population total. The Monte Carlo CV is the square root of the Monte Carlo variance, divided by the population total. The table distinguishes between short form characteristics, which corresponds to potential calibration constraints, and long form characteristics. No short form characteristic was studied for the Taylor method and only 1 out of the 12 pseudo-WAs was studied for PBRR$(\varepsilon)-2$. Only characteristics with a population total of 100 or more were retained in the analysis.

Table 1: Statistics on the difference between the root relative expected variance (%) and the Monte Carlo CV (%) for four variance estimation methods and for two types of variables of interest

Variable type	Method	Number of point estimates	Mean	1st Pctl	5th Pctl	Q1	Q2	Q3	95th Pctl	99th Pctl
Short form	Taylor	0	.	.	.	.	.	.	.	.
	PBRR(ε)-1	603	0.8	-1.1	-0.1	0.1	0.4	1.3	2.7	3.7
	PBRR(ε)-2	52	1.3	0.0	0.0	0.4	1.2	2.2	3.1	3.5
	PBRR(ε)-3	603	0.7	-1.5	-0.1	0.0	0.3	1.2	2.7	3.7
Long form	Taylor	2,564	-0.8	-4.3	-2.6	-1.2	-0.5	-0.2	0.0	0.0
	PBRR(ε)-1	2,564	0.0	-1.6	-0.8	-0.1	0.0	0.2	0.7	1.3
	PBRR(ε)-2	212	0.7	0.0	0.0	0.1	0.7	1.2	1.9	2.6
	PBRR(ε)-3	2,564	-0.1	-2.1	-1.1	-0.3	-0.0	0.1	0.5	0.9

Table 1 shows that the Taylor variance estimation method underestimates the variance of long form characteristics at least 75% of the time. At the other end of the spectrum, PBRR$(\varepsilon)-2$ overestimates the variance at least 75% of the time for both short form and long form characteristics. Not surprisingly the other two methods give rather similar results to one another. PBRR$(\varepsilon)-1$ tends to overestimate the variance for short form variables a bit more. As for long form variables, both methods are well centered, but PBRR$(\varepsilon)-3$ has a slight tendency to underestimate the variance. The PBRR(ε) results thus seem to be sensitive to the choice of ε.

Figures 1 to 4 of the Appendix present similar results, but in the form of scatter plots. The root relative expected variance (%) is plotted against the Monte Carlo CV (%) (both are called CVs on the axes for simplicity). Short form and long form variables are again studied independently. The graphs also show the 45 degrees line (full line) as well as the line of a regression with no intercept (dashed line). The Taylor-linearization underestimation for long form variables is again very clear on Figure 1. Interestingly, it seems that a common Monte Carlo adjustment to the variance for every characteristic would be appropriate for correcting the Taylor variance estimator's bias. Figure 3 shows the very frequent overestimation of PBRR$(\varepsilon)-2$. PBRR$(\varepsilon)-1$ and PBRR$(\varepsilon)-3$ again show very similar results to one another on Figures 2 and 4 respectively. There is an "S" shape around the 45 degrees line in both graphs for short form variables. Also the estimates of the slopes indicate that PBRR$(\varepsilon)-3$ tends to overestimate a little bit less than PBRR$(\varepsilon)-1$ for these variables. For long form variables, the graph of PBRR$(\varepsilon)-1$ is rather perfect and PBRR$(\varepsilon)-3$ shows a slight underestimation. For both methods, the differences between the two "CVs" tend to be minimal.

Figure 5 of the Appendix plots the difference between the root relative expected variance (%) and the Monte Carlo CV (%) against the population total for method PBRR$(\varepsilon)-3$ for both short form and long form variables. The short form variables graph was truncated to present only population totals of less than 1,000. On this graph, there is a very apparent pattern around a population total of 200. This corresponds more or less to the population equivalent of the limiting value used to define small constraints based on the sample distribution. For greater values of the

population total (including those not shown on the graph) there is a consistent and slight overestimation of the variance. For long form characteristics, the difference tends to vary a lot for smaller totals, otherwise the difference is very near 0. The difference between the PBRR$(\varepsilon)-3$ "CV" and the Monte Carlo CV is rather small, especially for large values of the population total.

In summary, the results of the Monte Carlo simulations show that the Taylor-linearization variance estimation method is, apart from not being an approach that could be implemented in the dissemination system, not appropriate for 2016 Census long form variance estimation. Given the calibration strategy, this method produces more or less systematically downward biased variance estimates. At the other end of the spectrum, PBRR$(\varepsilon)-2$ systematically overestimates the variance. This is probably due to the fact that ε_k is close to 1 for this method making it similar to the PBRR approach. The replicate weights before calibration d_{ak} are thus close to 0 or 2 times the weight d_k. This makes it difficult for the second step of calibration to be performed in a fashion that is similar to the second step of calibration of d_k, which in turn could increase the estimated variance. PBRR$(\varepsilon)-1$ and PBRR$(\varepsilon)-3$ behave very similarly to one another, with the former having a slight advantage for long form variables and the latter having a slight advantage for short form variables.

In addition, a more in depth study of the results of PBRR$(\varepsilon)-3$ was performed variable by variable. It was found that overestimation for short form variables is to a very large extent due to the definition of small constraints, and, to a lesser extent to elimination of constraints based on collinearity checks done at the sample level. Overestimation due to the definition of small constraints is in fact maximal for constraints that have an average number of households contributing to the constraint the closest to the value used in the definition of a small constraint. Overestimation is present in general for short form variables, except for some language variables. For long form variables, variables related to calibration constraints tended to show a slight overestimation, while the rest tended to show a slight underestimation.

4.3.2 Study of replication variance estimation methods using 2011 NHS data

Before the 2016 Census mandatory long form was reinstated in November 2015, several replication variance estimation methods were studied using 2011 NHS data in view of a voluntary 2016 NHS. For each method studied, 16 replicate weights were created. Depending on the method studied and except for the very first method considered, one of two types of replicate calibration was done. For the first type, the set of constraints used to calibrate the replicate weights was fixed to the set selected with the main weight. In this case, the estimated variance was compared directly to the HT variance estimator of the Taylor-linearized point estimator. For the second type, the constraint selection step was replicated and comparison was made with the published variance estimate, i.e. with the HT variance estimate of the Taylor-linearized point estimator multiplied by the Monte Carlo adjustment. In both cases, the comparisons were done under the assumptions made to derive the Taylor variance, that is 1) that unit non-response to the two-phase sample is equivalent to a third phase of sampling with the estimated probabilities of response treated as known and 2) ignoring surprise responses. National, provincial and territorial variance estimates were studied.

The development of the variance estimation methodology started with the study of the standard DRG method applied to the full sample calibrated weights (i.e. calibration was not replicated at the time). This method was first studied because of its ease of implementation. Groups were formed by dividing the first-phase sample in a stratified systematic fashion to mimic the first phase of sampling. The variance estimates with this approach were greater than the corresponding Taylor variance estimates, especially in the territories where the overall sampling fraction is the greatest since the NHS consists of a census in the majority of the areas composing the territories.

The method was then refined by applying a correction factor similar to the one given in (3), but with g_{ak} replaced with g_k and with f_h replaced with the overall sampling fraction of unit k, $f_k = 1/d_k$, where d_k is, in this study, the NHS weight before calibration (the design weight adjusted for non-response). The correction proved to be very beneficial in the territories. Replicating calibration of the DRG weight d_{ak} was then attempted to obtain g_{ak} in equation (3) and proved to be very challenging with sub-samples only 1/16th of the original sample. Calibration

often failed because there existed no feasible solution to the calibration minimisation problem when the calibrated weights were bounded to be positive.

In order to increase the size of the replicate sub-samples, PBRR was then considered instead. To form the half-samples, first-phase strata were divided systematically into 15 sub-strata and 2 systematic clusters of households of equal size (plus or minus one unit) were defined within each sub-stratum. Replicates were balanced within first-phase strata but not across first-phase strata. With this method, calibration could be performed. However, the resulting variances were 10% to 20% greater than the published variances depending on the province or territory.

Therefore PBRR(ε) with $\varepsilon < 1$ was considered to further increase the replicate sample sizes and to produce replicate weights closer to the full sample weight. The value of ε was also allowed to vary from unit to unit instead of using the correction given in (3). PBRR$(\varepsilon)-2$, with f_h replaced by f_k, was first tested with the replication of both steps of calibration. The resulting estimated variances were even greater than with PBRR: they were 40% to 60% greater than the published variances. For this reason, smaller values of replicate weight generation parameter ε_k and of variance estimation formula parameter ε were then considered. The last approach studied will be called PBRR$(\varepsilon)-4$. For this approach, the value of ε_k was set to $\sqrt{1-f_k}/2C$ to generate the replicate weights, where C is the national average of $\sqrt{1-f_k}$ of households with $f_k < 1$. This guaranties an average value of ε_k of 0.5 for these households. When estimating the variance with (1), ε was in turn set to $1/2C$.

Table 2: Slopes of a regression through the origin of the PBRR(ε)-4 CV on the Taylor CV

Province or Territory	Canada	Newfoundland and Labrador	Prince Edward Island	Nova Scotia	New Brunswick	Quebec	Ontario
Slope	1.02	1.05	1.05	1.05	1.03	1.03	0.99
MSE of errors	0.03	0.15	0.16	0.16	0.10	0.07	0.05

Province or Territory	Manitoba	Sasketchewan	Alberta	British Columbia	Yukon	Northwest Territories	Nunavut
Slope	1.05	1.02	1.05	1.00	1.02	1.03	0.98
MSE of errors	0.12	0.13	0.09	0.06	0.19	0.13	0.04

Table 2 gives, for national, provincial and territorial estimates of totals, the slope of a regression through the origin of the PBRR$(\varepsilon)-4$ CV on the Taylor CV, where calibration is applied to the replicate weights with a fixed set of constraints (the one selected with the main weight). A weight of 1 over the Taylor CV was used in the regression to make large CVs less influential on the slopes. This analysis is presented instead of an analysis comparing the published CV with the CV of PBRR$(\varepsilon)-4$ with replication of constraint selection, the reason being that the results are less difficult to interpret.

Most slopes are greater than unity and similar from one region to the next. The MSE of regression errors are also provided in the table and are all very small, showing that the data points are very close to the regression lines. Many reasons could explain the greater than unity slopes observed and it is not clear which of the two methods compared for estimating CVs should be the baseline. On the one hand, the PBRR$(\varepsilon)-4$ approach could be overestimating the CV because the finite population correction applied is an approximation of the appropriate correction. Another explanation for such an overestimation would be that $\varepsilon_k = \sqrt{1-f_k}/2C$ is still too close to 1 for calibration to be performed in a fashion similar to the main sample calibration. However, going below an average ε_k of 0.5 is in

practice not desirable when one is interested in estimating the variance of quantiles, which is the case with the census long form. On the other hand, the Taylor variance could be underestimating the variance. A similar result has been obtained by Stukel et al. (1996). In their paper, the underestimation was more pronounced when applying a given calibration strategy to a reduced sample. A similar phenomenon could be happening in the NHS given the great number of constraints selected for calibration by WA.

5. Discussion and next steps

The results obtained so far in the development of a variance estimation methodology for the 2016 Census long form give some insight on the estimation methods that should be adopted. Section 4.3.1 presented a Monte Carlo analysis of the impact of calibration on variance estimation in the context of the 2006 Census long form, which is similar to the 2016 context. The focus of the study was on capturing the variability due to both sampling and estimation, and in particular to the sample-dependent calibration constraint selection. The results indicated that the standard Taylor-linearization variance did not capture this variability properly. The method is in fact not designed to do so and is only valid under the assumption that the set of constraints is fixed.

Furthermore, Dr. Kott indicated in his discussion of the 2015 FCSM conference presentation of this paper that Taylor-linearization variance estimation might be showing downward bias even when the set of constraints used is fixed. He explained that it is because calculation of the HT variance estimator would require using the population residuals $E_k = y_k - \mathbf{x}_k^T\mathbf{B}$, where $\mathbf{B} = \left(\sum_U \mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\sum_U \mathbf{x}_j y_j$, but in practice they are replaced by their sample equivalent $e_k = y_k - \mathbf{x}_k^T\hat{\mathbf{B}}$, where $\hat{\mathbf{B}} = \left(\sum_s d_j\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\sum_s d_j\mathbf{x}_j y_j$, The variability of e_k is less than the variability of E_k since $e_k = E_k - \mathbf{x}_k\left(\sum_s d_j\mathbf{x}_j\mathbf{x}_j^T\right)^{-1}\sum_s d_j\mathbf{x}_j E_j$. Furthermore, this difference in variability is more pronounced as the number of calibration constraints is large compared to the sample size. Dr. Kott also pointed out that replication methods do not suffer from this, but that they may however suffer from the fact that replication of calibration can be difficult when the number of constraints is large.

For replication methods of the PBRR(ε) type, the Monte Carlo simulations showed that given the calibration methods used, the value of ε_k should be small enough for the second step of calibration to be replicated in a fashion representative of the second step of calibration of the main sample. Moreover, the variance of two types of variables was studied: calibration constraints (or short form variables) and variables of interest of the survey (or long form variables). For calibration constraints that are sure to be selected, an estimated variance of 0 would be desirable. A null variance estimate has a better chance of being obtained with an approach such that a value of $\varepsilon_k \propto \sqrt{1-f_h}$ is used for generating the replicate weights d_{ak} than with an approach where this value is fixed and where the large sampling fraction correction is applied after calibration of the replicate weights. In terms of long form variables, the two PBRR(ε) methods with the smallest values of ε_k performed well. However, all replication methods failed to properly capture the variability of calibration constraints that are not sure to be selected in every sample, in particular constraints for which their "small" status tends to vary greatly from one sample to the next or from one sub-sample to the next. Moreover, it is not clear if the between-sample variability due to constraint selection can be measured by a method designed to measure a within-sample variability (i.e. a replication method).

Section 4.3.2 presented the results of a study of replication methods in the more complex context of the 2011 NHS. The inclusion of the large sampling fraction correction in the PBRR(ε) showed to be very beneficial to avoid variance overestimation. This is especially true in areas where the overall sampling fraction is large such as in the territories. To this effect, Dr. Kott supported the use of the overall rate $f_k = 1/d_k$ rather than the use of the first-phase sampling fraction for its good model-based properties. It was also apparent in this analysis that with PBRR(ε) methods, the selected value for ε_k should not be too large in order for the second step of calibration to be replicated in a representative manner. Additionally, the analysis showed that there were some discrepancies between the variances of the Taylor linearization approach when compared to PBRR(ε) with an average value of ε_k of 0.5 and with a fixed set of constraints. Several factors could explain this difference, one of them being, as

discussed by Dr. Kott and outlined in the first paragraph of this section, the complexity of calibration, or in other words the number of constraints used in calibration given the sample size.

A natural solution for improving variance estimation in the 2016 Census long form and for obtaining better results in the Monte Carlo studies is in fact modifying the weighting methodologies. One possibility would be to make constraint selection more stable from one sample to the next, as well as from one replicate to the next. In other words, this means aiming towards having a fixed set of constraints, which is the classical assumption made when using the Taylor-linearization variance method or the other variance estimation methods studied. Another possibility would be to reduce the complexity of calibration by reducing the number of constraints relative to the sample size. In his discussion, Dr. Kott proposed a rule of thumb for the maximum number of calibration constraints. This number should not exceed the square root of the number of responses at the level where calibration is performed. In the context of the 2011 NHS, this corresponds to no more than 21 constraints used on average by WA, which is close to half the average number of constraints actually used. To make constraint selection less sample-dependent, a population-based and more conservative condition could be added to the definition of a small constraint. For example, the criterion could be that a constraint is considered to be small if there are fewer than 30 households contributing to the estimate or if the population total is less than 300. A criterion on the sample size would still be needed in practice in case a very small number of respondents with the characteristic of interest is observed for a given constraint. Furthermore, to reduce the complexity of calibration, more aggregate versions of the constraints could be considered. This could be done by collapsing the existing constraints, for example by combining 5 year age groups into 10 year age groups, or by performing calibration at a larger geographical level.

In fact, the next planned steps of development of the 2016 Census long form estimation methodologies are to study both of these types of collapsing. Simultaneous calibration on existing WA-level constraints and on aggregate versions of these constraints is being studied. Simultaneous calibration at the WA level and at the super-WA level is also being studied, a super-WA being the aggregation of 10 WAs on average. However, both of these changes increase the overall number of constraints sent to calibration. To reduce this number, one option considered is to be stricter on the definition of a small constraint by adding a criterion based on the population total similar to the one described in the previous paragraph. These additions will be evaluated in particular by revisiting the Monte Carlo simulations.

Variance estimation will also need to be developed further, in particular with regards to increasing the number of replicates from 16 to 32. Dr. Kott made the remark that variance of the variance estimators should be studied in addition to the bias of the variance estimators. In Figure 6 of the Appendix, the Monte Carlo CV of the variance estimators of method $\mathrm{PBRR}(\varepsilon)-3$ is plotted against the corresponding relative expected variance for both short form and long form variables. A horizontal line was also drawn on the graphs at 36.5% to represent the CV one would expect under the conditions outlined in Section 3 with 16 replicates. For short form variables, the CV of the variance estimate is close to 36.5% for "CVs" of the point estimate of 10% or more and tends to explode for small point estimate CVs. For long form variables, the 36.5% approximation is good for point estimate CVs between 2% and 20%. Again the variance estimate CV tends to explode for small point estimate CVs. For point estimate CVs over 20%, the variance estimate CV is somewhat greater than 36.5%. This will be further studied with the increase of the number of replicates to 32 and with the improvements made to calibration.

The choice of a variance estimation methodology will also be further studied. Dr. Kott suggested that the Jackknife method described in Kott (2001) be evaluated. Furthermore, a presentation similar to the one given at the 2015 FCSM conference was given to Statistics Canada's Advisory Committee on Statistical Methods (of which Dr. Kott is a member). Professor Jon Rao, the discussant of this presentation, suggested that the Jackknife of Kott (2001) be studied as well as bootstrap methods. He also remarked that one should choose a replication method for which one can justify linearization of the replicate estimates, which is the case of the Jackknife and bootstrap methods. We assume these methods should give good results, as long as the constraint selection is made more stable since methods favouring replicate estimate linearization could be detrimental to capturing the variability due to sample-dependent constraint selection. Professor Rao also expressed some concerns about the asymptotic properties of the replication methods used in the variance studies because the method chosen should have good asymptotic properties when the number of strata is fixed. However, although the method used in the studies does not correspond to method ABHS of Rao and Shao (1996), the two methods would correspond if there were no restriction on the number of replicates as in Rao and Shao's ABHS convergence theorem.

Finally other weighting steps will also be revisited before production starts in the summer of 2016. Non-response adjustments that are more standard and more directly based on statistical tests than the 2011 NHS adjustment will be evaluated. The estimation methodologies for Indian reserves and remote areas will also be revisited.

Acknowledgments

The authors would like to thank Yannick Bridé for his great contribution to the project during his internship, in particular in the Monte Carlo studies and in theoretical development.

References

Beaumont, J.-F. & Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80, 1, 127-148.

Dippo, C.S., Fay, R.E. & Morganstein, D.H. (1984). Computing variances from complex samples with replicate weights. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 489-494. Washington DC: American Statistical Association

Hansen, M.H. and Hurwitz, W.N. (1946). The problem of nonresponse in sample surveys. *Journal of the American Statistical Association*, 41, 517-429.

Kott, P. (2001). The delete-a-group jackknife. *Journal of Official Statistics*, 521–526.

Nascimento Silva, P.L.D. & Skinner, C.J. (1997). Variable selection for regression estimation in finite populations. *Survey Methodology*, 23, 1, 23-32.

Rao, J.N.K. & Shao, J. (1996). On balanced half-sample variance estimation in stratified random sampling. *Journal of the American Statistical Association*, 91, 343-348.

Rao, J.N.K. & Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

Stukel, D.M., Hidiroglou, M.A. & Särndal, C.-E. (1996). Variance estimation for calibration estimators: a comparison of jackknifing versus Taylor linearization. *Survey Methodology*, 22, 2, 117-125.

U.S. Census Bureau. (2009). Design and Methodology, American Community Survey. Washington, DC.

Verret, F. (2013). The estimation methodology of the 2011 National Household Survey. *Proceedings of the Survey Research Methods Section, American Statistical Association,* pp. 1876-1890. Montréal, Quebec, Canada: American Statistical Association.

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

Appendix

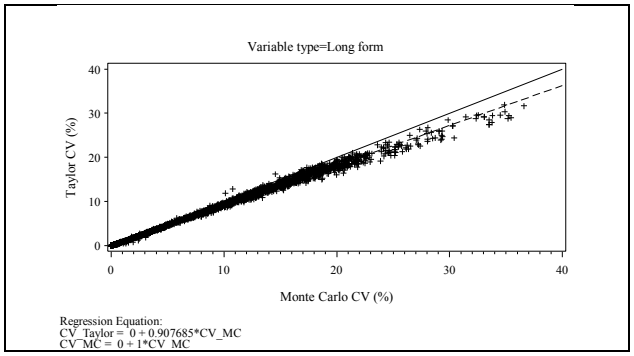

Figure 1: Plot of the root relative expected variance of the Taylor-linearization variance method (%) against the Monte Carlo CV (%) for long form variables

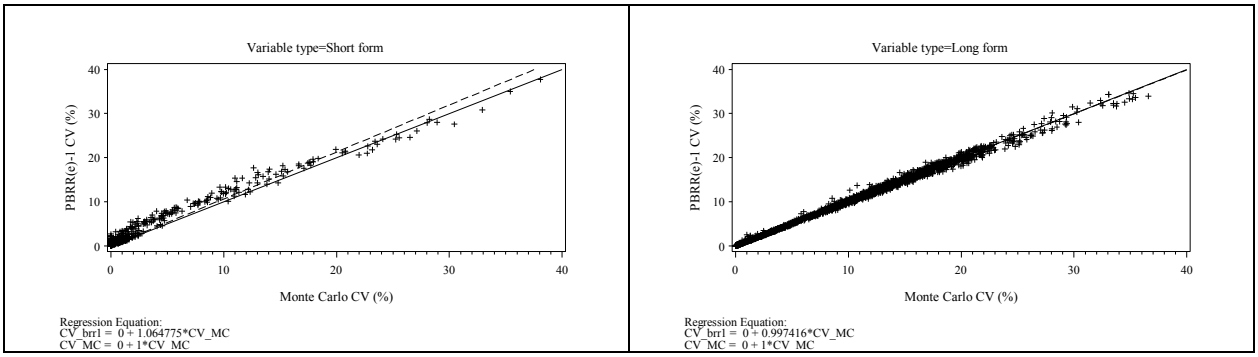

Figure 2: Plot of the root relative expected variance of PBRR(ε)-1 (%) against the Monte Carlo CV (%) for short form and long form variables

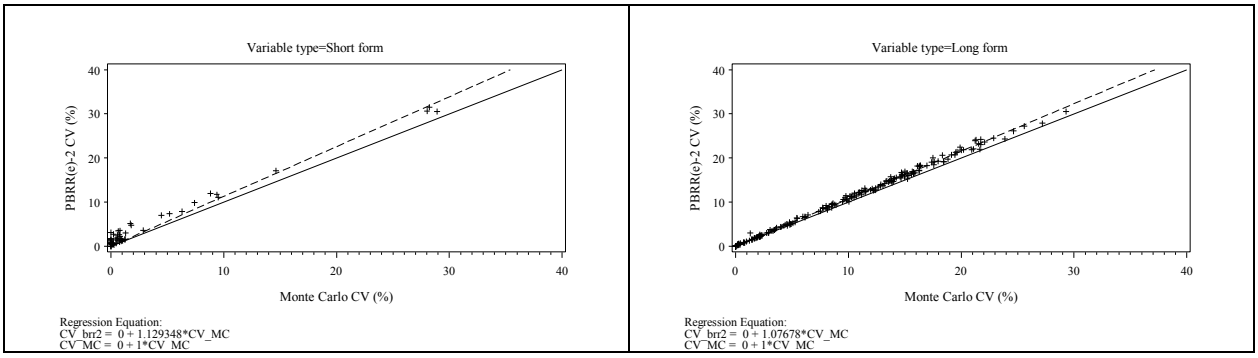

Figure 3: Plot of the root relative expected variance of PBRR(ε)-2 (%) against the Monte Carlo CV (%) for short form and long form variables

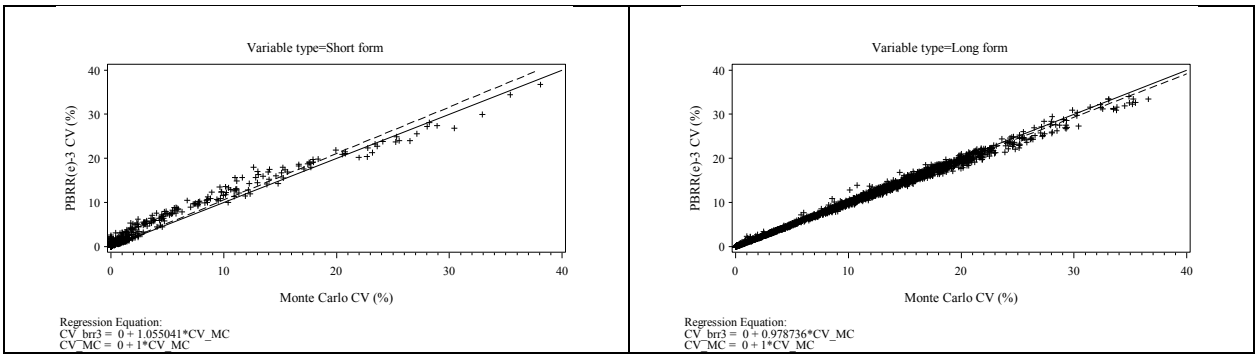

Figure 4: Plot of the root relative expected variance of PBRR(ε)-3 (%) against the Monte Carlo CV (%) for short form and long form variables

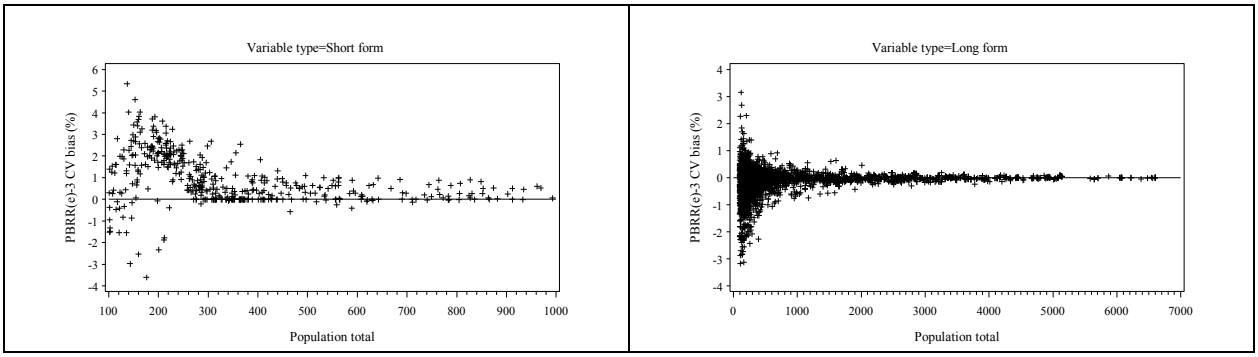

Figure 5: Plot of the difference between the root relative expected variance of PBRR(ε)-3 (%) and the Monte Carlo CV (%) against the population total for short form and long form variables

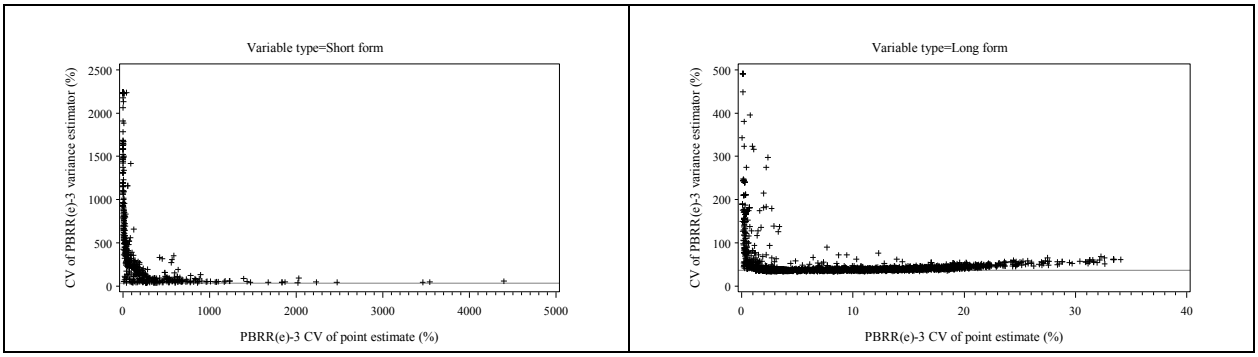

Figure 6: Plot of the Monte Carlo CV of the PBRR(ε)-3 variance estimator (%) against the root relative expected variance of PBRR(ε)-3 (%) for short form and long form variables