

## BAYES AND CENSUS UNDERCOVERAGE

Peter Dick<sup>1</sup> and Yong You<sup>2</sup>

### ABSTRACT

In 1991, Statistics Canada decided to benchmark the Population Estimates Program to the 1991 Census results adjusted for a survey estimate of missed persons derived from the Census coverage studies. In 1996, Statistics Canada is focusing on evaluating alternative methods that permit a compromise estimator between the extremes of the Census or the Census adjusted estimates that were used in 1991. This paper reports on methods of combining the Census results with the coverage studies to produce an accurate estimate of population shares as measured by weighted squared error loss functions.

A hierarchical Bayes model for estimating Census provincial undercoverage is proposed. Gibbs sampling methodology is used to overcome the computational difficulties with the hierarchical Bayes approach.

KEY WORDS: Hierarchical Bayes, Gibbs sampler

### RÉSUMÉ

En 1991, Statistique Canada a décidé d'ajuster le programme d'estimation de la population aux résultats ajustés du recensement de 1991 par une enquête sur le nombre de personnes manquantes à partir des études de recouvrement du recensement. En 1996, Statistique Canada a mis l'emphase sur l'évaluation de méthodes alternatives qui sont un compromis entre les valeurs du recensement de 1991 et les valeurs ajustées du même recensement. Cet article traite des méthodes utilisées pour combiner les résultats du recensement avec les études de recouvrement afin de produire des estimations de la population en utilisant une fonction de perte quadratique pondérée. Un modèle bayésien hiérarchique est proposé pour estimer le sous-dénombrement au niveau provincial. L'échantillonnage de Gibbs est utilisé pour résoudre les difficultés d'ordre calculatoire associées à l'approche bayésienne hiérarchique.

MOTS CLÉS: Approche bayésienne hiérarchique; échantillonnage de Gibbs.

### 1. INTRODUCTION

In 1991, in a major departure from the established procedure, it was decided to revise the population estimates to agree with the Census counts adjusted to account for the estimated difference between gross undercoverage and gross overcoverage, or net undercoverage, in the Census. The new base population was formed by adding the net provincial undercoverage estimate to the provincial Census count. This created an adjusted base upon which all the other population figures were derived using modelling and demographic methods.

Leading up to this decision, the focus of the research was on whether the provincial Census count adjusted for net undercoverage in the Census was an improvement on the provincial Census count alone. The result of this research was the development of a

procedure similar to a preliminary test (Royce, 1992). From this test - and from other considerations - Statistics Canada decided that the Census adjusted for the net undercoverage estimated from the coverage studies were the best estimates on which to base the population estimates program.

Up to the 1991 Census, Canadian population estimates were re-based solely on the Census results. In Census years, the annual population estimates were essentially the Census counts (with some minor adjustments for the Census reference date). Between Censuses, the provincial population estimates are obtained by taking into account the births, deaths, immigrants, emigrants and returning Canadians during the intercensal period. When the next Census results became available, the population estimates were revised to agree with the new counts.

In 1991, the population estimates were based on the Census counts adjusted for the estimated net

---

<sup>1</sup> Peter Dick, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario

<sup>2</sup> Yong You, Department of Mathematics and Statistics, Carleton University, Ottawa, Ontario

undercoverage in the Census. The technical criteria for adjustment resulted in a procedure known as the preliminary test. This test uses the results of the coverage studies to decide between a full adjustment and no adjustment. The results of this procedure indicated that the Census counts with the net undercoverage added in at the provincial level were an improvement on the Census counts alone with regard to, both, the estimates of the provincial populations and to the estimates of the provincial shares of the national population. The net undercoverage estimates were carried down to the Census Division level by sex and single year of age by combining an Empirical Bayes regression model (Dick, 1995) and demographic (synthetic) methods.

After the release of the 1991 coverage studies' results, a debate was started on examining various estimators of the provincial undercoverage. Rivest (1996) presented a composite estimator that used the national undercoverage rate as a synthetic estimate. The effect of this composite estimator was to shrink all provincial undercoverage rates to the national rate - with each province shrinking by a fixed ratio of the differential provincial undercoverage rate. Thus provinces close to the national rate moved relatively little while provinces far from the national rate moved by a more substantial amount.

The introduction of the composite estimator naturally leads to the question of exchangeability as discussed in Lindley and Smith (1972). If the true provincial undercoverage rate is actually a random draw from another distribution centred on the national undercoverage then the composite estimator could be recast in terms of a hierarchical Bayes model. From past experiences with the measurement of Census undercoverage and the estimates of population from the underlying population estimates programs (using purely demographic methods), this is not an assumption that is easily defensible. However, by creating a more flexible model that permits different levels of underlying provincial undercoverage, the exchangeability of provinces can be examined.

Up to now, however, the difficulty in using a hierarchical Bayes model has not been centred around possible underlying distribution. Instead the difficulty has been the lack of easily accessible method of estimating and assessing the posterior means and variances. Most prior distributions had been selected for their analytical convenience as opposed to their practical utility. With the advent of the Gibbs Sampler and associate software this difficulty has been overcome. As discussed in Gilks et al. (1995, page 16) "with the arrival of ... notably the Gibbs sampler program BUGS, we hope more applied statisticians will become familiar with Bayesian ideas, and apply them".

The objective of this paper is to do just that - to apply a hierarchical Bayesian methodology to the Census undercoverage problem. Section 2 presents a general model and some thoughts on adjusting the "fixed effects" component of the model to the Census situation. The hierarchical Bayes model is also presented in its general form and in the specific form used to address the undercoverage problem. Section 3 presents the Census undercoverage problem, the methods used to measure the undercoverage and the results of using the hierarchical Bayes model with the estimates from the 1991. Section 4 presents some conclusions and future directions the research will take.

## 2. BAYESIAN MODEL

### 2.1 General Model

Suppose there are  $n$  provinces and in the  $i$ -th province the Census has counted  $Y_i$  persons while an unknown number  $U_i$  persons were missed by the Census. The coverage studies provide an estimate,  $\hat{U}_i$ , of the net undercoverage along with an associated (**known**) variance,  $\xi_i^2$ . One objective of the Population Estimates Program is to estimate the true population of the  $i$ -th province on Census Day.

The true population of the  $i$ -th province is written as  $T_i = Y_i + U_i$ . Since the Census count is observed without sampling error most of the work in constructing a model centres around the estimate of missed persons. The coverage studies use a sample survey which through standard estimation procedures produce an estimate of the missed persons. The model generally used to describe this estimation situation is written as

$$\hat{U}_i = U_i + \epsilon_i, \quad i=1, \dots, n \quad (2.1.1)$$

where  $\epsilon_i \sim N(0, \xi_i^2)$ . This model assumes that the estimators  $\hat{U}_i$  are design unbiased, normally distributed with known sampling variances. These assumptions may be restrictive - in particular the estimates of missed persons,  $\hat{U}_i$ , are certainly subject to (unknown) bias. The assumed distribution (Normal) of the sampling errors, at the province level, seems quite reasonable because of the Central Limit Theorem.

Interpreting the differences in missed persons between provinces is difficult because of the large differences in provincial sizes, so a more meaningful measure is the undercoverage rate which is defined as  $r_i = U_i / (U_i + Y_i) = U_i / T_i$ : this is similar to the transformation used by Zaslavsky (1993) in his analysis of the undercount in the 1990 Census of the United States. This transformation implies that survey model can be written as

$$\hat{r}_i = r_i + e_i, \quad i=1, \dots, n \quad (2.1.2a)$$

where we take  $\hat{r}_i = \hat{U}_i / \hat{T}_i$  and we assume the sampling errors are  $e_i \sim N(0, \psi_i^2)$ . The sampling variance are related to the original variance  $\zeta_i^2$  by  $\psi_i^2 = \zeta_i^2 (1 - r_i)^2 / \hat{U}_i^2$  and are assumed known.

Suppose the true undercoverage rates are related to a number, say  $p$ , variables  $x_i = (x_{i1}, \dots, x_{ip})'$  specific to the  $i$ -th province (Fay and Herriot, 1979). In particular assume that the linear relationship

$$r_i = x_i' \beta + v_i, \quad i=1, \dots, n \quad (2.1.2b)$$

is true. Here  $\beta$  is the vector of regression coefficients and  $v_i$  are independent and identically distributed random variables with  $E(v_i) = 0$  and  $V(v_i) = \sigma_v^2$ .

Combining these two equations (2.1.2), we obtain the general mixed linear model

$$\hat{r}_i = x_i' \beta + v_i + e_i, \quad i=1, \dots, n \quad (2.1.3)$$

Note both design induced random variables,  $e_i$ , and model based random variables,  $v_i$ , are included in the model. An extensive discussion on this model can be found in Ghosh and Rao (1994). As in Ghosh and Rao, we assume that the sampling variance,  $\psi_i^2$ , is known: allowing for unknown sampling variance is briefly discussed in the conclusions to this paper.

## 2.2 Specifying the Model - Fixed Effects

The fixed effect part of the general model relates the true undercoverage rate to underlying set of auxiliary variables. The simplest model sets the regression component equal to some fixed value. A logical choice would be the national undercoverage rate,  $\bar{r} = \sum U_i / \sum T_i$ , hence we would equate the regression component to the national coverage rate,  $x_i' \beta = \bar{r}$ . This model can be made more flexible by the following approach.

For any province, the undercoverage rate can be written as  $r_i = r^t + \alpha_i$ , where  $r^t$  is an arbitrary fixed value and  $\alpha_i$  is a pre-specified value discussed below. However the national undercoverage rate can be written in terms of the fixed value,  $r^t$ , as

$$\bar{r} = \frac{\sum (r^t + \alpha_k) T_k}{\sum T_k} \quad (2.2.1)$$

which can be shown to be equivalent to

$$\bar{r} = r^t + \sum \alpha_k p_k \quad (2.2.2)$$

where  $p_i = T_i / \sum T_k$  is the share of the total population of the  $i$ -th province and is assumed known. Hence the undercoverage rate for any province can be written as

$$r_i = (\bar{r} - \sum \alpha_k p_k) + \alpha_i \quad (2.2.3)$$

This has some important properties. First, note that if  $\alpha_i = 0$  for all  $i=1, 2, \dots, n$  then  $r_i = \bar{r}$ . Secondly for any non-zero set of  $\alpha_i$  the national undercoverage rate is preserved. This can be seen by multiplying (2.2.3) by the true provincial population,  $T_i$ , which gives

$$U_i = p_i \sum U_k - (\sum \alpha_k p_k) T_i + \alpha_i T_i \quad (2.2.4)$$

Summing this quantity up for the  $n$  provinces will show that both sides total  $\sum U_k$ . This can be seen since  $p_i = T_i / \sum T_k$  and, consequently,  $\sum p_k = 1$ , then if we write  $T_i = p_i \sum T_k$ , this implies  $(\sum \alpha_k p_k) \sum p_k \sum T_k = T \sum \alpha_k p_k$  and hence both sides of (2.2.4) sum to the number of missed persons,  $\sum U_k$ .

The basic form (2.2.3) allows the pre-specified undercoverage rates,  $\alpha_i$ , to be written in terms of relative differences in undercoverage rates between provinces. For instance, for any two provinces, the undercoverage rates will differ by

$$r_i - r_j = \alpha_i - \alpha_j \quad (2.2.5)$$

This implies that the vector  $\alpha$  is just the difference in undercoverage between the  $i$ -th province and a fixed value. Thus by pre-specifying how the Census undercoverage is expected to differ between provinces we can specify the fixed component without regard for national undercoverage rate,  $\bar{r}$ .

This means that the general model (2.1.3) can be combined with (2.2.3) to write the model for undercoverage rates as

$$\sum \alpha_k p_k + \alpha_i + v_i + e_i \quad i=1, 2 \quad (2.2.6)$$

To fully specify the model only requires that prior distributions be applied to  $v_i$  and  $e_i$ .

## 2.3 A Hierarchical Bayes Model

Having determined the form of the fixed effects part of the model, the random effects will now be introduced. The basic framework of the Hierarchical Bayes model follows from Lindley and Smith discussion

of exchangeability. The first stage, assumes that the undercoverage rates for each province are unbiasedly estimated with known sampling variance. The second stage assumes that the true undercoverage rates for each province are unbiasedly estimated by the national undercoverage rate adjusted for differences between the provinces with an unknown variance. The errors associated with this stage are assumed to be exchangeable: that is the prior opinion on the unexplained portion of the undercoverage rate for Ontario would be the same as for Prince Edward Island. The final stage assumes that the national undercoverage rate has a known distribution.

More formally, this model can be written as:

(i) **The Sampling Model:**

$[\hat{r}_i | r_i] \sim N(r_i, \psi_i^2)$  for  $i = 1, \dots, n$   
and  $\psi_i^2$  are the known sampling variances;

(ii) **The Population Model:**

$[r_i | \bar{r}, c_i, \sigma_v^2] \sim t(\bar{r} + c_i, \sigma_v^2, \eta_i)$  where  $c_i = \alpha_i - \sum \alpha_k p_k$  is the fixed effect part discussed in Section 2.2,  $\sigma_v^2$  is the population variance and  $\eta_i$  are the (known) degrees of freedom associated with the t-distribution;

(iii) **Prior distributions:**

$\bar{r} \sim N(r_o, \Psi_o)$  where  $r_o$  and  $\Psi_o$  are known and  $\sigma_v^2 \sim \Gamma(\frac{1}{2}a, \frac{1}{2}b)$  where a and b are known.

The following comments can be made about the assumed Bayesian model. First, the assumed sampling model is typical of the standard models used in sampling surveys: the assumption of normality, given the large sample sizes used in coverage studies allows the Central Limit Theorem to be invoked. Secondly, the prior distribution are selected for convenience with known parameters. The vagueness that is assumed with these priors ensures they will have little impact on the final estimates.

Secondly, the usual assumption concerning the true undercoverage rates as being normally distributed is considered quite restrictive. In particular, the assumption that the variability is the same for all provinces is unlikely to be true. Datta and Lahiri (1994) suggested the t-distribution with different degrees of freedom for each area - essentially attempting to add an amount of "robustification" to the usual normality assumption. In addition, they recommend that the degrees of freedom,  $\eta_i$ , be set to one for the largest and smallest observations. This corresponds to a Cauchy distribution which falls between a uniform and a normal prior. They also suggest a moderate value for

the degrees of freedom for the other observations.

Finally, the prior distribution for the population variance,  $\sigma_v^2$ , is taken to be a Gamma distribution instead of the more standard Inverse Gamma as in Ghosh and Rao (1992). The Datta and Lahiri formulation assumes a Gamma prior distribution and is required to create proper conditional distributions. However, in the example, we essentially use parameter values that create a distribution equivalent to the Gamma.

The population model can be represented in a form more useful for implementing the Gibbs sampler. It can be shown that if  $r_i | \bar{r}, \sigma_v^2 \sim t(\bar{r} - \sum \alpha_k p_k + \alpha_i, \sigma_v^2, \eta_i)$  then this can be written in two stages as, first, as normal distribution:

$$r_i | \bar{r}, \sigma_v^2, \zeta_i \sim N(\bar{r} - \sum \alpha_k p_k + \alpha_i, 1/\zeta_i)$$

then the new variable  $\zeta_i$  is assumed to be a Gamma distribution:

$$\zeta_i \sim \Gamma\left(\frac{1}{2}\eta_i, \frac{1}{2}\sigma_v^2\eta_i\right)$$

Note, the degrees of freedom,  $\eta_i$ , are still assumed known.

The complete model can now be written as

$$\hat{r}_i | r_i \sim N(r_i, \psi_i^2) \tag{2.3.1.a}$$

$$r_i | \bar{r}, \sigma_v^2, \zeta_i \sim N(\bar{r} - \sum \alpha_k p_k + \alpha_i, 1/\zeta_i) \tag{2.3.1.b}$$

$$\zeta_i | \sigma_v^2 \sim \Gamma\left(\frac{1}{2}\eta_i, \frac{1}{2}\sigma_v^2\eta_i\right) \tag{2.3.1.c}$$

$$\bar{r} \sim N(r_o, \Psi_o) \tag{2.3.1.d}$$

$$\sigma_v^2 \sim \Gamma\left(\frac{1}{2}a, \frac{1}{2}b\right) \tag{2.3.1.e}$$

where the relative undercoverage rates  $\alpha_i$ , the provincial population shares  $p_i$  and the degrees of freedom  $\eta_i$  are pre-specified. The problem is now to calculate the posterior expectation and posterior variance for the true undercoverage rates.

## 2.4 Inference using the Gibbs Sampler

The Gibbs sampler is a technique for extracting the marginal distributions from the full conditional distributions instead of the full distribution. Instead of calculating the marginal distribution directly, the Gibbs sampler simulates the results of drawing from the appropriate target distribution. Hence, the posterior expectation and posterior variance can be calculated by, first, simulating a large number of draws from the distribution, and, secondly, calculating the expected value and variance of these draws from the target distribution. A straight forward explanation of the Gibbs Sampler can be found in Casella and George (1992).

To implement the Gibbs sampler we need the full conditionals. First set  $\alpha_i - \sum \alpha_k p_k = c_i$  and recall that differential undercoverage rates  $\alpha_i$  and the degrees of freedom  $\eta_i$  are both assumed known, then the full conditionals can be shown to be

$$r_i | \hat{r}_i, \bar{r}, \zeta_i, \sigma_v^2 \sim N \left( \frac{\psi_i^{-2} \hat{r}_i + (\bar{r} + c_i) \zeta_i}{\psi_i^{-2} + \zeta_i}, (\psi_i^{-2} + \zeta_i)^{-1} \right) \quad (2.4.1.a)$$

$$\bar{r} | r_i, \zeta_i, \psi_i^2 \sim N((\psi_i^{-2} + \sum \zeta_k)^{-1} (\psi_i^{-2} r_i + \sum \zeta_k (\bar{r} - c_i)), (\psi_i^{-2} + \sum \zeta_k)^{-1}) \quad (2.4.1.b)$$

$$\zeta_i | r_i, \bar{r} \sim \Gamma \left( \frac{1}{2}(\eta_i + 1), \frac{1}{2}(\sigma_v^2 \eta_i + (r_i - \bar{r} - c_i)^2) \right) \quad (2.4.1.c)$$

and

$$\sigma_v^2 | \zeta_i \sim \Gamma \left( \frac{1}{2}(a + \sum \eta_k), \frac{1}{2}(b + \sum \eta_k \zeta_k) \right) \quad (2.4.1.d)$$

Posterior inference about  $r_i$  is based on the above conditional distributions.

The Gibbs sampler used in the analysis of the model described in (2.3.1) was the Bayesian inference Under Bayes Sampling (Spiegelhalter et. al, 1996). The algorithm is relatively simple:

- (1) Draw the true undercoverage rate,  $r_i^{(1)}$ , using starting values  $\zeta_i^{(0)}$  and  $\bar{r}^{(0)}$  from the full conditional distribution in (2.4.1.a);

- (2) Draw the national undercoverage rate,  $\bar{r}^{(1)}$ , using starting values  $\zeta_i^{(0)}$  and  $r_i^{(1)}$  from the full conditional distribution in (2.4.1.b);
- (3) Draw the variable  $\zeta_i^{(1)}$  using starting values  $\sigma_v^{2(0)}$  and  $\bar{r}^{(1)}$  from the full conditional distribution in (2.4.1.c);
- (4) Finally, draw the variable  $\sigma_v^{2(1)}$  using starting values  $\zeta_i^{(1)}$  from the full conditional distribution in (2.4.1.d).

Running all four parts completes one cycle of the algorithm, the posterior expectation and posterior variance shown in the next section are the results of completing 12,000 cycles. In order to ensure that the distribution in which the inferences are made is the correct one, a "burn-in" of the first 2,000 cycles was discarded: in effect only the last 10,000 simulations are kept for the analysis. Convergence was checked by checking the autocorrelations within the chain of parameters. The convergence diagnostics, Geweke's Z-score methods was used. Details on this procedures can be found in Best, Cowles and Vines (1996). Finally 5 separate simulations were run using a different random seed: the results of these 5 simulations are used for calculating posterior expectation  $E(r_i | \hat{r}_i, \psi_i^2)$  and the posterior variance  $V(r_i | \hat{r}_i, \psi_i^2)$ .

## 3. CENSUS UNDERCOVERAGE ESTIMATION

### 3.1 Description of the Survey

Since 1966, the Reverse Record Check has been the survey vehicle used by Statistics Canada to measure gross number of persons missed by the Census. Starting in 1991, an Overcoverage Study, was conducted to measure the gross number of persons erroneously included in the Census. Together these coverage surveys provided an estimate of the net number of persons missed by the Census. Through the analysis of the results of these surveys, the collection methodology is adjusted in order to improve coverage in the succeeding Census.

The basic approach used by the coverage studies is to create an independent list of persons who should have been included in the Census. These lists are compiled from persons included in the previous Census, persons missed by that Census and persons new to Canada since the last Census. These new persons include the newborn, landed immigrants and non-permanent residents. Their are some minor exclusions to these lists, for instance Canadian living outside the country during the previous Census are not included.

permanent residents. There are some minor exclusions to these lists, for instance Canadian living outside the country during the previous Census are not included. More details on the coverage studies can be found in Germain and Julien (1993).

The survey had a sample of 56,000 in 1991 and is designed to estimate the number of missed persons in the Census. The sample was allocated to each province to ensure that the maximum standard error on the undercoverage rate would be less than 0.35%; the rest of the sample was allocated proportionally to population. The design was a stratified random sample with a disproportionate sample amongst young adults (20 to 29) - a group more prone to be missed. The sample allocation should be sufficient to give reliable estimates for the provinces and for national age and sex totals.

### 3.2 Estimation and Results

The model used is described in (2.3.1); only the values for the fixed effects part of the model  $\alpha_i$  and the degrees of freedom  $\eta_i$  in the population model need to be specified. Two different approaches were taken when pre-specifying the fixed effects. The first approach was to assume that all provinces have exchangeable errors when sampled from the national undercoverage rate. This model (denoted Model 1) implies that all  $\alpha_i=0$ .

The second approach was to assume that Ontario has an undercoverage rate that is 1% higher and Prince Edward Island has an undercoverage rate that is 1.5% lower than the other 8 provinces. This approach is accounting for the fact that it is becoming relatively difficult to conduct a Census in Ontario, and in particular in Toronto because of its size, diversity and complexity. On the other hand, this approach allows for the apparent ease in which a Census can be conducted in Prince Edward Island due to small size and a stable and homogeneous population. Hence the vector is set to

$$\alpha^t = (0, -0.015_{pei}, 0, \dots, 0.01_{ont}, 0, \dots, 0)$$

The population model also needs to have the degrees of freedom ( $\eta_i$ ) specified. Three separate assumptions were made concerning the degrees of freedom. First, in order to provide a benchmark series, we have assumed a "very large -  $\eta_i=200$ " number for the degrees of freedom for each province; this in effect assumes the population model has a normal distribution. Then, following the suggestion of Datta and Lahiri, we have assigned one degree of freedom (a Cauchy distribution) to the minimum and maximum observations - corresponding to PEI and Ontario; for the other provinces we have assumed 5 and 15 degrees of freedom.

For the prior distributions we follow closely the suggestion by Hobert and Casella (1996). They recommend avoiding improper posteriors by using

proper priors. They state that "ignorance can be modelled by using a normal prior with large variance and inverted gamma priors with small parameter values for the variance components". Hence, for the national undercoverage rate, we assumed that  $\bar{r} \sim N(0.02865, 1)$  which is the observed rate in 1991 with a large variance and for the population variance, we assumed  $\sigma_v^2 \sim \Gamma(0.0001, 0.0001)$  which under our model assumes that this quantity is distributed as an Inverse Gamma with small parameter values.

The results of the Gibbs sampler for the various models are displayed in Table 1. The direct estimates of undercoverage from the coverage studies are also displayed. The two models correspond to the different components to the  $\alpha$  vector. It should be noted that when the degrees of freedom ( $\eta_i$ ) are set at 5 and 15 this is only valid for the provinces other than PEI and Ontario; they have their degrees of freedom fixed at 1 (Cauchy distribution) for these simulations.

Model 1 shows that all provinces have moved toward the national undercoverage rate of 2.86%, the provinces close to this national rate, such as BC and Quebec move very little while the extreme provinces of PEI, New Brunswick and Ontario move the most. Examining the results of Model 1, the Normal ( $\eta_i=200$ ) moves the most toward the national rate; with more uncertainty in the population model - reflected in the t-distribution - the estimates change very little. It should be noted that when the population model permits 1 degree of freedom for PEI and Ontario and a normal distribution for the rest of the provinces the results are close to  $\eta_i=15$  assumption. In table 1, Model 2 shows little variability amongst the three possible population distributions. The estimates for PEI and Ontario are also much closer to the observed estimates. This seems to be due to the fixed effects component of the population model. Finally, looking at the estimates together, all the provinces have a consistent movement toward the national undercoverage rate for all models except for British Columbia and Quebec. Note that for Model 1, the estimates are toward the national rate while for Model 2 they are away from the national rate: a different model implies a different direction in the change of the estimates.

Table 2 shows the estimated efficiency gains; defined as the ratio of the observed variance to the estimated posterior variance. The most notable observation on the efficiencies from Model 1 is the loss experienced in PEI - the observed variance is smaller than the posterior variance except when the population model is assumed to be normal. This reflects the large bias introduced by Model 1. For Model 2 all provinces show a gain in efficiency; in particular, PEI and Ontario

now exhibit gains similar to the other provinces. This is a reflection of the small bias introduced by the population model. For Model 2, the exchangeable assumption seems reasonable, while for Model 1 this assumption is less plausible.

Table 3 shows the estimated coefficient of variation for the various estimates. Except for PEI, the CVs are remarkably constant. This would be a reflection of the large sample size and consequent reliability of the original observed estimates.

#### 4. CONCLUSIONS AND FUTURE RESEARCH

The Hierarchical Bayes model has shown that some improvement over the direct estimators is possible. The population model is flexible and permits an easy interpretation for the expected differences in provinces. Model 2 provides better estimates than Model 1 so we can conclude that the structure of the population model matters. The improvement in efficiency over the direct estimates points to the use of the Hierarchical Bayes model but the small change in coefficient of variation also shows the inherent reliability of the direct provincial estimates of undercoverage.

The suggestion by Datta and Lahiri to assume a  $t$ -distribution with 1 degree of freedom for the extreme observations is useful. In particular, the change in the efficiency from a gain when assuming a normal distribution to a loss with the Cauchy distribution points to the sensitivity of the population model to the underlying assumptions. The Datta and Lahiri approach allows for a more realistic set of underlying assumptions.

Future research will look at the assumption of known variance for the sampling model. Bell (1995) has suggested assuming a Wishart distribution for the sampling variance. Since the coverage studies use random groups for estimating the sampling variance Bell's suggestion fits in nicely with underlying estimation of the sampling variance. The adjustments to the assumed Hierarchical Model should be relatively straightforward. Methods of comparing the results from the various models will have to be developed. The population model should be placed on firmer ground by employing estimates from the Census on the expected differential undercoverage rates between the provinces. For instance, the observed non-response in the Census - a fairly good proxy to the ultimate undercoverage rates - could be examined for its potential in describing undercoverage in a Bayesian model. Another possible approach would be to use the population estimates series for the population model. Finally, the Hierarchical Bayes approach should be compared to other related approaches, such as Empirical Bayes and composite

estimators.

#### REFERENCES

- Bell, W. R. (1995) Bayesian sampling error modelling with application. *Proceedings of the Survey Methods Section, Statistical Society of Canada Annual Meeting: 19 - 31.*
- Best, N., Cowles, M.K. and Vines, K. (1996) *CODA, Convergence diagnosis and output analysis software for Gibbs sampling output, Version 0.30* MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, England CB2 2SR.
- Casella, G. and George, E.I. (1992) Explaining the Gibbs Sampler. *American Statistician* 46: 167 - 174.
- Datta, G.S. and Lahiri, P. (1994) Robust hierarchical Bayes estimation of small characteristics in presence of covariates. *Private communication to JNK Rao.*
- Fay, R. E. and Herriot, R. A. (1979) Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association* 74: 269 - 277.
- Germain, M.-F. and Julien, C. (1993) Results of the 1991 Census coverage error measurement program. *Proceedings of Seventh Annual Research Conference.* United States Bureau of the Census. 55 - 70.
- Gilks, W.R., Richardson, S. and Spiegelhalter, D.J. (1995) Introducing Markov chain Monte Carlo. In *Markov Chain Monte Carlo in Practice* (editors W.R. Gilks, S. Richardson and D.J. Spiegelhalter, pages 1 - 19. London: Chapman & Hall
- Ghosh, M. and Rao, J.N.K. (1994) Small area estimation: an appraisal. *Statistical Science* 9: 55 - 93.
- Hobert, J.P. and Casella, G. (1996) The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* 91: 1461 - 1473.
- Lindley, D.V. and Smith, A.F.M. (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society Series B: 1 - 41.*
- Rivest, L.-P. (1996) Some shrinkage estimators for Census undercoverage. *Technical Report 96-01, Université Laval.*
- Royce, D. (1992) A comparison of some estimators of a set of population totals. *Survey Methodology* 18: 109 -

125.

Spiegelhalter, D., Thomas, A., Best, N. and Gilks, W. (1996) *BUGS 0.5, Bayesian inference using Gibbs sampling manual*. MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge, England CB2 2SR.

Zaslavsky, A. M. (1993) Combining Census, dual-system and evaluation study data to estimate population shares. *Journal of the American Statistical Association* 423: 1092 - 1105.

**Table 1: Observed and Estimated Undercoverage Rate**

Province	Observed	Model 1			Model 2		
		$\eta_i=200$ (Normal)	$\eta_i=5$	$\eta_i=15$	$\eta_i=200$ (Normal)	$\eta_i=5$	$\eta_i=15$
Nfld	1.99	2.06	2.08	2.08	2.08	2.08	2.08
PEI	0.93	1.07	1.00	1.00	0.93	0.94	0.94
NS	1.89	1.99	2.02	2.02	2.03	2.02	2.02
NB	3.25	3.17	3.15	3.15	3.01	3.02	3.02
Quebec	2.60	2.61	2.61	2.61	2.59	2.59	2.59
Ontario	3.64	3.57	3.55	3.55	3.60	3.60	3.60
Manitoba	1.86	1.98	2.00	2.01	2.01	2.00	2.01
Sask	1.80	1.90	1.92	1.93	1.93	1.93	1.93
Alberta	2.00	2.05	2.07	2.07	2.07	2.07	2.07
B.C.	2.73	2.73	2.74	2.74	2.69	2.69	2.69

Note: Note that PEI and Ontario are assumed to be drawn from a t-distribution with 1 degree of freedom (Cauchy distribution) **only** when the other provinces are assumed to be drawn from a t-distribution with  $\eta_i=5$  and  $\eta_i=15$ .

**Table 2: Estimated Efficiency Gains**

Province	Model 1			Model 2		
	$\eta_i=200$ (Normal)	$\eta_i=5$	$\eta_i=15$	$\eta_i=200$ (Normal)	$\eta_i=5$	$\eta_i=15$
Nfld	1.08	1.08	1.10	1.19	1.18	1.19
PEI	1.04	0.98	0.96	1.21	1.24	1.24
NS	1.10	1.10	1.14	1.26	1.25	1.25
NB	1.18	1.20	1.20	1.34	1.31	1.30
Quebec	1.06	1.06	1.06	1.11	1.11	1.11
Ontario	1.07	1.03	1.03	1.20	1.25	1.25
Manitoba	1.13	1.11	1.14	1.29	1.25	1.26
Sask	1.12	1.07	1.10	1.25	1.20	1.21
Alberta	1.08	1.07	1.09	1.17	1.16	1.16
B.C.	1.07	1.09	1.11	1.15	1.14	1.15

Note: Note that PEI and Ontario are assumed to be drawn from a t-distribution with 1 degree of freedom (Cauchy distribution) **only** when the other provinces are assumed to be drawn from a t-distribution with  $\eta_i=5$  and  $\eta_i=15$ .

**Table 3: Estimated Coefficient of Variation**

Province	Observed	Model 1			Model 2		
		$\eta_i=200$ (Normal)	$\eta_i=5$	$\eta_i=15$	$\eta_i=200$ (Normal)	$\eta_i=5$	$\eta_i=15$
Nfld	15.6	14.5	14.4	14.2	13.7	13.8	13.7
PEI	29.7	25.2	27.9	28.1	26.7	26.5	26.5
NS	19.6	17.7	17.5	17.2	16.4	16.4	16.4
NB	13.3	12.5	12.5	12.5	12.4	12.5	12.5
Quebec	8.1	7.9	7.8	7.8	7.7	7.7	7.7
Ontario	8.2	8.0	8.2	8.2	7.5	7.4	7.4
Manitoba	20.4	18.1	18.1	17.8	16.7	17.0	16.9
Sask	18.5	16.5	16.8	16.5	15.4	15.8	15.7
Alberta	14.2	13.3	13.3	13.1	12.7	12.8	12.8
B.C.	9.6	9.3	9.8	9.1	9.1	9.1	9.1

Note: Note that PEI and Ontario are assumed to be drawn from a t-distribution with 1 degree of freedom (Cauchy distribution) **only** when the other provinces are assumed to be drawn from a t-distribution with  $\eta_i=5$  and  $\eta_i=15$ .