Proceedings of Statistics Canada Symposium 2002
Modelling Survey Data for Social and Economic Research

2001 CANADIAN CENSUS WEIGHTING

Michael Bankier[1]

ABSTRACT

This paper describes two changes made to the sample weighting methodology for the 2001 Canadian Census. First, a decision was made to use a pseudo-optimal regression estimator rather than a projection GREG estimator. Second, the processing was done on PCs rather than on the mainframe. This allowed multiple production runs, using different parameters, to be carried out. The "best" production run (in terms of objective (a) of Section 2.2), for each small area, was then selected and retained. These two changes allowed more auxiliary variables to be retained in the regression estimators while at the same time allowing all Census weights to be at least one.

KEY WORDS: Discarding Auxiliary Variables; Superpopulation Models; Calibration Estimation.

1. OUTLINE OF PAPER

Section 2 describes the Census sample design plus the objectives for Census weighting. Section 3 discusses the Generalized Regression (GREG) Estimator, the optimal regression estimator and the situation where the two are the same estimator. Regression estimators are used in the Census because they allow estimate/population consistency to be achieved for a number of auxiliary variables simultaneously while reducing the variance. In addition, Section 3 shows why it could be advantageous to use an optimal regression estimator rather than the projection GREG. Section 4 provides details on the two step pseudo-optimal regression estimator used in the 2001 Census including the method used to discard auxiliary variables to ensure the weights are at least one. Section 5 examines the 2001 Census weights processing including the analysis done to determine which combinations of parameters should be used for the ten production runs. In addition, the method used, for each small geographical area, to select the best of the ten production runs is described. Finally, in Sections 6 some conclusions are provided.

2. BACKGROUND

2.1 Census Sample Design

In the Canadian Census, basic person and dwelling information is gathered on a 100% basis. This will be called 2A information after the 2A Census short form. For a 1 in 5 sample of private households (2.2 million sampled households in 2001) stratified by 35,885 Enumeration Areas (EAs), additional questions are asked. These will be called 2B information after the 2B Census long form. A uniform 1 in 5 sampling fraction is used for each province (except for 100% sampling in a few special EAs) so that sub-provincial estimates of equal size are of equal reliability in all parts of the country.

2.2 Objectives For 2001 Census Weighting

A weight for each sampled household is calculated. This single weight is used to produce all published household and person characteristic estimates. A single weight is used in the interests of simplicity and consistency. Published 100% counts of 2A information should agree closely with published estimates of 2A information based on the 20% sample since large differences between the 100% counts and the 20% estimates cause concern to users of Census data. The Census estimation methodology, therefore, aims to reduce or eliminate such estimate/population differences for small

[1]Statistics Canada, R.H. Coats Building 15th floor, Ottawa, Ontario, K1A 0T6, Canada, bankier@statcan.ca

geographical areas. At the same time, the standard errors of the Census estimators are also reduced. The estimation methodology was designed to perform well for the thousands of published estimates generated with a minimum of manual intervention during the processing of Census data. Characteristics for which consistency is required between the sample estimate and the population count will be called auxiliary variables or, alternatively, constraints on the weights. The objectives for 2001 Census weighting are outlined in more detail below.

For five year age ranges, marital status, common-law status, sex and household size (32 auxiliary variables), the objectives are:
(a) To have exact estimate/population agreement at the Weighting Area (WA) level for as many of the 32 auxiliary variables as possible. There are 6142 WAs subject to sampling which are frequently small municipalities or Census Tracts. A WA is made up of, on average, 8 whole Dissemination Areas (DAs). Canada is partitioned into 47,933 sampled DAs with, on average, 239 private occupied households in each.
(b) To have approximate estimate/population agreement for the larger DAs for the 32 auxiliary variables.

In addition, it is required that:
(c) There should be exact estimate/population agreement for 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. In 1996, the final census weights were allowed to be in the range 0.01 to 25 inclusive.
(e) The method to generate weights should be highly automated since the 6142 WAs must be processed in a short period of time. This method must also adjust automatically for the different patterns of responses in WAs across the country.
(g) For 2001, it was desired to improve on how well objective (a) was satisfied. The software used in the 1991 and 1996 Censuses, however, had to be used with few changes because of budget and staff shortages.

3. REGRESSION ESTIMATORS UNDER STRATIFIED SAMPLING

An excellent review on the subject of regression estimation for survey samples is given by Fuller (2002). This section compares the projection Generalized Regression (GREG) estimator to the optimal regression estimator.

For simplicity, estimators for a single WA made up of H EAs 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 h^{th} EA, h = 1 to H and that $n = \sum_h n_h$ and $N = \sum_h N_h$. It will also be assumed in this section (though this is not the case in the Census) that the sampling fraction n_h/N_h can vary considerably by EA. The reason for this assumption being made in the Census context will become clear in Section 4.

The simplest estimator possible is the Horvitz-Thompson estimator $\hat{Y}^{(0)} = \sum_i w_i^{(0)} y_i$ where $w_i^{(0)} = N_h/n_h$ if the i^{th} sampled household is in the h^{th} EA. Generally, however, there is no guarantee that objective (a) above will be achieved for any of the 32 auxiliary variables with the Horvitz-Thompson estimator. It is for this reason that various types of regression estimators are considered below.

3.1 GREG

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, $X = [x_{pi}]$ is a P x N matrix, x_{pi} represents the value for the p^{th} auxiliary variable for the i^{th} 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 X which correspond to the sampled households, $\underset{\sim}{W}^{(0)} = [w_i^{(0)}]$ is a n x 1 vector of the initial weights and $\tilde{diag}(\underset{\sim}{W}^{(0)})$ is a n x n matrix with $\underset{\sim}{W}^{(0)}$ running down the diagonal with zeros elsewhere.

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

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

where $\hat{\underset{\sim}{V}}$ 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.

It is possible to write $\hat{Y}$ in the standard form of a regression estimator as

$$\begin{aligned}\hat{Y} &= \hat{Y}^{(0)} + \hat{\underset{\sim}{B}}'(\underset{\sim}{X}\underset{\sim}{1}_N - \hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{1}_n) \\ &= \hat{\underset{\sim}{B}}'\underset{\sim}{X}\underset{\sim}{1}_N + \hat{\underset{\sim}{e}}^{(0)}\underset{\sim}{1}_n \qquad (2)\end{aligned}$$

where $\hat{\underset{\sim}{B}} = [\hat{B}_p] = (\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'})^{-1}\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{Y}}^{(0)'}$ is a P x 1 vector, $\hat{\underset{\sim}{e}}^{(0)} = [w_i^{(0)}e_i] = \hat{\underset{\sim}{Y}}^{(0)} - \hat{\underset{\sim}{B}}'\hat{\underset{\sim}{X}}^{(0)}$ is a 1 x n vector of residuals and $e_i = y_i - \hat{\underset{\sim}{B}}'\underset{\sim}{x}_i$. It can be shown that $\hat{\underset{\sim}{B}}$ minimizes the loss function $L = \hat{\underset{\sim}{e}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{e}}^{(0)'}$.

Särndal , Swensson and Wretman (1992) use models to help justify the choice of $\hat{\underset{\sim}{V}}$. They assume that the population quantities $\underset{\sim}{Y}$ were generated by a model as follows. Assume, under a model ξ, that

$$\underset{\sim}{Y} = \underset{\sim}{\beta}'\underset{\sim}{X} + \underset{\sim}{e} \qquad (3)$$

where $E_\xi(\underset{\sim}{e}) = \underset{\sim}{0}$ and $Cov_\xi(\underset{\sim}{e}, \underset{\sim}{e}') = \underset{\sim}{V}$ while $\underset{\sim}{Y}$ and $\underset{\sim}{X}$ are respectively 1 x N and P x N population level matrices. E_ξ, Cov_ξ and V_ξ denote the expected value, covariance and variance with respect to the model ξ while $\underset{\sim}{\beta}$ (a P x 1 vector) and $\underset{\sim}{V}$ (a N x N symmetric matrix) are model parameters. It is desired to find an estimator $\underset{\sim}{B}$ such that for an arbitrary P x 1 vector $\underset{\sim}{t}$, $\underset{\sim}{t}'\underset{\sim}{B}$ is the best linear unbiased estimator (b.l.u.e) of $\underset{\sim}{t}'\underset{\sim}{\beta}$ where

- by "best" we mean that $V_\xi(\underset{\sim}{t}'\underset{\sim}{B})$ is minimized,
- by linear we mean that $\underset{\sim}{t}'\underset{\sim}{B} = \underset{\sim}{\gamma}'\underset{\sim}{Y}$ for some $\underset{\sim}{\gamma}$ (i.e. it is a linear function in terms of y_i, i = 1 to N) and
- by unbiased we mean that $E_\xi(\underset{\sim}{t}'\underset{\sim}{B}) = \underset{\sim}{t}'\underset{\sim}{\beta}$

These objectives are achieved, if the assumed model ξ is correct, by choosing

$$\underset{\sim}{B} = (\underset{\sim}{X}\,\underset{\sim}{V}^{-1}\underset{\sim}{X}')^{-1}\underset{\sim}{X}\,\underset{\sim}{V}^{-1}\underset{\sim}{Y}' \qquad (4)$$

Särndal et al (1992) indicate that the role of the model ξ is to describe the finite population point scatter. The model serves as a vehicle for finding an appropriate $\hat{\underset{\sim}{B}}$ to put into the regression estimator formula. Given that $\underset{\sim}{t}'\underset{\sim}{B}$ is the b.l.u.e of $\underset{\sim}{t}'\underset{\sim}{\beta}$, it seems reasonable, if the model ξ is correct, to estimate $\underset{\sim}{B}$ and hence $\underset{\sim}{\beta}$ with the approximately unbiased estimator $\hat{\underset{\sim}{B}}$ defined earlier in this section.

Typically, Särndal et al (1992) assume that $\underset{\sim}{V}$ is a diagonal matrix with elements $v_i = \sigma_i^2$ running down the diagonal and zeros elsewhere. They provide some simple examples with a single auxiliary variable where, for example, it is

assumed that $\sigma_i^2 = \sigma^2 x_i$. In practice, however, σ_i^2, i = 1 to N, are not known and are difficult to estimate. Errors in their estimation degrade the quality of the estimator $\underset{\sim}{B}$ of $\underset{\sim}{\beta}$. The σ_i^2 are also usually different for each y characteristic being considered. This causes different calibrated weights to be needed for different y characteristics which is not convenient with a multi-characteristic survey. Finally Särndal et al (1992) generally assume in practice, if there are two or more auxiliary variables, that $\sigma_i^2 = \sigma^2$ so that σ_i^2 disappears from the variance formula for $\hat{\underset{\sim}{B}}$. With the optimal regression estimator described below, no appeal is made to a superpopulation model. Thus, rather than use σ_i^2, the more general notation v_i will be used here when it is assumed that $\hat{\underset{\sim}{V}}$ is a diagonal matrix.

3.2 Optimal Regression Estimator

Cochran (1942) and Rao (1994) recommend the use of the optimal regression estimator. It is called this since the variance of

$$\hat{Y}_{opt} = \hat{Y}^{(0)} + \hat{\underset{\sim}{B}}_{opt}(\underset{\sim}{X}\underset{\sim}{1}_N - \hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{1}_n) \qquad (5)$$

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 $\hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{1}_n$ and the P x 1 vector of covariances $Cov(\hat{Y}^{(0)}, \hat{X}_p^{(0)})$ with $\hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{1}_n = [\hat{X}_p^{(0)}]$. The standard estimator of $\underset{\sim}{B}_{opt}$ (which is approximately unbiased) is $\hat{\underset{\sim}{B}}_{opt} = \hat{\underset{\sim}{\Sigma}}_{xx}^{-1}\hat{\underset{\sim}{\sigma}}_{yx}$ where $\hat{\underset{\sim}{\Sigma}}_{xx}$ and $\hat{\underset{\sim}{\sigma}}_{yx}$ are unbiased estimators of $\underset{\sim}{\Sigma}_{xx}$ and $\underset{\sim}{\sigma}_{yx}$.

3.3 Situation When GREG and Optimal are the Same Estimator

Assume that the set of constraints $\hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{g} = \underset{\sim}{X}\underset{\sim}{1}_N$ includes the constraints $\hat{X}_h = X_h$, h = 1 to H, where $X_h = N_h$ is the number of households in the h[th] EA and $\hat{X}_h$ is the corresponding estimate after calibration. It is then possible to write (subject to the appropriate choice of $\hat{\underset{\sim}{V}}$ for $\hat{Y}$) :

$$\begin{aligned} \hat{Y} &= \hat{Y}^{(0)} + \hat{\underset{\sim}{B}}(\underset{\sim}{X}\underset{\sim}{1}_N - \hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{1}_n) \\ &= \hat{Y}_{opt} = \hat{Y}^{(0)} + \hat{\underset{\sim}{B}}_{\_opt}(\underset{\sim}{X}_{\&}\underset{\sim}{1}_N - \hat{\underset{\sim}{X}}_{\&}^{(0)}\underset{\sim}{1}_n) \end{aligned} \qquad (6)$$

where the first H rows of $\hat{\underset{\sim}{X}}^{(0)}$ and $\underset{\sim}{X}$ relate to the EA level constraints $\hat{X}_h = X_h$ and can be represented by the matrices $\hat{\underset{\sim}{X}}_H^{(0)}$ and $\underset{\sim}{X}_H$ respectively while the last P - H rows of $\hat{\underset{\sim}{X}}^{(0)}$ and $\underset{\sim}{X}$ relate to the other constraints and can be represented by the matrices $\hat{\underset{\sim}{X}}_{\&}^{(0)}$ and $\underset{\sim}{X}_{\&}$ respectively. $\hat{\underset{\sim}{B}}_{\_opt} = \hat{\underset{\sim}{\Sigma}}_{xx\&}^{-1}\hat{\underset{\sim}{\sigma}}_{yx\&}$ is a (P - H) x 1 vector where the covariances used to calculate $\hat{\underset{\sim}{B}}_{\_opt}$ are restricted to the auxiliary variables represented by the rows of $\hat{\underset{\sim}{X}}_{\&}^{(0)}$. $\hat{\underset{\sim}{B}} = [\hat{B}_p] = (\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)})^{-1}\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{Y}}^{(0)}$ is a P x 1 vector where, to achieve $\hat{Y} = \hat{Y}_{opt}$, it is necessary that $\hat{\underset{\sim}{V}}$ be a n x n diagonal matrix with $\hat{V}_{hi} = W_h^{(0)} v_{hi} > 0$ running down the diagonal and zeros elsewhere and with $v_{hi} = (n_h - 1)/(n_h(W_h^{(0)} - 1))$ where $W_h^{(0)} = N_h/n_h$. $\hat{V}_{hi}$ corresponds to the i[th] sampled household from the h[th] EA.

In the 1991 and 1996 Censuses, however, $\hat{\underset{\sim}{V}} = diag(\hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{1}_P)$,where $\underset{\sim}{1}_P$ is a P x 1 vector of 1's. As a result, households with more persons usually had larger values of $\underset{\sim}{v}_{hi}$ which meant that their weighting adjustment factors g_{hi} tended to be smaller, all other things being equal. This choice of $\hat{\underset{\sim}{V}}$ is consistent with the suggestion of Särndal et al (1992) that $\hat{\underset{\sim}{V}} = diag(\hat{\underset{\sim}{V}}\underset{\sim}{\gamma})$ where $\hat{\underset{\sim}{V}}\underset{\sim}{\gamma} = \hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{\gamma} = [W_h^{(0)} v_{hi}]$ is a n x 1 vector and $\underset{\sim}{\gamma}$ is an P x 1 vector which does not result in any of the elements of $\hat{\underset{\sim}{X}}^{(0)}\underset{\sim}{\gamma}$ becoming zero. They suggest $\hat{\underset{\sim}{V}}$ matrices of this form because then $\hat{Y}$ takes the simple projection form $\hat{Y} = \hat{\underset{\sim}{\beta}}\underset{\sim}{X}\underset{\sim}{1}_N$. It should be noted that, except in special cases, the projection GREG is not the optimal regression estimator.

Särndal et al (1992) often assume that $\sigma_i^2 = \sigma^2$ which is equivalent to assuming that $v_{hi} = 1$ and $\hat{V}_{hi} = W_h^{(0)} N_h / n_h$. This value of v_{hi} can be achieved for the projection GREG by assuming that the first H rows of $\hat{X}^{(0)}$ represents the EA level constraints $\hat{N}_h = N_h$ while $\underset{\sim}{Y}$ has $Y_p = 1$ for p = 1 to H and $Y_p = 0$ otherwise.

Särndal (1996) shows that the variance of the Taylor Series approximation of the GREG is minimized under stratified s.r.s.w.o.r. (assuming that $\hat{X}_h = X_h$, h = 1 to H are among the constraints used) if $v_{hi} = n_h(N_h - 1)/(N_h(N_h - n_h))$. This is approximately equal to the $v_{hi} = (n_h - 1)/(n_h(W_h^{(0)} - 1))$ derived above where the GREG was made identical to the optimal regression estimator. Särndal (1996) in Remark 3.2 states that "there may some (although usually modest) advantage in taking" $v_{hi} = n_h(N_h - 1)/(N_h(N_h - n_h))$ rather than $v_{hi} = 1$. In the remainder of this section, it is shown that if the sampling fraction varies significantly by stratum, there may be significant benefits to have v_{hi} vary by stratum.

Assume that the sampling fraction varies considerably by EA within a WA. Table 1 below gives, for the optimal estimator, the values of $\hat{V}_{hi} = (n_h - 1) W_h^{(0)} / (n_h(W_h^{(0)} - 1))$ and for the projection GREG , the values of $\hat{V}_{hi} = W_h^{(0)}$ for $N_h = 400$ private households and for various values of n_h. The loss function being minimized for both the optimal estimator and the projection GREG (but with different values of $\hat{V}_{hi}$) is $L = \sum_h \sum_i \hat{V}_{hi}(g_{hi} - 1)^2$ where g_{hi} is the weighting adjustment factor for the ith sampled unit from the hth EA.

For the optimal estimator, as the sample fraction increases from 5% to 94%, $\hat{V}_{hi}$ rises from 1 to 16 in Table 1. This indicates that EAs with larger sampling fractions will tend to have g_{hi} close to 1, all other things being equal, because of the loss function being minimized. This makes sense because estimates from EAs with larger sampling fractions should be more reliable and hence their estimates should be adjusted less than estimates from EAs with smaller sampling fractions.

With the projection GREG , as the sample fraction increases from 5% to 94%, $\hat{V}_{hi}$ decreases from 20 to 1.1 in Table 1. This indicates that EAs with larger sampling fractions will tend to have g_{hi} far from 1, all other things being equal, because of the loss function being minimized. This is counterintuitive. It is also undesirable because if $W_h^{(0)}$ is close to 1 and if $g_{hi} < 1$, there is a distinct possibility that the adjusted weight $g_{hi} W_h^{(0)}$ will be less than 1 or even negative. Under the current census estimation methodology (to be described in Section 4), constraints are dropped if this occurs. Thus using $v_{hi} = (n_h - 1)/(n_h(W_h^{(0)} - 1))$, as required for the optimal regression estimator, seems preferrable since besides minimizing the variance, it may also allow more constraints to be retained under the requirement that the adjusted weights not be less than 1.

Table 1: $\hat{V}_{hi}$ for Various Sampling Fractions Assuming $N_h = 400$

n_h	$100 n_h / N_h$	$\hat{V}_{hi}$ Optimal	$\hat{V}_{hi}$ GREG
20	5	1.0	20.0
40	10	1.1	10.0
80	20	1.2	5.0
120	30	1.4	3.3
200	50	2.0	2.0
300	75	4.0	1.3
375	94	16.0	1.1

4. REGRESSION ESTIMATOR USED IN 2001 CENSUS

More details on the methodology described below can be found in Bankier, Rathwell and Majkowski (1992).

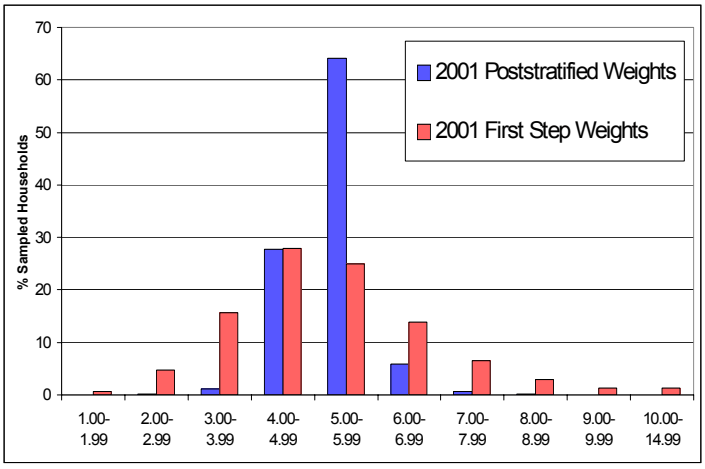
4.1 Two Step Regression Estimator

In this section, for simplicity, we will ignore the fact that constraints are dropped for a variety of reasons (see Section 4.2). Weights are calculated separately in each WA. The 2001 Census initial EA level weights $w_i^{(0)} = N_h/n_h$ have either two or three weighting adjustment factors applied. First households are sometimes poststratified at the WA level on household size because small and very large households tend to be under-represented in the sample. A second adjustment to the weights is then done to try to achieve approximate estimate/population agreement at the DA level as described in objective (b) of Section 2.2. Finally, a third adjustment to the weights is done to achieve exact estimate/population agreement at the WA and DA levels as described in objectives (a) and (c) of Section 2.2. These three adjustments are described in more detail in the following paragraphs.

First, the households are sometimes **poststratified** on household size (1,2,3,4,5,6+ persons) at the WA level and then the poststratified weights $w_i^{(1)} = g_i^{(1)} w_i^{(0)}$ are calculated. Very occasionally, $w_i^{(1)}$ is truncated to ensure that it lies within the range 1 to 20 inclusive. An upper limit of 20 rather than 25 is used to give some "room" for further adjustment.

Next, a **first step** regression weighting adjustment factor is calculated at the DA level. The 32 auxiliary variables (age, sex, marital status, household size) that are to be applied at the WA level in the second step 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 16 constraints go into a second group. The weighting adjustment factors resulting for each group of constraints (labeled $g_i^{(A1)}$ and $g_i^{(A2)}$) are averaged together to create $g_i^{(A)} = (g_i^{(A1)} + g_i^{(A2)})/2$ which then generates the first step weights $w_i^{(A)} = g_i^{(A)} w_i^{(1)}$. Estimate/population differences at the DA level for the 32 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 32 constraints are applied at the WA level along with 2 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 $w_i = g_i w_i^{(A)}$.



In the first step $\hat{V}_i = w_i^{(1)}/(w_i^{(1)}-1)$ 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 (see chart on left for distribution of these weights in the 2001 Census) 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**. 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).

4.2 Discarding Constraints

See Silva and Skinner (1997) and Fuller (2002) for a discussion of the rationale behind discarding constraints. The Silva and Skinner paper was partially motivated by the methodology described in this section. 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, equalled 20, 30 or 40 in 2001) is discarded because estimates, for the small constraints, tend to be very unstable. Then, since the matrix $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'}$ has to be inverted to calculate $\underset{\sim}{g}$ (see Section 3.1), 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 $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'}$ (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 $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'}$. High condition numbers indicate near colinearity among the constraints. To lower the condition number, a forward selection approach is used. The matrix $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'}$ 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 $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'}$ 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 $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'}$ 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 cth DA (c = 1 to C), the remaining constraints are dropped as necessary because they are small for the cth DA. The constraints which remain are partitioned into two groups as described in Section 4.1. 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 $\hat{\underset{\sim}{X}}^{(0)}\hat{\underset{\sim}{V}}^{-1}\hat{\underset{\sim}{X}}^{(0)'}$. 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.

5. 2001 CENSUS WEIGHTS PROCESSING

The Census weights are calculated using the SAS interactive matrix language. For 1996, processing the whole country once took approximately two weeks on the mainframe computer. In 2001, six Pentium IV 1.7 Ghz PCs processed the whole country in under 24 hours.

PCs allowed extensive testing to be done with two samples of WAs (121 and 616 WAs respectively) to determine the ten "best" combinations of weighting system parameters. "Best" was defined as those parameters which minimized -ABSDIFF3 which equalled the sum of the absolute value of the estimate/population differences for the 32 auxiliary variables where the differences were based on the estimate and population counts totaled across the sampled WAs. This served as a large area measure of estimate/population consistency for these variables.

Then the whole country was processed ten times using the best combinations of the weighting parameters. The values of ABSDIFF3 at the Canada level from the initial production runs helped guide the choice of parameters to be used with the samples of WAs and in the later production runs.

After all ten production runs were completed, the "best" production run for each WA was retained. The "best" production run was defined as that which minimized the **sum** of
- ABSDIFF2 which equalled the sum of the absolute value of the estimate/population differences at the WA level for the 32 auxiliary variables which were then summed over all WAs (a small area measure of estimate/population consistency) **and**
- ABSDIFF1 which equalled the sum of the absolute value of the estimate/population differences for the 2 auxiliary variables at the DA level (number of households and persons) which were then summed over all DAs (another small area measure of estimate/population consistency).

This "cherry-picking" of the production runs allowed smaller estimate/population differences to be achieved at the Canada level than was possible in earlier censuses where the same combination of parameters was used for all WAs.

The priorities implicit in this approach were to have the best estimate/population consistency possible for the 32 auxiliary variables at the Canada level as well as very good estimate/population consistency for these variables at the WA level.

The parameters used with the ten production runs and the number of WAs which used these parameters after "cherry-picking" are listed in Table 2. To insure that certain important constraints were always retained, two WAs were run with "customized" parameters. These are listed at the end of Table 2. MAXC and SMALL were defined in Section 4.2. POST = 1 indicates that poststratification by household size (as described in Section 4.1) was done while POST = 0 indicates that it was not. The parameter COND is not listed in Table 2 because COND always equalled MAXC/10. This was the ratio between these two parameters in 1996 and there was insufficient time to experiment with other values.

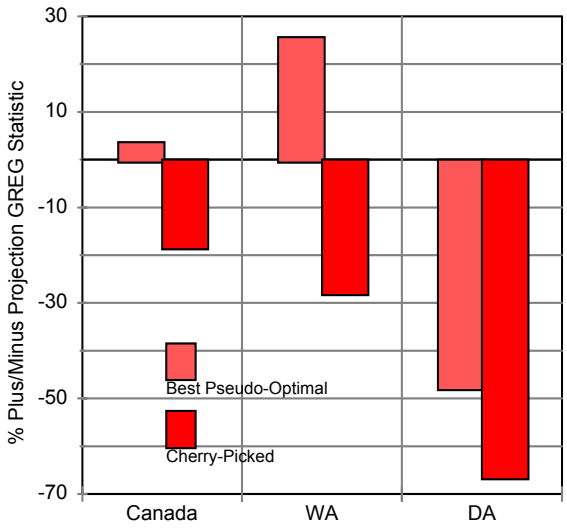
For all ten production runs, the pseudo-optimal estimator was used. This was because, with a sample of 121 WAs and for different combinations of parameters, ABSDIFF3 for the pseudo-optimal estimator was on average 46% smaller than the projection GREG when both had their weights restricted to the range 1 to 25. This can be explained by the fact that, on average, 2.2 constraints were dropped per WA for causing outlier weights with the projection GREG compared with 0.9 constraints with the pseudo-optimal. With the projection GREG, $\hat{\underset{\sim}{V}}\,' \; diag(\hat{\underset{\sim}{X}}^{(0)})\underset{\sim}{1}_P$ at the first step and $\hat{\underset{\sim}{V}}\,' \; diag(\hat{\underset{\sim}{X}}^{(A)})\underset{\sim}{1}_P$ at the second step so as to replicate the loss functions from previous Censuses.

In addition, to serve as a baseline, a production run was done using the projection GREG with weights > 0, MAXC = 10000, SMALL = 20 and POST = 1 to replicate the 1996 Census parameters. In Figure 2, the Cherry-Picked production run plus the best pseudo-optimal production run (with MAXC = 80,000, SMALL = 20 and POST = 1 which resulted in the smallest ABSDIFF3 statistic) are compared to the projection GREG. Figure 2 shows that the best pseudo-optimal production run (with weights of 1 or more) compared to the projection GREG (with weights > 0) does 4% worse at the Canada level (ABSDIFF3), 26% worse at the WA level (ABSDIFF2) but 49% better at the DA level (ABSDIFF1). Figure 2 also shows that the Cherry-Picked production run does 19%, 28% and 66% better than the projection GREG

Table 2: Parameters Used in Census Production Runs

Number of WAs	Percent	MAXC	SMALL	POST
1300	21.2	160,000	20	0
1135	18.5	80,000	20	1
903	14.7	80,000	30	1
725	11.8	80,000	30	0
539	8.8	40,000	40	1
436	7.1	20,000	40	1
363	5.9	40,000	20	1
255	4.2	40,000	30	1
251	4.1	10,000	40	1
233	3.8	20,000	30	1
1	0.0	10,000	95	0
1	0.0	160,000	50	0
6142	100.0			

Figure 2: Comparison of ABSDIFF1,2 and 3



at the Canada, WA and DA levels respectively. This demonstrates the significant benefits of cherry-picking the parameters at the WA level.

To explain the results in Figure 2, it is useful to study the average number of constraints dropped at the WA and DA level. Figure 3 shows that the best pseudo-optimal production run (with weights of one or more) and the GREG (with weights greater than zero) drop the same average number of constraints for being small and LD. This is not surprising given that SMALL = 20 for both these runs. The Cherry-Picked run drops somewhat more constraints for being small (some of which are LD) but this is counterbalanced by it dropping somewhat few constraints for being LD. The larger number of constraints dropped for being small is explained by some of the WAs having SMALL > 20 for the Cherry-Picked run as is shown in Table 2. The Cherry-Picked run and the best pseudo-optimal production run drop a similar number of constraints for being NLD. The best pseudo-optimal production run had MAXC = 80,000 while the WAs in the Cherry-Picked run used a range of MAXC values with the majority being MAXC $ 80,000. The Projection GREG in comparison dropped more constraints for being NLD. This is not surprising given that MAXC = 10,000 for this run. Larger values of MAXC were used in 2001 based on the advice of Press (1992, Section 2.6) that matrices can

Figure 3: Constraints Dropped WA Level

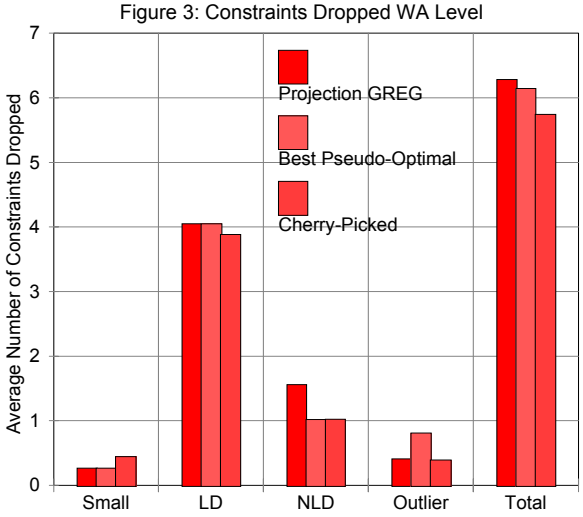
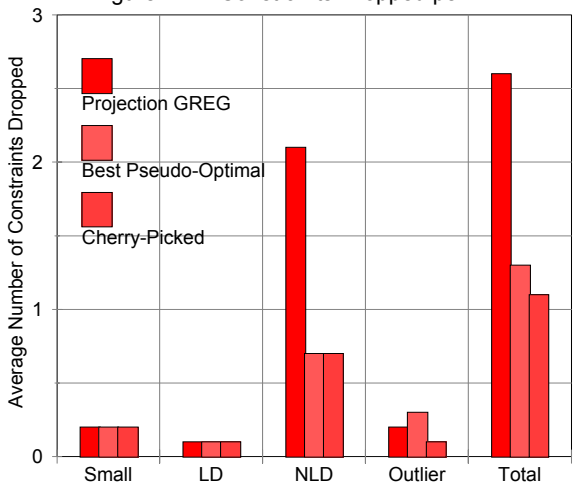


Figure 4:DA Constraints Dropped per WA



be inverted with reasonable precision as long as the inverse of the condition number does not approach the computer's floating point precision. Since the calculations are carried out in double precision, this suggests that matrices whose condition numbers do not approach 10^{12} can be inverted with some confidence. Finally, it can be seen in Figure 3 that the best pseudo-optimal production run drops more constraints for generating outlier weights than the Projection GREG. This is not surprising given that weights less than 1 are not tolerated for the pseudo-optimal estimator while they are with the Projection GREG. It appears, based on the above analysis, that the Cherry-Picked run outperforms the Projection GREG at the Canada and WA levels by discarding fewer constraints for NLD (because of higher MAXC values) and matches the Projection GREG in terms of the number of constraints dropped for generating outlier weights. Overall, the Projection GREG drops the highest number of constraints while the Cherry-Picked run drops the fewest.

Figure 4 shows the average number of DA level constraints (number of persons plus number of households) dropped in a WA. It shows that the superior performance (as seen in Figure 2) of the Cherry-Picked Run compared to the Projection GREG run at the DA level is the result of many fewer constraints being dropped for NLD. This is because of the higher values of MAXC used in general in the Cherry-Picked run. In addition, somewhat fewer constraints are dropped for causing outlier weights in the Cherry-Picked run.

6. CONCLUSION

The analysis performed in Section 3 suggested that switching in the Census from the projection GREG to the pseudo-optimal regression estimator would result in fewer constraints being discarded for causing the adjusted weights to be less than 1. This was confirmed numerically in Section 5 based on a sample of 121 WAs. When the best pseudo-optimal estimator with weights of 1 or more was compared to the projection GREG with weights greater than 0, the estimate/ population differences were slightly worse at the Canada level and were significantly worse at the WA level for the pseudo-optimal estimator. Doing ten production runs of the pseudo-optimal estimator with different parameters and then cherry-picking the best production run for each WA, however, resulted in estimate/population differences being much smaller than the projection GREG differences at the Canada, WA and DA levels. The ability to do ten production runs and then cherry-pick the best run was made possible from a timing and cost viewpoint by the switch from mainframe processing to processing on PCs for the 2001 Census.

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 D., Rathwell, Stephen and Majkowski, Mark (1992), "Two Step Generalized Least Squares Estimation in the 1991 Canadian Census", Methodology Branch Working Paper, August 1992.

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.

Press, William H. (1992), *Numerical Recipes in C: The Art of Scientific Computer Programming*, Cambridge University Press.

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.

Särndal, C.E. (1996), "Efficient Estimators with Simple Variance in Unequal Probability Sampling", *Journal of the American Statistics Association*, **91**, pp. 1289-1300.

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

Silva, P.L.D.N. and Skinner, C.J. (1997), "Variable Selection for Regression Estimation in Finite Populations", *Survey Methodology*, **23**, No. 1, pp. 23-32.