

BENCHMARKING HIERARCHICAL BAYES SMALL AREA ESTIMATORS WITH APPLICATION IN CENSUS UNDERCOVERAGE ESTIMATION

Yong You, J.N.K. Rao and Peter Dick¹

ABSTRACT

Linear mixed effects models such as the Fay-Herriot model (1979) and non-linear mixed effects models such as the unmatched area level models proposed by You and Rao (2002) have been used in small area estimation to obtain efficient model-based small area estimators. It is often desirable to benchmark the model-based estimates so that they add up to the direct survey estimates for large areas to protect against possible model mis-specification and possible overshrinkage. In this paper, hierarchical Bayes (HB) unmatched area level models are considered. Posterior means and posterior variances of parameters of interest are first obtained using the Gibbs sampling method. Then we benchmark the HB estimators (posterior means) to obtain the benchmarked HB (BHB) estimators. Posterior mean squared error (PMSE) is then used as a measure of uncertainty for the BHB estimators. The PMSE can be represented as the sum of the usual posterior variance and a bias correction term. We evaluate the HB and the BHB estimators in the application of Canadian census undercoverage estimation. The sum of the provincial BHB census undercount estimates is equal to the direct survey estimate of the census undercount for the whole nation.

KEY WORDS: Benchmarking, Census undercoverage, Hierarchical Bayes, Posterior mean squared error, Unmatched model, Small area.

RÉSUMÉ

Les modèles linéaires à effets mixtes tels le modèle de Fay-Herriot (1979) et les modèles non-linéaires à effets mixtes comme les modèles non appariés proposés par You et Rao (2002), ont été utilisés afin d'obtenir des estimations efficaces. Il est souvent désirable de faire de l'étalonnage sur les estimations obtenues de manière à ce que leur somme coïncide avec les estimations directes pour les grandes régions, ce qui assurera une certaine protection si le modèle est mal spécifié ou surrétrécissement. Dans cet article, des modèles bayésiens hiérarchiques non-appariés sont considérés. Les moyennes et les variances a posteriori des paramètres d'intérêt sont d'abord obtenus à l'aide de l'échantillonneur de Gibbs. Ensuite, nous faisons de l'étalonnage sur les estimateurs HB (moyennes a posteriori) de manière à obtenir les estimateurs BHB. L'erreur quadratique moyenne a posteriori (PMSE) est ensuite utilisée comme mesure d'incertitude pour les estimateurs BHB. Le PMSE peut être exprimé comme la somme de la variance a posteriori habituelle et d'une correction pour le biais. Nous évaluons les estimateurs HB et BHB dans le cadre d'une application à l'estimation de la sous-couverture dans le recensement Canadien. La somme des estimateurs BHB du sous-dénombrement au niveau des provinces est égal à l'estimation directe du sous-dénombrement au niveau du pays.

MOTS CLÉS : Bayésienne hiérarchique, Erreur quadratique moyenne a posteriori, Étalonnage, Modèle non apparié, Petit domaine, Sous-converture au recensement.

1. INTRODUCTION

Sample surveys are used to provide estimates not only for the total population but also for a variety of sub-populations (domains). Direct survey estimators, based only on the domain-specific sample data, are typically used to estimate parameters for large domains. But sample sizes in small domains, particularly small geographical areas, are rarely large enough to provide reliable direct estimates for specific small domains. In making estimates

for small areas, it is necessary to “borrow strength” from related areas to form indirect estimators that increase the effective sample size and thus increase the precision. Such indirect estimators are based on either implicit or explicit models that provide a link to related small areas through supplementary data such as recent census counts and current administrative records. It is now generally accepted that when indirect estimates are to be used they should be based on explicit models that relate the small areas of interest through supplementary data. Small area models may be broadly classified into two types: area level and

¹ Yong You, Household Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6, J.N.K. Rao, School of Mathematics and Statistics, Carleton University, Ottawa, Canada, K1S 5B6, Peter Dick, Social Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6, yongyou@statcan.ca, jrao@math.carleton.ca, dickpet@statcan.ca.

unit level models. Ghosh and Rao (1994) and Rao (1999) presented a comprehensive overview and appraisal of models and methods for small area estimation. In this paper, we focus on area level models for small area estimation. In particular, we apply hierarchical Bayes (HB) approach to general area level models or unmatched sampling and linking models defined in You and Rao (2002) to obtain model-based HB estimators for parameters of interest in small areas.

A basic area level model is the well-known Fay-Herriot model (1979) that includes a linear sampling model for direct survey estimates and a linear linking model for the parameters of interest. However, nonlinear linking models are often needed in practice to provide better model fit to the data. For example, if the parameter of interest is a probability or a rate within the range of 0 and 1, a linear linking model with normal random effects may not be appropriate. A customary linking model could be a log-linear model or a logistic regression model. In Section 2, we will consider this kind of general area level models for small area estimation.

Another important problem is that the model-based estimators do not benchmark to the direct survey estimate for large areas. In order to protect against possible model mis-specification as well as possible overshrinkage, we benchmark the model-based HB estimates so that the benchmarked HB (BHB) estimates add up to the direct large area estimate. In particular, we benchmark the small area HB estimates such that the sum of the small area BHB estimates is equal to the direct estimate of the whole. To measure the variability of the BHB estimators, we use the posterior mean squared error (PMSE), similar to the posterior variance associated with the HB estimators. It can be shown that the PMSE is simply equal to the sum of the posterior variance and a bias correction term, provided that the BHB estimator is a known function of the HB estimators.

In Section 2, we present the unmatched sampling and linking model as well as the BHB estimators for small area estimation. In Section 3, we apply the proposed method to the census undercoverage estimation and obtain the BHB estimates for the provincial level census undercoverage estimates. And in Section 4 we give a summary and direction for further research.

2. INFERENCE BASED ON AREA LEVEL MODELS

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 consider the following sampling model for y_i :

$$y_i = \theta_i + \varepsilon_i, \quad i = 1, \dots, m, \quad (1)$$

with $E(\varepsilon_i | \theta_i) = 0$, 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 $V(\varepsilon_i | \theta_i) = \sigma_i^2$. The sampling variance is usually assumed to be known in the model, but it may depend on the unknown parameter θ_i (You and Rao, 2002).

The unknown parameter 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

$$g(\theta_i) = x_i' \beta + v_i, \quad i = 1, \dots, m, \quad (2)$$

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.

The sampling model (1) and the linking model (2) are unmatched in the sense that they cannot be combined directly to produce a linear mixed effects model for small area estimation if the linking function g is a non-linear function.

2.2 Log-linear Unmatched Models

A very useful and important linking model is the log-linear model, i.e.,

$$\log(\theta_i) = x_i' \beta + v_i, \quad i = 1, \dots, m. \quad (3)$$

The sampling model (1) and the linking model (3) can be presented in a hierarchical Bayes framework as follows:

$$y_i | \theta_i \sim N(\theta_i, \sigma_i^2), \quad i = 1, \dots, m; \quad (4)$$

and

$$\log(\theta_i) | \beta, \sigma_v^2 \sim N(x_i' \beta, \sigma_v^2), \quad i = 1, \dots, m. \quad (5)$$

The linking model (5) implies that the small area mean θ_i conditionally has a log-normal distribution with density function given by

$$f(\theta_i | \beta, \sigma_v^2) = \frac{1}{\sqrt{2\pi\sigma_v\theta_i}} \exp\left\{-\frac{1}{2\sigma_v^2}(\log\theta_i - x_i^T\beta)^2\right\}.$$

We are interested in making inference on the small area means θ_i given the direct survey estimates y_i . By using a complete HB approach, we can obtain the posterior mean as the HB estimator and the posterior variance as the measure of uncertainty for the estimator. Gibbs sampling method (Gelfand and Smith, 1990) with Metropolis-Hastings algorithm (Chip and Greenberg, 1995) can be used to find the posterior means and posterior variances. Details can be found in You and Rao (2002).

2.3 Benchmarked HB Estimators

Let $\hat{\theta}_i^{HB}$ denote the HB estimator of θ_i and $\hat{V}(\theta_i)$ the posterior variance of θ_i . Let $\hat{\theta}_i^{BHB}$ denote the benchmarked HB (BHB) estimator of θ_i such that $\hat{\theta}_i^{BHB}$ is a function of the HB estimators $\hat{\theta}_i^{HB}$, $i = 1, \dots, m$, i.e., $\hat{\theta}_i^{BHB} = f(\hat{\theta}_1^{HB}, \dots, \hat{\theta}_m^{HB})$ for some function $f(\cdot)$, and satisfies the benchmark property:

$$\sum_{i=1}^m \hat{\theta}_i^{BHB} = \sum_{i=1}^m y_i.$$

For example, a ratio BHB (RBHB) estimator can be obtained as

$$\hat{\theta}_i^{RBHB} = \hat{\theta}_i^{HB} \frac{\sum_{k=1}^m y_k}{\sum_{k=1}^m \hat{\theta}_k^{HB}}.$$

To obtain a measure of variability associated with the BHB estimator $\hat{\theta}_i^{BHB}$, we use the following posterior mean squared error (PMSE),

$$\text{PMSE}(\hat{\theta}_i^{BHB}) = E[(\hat{\theta}_i^{BHB} - \theta_i)^2 | y],$$

which is similar to the posterior variance associated with the HB estimator $\hat{\theta}_i^{HB}$. It can be shown (see Appendix) that the PMSE of $\hat{\theta}_i^{BHB}$ is given by

$$\text{PMSE}(\hat{\theta}_i^{BHB}) = (\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2 + V(\theta_i | y). \quad (6)$$

Thus the PMSE of $\hat{\theta}_i^{BHB}$ is simply the sum of the posterior variance $V(\theta_i | y)$ and a bias correction term $(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2$. The PMSE is readily obtained from the posterior variance and the estimators $\hat{\theta}_i^{HB}$ and $\hat{\theta}_i^{BHB}$. For the ratio benchmarked estimator $\hat{\theta}_i^{RBHB}$, the $\text{PMSE}(\hat{\theta}_i^{RBHB})$ is given as

$$\text{PMSE}(\hat{\theta}_i^{RBHB}) = [\hat{\theta}_i^{HB} (\frac{\sum_{k=1}^m y_k}{\sum_{k=1}^m \hat{\theta}_k^{HB}} - 1)]^2 + V(\theta_i | y).$$

3. CENSUS UNDERCOVERAGE ESTIMATION

3.1 Background

In Canada, a census is conducted every five years. However, the census does not enumerate all the inhabitants that should fill a census form on Census Day. In the 1991 Canadian census, it is estimated that about 3% of the population were not enumerated. Thus the census needs to be adjusted for undercoverage in order to properly represent the demographic picture of the country on

Census Day. Since 1966, the Reverse Record Check (RRC) has been used by Statistics Canada to measure the gross number of persons missed by the census. The RRC is a sample survey with a sample size of 60,000 persons, estimating the two types of census 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. Starting 1991, an Overcoverage Study was conducted to measure the gross number of persons erroneously included in the census. In 1991, for the first time, the RRC results were combined with those of the Overcoverage Study to produce the direct survey estimates of the net undercoverage for the nation and all provinces. Through the analysis of the results of these coverage studies, the census collection methodology is adjusted in order to improve coverage in the succeeding census. In 1991, the population estimates were based on the census counts adjusted for the estimated net undercoverage in the census. The 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. Rivest (1995) proposed a composite estimator to estimate the provincial undercoverage using the national undercoverage rate as a synthetic estimate. The effect of the composite estimator is to shrink all provincial rates to the national rate. The composite estimator performs poorly at extreme provinces, namely, P.E.I. and Ontario, the smallest and the largest provinces of Canada (Rivest, 1995; Rivest and Belmonte, 2000).

In recent years, modelling techniques have been applied widely in practice to obtain reliable model-based estimates and therefore to improve the direct survey estimates from sample surveys. In this paper, along the line of You and Rao (2002), we apply the model-based approach to improve the provincial level census undercoverage estimates.

3.2 Models and Inference

Following You and Rao (2002), we consider the following unmatched sampling and linking models to obtain HB and BHB estimates of provincial census undercoverage:

$$y_i = u_i + \varepsilon_i, \quad i = 1, \dots, 12, \quad \varepsilon_i \sim N(0, \sigma_i^2) \quad (7)$$

and

$$\log(u_i / (u_i + c_i)) = x_i' \beta + v_i, \quad i = 1, \dots, 12, \quad (8)$$

where u_i is the true undercoverage count for the i th province, y_i is the direct estimate of u_i , the sampling variances σ_i^2 are known. The linking model (8) is a log-

linear random effects model for the undercoverage rate, which is defined as $\theta_i = u_i / (u_i + c_i)$. Thus the rate θ_i is expressed as a function of count u_i . The linking model (8) is a more complex model than the regular log-linear model (3). Following You and Rao (2002) and by using a complete HB approach with the Gibbs sampling method, we can find the posterior estimates of the undercoverage counts u_i , and the associated posterior variances. Let \hat{u}_i^{HB} denote the HB estimator of u_i . By using the simple ratio benchmarking approach given in subsection 2.3, we can obtain the ratio BHB estimator \hat{u}_i^{BHB} .

3.3 Application and Results

We used the 1991 and 1996 Canadian census undercoverage data in our analysis. We used log-transformation of the census count as the auxiliary variable, that is, $x_{1i} = \log(c_i)$, and the linking model for u_i is $\log(u_i / (u_i + c_i)) = \beta_0 + x_{1i}\beta_1 + v_i$. To implement and monitor the convergence of the Gibbs sampler, we followed the basic approach given in Gelman and Rubin (1992). We independently simulated L=8 sequences, each of length t=2d, with d=5000. The first 5000 iterations of each sequence were deleted. To reduce the autocorrelation in the sequence, we took every 10th iteration of the remaining 5000 iterations, leading to 500 samples for each sequence. Thus we finally have L=8 sequences with sample size n=500 for each sequence.

Tables 1 and 2 present the direct, HB and BHB undercoverage estimates for the 10 provinces, together with the corresponding standard errors and coefficients of variations (CVs). The standard error of the HB estimate is the squared root of the posterior variance, and the standard error of the BHB estimate is the squared root of the corresponding PMSE. The HB and BHB estimates have smaller CVs than the direct estimates, especially for some smaller provinces. The BHB estimates add up to the total of the direct survey estimates. The BHB estimates have slightly larger standard errors than the HB estimates due to the benchmarking property. But the BHB estimates have the same CVs as the HB estimates in this application.

3.4 Test of Model Fit

Following Datta et al. (1999) and You and Rao (2002), we use the method of posterior predictive p value for model fit analysis. The posterior predictive p value is defined as $p = \Pr(T(y^*, \theta) > T(y_{obs}, \theta) | y_{obs})$, where y^* is a sample from the posterior predictive distribution $f(y | y_{obs})$, and $T(y, \theta)$ is a discrepancy measure depending on the data y and on parameters θ . Let θ^* represent a draw from the posterior distribution of θ and let y^* represent a draw from $f(y | \theta^*)$, then marginally $y^* \sim f(y | y_{obs})$. Note that the probability is with respect to the posterior distribution given the observed data. If a model fits the observed data, then the two values of the discrepancy measure are similar. In other words, if the given model adequately fits the observed model, then $T(y_{obs}, \theta)$ should be near the central part of the histogram of the $T(y^*, \theta)$ values if y^* is generated repeatedly from the posterior predictive distribution. Consequently, the posterior predictive p value is expected to be near 0.5 if the model adequately fits the data. Extreme p values (near 0 or 1) suggest poor fit. Computing the p value is relatively easy using the posterior simulation from the Gibbs sampler. For each simulated value θ^* , we can simulate y^* from the model and compute $T(y^*, \theta^*)$ and $T(y_{obs}, \theta^*)$. Then the p value is estimated by the proportion of times that $T(y^*, \theta^*)$ exceeds $T(y_{obs}, \theta^*)$. In the present context of census undercoverage, the discrepancy measure that we used for overall fit is $T(y, u) = \sum_i (y_i - u_i)^2 / \sigma_i^2$. Similar discrepancy measure is also used, for example, in Datta, et al. (1999). For the 1991 census undercoverage data, the estimated p value is 0.383; whereas for the 1996 data, the estimated p value is 0.493. Thus we have no indication of any lack of overall fit of the model for both the 1999 and 1996 data. The p value strongly suggests the adequacy of the model.

Table 1 – 1991 Census undercoverage estimation

Province	Estimate			Standard Error			CV		
	Direct	HB	BHB	Direct	HB	BHB	Direct	HB	BHB
NFLD	11566	10782	10925	1846	1471	1478	0.16	0.14	0.14
PEI	1220	1486	1506	366	289	290	0.30	0.19	0.19
NS	17329	17412	17643	3475	2474	2485	0.20	0.14	0.14
NB	24280	18948	19200	3333	3294	3304	0.14	0.17	0.17
QUE	184473	189599	192119	15400	15105	15314	0.08	0.08	0.08
ONT	381104	368424	373321	32260	31316	31697	0.08	0.08	0.08
MAN	20691	21504	21790	4310	3077	3090	0.21	0.14	0.14
SASK	18106	18822	19072	3416	2550	2562	0.19	0.14	0.13
ALTA	51825	55392	56128	7553	6591	6632	0.15	0.12	0.12
BC	92236	89929	91124	9096	8109	8197	0.10	0.09	0.09
Total	802830	792298	802830						

Table 2 – 1996 Census undercoverage estimation

Province	Estimate			Standard Error			CV		
	Direct	HB	BHB	Direct	HB	BHB	Direct	HB	BHB
NFLD	9424	9301	9308	1759	1543	1543	0.19	0.17	0.17
PEI	1149	1419	1420	437	364	364	0.38	0.26	0.26
NS	20821	19999	20013	2580	2357	2357	0.12	0.12	0.12
NB	14225	13835	13845	2354	2102	2102	0.17	0.15	0.15
QUE	116750	124264	124354	14963	14659	14660	0.13	0.12	0.12
ONT	301368	301422	301641	21265	20816	20817	0.07	0.07	0.07
MAN	18881	19379	19393	3875	3364	3364	0.21	0.17	0.17
SASK	28051	26023	26042	3521	3411	3411	0.13	0.13	0.13
ALTA	66327	65038	65085	7555	7258	7258	0.11	0.11	0.11
BC	142443	138236	138336	9967	10337	10337	0.07	0.07	0.07
Total	719439	718915	719439						

4. SUMMARY

In this paper, we have studied benchmarked HB (BHB) estimators for small area estimation based on unmatched sampling and linking models proposed by You and Rao (2002). The BHB estimates add up to the direct survey estimates for large areas to protect against possible model mis-specification and possible overshrinkage for the direct survey estimates. This is very appealing to survey practitioners. In particular, we also developed posterior MSE (PMSE) as a measure of uncertainty for the BHB estimators. The PMSE is very easy to compute using the HB, BHB estimates and the posterior variance. We applied the proposed BHB approach to the Canadian census undercoverage estimation and obtained the BHB undercoverage estimates for the 10 provinces across Canada. The proposed BHB estimation approach can be used for all model-based small area estimation under the hierarchical Bayes framework.

For the future study, we will use the proposed BHB estimation approach to produce the small domain estimates used in Dick (2001). Dick used an empirical Bayes (EB) approach with the estimates calibrated to the known survey totals. However the MSE approximation used in Dick (2001) did not account for this calibration. This research should also provide a useful data set to compare the results of the EB approach with the BHB approach.

APPENDIX

$$\text{Proof of } \text{PMSE}(\hat{\theta}_i^{BHB}) = (\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2 + V(\theta_i | y):$$

We have

$$\begin{aligned} \text{PMSE}(\hat{\theta}_i^{BHB}) &= E[(\hat{\theta}_i^{BHB} - \theta_i)^2 | y] \\ &= E[(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB} + \hat{\theta}_i^{HB} - \theta_i)^2 | y] \\ &= E[(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2 | y] + \\ &\quad E[(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})(\hat{\theta}_i^{HB} - \theta_i) | y] + E[(\hat{\theta}_i^{HB} - \theta_i)^2 | y] \\ &= (\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2 + V(\theta_i | y) \end{aligned}$$

by noting that the cross-product term is equal to 0, that is,

$$E[(\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})(\hat{\theta}_i^{HB} - \theta_i) | y] \\ = (\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})E[(\hat{\theta}_i^{HB} - \theta_i) | y] = 0.$$

REFERENCES

- Chip, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician*, 49, 327-335.
- Datta, G.S., Lahiri, P., Maiti, T. and Lu, K.L. (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94, 1074-1082.
- Dick, J.P. (2001). Small domain estimation of missed persons in the 2001 census. *Proceedings of the Survey Method Section, Statistical Society of Canada*, 37-46.
- 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.
- Gelman, A. and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.
- Ghosh, M. and Rao, J.N.K. (1994). Small area estimation: an appraisal (with discussion). *Statistical Science*, 9, 55-93.
- Rao, J.N.K. (1999). Some recent advances in model-based small area estimation. *Survey Methodology*, 25, 175-186.
- Rivest, L.P. (1995). A composite estimator for provincial undercoverage in the Canadian census. *Proceedings of the Survey Method Section, Statistical Society of Canada*, 33-38.
- Rivest, L.P. and Belmonte, E. (2000). A conditional mean squared error of small area estimators. *Survey Methodology*, 26, 67-78.
- 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.