Hierarchical Bayes Small Area Inference to the 2001 Census Undercoverage Estimation

Yong You and Peter Dick, Statistics Canada
Yong You, HSMD, RHC-16, Statistics Canada, Ottawa, ON, Canada, K1A 0T6

Key Words: Benckmarking, Census, CV reduction, Gibbs sampling, PMSE, Sampling variance.

1. Introduction

The Census of Canada is conducted every 5 year. The last census was conducted on May 15, 2001. One objective is to provide the Canadian Population Estimates Program with accurate baseline counts of the number of persons by age and sex for specified geographic areas. The count of persons includes usual residents, immigrants and non-permanent residents; excluded are all foreign visitors and non Canadian residents without a permit. Unfortunately, not all persons are correctly enumerated by the Census. Two errors that occur are undercoverage - exclusion of eligible persons - and overcoverage - erroneous inclusion of persons.

The main coverage vehicle used by Statistics Canada is the Reverse Record Check (RRC). This sample survey, with a sample size of 60,000 persons, estimates the net number of persons missed by the Census. This estimate is the combined total of the two types of coverage errors, the gross number of persons missed by the Census and the gross number of persons erroneously included in the final Census count. Once these estimates are adjusted for the coverage errors of persons living in collective dwellings, the final net number of people missed by the Census can be produced. The RRC sample size produces reliable direct estimates for large areas, such as provinces, and for large domains, such as broad age - sex combinations at the national level. However, the Population Estimates Program requires estimates of missed persons for single year of age for both sexes for each province and territory - over 2,000 estimates. Clearly the direct survey estimate would result in estimates having either unacceptably high standard errors due to insufficient sample in the small domain or having no estimate at all due to no sample in the domain. In addition, estimates have to be produced for the 288 Census Divisions and 4 different types of marital status. Altogether over 2.5 million estimates have to be created.

The current methodology used to generate these estimates has essentially been in place since 1991. One component of the procedure is to use the basic small area estimation model, as in Fay and Herriot (1979). However some modifications have been made to this basic model that needs to be evaluated. Specifically, the usual basic area model assumes that the sampling variances are known. The Census undercoverage model has to smooth the observed sampling variances before they can be used in the model. The final model-based standard error, the squared root of mean square error (MSE), does not take into account this estimation so clearly this approach has underestimated the uncertainty. Another drawback to the current methodology concerns the constraints that are imposed on the final estimates. Again the impact of this approach is to underestimate the MSE. The proper evaluation and impact of these two approaches is addressed in this paper. The chosen method is to adjust the model fit it into a hierarchical Bayes (HB) framework. With this approach we can use the machinery developed over the last 10 years for evaluating this HB model and observe if the measures of uncertainty are comparable.

An advantage of the HB approach is that it is relatively straightforward and the inferences about the level parameters are "exact" unlike the empirical best linear unbiased prediction (EBLUP) approach. The HB approach will automatically take into account the uncertainties associated with unknown parameters. However, it does require the specification of prior distributions. Fortunately the Census provides a case in which specifying the model is, again, relatively straightforward. The main purpose of this paper is to introduce a HB model for the census undercoverage estimation and to provide a HB benchmarking method for the marginal constrains.

The paper is organised as follows. Section 2 presents various small area models considered for the study. Section 3 presents the hierarchical Bayes estimation and the benchmarked HB estimators for the census undercoverage. In section 4, we present some results based on the year 2001 census undercoverage data. And finally in section 5, we offer some concluding remarks.

2. Model Specification

2.1 General Area Level Models

Let y_i denote the direct survey estimator of the i-th small area parameter of interest θ_i. Following You and Rao (2002), we may consider a sampling model for y_i: $y_i = \theta_i + \varepsilon_i$, $i = 1,...,m,$ with $E(\varepsilon_i \mid \theta_i = 0)$, that is, the

direct survey estimator y_i is design-unbiased for the small area parameter θ_i. The sampling variance of y_i is $V(\varepsilon_i \mid \theta_i) = \sigma_i^2$. The sampling variance is usually assumed to be known in the model, but it may be unknown. The unknown parameter of interest θ_i is assumed to be related to area level auxiliary variable x_i through a linking function g with random area effects v_i as $g(\theta_i) = x_i'\beta + v_i$, $i = 1,...,m$, where β is a vector of unknown regression parameters, and the v_i's are uncorrelated with $E(v_i) = 0$ and $V(v_i) = \sigma_v^2$, where σ_v^2 is unknown. Normality of v_i is also assumed. If the linking function g is a non-linear function, then the sampling model and the linking model are unmatched in the sense that they cannot be combined directly to produce a linear mixed effects model for small area estimation (You and Rao, 2002).

2.2 Fay-Herrit model under HB framework

The Fay-Herriot model is a special case of the general model given in the Section 2.1. In the Fay-Herriot model, the linking function $g(\theta_i) = \theta_i$ and the sampling variance σ_i^2 is replaced by a smoothed estimator $\tilde{\sigma}_i^2$ and then treated as known in the model. Under the HB framework, the Fay-Herriot model is given as (1) $y_i|\theta_i \sim ind\ N\left(\theta_i, \tilde{\sigma}_i^2\right)$, $i = 1,\ldots,m$; (2) $\theta_i|\beta,\sigma_v^2 \sim ind\ N\left(x_i'\beta, \sigma_v^2\right)$, $i = 1,\ldots,m$; (3) Priors: $\pi(\beta) \propto 1$, $\pi(\sigma_v^2) \sim IG(a_0, b_0)$. Inference based on the Gibbs sampling approach can be found in You and Rao (2002).

2.3 Unknown sampling variances

The Fay-Herriot model assumes that the sampling variances σ_i^2 are known in the model. This is a very strong assumption. Usually a smoothed estimator of σ_i^2 is used in the model and then treated as known. In practice, the sampling variances σ_i^2 are usually unknown and are estimated directly by unbiased estimators s_i^2. The estimators s_i^2 are independent of the direct survey estimators y_i. Following Wang (2000), Rivest and Vandal (2002) and Wang and Fuller (2003), we also assume that $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$ and n_i is the sample size for the *i*-th area. For example, suppose we have n_i observations from small area *i* and these observations are iid $N(\mu_i, \sigma^2)$. Let y_i be the sample mean of the n_i observations. Then $y_i \sim N(\mu_i, \sigma_i^2)$ and $\sigma_i^2 = \sigma^2 / n_i$. Then we can obtain an estimator of σ_i^2 as $s_i^2 = s^2 / n_i$, where s^2 is the sample variance of the n_i observations. Also y_i and s_i^2 are independent and $(n_i - 1)s_i^2 \sim \sigma_i^2 \chi_{n_i-1}^2$. We now present the Fay-Herriot model with the estimated sampling variances s_i^2 in a HB framework as follows: (1) $y_i|\theta_i,\sigma_i^2 \sim ind\ N\left(\theta_i, \sigma_i^2\right)$, $i = 1,\ldots,m$; (2) $d_i s_i^2 \left|\sigma_i^2 \right. \sim ind\ \sigma_i^2 \chi_{d_i}^2$, $d_i = n_i - 1$, $i = 1,\ldots,m$; (3) $\theta_i|\beta,\sigma_v^2 \sim ind\ N\left(x_i'\beta, \sigma_v^2\right)$, $i = 1,\ldots,m$; (4) Priors for the parameters: $\pi(\beta) \propto 1$, $\pi(\sigma_i^2) \sim IG(a_i, b_i)$, $i = 1,\ldots,m$, $\pi(\sigma_v^2) \sim IG(a_0, b_0)$, where a_i, b_i $(0 \le i \le m)$ are chosen to be very small known constants to reflect vague knowledge on σ_i^2 and σ_v^2. *IG* denotes the inverse gamma distribution.

3. Hierarchical Bayes Estimation

Let c_i denote the census count for the i-th area (domain), m_i denote the missed persons by the census. We define the census undercoverage ratio as $\theta_i = m_i / c_i$. Let $\hat{m}_i$ be the direct estimator of m_i. The direct estimator of θ_i is given as $\hat{\theta}_i = \hat{m}_i / c_i$. We then apply the Fay-Herriot model with unknown sampling variances given in section 2.3 to the census undercoverage ratio estimation by letting $y_i = \hat{\theta}_i$ and $\theta_i = m_i / c_i$.

3.1. HB estimators

For a complete HB inference about θ_i, the Gibbs sampling method will be used. The full conditional distributions for the Gibbs sampler are given as follows:

- $\left[\theta_i \left| \hat{\boldsymbol{\theta}}, \boldsymbol{\beta}, \sigma_i^2, \sigma_v^2 \right.\right] \sim N(\gamma_i \hat{\theta}_i + (1-\gamma_i)\mathbf{x}_i'\boldsymbol{\beta}, \gamma_i \sigma_i^2)$
- $\left[\boldsymbol{\beta} \left| \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}, \sigma_i^2, \sigma_v^2 \right.\right] \sim N_p((\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i')^{-1}(\sum_{i=1}^{m} \mathbf{x}_i \theta_i), \sigma_v^2 (\sum_{i=1}^{m} \mathbf{x}_i \mathbf{x}_i')^{-1})$
- $\left[\sigma_i^2 \left| \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2 \right.\right] \sim IG(a_i + \frac{d_i + 1}{2}, b_i + \frac{(\hat{\theta}_i - \theta_i)^2 + d_i s_i^2}{2})$
- $\left[\sigma_v^2 \left| \hat{\boldsymbol{\theta}}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_i^2 \right.\right] \sim IG(a_0 + \frac{m}{2}, b_o + \frac{1}{2}\sum_{i=1}^{m} (\theta_i - \mathbf{x}_i'\boldsymbol{\beta})^2)$

We are interested in estimating the undercoverage ratio θ_i. The HB estimator of θ_i, based on the Gibbs sampler, is given as

$$\hat{\theta}_i^{\mathrm{HB}} = G^{-1}\sum_{k=1}^{G}\left(\gamma_i^{(k)}\hat{\theta}_i + (1-\gamma_i^{(k)})\mathbf{x}_i'\boldsymbol{\beta}^{(k)}\right), \qquad (1)$$

where $\gamma_i^{(k)} = \sigma_v^{2(k)}/(\sigma_v^{2(k)} + \sigma_i^{2(k)})$. The posterior variance is used as the measure of uncertainty, and is estimated by

$$\begin{aligned}\hat{\mathrm{V}}(\theta_i \mid \hat{\boldsymbol{\theta}}) &= G^{-1}\sum_{k=1}^{G}(\gamma_i^{(k)}\sigma_i^{2(k)}) \\ &+ G^{-1}\sum_{k=1}^{G}(\gamma_i^{(k)}\hat{\theta}_i + (1-\gamma_i^{(k)})\mathbf{x}_i'\boldsymbol{\beta}^{(k)})^2 \\ &-[(G^{-1}\sum_{k=1}^{G}(\gamma_i^{(k)}\hat{\theta}_i + (1-\gamma_i^{(k)})\mathbf{x}_i'\boldsymbol{\beta}^{(k)}).\end{aligned}$$

Then we can obtain the HB estimator of uncercoverage count as $\hat{m}_i^{\mathrm{HB}} = c_i \times \hat{\theta}_i^{\mathrm{HB}}$. The measure of uncertainty is estimated by $\hat{V}(m_i \mid \hat{\boldsymbol{\theta}}) = c_i^2 \times \hat{V}(\theta_i \mid \hat{\boldsymbol{\theta}})$.

3.2. Benchmarked HB estimators

The HB estimators of undercoverage counts $\hat{m}_i^{\mathrm{HB}}$ are no longer consistent with the total of the direct survey estimates. However, the national total of direct estimate is respected. Also, in order to protect possible model mis-specification and possible over shrinkage, we need to benchmark the HB estimators so that the benchmarked HB estimators add up to the direct total estimate. You, Rao and Dick (2002, 2004) constructed benchmarked HB estimators for small areas. Let $\hat{m}_i^{\mathrm{BH}B}$ denote the benchmarked HB (BHB) estimator of m_i such that $\hat{m}_i^{\mathrm{BH}B}$ is a function of the HB estimators $\hat{m}_i^{\mathrm{HB}}$, $i = 1,...,m$, i.e., $\hat{m}_i^{\mathrm{BH}B} = f(\hat{m}_1^{\mathrm{HB}},...,\hat{m}_m^{\mathrm{HB}})$ for some known function $f(\cdot)$, and satisfies the benchmark property: $\sum_{i=1}^{m}\hat{m}_i^{\mathrm{BHB}} = \sum_{i=1}^{m}\hat{m}_i$, where $\sum_{i=1}^{m}\hat{m}_i$ is the total of the direct estimates. For example, a ratio BHB (RBHB) estimator can be obtained as

$$\hat{m}_i^{\mathrm{RBHB}} = \hat{m}_i^{\mathrm{HB}}\frac{\sum_{j=1}^{t}\hat{m}_j}{\sum_{j=1}^{t}\hat{m}_j^{\mathrm{HB}}}. \qquad (2)$$

To obtain a measure of variability associated with the BHB estimator $\hat{m}_i^{\mathrm{BH}B}$, we used the posterior mean squared error (PMSE), given as

$$\mathrm{PMSE}(\hat{m}_i^{\mathrm{BHB}}) = \mathrm{E}\left[(\hat{m}_i^{\mathrm{BHB}} - m_i)^2 \mid \hat{\boldsymbol{\theta}}\right],$$

which is similar to the posterior variance associated with the HB estimator $\hat{m}_i^{\mathrm{HB}}$. It can be shown (You, Rao and Dick, 2002, 2004) that the PMSE of $\hat{m}_i^{\mathrm{BH}B}$ is given by

$$\mathrm{PMSE}(\hat{m}_i^{\mathrm{BHB}}) = (\hat{m}_i^{\mathrm{BHB}} - \hat{m}_i^{\mathrm{HB}})^2 + V(m_i \mid \hat{\boldsymbol{\theta}}).$$

Thus the PMSE of $\hat{m}_i^{\mathrm{BH}B}$ is simply the sum of the posterior variance $V(m_i \mid \hat{\boldsymbol{\theta}})$ and a bias correction term $(\hat{m}_i^{\mathrm{BHB}} - \hat{m}_i^{\mathrm{HB}})^2$. The PMSE is readily obtained from the posterior variance and the estimators $\hat{m}_i^{\mathrm{HB}}$ and $\hat{m}_i^{\mathrm{BH}B}$. The advantage of the BHB estimator and the PMSE is well-defined and easy to compute, unlike the benchmarked EB or EBLUP approaches.

4. Application to the 2001 census data

We applied the Fay-Herriot model with unknown sampling variances and the BHB approach to the 2001 census undercoverage data. The HB model requires auxiliary variables for the small area estimation. Previous studies have shown that the undercoverage varies by age, sex, tenure, marital status and immigration status. Initially 48 variables were selected. After variable selection and model analysis, finally the auxiliary variables in the linking model for the undercoverage ratio θ_i reduce to eight variables and an intercept term. The eight variables are Yukon, Nunavet, Male 20 to 29, Male 30 to 44, Female 20 to 29, BC renters, ON renters and NWT renters. To implement the Gibbs sampling, we considered L=10 parallel chains, each of length $2d$=2000. For each chain, the first d=1000 "burn-in" iterations were deleted. Table 1 gives HB estimates of the model components. The "t-value" is simply the ratio of estimate over standard error (SD). The SD is the squared root of the posterior variance.

Table 1: Estimation of fixed effects.

Variable	Estimate	SD	"t-value"
Mean	0.0084	0.0019	4.42
Yukon	0.0291	0.0111	2.62
Nunavut	0.0251	0.0112	2.24
Male 20 to 29	0.0856	0.0047	18.21
Male 30 to 44	0.0416	0.0041	10.15
Female 20 to 29	0.0425	0.0047	9.04
BC Renters	0.0946	0.0151	6.26
ON Renters	0.0752	0.0141	5.37
NWT Renters	0.1733	0.0233	7.43

To estimate the undercoverage counts, we first obtained the HB estimators of undercoverage ratios $\hat{\theta}_i^{\mathrm{HB}}$ using equation (1). Then we obtained $\hat{m}_i^{\mathrm{HB}}$ using

$\hat{m}_i^{HB} = c_i \times \hat{\theta}_i^{HB}$ and the BHB estimator $\hat{m}_i^{RBHB}$ using equation (2). Transforming back to undercoverage ratio, we can obtain the benchmarked undercoverage ratio estimator as $\hat{\theta}_i^{RBHB} = \hat{m}_i^{RBHB} / c_i$. Figure 1 displays the direct and HB estimates of undercoverage ratios by the domain sample sizes. Figure 2 displays the corresponding coefficients of variation (CV) of the direct and HB estimates.

Figure 1. Comparison of Direct and HB Estimates

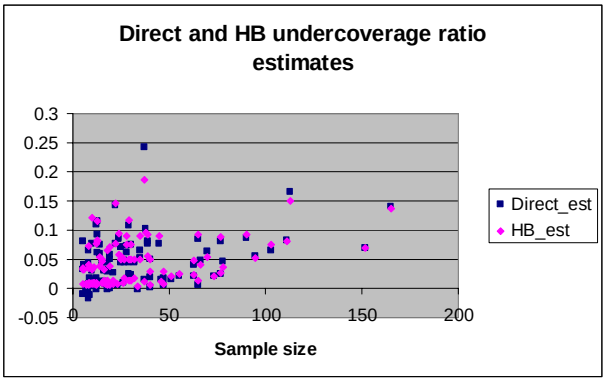
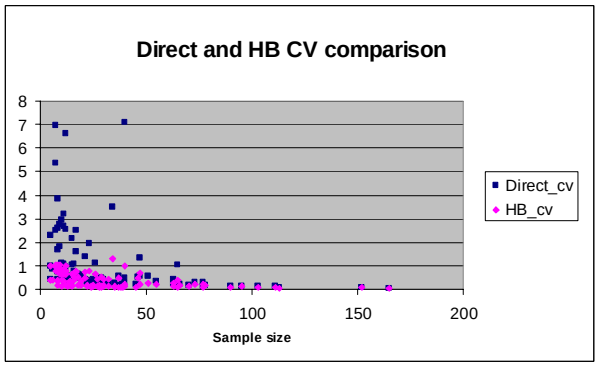


Figure 2. Comparison of Direct and HB CVs



From figure 1, the HB approach leads to smoothed estimates, particularly for the domains with relatively small sample sizes. When sample size is small, some direct net undercoverage estimates are negative due to the fact that the overcoverage estimates are larger then the undercoverage estimates. The HB method "corrected" the negative values. All the HB net undercoverage estimates are positive. For the CV comparison given in figure 2, when the sample sizes are small, the HB approach has achieved a large CV reduction as expected. As sample size increases, the CV reduction becomes smaller. When sample size is large, both the direct and HB estimates and CVs are about the same.

Figure 3. Comparison of HB and BHB Estimates

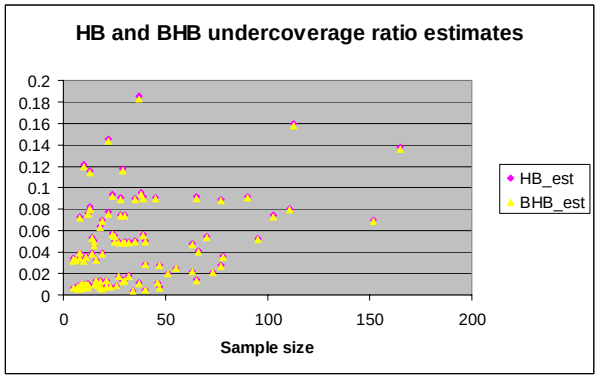


Figure 4. Comparison of HB and BHB CVs

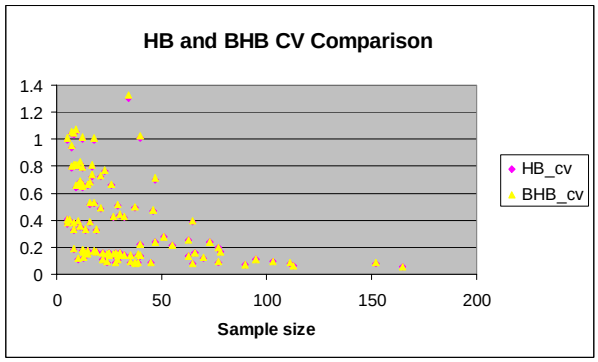


Figure 3 and figure 4 compare the HB and the BHB estimates and the corresponding CVs. It is clear that the benchmarking only makes slight change to the HB estimates in this case. The BHB CV is slightly larger than the HB CV as expected.

5. Concluding Remarks

In this paper we have presented a hierarchical Bayes cross-sectional model for the census undercoverage small domain estimation. The proposed model is an extension of the Fay-Herriot model with unknown sampling variances. The unknown sampling variances are estimated directly by direct survey estimators. We have modelled both the small area parameters (undercoverage) and the sampling variances. We also benchmarked the HB estimators using a simple method proposed by You, Rao and Dick (2004). This paper is an extension of You, Rao and Dick (2004) to within province small domains with a model for the sampling variances.

REFERENCES

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.

Gelfand, A.E. and Smith, A.F.M. (1990) Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 398-409.

Rivest, L. P. and Vandal, N. (2002), Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, July 10-13, 2002, Ottawa, Canada.

Wang, J. (2000), Topics in Small Area Estimation with Applications to the National Resources Inventory, Ph.D. dissertation, Iowa State University.

Wang, J. and Fuller, W. A. (2003), The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y., Rao, J.N.K. and Gambino, J. (2000) Hierarchical Bayes estimation of unemployment rates for sub-provincial regions using cross-sectional and time series data. *American Statistical Association, Proceedings of the Section on Government Statistics and Section on Social Statistics*,160-165.

You, Y. and Rao, J.N.K. (2002) Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.

You, Y., Rao, J.N.K. and Dick, P. (2002) Benchmarking hierarchical Bayes small area estimators with application in census undercoverage estimation. *Statistical Society of Canada 2002 Proceedings of the Survey Methods Section [CD-ROM]*, 81-86.

You, Y., Rao, J.N.K. and Dick, P. (2004) Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.