Results from the 2006 Canadian Census Weighting Process

Wesley Benjamin[1], Darryl Janes[1], Mike Bankier[1]
[1]Statistics Canada, Ottawa, ON, K1A 0T6

Abstract

In the Canadian Census, basic person and dwelling information is gathered on a 100% basis, and additional information is collected from a 20% random sample of private households. Initial weights of approximately five are applied to the sampled population but are then calibrated independently for approximately 6,600 geographical regions through a multi-step regression estimation process to produce the final weights. Within each region, auxiliary variables are discarded from the weighting process for being collinear, nearly collinear or causing outlier weights. The weighting process is evaluated by comparing the estimates based on the final weights with the 100% totals for all the auxiliary variables. In addition, biases in the census sample for the auxiliary variables are studied by comparing estimates based on the initial weights to the 100% totals at both the national and regional levels.

Key Words: census, household weights, sampling bias, discrepancies, calibration, regression estimation

1. 2006 Weighting Process Overview

The 2006 Canadian Census collected data from the entire population of over 31 million persons. 100% of the population responds to basic demographic and household questions. This is referred to as 2A information, after the 2A short questionnaire that approximately 80% of the population receives. A 20% sample of private households responds to additional questions on education, labour force status, ethnic origin, shelter cost and others. This is referred to as 2B information, after the 2B long questionnaire received by the sampled population. The 1 in 5 sample is subject to small but significant biases as will be displayed in the results from 2006.

A single weight is calculated for each sampled household. These weights are used to produce all published 2B person and household characteristic estimates. Published estimates of 2A information from the sampled households should agree closely with the published 100% counts of the same 2A information. The Census weighting methodology is designed to reduce or eliminate such population/estimate differences at the national level as well as for smaller geographical areas (there are some inherent minor differences, such as the 100% counts include institutional residents while the estimates do not). At the same time, the standard errors of the Census estimators are reduced. 2A characteristics for which population count and sample estimate consistency is desired will be called auxiliary variables, or alternatively, constraints on the weights. 34 auxiliary variables were used in 2006, consisting of five year age ranges, marital status, common-law status, sex, household sizes, dwelling type, population count and household count. Certain WAs only used 32 of the 34 constraints, as two constraints related to dwelling type were only applied to a subset of the WAs where they showed a positive impact. For simplicity, the discussion below will always refer to 34 constraints.

The sampled population is grouped into weighting areas (WAs), and the weighting system is run independently for each WA. There were 6602 WAs subject to sampling, with the majority of WAs containing between 1,000 and 3,000 dwellings often representing small municipalities or neighbourhoods within larger municipalities. WAs are formed by grouping together approximately eight smaller geographic areas called dissemination areas (DA). Canada is partitioned into 53,654 DAs with, on average, 233 private occupied households in each. Although the 1 in 5 sample is taken at a similar geographic level, the collection unit, this paper will assume the sample is taken at the DA level for simplicity.

The principle objectives of the weighting process are as follows:

(a) To have exact estimate/population agreement at the WA level for the two auxiliary variables total number of households and total number of persons and as many of the remaining 32 auxiliary variables as possible.
(b) To have approximate estimate/population agreement at the DA level for the 34 auxiliary variables.

In addition, it is required that:
(c) There should be exact estimate/population agreement for the total number of households and total number of persons for as many DAs as possible.
(d) Final census weights should be in the range 1 to 25 inclusive to ensure that every sampled household represents at least itself in the estimate but does not have too great an effect.
(e) The method to generate weights should be highly automated since the 6602 WAs must be processed in a short period of time.

The final census weights are calculated through a four step process. For simplicity, estimators for a single WA made up of H DAs will be discussed. It will be assumed that a simple random sample without replacement (s.r.s.w.o.r.) of size n_h has been selected from the population of N_h households in the hth DA, h=1 to H and that $n = \sum_h n_h$ and $N = \sum_h N_h$. The **initial simple weights** are calculated as the inverse of the achieved sampling fraction within each DA ($W_i^{(0)} = N_h / n_h$ if the ith sampled household is in the hth DA). In 2006, senior residences were 20% sampled instead of being treated as a 100% sampled collective as was done in the past. There are notable differences in the characteristics of this sub-group from the rest of the population. Due to this, DAs with a significant number of both senior and non-senior dwellings were split into subpopulations for the creation of initial weights. As an optional step, the initial weights are recalculated after households are **poststratified** on household size (1,2,3,4,5,6+ persons) at the WA level ($W_i^{(*)} = g_i^{(*)} W_i^{(0)}$ where $g_i^{(*)}$ represents an adjustment factor of the type described in section 1.1). This is done because small and very large households tend to be under-represented in the sample due to response bias. In 2006, poststratification was performed for approximately 66% of the WAs. $W_i^{(*)} = W_i^{(0)}$ when poststratification was not performed.

Next, a **first step** regression weighting adjustment factor is calculated at the DA level. The 34 auxiliary variables are sorted in descending order based on the number of households they apply to in the population at the DA level. The first, third, etc. constraints on this ordered list go into one group while the other 17 constraints go into a second group. Weighting adjustment factors are calculated separately for each group of constraints for each DA (labeled $g_i^{(A1)}$ and $g_i^{(A2)}$). The two adjustment factors are averaged together to generate $g_i^{(A)} = (g_i^{(A1)} + g_i^{(A2)})/2$ which then generates the first step weights $W_i^{(A)} = g_i^{(A)} W_i^{(*)}$. Estimate/population differences at the DA level for the 34 constraints are usually reduced but not eliminated using the first step weights.

Finally, a **second step** regression weighting adjustment factor is calculated at the WA level. The 34 constraints are applied at the WA level along with two constraints (number of households and number of persons) for each DA in the WA to determine the final weighting adjustment factor g_i. These then generate the second step weights $W_i = g_i W_i^{(A)}$. These are the final household weights.

1.1 Optimal Regression Estimator

An excellent review on the subject of regression estimation for survey samples is given by Fuller (2002). The simplest estimator possible is the Horvitz-Thompson estimator $\hat{Y}^{(0)} = \sum_i W_i^{(0)} y_i$ where $W_i^{(0)} = N_h / n_h$. Generally, however, there is no guarantee that objective (a) above will be achieved for any of the 34 auxiliary variables with the Horvitz-Thompson estimator. It is for this reason that the optimal regression estimator is considered below, as recommended by Cochran (1942) and Rao (1994). This estimator is considered optimal because it minimizes the variance of the estimate.

Calibration estimators take the form $\hat{Y}=\underset{\sim}{\hat{Y}}^{(0)}\underset{\sim}{g}=\sum_i g_i W_i^{(0)} y_i$ where the n x 1 vector $\underset{\sim}{g}=[g_i]$ of weighting adjustment factors (otherwise known as g-weights) is chosen such that some loss function L is minimized subject to constraints $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{g}=\underset{\sim}{X}\underset{\sim}{1}_N$ where $\underset{\sim}{\hat{Y}}^{(0)}=[W_i^{(0)}y_i]$ is a 1 x n matrix, $\underset{\sim}{X}=[x_{pi}]$ is a P x N matrix, x_{pi} represents the value for the pth auxiliary variable for the ith household in the WA, $\underset{\sim}{\hat{X}}^{(0)}=\underset{\sim}{x}\,diag(\underset{\sim}{W}^{(0)})=[W_i^{(0)}x_{pi}]$, $\underset{\sim}{x}$ is a P x n matrix which contains the n columns from $\underset{\sim}{X}$ which correspond to the sampled households, $\underset{\sim}{W}^{(0)}=[W_i^{(0)}]$ is a nx1 vector of the initial weights and $diag(\underset{\sim}{W}^{(0)})$ is an n x n matrix with $\underset{\sim}{W}^{(0)}$ running down the diagonal with zeros elsewhere.

With the optimal estimator in its most general form, the loss function takes the form $L=(\underset{\sim}{g}-\underset{\sim}{1}_n)'\underset{\sim}{\hat{V}}(\underset{\sim}{g}-\underset{\sim}{1}_n)$ and the vector $\underset{\sim}{g}$ which minimizes L is

$$\underset{\sim}{g}=\underset{\sim}{1}_n+\underset{\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{X}}^{(0)'}(\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{X}}^{(0)'})^{-1}(\underset{\sim}{X}\underset{\sim}{1}_N-\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{1}_n)$$

where $\underset{\sim}{\hat{V}}$ is assumed to be a symmetric n x n matrix which has to be positive definite (which in turn implies that it is nonsingular) to ensure that the loss function L is non-negative.

In the first step $\hat{V}_i=W_i^{(*)}/(W_i^{(*)}-1)$ while in the second step $\hat{V}_i=W_i^{(A)}/(W_i^{(A)}-1)$. These choices of $\hat{V}_i$ in the first and second steps make the loss function being minimized resemble that used with the optimal regression estimator. They also encourage the generation of first and second step weighting adjustment factors close to 1 for the smaller poststratified and first step weights and hence discourage the creation of adjusted weights less than 1. Because of this choice of $\hat{V}_i$, the estimator used in the Census will be called a **two step pseudo-optimal estimator**. Bankier and Janes (2003) provide further information on this choice of estimator. The variance of this two step regression estimator can be estimated by using Taylor Series to numerically linearize the data two or three times in a fashion similar to that proposed for raking ratio estimation in Bankier (1986).

It is possible to write $\hat{Y}$ in the standard form of a regression estimator as $\hat{Y}_{opt}=\hat{Y}^{(0)}+\underset{\sim}{B}_{opt}(\underset{\sim}{X}\underset{\sim}{1}_N-\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{1}_n)$. The variance of $\hat{Y}_{opt}$ is minimized if $\underset{\sim}{B}_{opt}=\underset{\sim}{\Sigma}_{xx}^{-1}\underset{\sim}{\sigma}_{yx}$ where $\underset{\sim}{\Sigma}_{xx}$ and $\underset{\sim}{\sigma}_{yx}$ represent respectively the P x P covariance matrix of $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{1}_n=[\hat{X}_p^{(0)}]$ and the P x 1 vector of covariances $Cov(\hat{Y}^{(0)},\hat{X}_p^{(0)})$. The standard estimator of $\underset{\sim}{B}_{opt}$ (which is approximately unbiased) is $\underset{\sim}{\hat{B}}_{opt}=\underset{\sim}{\hat{\Sigma}}_{xx}^{-1}\underset{\sim}{\hat{\sigma}}_{yx}$ where $\underset{\sim}{\hat{\Sigma}}_{xx}$ and $\underset{\sim}{\hat{\sigma}}_{yx}$ are unbiased estimators of $\underset{\sim}{\Sigma}_{xx}$ and $\underset{\sim}{\sigma}_{yx}$.

1.2 Discarding Constraints

Constraints are discarded for being small, linearly dependent (LD), nearly linearly dependent (NLD) or causing outlier weights (those outside the range 1 to 25) during the calculation of the weights. Initially, a check is done for small, LD and NLD constraints at the WA level as follows. The size of a constraint is defined as the number of households in the population to which it applies. Initially, any constraint whose size is SMALL or less (SMALL, a parameter, equaled 20, 30 or 40 in 2006) is discarded because estimates, for the small constraints, tend to be very unstable. Then, since the matrix $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{X}}^{(0)'}$ has to be inverted to calculate $\underset{\sim}{g}$, linearly dependent sets of constraints, which cause this matrix to be singular, are identified and the smallest constraint in each set is discarded. Next, the condition number of $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{X}}^{(0)'}$ (which is generally relatively large in the Census) is lowered by discarding what are called NLD constraints. The condition number is the ratio of the largest eigenvalue to the smallest eigenvalue of $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{X}}^{(0)'}$. High condition numbers indicate near colinearity among the constraints. To lower the condition number, a forward selection approach is used. The matrix $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{X}}^{(0)'}$ is recalculated based only on the two largest constraints. If the condition number exceeds the parameter COND (which, for example, could equal 1,000), the second largest constraint is discarded. Then the next largest constraint is added, the matrix $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{X}}^{(0)'}$ is recalculated and its condition number is determined. If the condition number increases by more than COND, the constraint just added is discarded. This process continues until all constraints have been checked in this fashion. If, after dropping these NLD constraints, the condition number exceeds the parameter MAXC (which, for example, could equal 10,000), additional constraints are dropped. Constraints are dropped in descending order of the amount by which they increased the condition number when they were initially included in the matrix. The condition number of the matrix $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{X}}^{(0)'}$ is recalculated every time a constraint is dropped. When the condition number drops below MAXC, no more constraints are dropped. Any constraints dropped up to this point are not used in the weighting calculations.

Before calculating the first step weighting adjustment factors $g_i^{(A)}$ for the h^th DA (h = 1 to H), the remaining constraints are dropped as necessary because they are small for the h^th DA. The constraints which remain are partitioned into two groups. Then for each group of constraints, linearly dependent constraints are identified and dropped (constraints which are linearly dependent at the DA level may not be linearly dependent at the WA level). Based on the remaining constraints, the first step weighting adjustment factors $g_i^{(A1)}$ and $g_i^{(A2)}$ are calculated. If any of the first step adjusted weights fall outside the range 1 to 25 inclusive, additional constraints are dropped. A method similar to that used to discard NLD constraints is applied here except that a constraint is discarded if it causes outlier weights. In the interests of computational efficiency, however, the bisection method is used to identify which constraints should be dropped.

Next, the second step weighting adjustment factors $\underset{\sim}{g_i}$ are calculated based on those constraints that were not discarded for being small, linearly dependent or nearly linearly dependent based on the initial analysis of the matrix $\underset{\sim}{\hat{X}}^{(0)}\underset{\sim}{\hat{V}}^{-1}\underset{\sim}{\hat{X}}^{(0)'}$. If any of the second step adjusted weights fall outside the range 1 to 25 inclusive, then additional constraints are dropped using the method outlined for the first step adjustment.

1.3 2006 Census Weights Processing

The Census weights were calculated using the SAS interactive matrix language on personal computers. Six PCs processed a set of weights for the entire country in approximately eight hours. Due to these efficient calculations, 20 sets of weights were calculated based on different parameter combinations. These parameters included MAXC, COND and SMALL as described in section 1.2, as well as POST (indicating if the initial weights should be poststratified) and DTYPE (indicating if the two dwelling type constraints should be used) as described in section 1.

The quality of the results from each set of final weights was determined by a summary statistic. The absolute value of the estimate/population differences for the 34 auxiliary variables was summed across each WA. The two dwelling type auxiliary variables were included in this calculation for all WAs, including those where they were not used as constraints. The set of weights that minimized the value of this statistic were used as the final weights for that particular WA. This “cherry-picking” of the production runs allowed smaller estimate/population differences to be achieved at the Canada level than was possible when the same combination of parameters was used for all WAs.

2. Results from 2006

2.1 Initial and Final Weight Discrepancies

Figure 1 displays the discrepancy between the population counts and the sample estimates at the national level for the 34 auxiliary variables for both the initial and the final (second step) weights. This discrepancy is defined as 100*[estimate-count]/count. The discrepancy in the initial weights demonstrates bias in the sample that can originate from a variety of sources. This includes census representative errors (e.g., not selecting the sample according to specifications), non-response bias (e.g., young adult males are less likely to complete a long questionnaire than a short questionnaire), response bias (e.g., respondents answering differently on a long questionnaire than on a short questionnaire), processing errors, and so on.

The initial weights in Figure 1 demonstrate a downward bias for the person level constraints males, males ≥ 15, total population, persons aged 0-9 and 15-44, single, widowed, divorced, separated and common-law equal to yes. There is a downward bias observed for household level constraints 1, 3 and 6+ person households and the dwelling type constraint apartments with less than 5 floors. There is an upward bias observed for person level constraints persons ≥ 15, ages 45+ and married, as well as household level constraints 2, 4 and 5 person households and single detached dwelling type.

The large biases in the 2006 sample for 5 person and 6+ person households were the result of reducing the number of persons on the long questionnaire from six persons to five persons because more space was required to allow the automated data capture of write-in responses. The number of persons on the short questionnaire remained at six. In 2006, sometimes households with more than five persons who received a long questionnaire did not request a second long questionnaire and only listed five persons as living in the household. This caused the large increase in the upward bias in the sample for 5 person households and a corresponding large increase in the downward bias in the sample for 6+ person households. The weighting calibration process was only able to partially correct for these biases and these biases also made it more difficult for the calibration to correct for other biases. The discrepancies for these two constraints are still significantly reduced with the final weights compared to those with the initial weights.

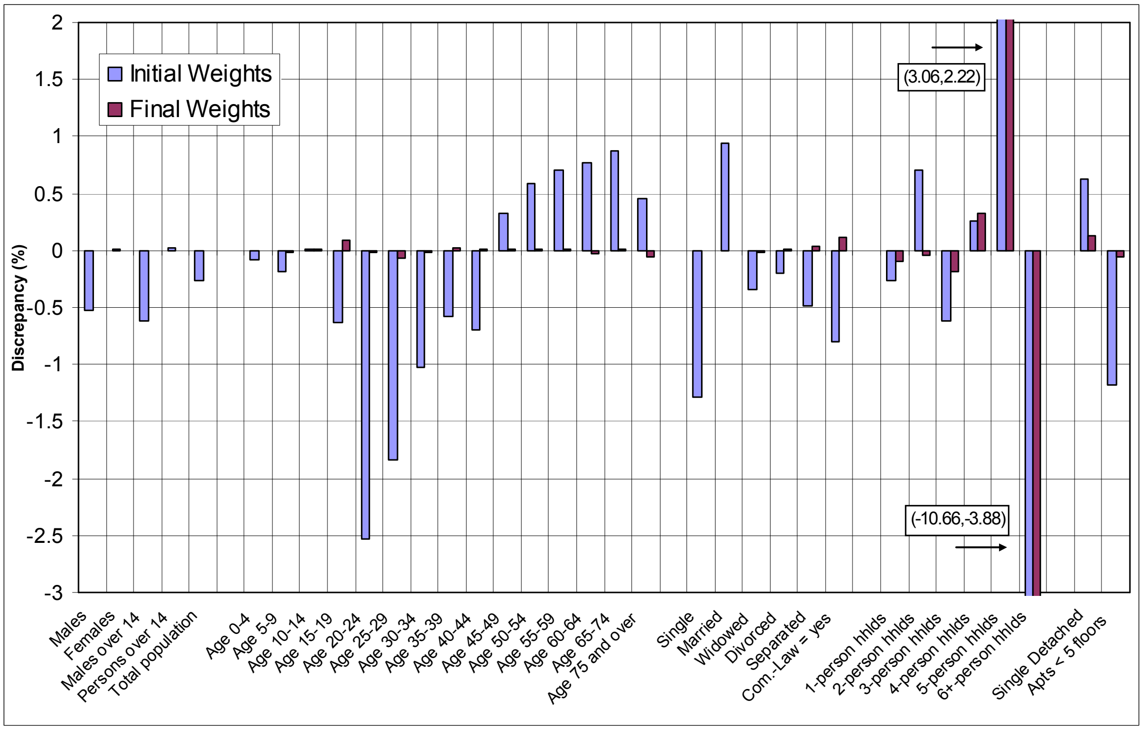

Figure 1: 2006 National Population/Estimate Discrepancies, Initial vs. Final Weights

For the remaining constraints, the biases resulting from the final weights in Figure 1 have been significantly reduced from those present with the initial weights. The majority of constraints had their discrepancies reduced to near zero, although there are still noticeable biases remaining for persons aged 15-19, 25-29, 60-64, 75+, separated, common-law equal yes and all household level constraints. It should be noted that the discrepancies based on the final weights for the two dwelling type constraints (Single Detached and Apartments < 5 floors) have been noticeably reduced from those based on the initial weights despite the fact that these were not controlled on in all WAs. The reduction in the discrepancy for these constraints likely resulted in an increase in the discrepancies for other constraints that were dropped in their stead. The exact impact on the other constraints could not be observed due to the many factors at play.

2.2 Weight Distributions

Figure 2 compares the distributions of the 2006 Census initial weights, poststratified weights, first step weights and second step (final) weights. The initial weights are tightly clustered around 5 as a result of a one-in-five sample of households being selected. The poststratified, first step and final weight distributions become progressively more spread out as the constraints become more restrictive.

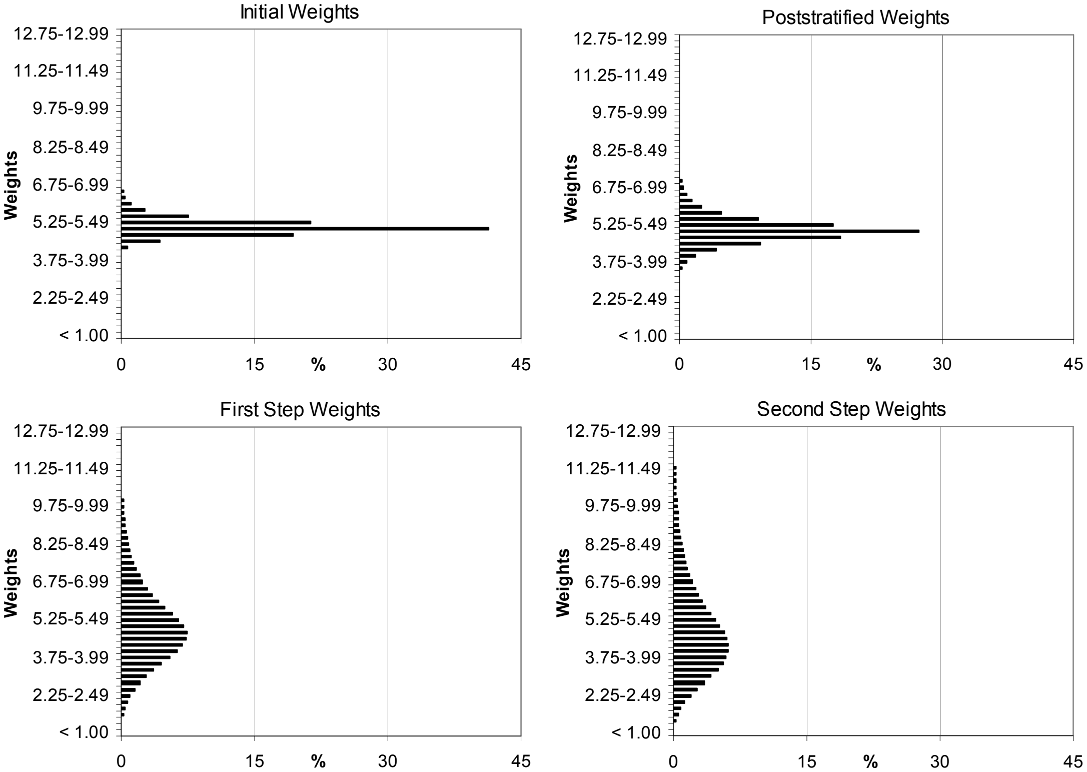

Figure 2: 2006 Weight Distributions: Initial, Poststratified, First Step and Second Step Weights

3. Two-Pass Processing

For the 2006 census, short questionnaire write-in responses to the relationship question were not captured due to budgetary constraints. Instead they were coded under the generic value 'Other'. Long questionnaire write-in responses to the relationship question were still captured and coded in the normal fashion.

During two-pass processing, the long questionnaire data are processed in two stages. In the first stage, called Pass 1, the long and short questionnaires are processed together, representing 100% of the data. The captured long questionnaire write-in responses for relationship are ignored and assigned the generic value 'Other' to coincide with the short questionnaire write-in responses. Editing and imputation is performed the same way for both the long and short questionnaires. In the second stage, called Pass 2, only the long questionnaires are processed; the short questionnaires are not available during imputation. The captured long questionnaire write-in responses for relationship are used rather than the ‘Other’ responses. Because of the availability of the write-in responses, the quality of the results is assumed to be higher in Pass 2 than in Pass 1.

The weighting system uses the Pass 1 results for all households to calculate the household weights. While it might be possible to use the Pass 1 results for the short questionnaires and Pass 2 results for the long questionnaires, this method could bias the census estimates. This is because of differences in the distribution of the responses for the demographic variables between Pass 1 and Pass 2 as a result of the write-in responses for relationship being present in Pass 2. Published census estimates were produced using Pass 1 weights applied to Pass 2 long questionnaire imputed results. The difference between the population counts (based on Pass 2 data for the sampled population and Pass 1 results for the remaining 80% of the population) and Pass 2 estimates are slightly larger than the Pass 1 estimates but remain small for most constraints.

4. Whole household imputation

In the 2006 census, occupied dwellings with total non-response had the number of usual residents (if not known) and all of the responses to the census questions imputed by borrowing the unimputed responses from another household usually within the same DA. This process was called Whole Household Imputation (WHI). In 2001, total non-response long questionnaires were converted to total non-response short questionnaires in a process called Document Conversion, effectively reducing the sample size. The total non-response households were then imputed as part of the main edit and imputation (E&I) process, and the weighting process accounted for the reduction in sample size. In 2006, WHI imputed 96% of the total non-response households while the other 4%, where no donor household was found under the WHI process, were imputed as part of the main E&I process. Utilizing a single donor under WHI was more efficient computationally and was less likely to produce implausible results than using several donors as part of the main E&I process, as was done in 2001.

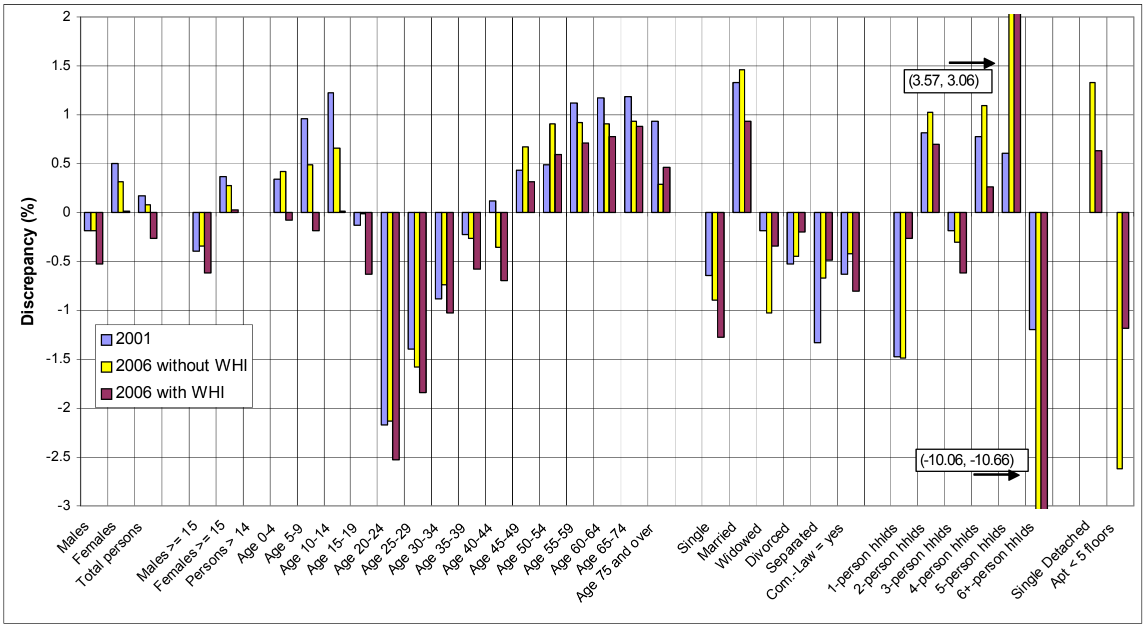

Figure 3: 2006 National Population/Estimate Discrepancies based on Initial Weights, 2001 vs. 2006 without Whole Household Imputation vs. 2006 with Whole Household Imputation

Figure 3 shows what the sampling bias would have been in 2006 if the 2001 approach of applying Document Conversion rather than Whole Household Imputation had been used to deal with total non-response long questionnaires. To determine this, long questionnaires with Whole Household Imputation applied were treated as short questionnaires and the initial weights were recalculated at the DA level to reflect this. The recalculated initial weights were applied to the reduced sample to generate new population/estimate discrepancies that appear in the column '2006 without WHI'. The population/estimate discrepancies under Whole Household Imputation using the original initial weights and the unreduced sample are placed in the column labelled '2006 with WHI'.

In general, it can be seen that the '2006 without WHI' discrepancies in Figure 3 are much more like the 2001 discrepancies than the '2006 with WHI' discrepancies. Also, the population/estimate discrepancies are frequently smaller for '2006 with WHI' than for '2006 without WHI', (e.g. this is the case for Female, persons aged < 15 or age 45+, marital status married, widowed, divorced, separated, households of size 1, 2, 4 and 5, and dwelling types singled detached and apartments less than 5 stories). Thus the introduction of Whole Household Imputation in 2006 to deal with total non-response households was generally beneficial.

4. Research for the 2011 Census

Regression estimation was used in 2006 and previous censuses because its methodology is effective, well known and well accepted. In addition, regression estimation has a non-iterative solution so there are no problems with lack of convergence. Without changing the overall strategy, there are still numerous enhancements that can be made, as are discussed below.

Figure 1 shows that the population/estimate discrepancies for 5 year age ranges were significantly reduced with the final weights. Near consistency for 5 year age ranges is often achieved by underestimates and overestimates of single years within that range cancelling each other out. For 2011, one option that will be investigated is adding another weighting step between the first and second steps where approximate agreement of the single year of age constraints would be attempted, similar to what is done in the first step for the DA level constraints. This approximate agreement would help reduce the underestimates and overestimates that would still exist after the second step weights are calculated for age ranges.

The 2011 questionnaire will be modified such that write-in responses for the relationship question will be captured for the short questionnaire as well as the long questionnaire. Having these responses captured and coded instead of being coded to 'Other' will remove the need for two-pass processing, which will in turn reduce the discrepancies observed with the final weights. This change will also result in family data being available for the entire population. This will require adding additional constraints to the weighting system in order to reduce the population/estimate discrepancies for this new information. These constraints would involve family size (similar to what is done for household size currently) and controlling on the number of children in families. Introducing new constraints to the system will have a negative impact on the 34 constraints already controlled upon. One option to address this issue would be to modify the system such that approximate agreement would be desired for all constraints rather than attempting to get exact agreement for certain constraints while discarding the others. Ridge regression will be explored as a method of obtaining this approximate agreement along with the method currently used in the first step for the DA level constraints.

References

Bankier, M. D. (1986), "Estimators Based on Several Stratified Samples With Applications to Multiple Frame Surveys", *Journal of the American Statistical Association*, **81**, pp. 1074-1079.

Bankier, Michael and Janes, Darryl (2003), "2001 Canadian Census Weighting", *2003 ASA Proceedings*, pp. 442-449.

Cochran, W.G. (1942), "Sampling Theory When the Sampling Units are of Unequal Sizes", *Journal of the American Statistical Association*, **37**, pp. 199-212.

Fuller, Wayne A. (2002), "Regression Estimation for Survey Samples", *Survey Methodology*, **28**, No. 1, pp. 5-23.

Rao, J.N.K. (1994), "Estimating Totals and Distribution Functions Using Auxiliary Information at the Estimation Stage", *Journal of Official Statistics*, **10**, pp. 153-165.