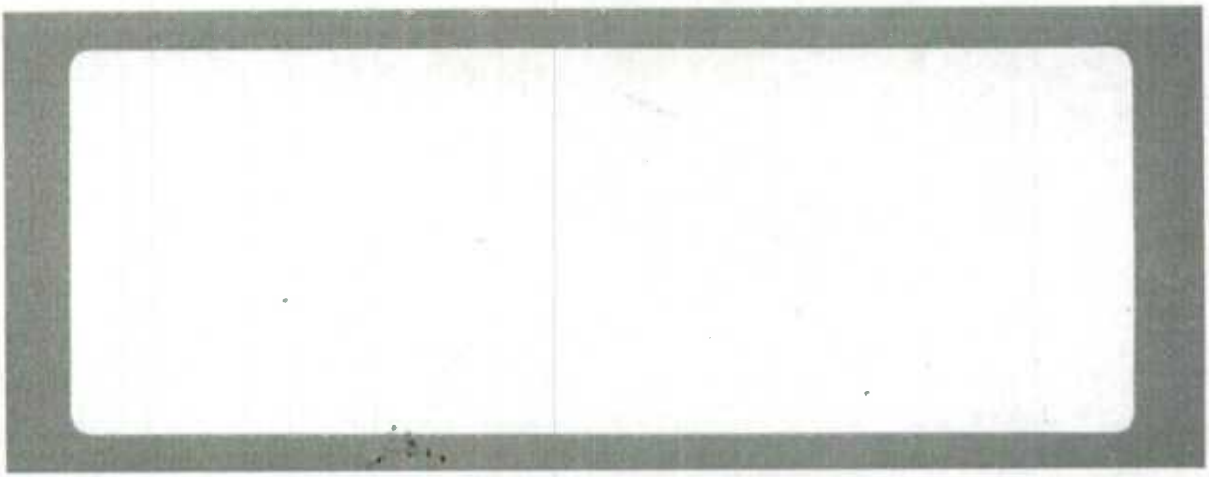
Statistics Canada Statistique Canada



Methodology Branch

Household Survey
Methods Division

Direction de la méthodologie

Division des méthodes
d'enquêtes des ménages

Canada

WORKING PAPER
METHODOLOGY BRANCH

Model-based Approach to Small Area Census Undercoverage Estimation with Application to 2001 Census Data

HSMD – 2008-005E

Yong You and Peter Dick

Household Survey Methods Division
Statistics Canada

July 2008

The work presented in this paper is the responsibility of the author and does not necessarily represent the views or policies of Statistics Canada

Model-based Approach to Small Area Census Undercoverage Estimation with Application to 2001 Census Data

Yong You and Peter Dick[1]

Abstract

In this paper we consider the well known Fay-Herriot model for census undercoverage small area estimation across Canada. In general sampling variances are assumed to be known in the Fay-Herriot model. However in this paper we consider the case of unknown sampling variances. Direct estimators are available for the sampling variances. We construct sampling models for both the direct survey estimators of parameters of interest and the direct estimators of sampling variances. For inference we have considered the empirical best linear unbiased prediction (EBLUP) approach and a full hierarchical Bayes (HB) approach. For EBLUP approach, we obtain the EBLUP estimators and the corresponding MSE estimators with an extra term to account for the uncertainty associated with the estimation of the sampling variances (Wang and Fuller, 2003). For HB approach, we apply the Gibbs sampling method and obtain the HB estimators and the corresponding posterior variances (You and Chapman, 2006). In particular, the HB estimators are benchmarked to agree with direct total estimate using the benchmarking HB method of You, Rao and Dick (2004). Posterior mean squared error (PMSE) is used as a measure of uncertainty for the benchmarked HB estimators. We compare the EBLUP and HB estimators in the data analysis with the direct survey estimators. Model validation and model fit analysis are also provided.

KEY WORDS: Benchmarking, Census undercoverage, Hierarchical Bayes, Sampling variance, Small area estimation.

[1] Yong You, Household Survey Methods Division; Peter Dick, Social Survey Methods Division, Statistics Canada, Ottawa, Canada, K1A 0T6.

Approche basée sur un modèle pour l'estimation sur petits domaines du sous-dénombrement au recensement s'appliquant aux données du Recensement de 2001

Yong You et Peter Dick

Résumé

Dans la présente étude, nous examinons le modèle bien connu de Fay-Herriot pour l'estimation sur petits domaines du sous-dénombrement au recensement dans l'ensemble du Canada. En général, dans le modèle de Fay-Herriot, il est supposé que les variances d'échantillonnage sont connues. Dans la présente étude, toutefois, nous examinons le cas où les variances d'échantillonnage sont inconnues. Il existe des estimateurs directs pour les variances d'échantillonnage. Nous construisons des modèles d'échantillonnage tant pour les estimateurs par sondage directs des paramètres d'intérêt que pour les estimateurs directs des variances d'échantillonnage. Pour faire des inférences, nous avons examiné l'approche empirique du meilleur prédicteur linéaire sans biais (EBLUP) et une approche hiérarchique bayésienne (HB) complète. Pour l'approche EBLUP, nous obtenons les estimateurs EBLUP et les estimateurs de l'erreur quadratique moyenne correspondants avec un terme supplémentaire permettant de tenir compte de l'incertitude associée à l'estimation des variances d'échantillonnage (Wang et Fuller, 2003). Pour l'approche HB, nous utilisons la méthode d'échantillonnage de Gibbs pour obtenir les estimateurs HB et les variances *a posteriori* correspondantes (You et Chapman, 2006). Notamment, les estimateurs HB sont calés de manière à correspondre aux estimations directes totales calculées à l'aide de la méthode d'étalonnage HB de You, Rao et Dick (2004). L'erreur quadratique moyenne *a posteriori* (EQMP) sert à mesurer l'incertitude des estimateurs HB calés. Dans l'analyse des données, nous comparons les estimateurs EBLUP et HB avec les estimateurs par sondage directs. Nous présentons également une analyse de la validation du modèle et de l'ajustement du modèle.

MOTS CLÉS : calage, sous-dénombrement au recensement, approche hiérarchique bayésienne, variance d'échantillonnage, estimation sur petits domaines.

1. Introduction

The Census of Canada is conducted every 5 year. The last census was conducted on May 16, 2006. One objective of the census 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 at sub-provincial level across Canada. The count of persons includes usual residents, immigrants and non-permanent residents; excluded are all foreign visitors and non permant residents without a valid permit. Unfortunately, not all persons are correctly enumerated by the census. The census data needs to be adjusted for undercoverage in order to properly represent the demographic picture on census day. Two errors that occur in the census are undercoverage – exclusion of eligible persons – and overcoverage – erroneous inclusion of persons. The direct net undercoverage estimates are obtained by subtracting the overcoverage estimates from the direct undercoverage estimates.

The main coverage study conducted by Statistics Canada is a survey called Reverse Record Check (RRC). The RRC is a sample survey, with a sample size of 60,000 persons, that 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 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, such as the well known Fay and Herriot (1979), to obtain model-based undercoverage estimates (e.g., Dick, 1995). 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 smoothing of the sampling variance requires external variables or models such as the generalized variance function (GVF) method (e.g., Bell and Otto, 1995; Dick, 1995; Dick and You, 2003). 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 mean squared error (MSE) using the empirical best linear unbiased prediction (EBLUP) approach if a final benchmarking step is imposed on the EBLUP estimates. Hierarchical Bayes (HB) approach has been studied extensively in recent years in small area estimation to account for complex models (e.g., Rao, 2003). The HB approach also has been used for census undercoverage estimation (e.g., You and Rao, 2002; You and Dick, 2004). The proper comparison of these two approaches is addressed in this paper. The chosen method is to adjust the model fit into a hierarchical Bayes framework. With this approach we can use the methods developed recently for evaluating the 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 area level parameters are exact in the sense that the posterior means and posterior variances are computed exactly, unlike the EBLUP approach where approximation is needed when estimating the MSE. 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 illustrate both the EBLUP and HB inference methods using the Fay-Herriot type model with sampling variances modeled by the direct estimates 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 the small area model considered for the census undercoverage estimation and inference methods. We provide formulas for both the EBLUP and HB methods. In Section 3 we present data analysis results based on the year 2001 census undercoverage data at provincial level age-sex domains and some related model diagnostics. And finally in section 4, we offer some concluding remarks on the EBLUP and HB methods.

2. Model Specification and Inference

2.1 Small area model

Let y_i denote the direct survey estimator of the i-th small area parameter of interest θ_i. The sampling model for y_i can be expressed as

$$y_i = \theta_i + \varepsilon_i, \quad i = 1,...,m, \tag{1}$$

where ε_i is the sampling error associated with the direct estimator y_i 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 in applications. 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).

The Fay-Herriot model (Fay and Herriot, 1979) is a special case of the general area level model. In the Fay-Herriot model, the linking function is given as $g(\theta_i) = \theta_i$ 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. The Fay-Herriot model assumes that the sampling variances σ_i^2 are known in the model. This is a very strong assumption. Usually external variables and models are needed to obtain a smoothed estimate of σ_i^2 and then the smoothed estimate is treated as known in the Fay-Herriot model. 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 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 $(n_i - 1)s_i^2 \sim \sigma_i^2 \chi^2_{n_i - 1}$. In this paper, we consider the following area level model specification for the census undercoverage estimation:

- Sampling model for y_i: $y_i = \theta_i + \varepsilon_i$, where $V(\varepsilon_i) = \sigma_i^2$ and σ_i^2 unknown.
- Sampling model for s_i^2: $(n_i - 1)s_i^2 \sim \sigma_i^2 \chi^2_{n_i - 1}$.
- Linking model for θ_i: $\theta_i = x_i'\beta + v_i$, where $V(v_i) = \sigma_v^2$, β and σ_v^2 unknown.

Normal distribution is also assumed for sampling error ε_i and model error v_i.

Let c_i denote the census count for the i-th small 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. Then the direct estimator of θ_i is given by $\hat{\theta}_i = \hat{m}_i / c_i$. We then apply the proposed model with unknown sampling variances to the census undercoverage ratio estimation by letting $y_i = \hat{\theta}_i$ and $\theta_i = m_i / c_i$. In the following sections, we consider the empirical best linear unbiased prediction (EBLUP) approach and the hierarchical Bayes (HB) approach to obtain model-based estimators of θ_i.

2.2 EBLUP approach

In this section we consider the EBLUP method to estimate θ_i. Combining the sampling model for y_i and the linking model for θ_i, we obtain the model

$$y_i = x_i'\beta + v_i + \varepsilon_i, \qquad (2)$$

which is a mixed effects linear model. By assuming σ_i^2 and σ_v^2 to be known in the model, we can obtain the best linear unbiased prediction (BLUP) estimator of θ_i as

$$\tilde{\theta}_i = \gamma_i y_i + (1 - \gamma_i) x_i' \tilde{\beta}, \qquad (3)$$

where $\gamma_i = \sigma_v^2/(\sigma_v^2 + \sigma_i^2)$ and

$$\tilde{\beta} = \left[\sum_{i=1}^{m}(\sigma_i^2 + \sigma_v^2)^{-1} x_i x_i'\right]^{-1}\left[\sum_{i=1}^{m}(\sigma_i^2 + \sigma_v^2)^{-1} x_i y_i\right]. \qquad (4)$$

To estimate the variance component σ_v^2, we have to assume σ_i^2 to be known in the model. Replacing σ_i^2 by its estimate s_i^2, we use the Fay-Herriot iterative (FHI) method (Fay and Herriot, 1979) to estimate σ_v^2. The FHI method is evaluated and preferred by Datta, Rao and Smith (2005). The FHI method is as follows: starting with $\sigma_v^{2(0)} = 0$, solve iteratively,

$$\sigma_v^{2(a+1)} = \sigma_v^{2(a)} + \frac{1}{h_*'(\sigma_v^{2(a)})}\left[m - p - h(\sigma_v^{2(a)})\right]$$

constraining $\sigma_v^{2(a+1)} \geq 0$, where $h(\sigma_v^2) = \sum_{i=1}^{m}(y_i - x_i'\tilde{\beta})^2/(s_i^2 + \sigma_v^2)$ and

$$h'_*(\sigma_v^2) = -\sum_{i=1}^{m}(y_i - x_i'\tilde{\beta})^2/(s_i^2 + \sigma_v^2)^2 .$$

$\tilde{\beta}$ is given by (4). Convergence of the iteration is rapid. The FHI method does not require normality and like the simple moment estimator leads to consistent estimators as $m \to \infty$ (Rao, 2003). The asymptotic variance of the FHI estimator σ_v^2 was obtained by Datta, Rao and Smith (2005) as

$$V(\sigma_v^2) = 2m\left[\sum_{i=1}^{m}\frac{1}{(\sigma_v^2 + s_i^2)}\right]^{-2}.$$

Replacing σ_v^2 and σ_i^2 by $\hat{\sigma}_v^2$ and s_i^2 in (3), we obtain the EBLUP estimator of θ_i as

$$\hat{\theta}_i = \hat{\gamma}_i y_i + (1 - \hat{\gamma}_i)x_i'\hat{\beta}, \qquad (5)$$

where $\hat{\gamma}_i = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + s_i^2)$.

An estimator of the mean squared error (MSE) of θ_i is given as

$$mse(\hat{\theta}_i) = g_{0i} + g_{1i} + g_{2i} + 2g_{3i} + g_{4i}, \qquad (6)$$

where g_{1i}, g_{2i} and g_{3i} are the terms obtained by Prasad and Rao (1990) in the MSE estimation. The term $g_{1i} = \hat{\gamma}_i s_i^2$ is the leading term, g_{2i} is due to the estimation of β and given as

$$g_{2i} = \hat{\sigma}_v^2 (1-\hat{\gamma}_i)^2 x_i' \left(\sum_{i=1}^{m} \hat{\gamma}_i x_i x_i' \right)^{-1} x_i .$$

Term $g_{3i} = s_i^4 (\hat{\sigma}_v^2 + s_i^2)^{-3} V(\hat{\sigma}_v^2)$ is due to estimation of σ_v^2, $g_{0i} = -(1-\hat{\gamma}_i)^2 b(\hat{\sigma}_v^2)$ is an extra term due to estimation of σ_v^2 using the FHI method (Rao, 2003; Datta, Rao and Smith, 2005), where

$$b(\hat{\sigma}_v^2) = 2\{m \sum_{i=1}^{m} (s_i^2 + \hat{\sigma}_v^2)^{-2} - (\sum_{i=1}^{m} (s_i^2 + \hat{\sigma}_v^2)^{-1})^2\} / \{\sum_{i=1}^{m} (s_i^2 + \hat{\sigma}_v^2)^{-1}\}^3 .$$

Finally g_{4i} is a new term due to unknown σ_i^2 in the sampling model (1). Rivest and Vandal (2002) and Wang and Fuller (2003) obtained the g_{4i} term to account for the extra uncertainty associated with the estimation of σ_i^2 by s_i^2. The g_{4i} term is given as

$$g_{4i} = \frac{4}{n_i - 1} \frac{\hat{\sigma}_v^4 s_i^4}{(\hat{\sigma}_v^2 + s_i^2)^3} .$$

In the data analysis section, we will compute the EBLUP and the MSE estimators and compare them with the HB estimators empirically.

2.3 Hierarchical Bayes approach

Following You and Chapman (2006), we now present the proposed model in Section 2.1, i.e., the Fay-Herriot type model with unknown sampling variances estimated by the direct estimators s_i^2 in a HB framework as follows:

(1) $y_i | \theta_i, \sigma_i^2 \sim ind\ N(\theta_i, \sigma_i^2), I = 1,\ldots,m;$

(2) $d_i s_i^2 | \sigma_i^2 \sim ind\ \sigma_i^2 \chi_{d_i}^2,\ d_i = n_i - 1, I = 1,\ldots,m;$

(3) $\theta_i | \beta, \sigma_v^2 \sim ind\ N(x_i'\beta, \sigma_v^2), I = 1,\ldots,m;$

(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.

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:

- $\left[\theta_i \mid y_i, \beta, \sigma_i^2, \sigma_v^2\right] \sim N(\gamma_i y_i + (1-\gamma_i)x_i'\beta, \gamma_i \sigma_i^2)$
- $\left[\beta \mid y_i, \theta_i, \sigma_i^2, \sigma_v^2\right] \sim N_p((\sum_{i=1}^{m} x_i x_i')^{-1}(\sum_{i=1}^{m} x_i \theta_i), \sigma_v^2 (\sum_{i=1}^{m} x_i x_i')^{-1})$
- $\left[\sigma_i^2 \mid y_i, \theta_i, \beta, \sigma_v^2\right] \sim IG(a_i + \frac{d_i + 1}{2}, b_i + \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2})$
- $\left[\sigma_v^2 \middle| y_i, \theta_i, \beta, \sigma_i^2\right] \sim IG(a_0 + \frac{m}{2}, b_o + \frac{1}{2}\sum_{i=1}^{m}(\theta_i - x_i'\beta)^2)$

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

$$\hat{\theta}_i^{\mathrm{HB}} = G^{-1} \sum_{k=1}^{G} \left(\gamma_i^{(k)} y_i + (1-\gamma_i^{(k)}) x_i' \beta^{(k)} \right), \qquad (7)$$

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

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

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 $\hat{V}(m_i \mid y_i) = c_i^2 \times \hat{V}(\theta_i \mid y_i)$. 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 direct estimate of the national total is respected. Also, in order to protect possible model mis-specification and possible over shrinkage, we may consider to benchmark the HB estimators so that the benchmarked HB estimators add up to the direct total estimate. You, Rao and Dick (2004) constructed benchmarked HB estimators for small areas. Let $\hat{m}_i^{\mathrm{BHB}}$ 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

$$\hat{m}_i^{\mathrm{RBHB}} = \hat{m}_i^{\mathrm{HB}} \frac{\sum_{j=1}^{t} \hat{m}_j}{\sum_{j=1}^{t} \hat{m}_j^{\mathrm{HB}}}. \qquad (8)$$

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

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

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, 2004) that the PMSE of $\hat{m}_i^{\mathrm{BHB}}$ is given by

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

Thus the PMSE of $\hat{m}_i^{\mathrm{BHB}}$ is simply the sum of the posterior variance $V(m_i \mid y_i)$ 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 $\hat{m}_i^{\mathrm{BHB}}$. The advantage of the BHB estimator and the PMSE is that the benchmarking estimation procedure is well-defined and very easy to compute, unlike the benchmarked EB or EBLUP approaches (e.g., see Pfeffermann, 2006).

3. Application to the 2001 Census Data

We applied the proposed area level model with the estimated sampling variances to the 2001 Census undercoverage data for small domains across Canada; for more discussion on related data and methods, see Dick and You (2003). We have direct survey estimates for net undercoverage for 104 domains across Canada. The domains are defined as age (0-19, 20-29, 30-44, 45+) , sex (2 groups) and province/territory (13 groups). The model requires domain level 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 were reduced to eight variables and an intercept term (Dick and You, 2003). The eight variables are Yukon, Nunavut, Male 20 to 29, Male 30 to 44, Female 20 to 29, British Columbia (BC) renters, Ontario (ON)

renters and Northwest Territory (NWT) renters. We have implemented both the EBLUP and HB estimation approaches to the data. For the HB approach, to implement the Gibbs sampling method, we considered L=10 parallel chains, each of length $2d$=2000. For each chain, the first d=1000 iterations were treated as "burn-in" period and deleted from final computation.

3.1 EBLUP and HB estimation

In this section we present EBLUP and HB estimates. Table 1 gives both EBLUP and HB estimates with standard errors of the fixed effects for the variables in the model. The t-value is simply the ratio of estimate over standard error. The EBLUP standard error is the squared root of the MSE and the HB standard error is the squared root of the posterior variance. As we can see from Table 1, both EBLUP and HB give similar estimates for the fixed effects. The HB approach in general has slightly smaller standard errors than the EBLUP approach. The t-values under the HB approach are slightly larger than the EBLUP approach.

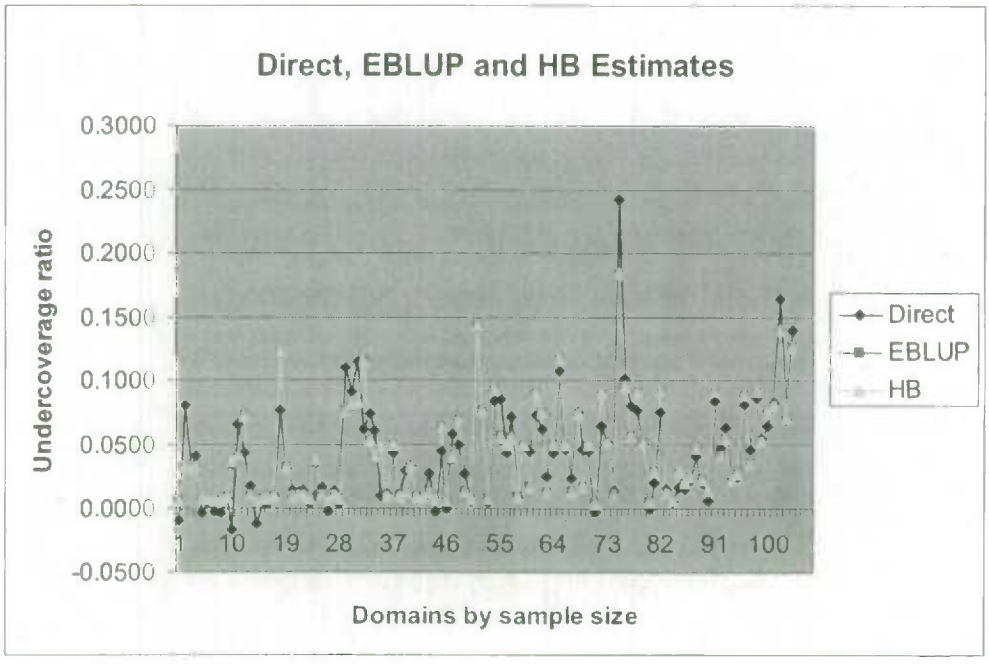
Table 1: Estimation of Fixed Effects.

Variable	Estimates		Standard Errors		t-values	
	EBLUP	HB	EBLUP	HB	EBLUP	HB
Mean	0.0089	0.0084	0.0020	0.0019	4.43	4.42
Yukon	0.0279	0.0291	0.0113	0.0111	2.46	2.62
Nunavut	0.0235	0.0251	0.0116	0.0112	2.03	2.24
Male 20 to 29	0.0848	0.0856	0.0050	0.0047	16.88	18.21
Male 30 to 44	0.0417	0.0416	0.0046	0.0041	9.10	10.15
Female 20 to 29	0.0429	0.0425	0.0050	0.0047	8.53	9.04
BC Renters	0.0919	0.0946	0.0154	0.0151	5.95	6.26
ON Renters	0.0732	0.0752	0.0151	0.0141	4.85	5.37
NWT Renters	0.1728	0.1733	0.0261	0.0233	6.61	7.43

Our main interest is to estimate the small domain undercoverage ratio and compare the model-based estimates with the direct survey estimates. Figure 1 presents the comparison of the point estimates by domains listed by their sample sizes from smallest (left hand side) to largest (right hand side). It is clear that the direct estimates tend to have more variation and more extreme values, whereas the EBLUP and

HB estimates are similar to each other and both lead to moderate smooth of the direct estimates. We also note that some direct undercoverage estimates are negative, due to the fact the overcoverage estimates are larger than the undercoverage estimates in those domains which usually have relatively small sample sizes. The EBLUP and HB methods "correct" the negative values for those small domains. All the model-based undercoverage estimates are positive as we expect.

Figure 1. Comparison of Direct, EBLUP and HB Estimates.



To compare the standard errors, we compute the efficiency of the model-based estimates. The efficiency is defined as the ratio of the sampling variance to the estimated MSE or the estimated posterior variance. Figure 2 presents the EBLUP and HB efficiency gains over the direct estimates. It is clear from Figure 2 that both the EBLUP and HB estimates lead to large efficiency gains over the direct estimates. As the sample sizes increase, the direct estimates tend to be more reliable and the model-based efficiency gain decreases as expected. Table 2 gives the efficiency gain of the EBLUP and HB estimates at the small domain level, summarized by domain sample sizes. For example, for domains with sample sizes less than 10, the average efficiency gain is 5.93 for EBLUP and 6.36 for HB. For domains with sample sizes larger than 100, the average efficiency gain is only 1.64 for EBLUP and 1.72 for HB. Also we note that in general the HB approach has slightly larger efficiency gain than the EBLUP approach.

Figure 2. Comparison of EBLUP and HB Efficiency Gains Over Direct Estimates.

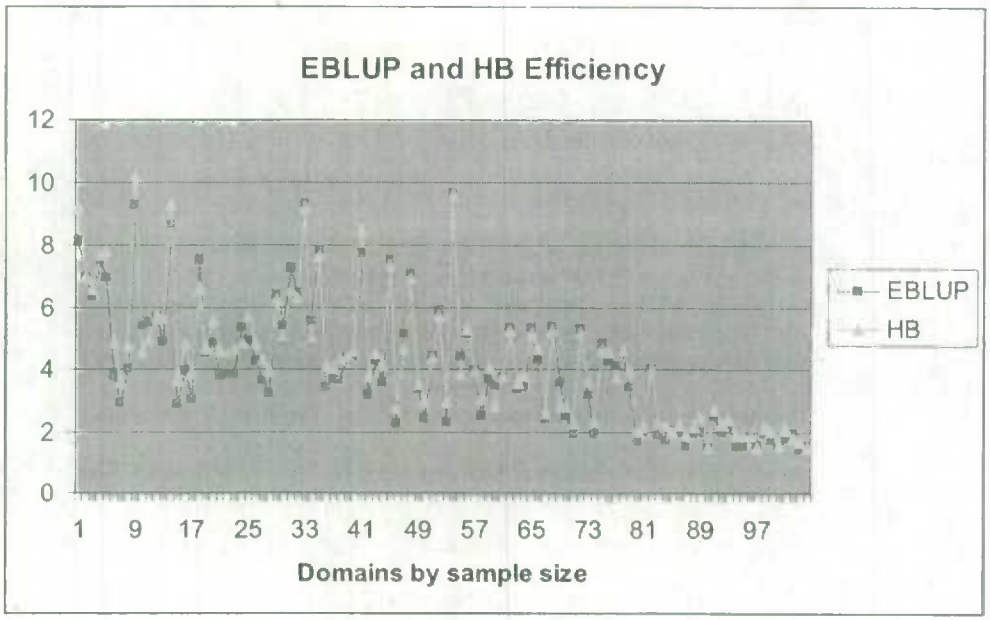


Table 2. EBLUP and HB Efficiency Comparison

	EBLUP Approach			HB Approach		
Sample size	average	min	max	average	min	max
< 10	5.93	2.88	9.26	6.36	3.60	10.19
10 – 19	5.01	2.25	9.33	5.25	2.77	9.17
20 – 49	3.73	1.67	9.62	3.84	2.01	9.57
50 – 99	1.85	1.51	2.51	2.08	1.47	2.71
≥ 100	1.64	1.44	1.94	1.72	1.48	2.09

One advantage of the HB approach is that we can easily compute the benchmarked undercoverage counts for domains as shown in Section 2.3. Posterior MSE is then used as measure of uncertainty for the benchmarked estimates. You and Dick (2004) obtained the BHB estimates and compared with the HB estimates. It is shown in You and Dick (2004) that the BHB estimates and HB estimates are very close to each other. The benckmarking only makes slight change to the HB estimates and the BHB CV is only slightly larger than the HB CV as we hope. The results have shown that the proposed model-based estimates do not lead to over shrinkage or under shrinkage of the direct survey estimates.

3.2 Bias diagnostic using regression analysis

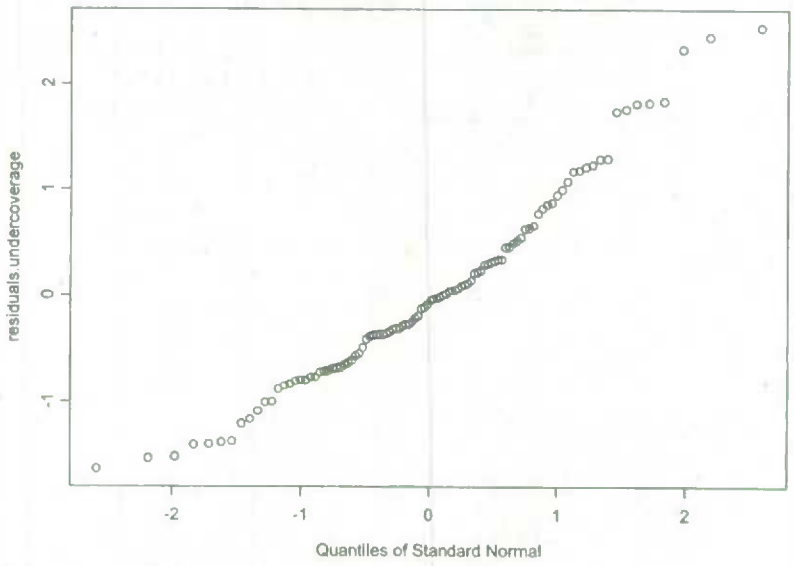
To evaluate the possible bias introduced by the model, we use a simple method of ordinary least squares regression analysis for the direct estimates and the EBLUP and HB estimates. The regression method is suggested by Brown, Chambers, Heady and Heasman (2001). If the model-based estimates are close to the true values, then the direct estimators should behave like random variables whose expected values correspond to the values of the model-based estimates. We plot the model-based estimates as X and the direct estimates as Y, and assess how close the regression line is to Y=X. In terms of regression, basically we fit the regression model $Y = \alpha X$ to the data and estimate the coefficient α. Less biased model-based estimates should lead to the value of α close to 1. For the 2001 census undercoverage data, let Y be the direct undercoverage estimates, and X be the model-based estimates. For the EBLUP estimates, we obtain the estimated α value as 1.0010 with standard error 0.0291. For the HB estimates, the estimated α value is 0.9977 with standard error 0.0296. Thus the regression results show no difference of the fitted line from Y=X. Therefore, we conclude that the model-based estimates derived from the proposed model are consistent with the direct estimates with no extra possible bias included. The result may also indicate no evidence of any bias due to possible model misspecification.

3.3 The marginal model checking

We assume a normal sampling model for the direct estimates y_i's and a normal linking model for the parameters of interest θ_i's. However, as θ_i's are not directly observable, the normal distribution is often chosen as an approximation. By combining the sampling model $y_i|\theta_i,\sigma_i^2 \sim N\left(\theta_i,\sigma_i^2\right)$ and the linking model $\theta_i|\beta,\sigma_v^2 \sim N\left(x_i'\beta,\sigma_v^2\right)$, we can obtain a marginal model as $y_i|\beta,\sigma_i^2,\sigma_v^2 \sim N\left(x_i'\beta,\sigma_i^2+\sigma_v^2\right)$ to check if the normality assumption is valid. An obvious model checking diagnostic is a normal probability plot of the standardized residuals $z_i = (y_i - x_i'\beta)/\sqrt{\sigma_i^2+\sigma_v^2}$. Computing z_i requires point estimates of β, σ_i^2 and σ_v^2. We can simply use the direct estimates s_i^2 to estimate the sampling variance σ_i^2. As in the EBLUP approach, we use the Fay-Herriot iterative (FHI) method to estimate σ_v^2. We denote the FHI estimator of σ_v^2 as

$\hat{\sigma}^2_{v(FHI)}$. Using s_i^2 and $\hat{\sigma}^2_{v(FHI)}$, we can obtain $\hat{\beta} = \tilde{\beta}(s_i^2 + \hat{\sigma}^2_{v(FHI)})$. The standardized residuals z_i can be calculated as $\hat{z}_i = (y_i - x_i'\hat{\beta})/\sqrt{s_i^2 + \hat{\sigma}^2_{v(FHI)}}$. Figure 2 gives the normal probability plot of estimated residuals $\hat{z}_i$'s. The plot indicates that the normal model assumption on the sampling and linking models is a good approximation and reasonably valid.

Figure 2: QQ Plot of Standardized Residuals



We can also use the posterior estimates of β, σ_i^2 and σ_v^2 to compute the standardized residuals. That is, let $\hat{z}_{i(HB)} = (y_i - x_i'\hat{\beta}_{(HB)})/\sqrt{\hat{\sigma}^2_{i(HB)} + \hat{\sigma}^2_{v(HB)}}$. The normal probability plot of $\hat{z}_{i(HB)}$'s is similar to the plot of $\hat{z}_i$'s. An alternative is to conceptualize a series of normal probability plots, one for each posterior simulation of β, σ_i^2 and σ_v^2 from the Gibbs sampler, for more comprehensive analysis of the plots based on HB estimates.

3.4 Posterior predictive model checking

To check the overall fit of the proposed model, we use the method of posterior predictive distribution. Let y_{rep} denote the replicated observation under the model. The posterior predictive distribution of y_{rep} given the observed data y_{obs} is defined as

$$f(y_{rep} \mid y_{obs}) = \int f(y_{rep} \mid \theta) f(\theta \mid y_{obs}) d\theta .$$

In this approach, a discrepancy measure $D(y,\theta)$ that depends on the data y and the parameter θ can be defined and the observed value $D(y_{obs},\theta \mid y_{obs})$ compared to the posterior predictive distribution of $D(y_{rep},\theta \mid y_{obs})$ where any significant difference indicates a model failure. Gelman, Meng and Stern (1996) proposed the posterior predictive p-value as

$$p = P(D(y_{rep},\theta) \geq D(y_{obs},\theta) \mid y_{obs}) .$$

This is a natural extension of the usual p-value in a Bayesian context. 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 data, then $D(y_{obs},\theta \mid y_{obs})$ should be near the central part of the histogram of the $D(y_{rep},\theta \mid y_{obs})$ values if y_{rep} 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.

The posterior predictive p-value can be estimated as follows: Let θ^* represent a draw from the posterior distribution $f(\theta \mid y_{obs})$, and let y^*_{rep} represent a draw from $f(y_{rep} \mid \theta^*)$. Then marginally y^*_{rep} is a sample from the posterior predictive distribution $f(y_{rep} \mid y_{obs})$.Computing the p-value is relatively easy using the simulated values of θ^* from the Gibbs sampler. For each simulated value θ^*, we can simulate y^*_{rep} from the model and compute $D(y^*_{rep},\theta^*)$ and $D(y_{obs},\theta^*)$. Then the p-value is estimated by the proportion of times $D(y^*_{rep},\theta^*)$ exceeds $D(y_{obs},\theta^*)$.

For the proposed HB model, the discrepancy measure used for overall fit is given by $D(y,\theta) = \sum_{i=1}^{m} (y_i - \theta_i)^2 / \sigma_i^2$. We computed the p-value by combining the simulated θ^* and y^* from all 10 parallel runs. We obtained an estimated p-value equal to 0.466. Thus we have no indication of lack of overall model fit.

The posterior predictive p-value model checking has been criticized for being conservative due to the double use of the observed data y_{obs}. The double use of the data can induce unnatural behaviour, as demonstrated by Bayarri and Berger (2000). They proposed alternative model checking p-value

measures, named the partial posterior predictive p-value and the conditional predictive p-value. However, their methods are more difficult to implement and interpret (Rao, 2003; Sinharay and Stern, 2003). As noted in Sinharay and Stern (2003), the posterior predictive p-value is especially useful if we think of the current model as a plausible ending point with modifications to be made only if substantial lack of fit is found.

4. Concluding Remarks and Future Work

In this paper, we have presented model-based estimates for census undercoverage in small domains across Canada using both the EBLUP and HB approaches. Both the EBLUP and HB estimates improve the direct undercoverage estimates significantly and have achieved large efficiency gains, especially for domains with smaller sample sizes. The proposed model takes into account the uncertainty of estimating the sampling variances by modeling the sampling variances directly as in Wang and Fuller (2003) for EBLUP approach and You and Chapman (2006) for HB approach. The model used with estimated sampling variances is an extension of the well-known Fay-Herriot model. The advantage of the proposed model is that it does not require the smoothing of the direct estimates of the sampling variances. In our study, both the EBLUP and HB approaches work very well in terms of efficiency gains over the direct estimates. However, the EBLUP approach may lead to some underestimation of the MSE when area specific sample sizes are very small, as shown in You and Chapman (2003) using aggregated area level data obtained from the unit level crop data of Battese, Harter and Fuller (1988), where the sample sizes are in the range of 3-5 for each county.

We also discussed benchmarking the HB estimates to obtain the BHB estimates so that these BHB estimates add up to the total of the direct estimates for the whole nation (large areas). The property of benchmarking is important in practice. First, it can provide consistency of the benchmarked model-based estimates in the sense that they add up to the sum of direct survey estimates. Second, it can provide some protection against possible model failure. We studied the method of BHB estimation proposed by You, Rao and Dick (2004). It has the advantage of easy implementation for practical applications. For the EBLUP approach, it is more difficult to obtain the benchmarked estimates, particularly to obtain the MSE estimation of the benchmarked EBLUP estimators. Some resampling method such as bootstrap method may be needed (Pfeffermann, 2006). For the EBLUP approach,

Pfeffermann and Barnard (1991), Pfeffermann and Tiller (2005), and Wang, Fuller and Qu (2006) proposed benchmarking method for the EBLUP estimators. It will be interesting to explore the EBLUP benchmarking method for the census undercoverage estimation and compare it with the BHB method of You, Rao and Dick (2004).

In this paper we consider both the EBLUP and HB approaches for small area census undercoverage estimation. The use of the HB approach enables simple computation of the posterior variance of the small area predictors, without the need to rely on large sample approximation for MSE estimation, especially when the sampling variances are estimated directly and the sample sizes and number of small areas are small (You and Chapman, 2003). It is also much easier to compute the benchmarked HB estimates and obtain the corresponding posterior MSE than the EBLUP approach. Of course, application of the HB approach requires specification of prior distributions for the unknown model hyper-parameters. In most cases, vague proper priors or noninformative priors are used in the HB models; and different priors, particularly for the variance components, may lead to different HB estimators in some cases, especially when the number of small areas is small. In practice, it is better to evaluate the sensitivity of the HB estimates to the choice of priors for publication of the HB estimates (You and Chapman, 2006).

We evaluated the EBLUP and HB approaches using the 2001 census data. We will use the proposed model and both the EBLUP and HB approaches to produce the model-based census undercoverage estimates using the 2006 census data for age-sex domains across Canada once the 2006 undercoverage data becomes available.

REFERENCES

Battese, G.E., Harter, R.M. and Fuller, W.A. (1988) An error components model for prediction of county crop area using survey and satellite data. *Journal of American Statistical Association*, 83, 28-36.

Bayarri, M.J. and Berger, J.O. (2000) P values for composite null models. *Journal of the American Statistical Association,* 95, 1127-1142.

Bell, W. and Otto, M. (1995) Sampling error modeling of poverty and income statistics for States. Proceedings of the Section on Survey Research Methods, Washington D.C., American Statistical Association.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001) Evaluation of small area estimation methods – An application to unemployment estimates from the UK LFS. *Proceedings of Statistics Canada Symposium 2001 Achieving Data Quality in a Statistical Agency: A Methodological Perspective*, CD-ROM.

Datta, G.S., Rao, J.N.K. and Smith, D.D. (2005) On measuring the variability of small area estimators under a basic area level model. *Biometrika*, 92, 183-196.

Dick, P. (1995) Modeling net undercoverage in the 1991 Canadian census. Survey Methodology, 21, 45-54.

Dick, P. and You, Y. (2003) Methods used for small domain estimation of census net undercoverage in the 2001 Canadian census. *Proceedings of 2003 Federal Committee on Statistical Methodology Research Conference*, 43-48: Washington DC.

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.

Gelman, A., Meng, X.-L. and Stern, H. (1996) Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6, 733-807.

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.

Pfeffermann, D. (2006) Mixed model prediction and small area estimation. *Test*, 15, 1-96.

Pfeffermann, D. and Barnard, C.H. (1991) Some new estimators for small area means with application to the assessment of farmland values. *Journal of Business and Economic Statisitcs*, 9, 73-84.

Pfeffermann, D. and Tiller, R. (2005) Small area estimation with state-space models subject to benchmark constraints. Working Paper Series M05/14, Southampton Statistical Sciences Research Institute, University of Southampton, U.K.

Prasda, N.G.N. and Rao, J.N.K. (1990) The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association,* 85, 163-171.

Rao, J.N.K. (2003) *Small Area Estimation.* New York: Wiley.

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.

Sinharay, S. and Stern, H.S. (2003) Posterior predictive model checking in hierarchical models. *Journal of Statistical Planning and Inference*, 111, 209-221.

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.

Wang, J., Fuller, W.A., and Qu, Y. (2006) Small area estimation under a restriction. Unpublished manuscript.

You, Y. (2006) Model-based small area estimation in the Canadian Labour Force Survey. Methodology Branch Working Paper, HSMD-2006-004E, Statistics Canada.

You, Y. and Chapman, B. (2003) Small area estimation using area level models and estimated sampling variances. Methodology Branch working paper, HSMD-2003-005E, Statistics Canada.

You, Y. and Chapman, B. (2006) Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.

You, Y. and Dick, P. (2004) Hierarchical Bayes small area inference to the 2001 census undercoverage estimation. *2004 Proceedings of the American Statistical Association, Section on Government Statistics [CD-ROM]*, 1836-1840, Alexandria, VA: American Statistical Association.

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. (2004) Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.