Component of Statistics Canada
Catalogue no. 12-001-X Business Survey Methods Division

Article

An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada

by Yong You

June 2008

Statistics Canada Statistique Canada

Canada

Survey Methodology, June 2008
Vol. 34, No. 1, pp. 19-27
Statistics Canada, Catalogue No. 12-001-X

An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada

Yong You [1]

Abstract

The Canadian Labour Force Survey (LFS) produces monthly estimates of the unemployment rate at national and provincial levels. The LFS also releases unemployment estimates for sub-provincial areas such as Census Metropolitan Areas (CMAs) and Urban Centers (UCs). However, for some sub-provincial areas, the direct estimates are not reliable since the sample size in some areas is quite small. The small area estimation in LFS concerns estimation of unemployment rates for local sub-provincial areas such as CMA/UCs using small area models. In this paper, we will discuss various models including the Fay-Herriot model and cross-sectional and time series models. In particular, an integrated non-linear mixed effects model will be proposed under the hierarchical Bayes (HB) framework for the LFS unemployment rate estimation. Monthly Employment Insurance (EI) beneficiary data at the CMA/UC level are used as auxiliary covariates in the model. A HB approach with the Gibbs sampling method is used to obtain the estimates of posterior means and posterior variances of the CMA/UC level unemployment rates. The proposed HB model leads to reliable model-based estimates in terms of CV reduction. Model fit analysis and comparison of the model-based estimates with the direct estimates are presented in the paper.

Key Words: Design effect; Hierarchical Bayes; Log-linear mixed effects model; Model checking; Sampling variance; Small area.

1. Introduction

The unemployment rate is generally viewed as a key indicator of economic performance. In Canada, the unemployment rate estimates are produced monthly by the Labour Force Survey (LFS) of Statistics Canada. The LFS is a monthly survey of 53,000 households selected using a stratified, multistage design. Each month, one-sixth of the sample is replaced. Thus five-sixths of the sample is common between two consecutive months. This sample overlap induces correlations which can be exploited to produce better estimates by alternative methods such as model-based methods to borrow strength over time; more details will be discussed in Section 2. For a detailed description of the LFS design, see Gambino, Singh, Dufour, Kennedy and Lindeyer (1998). The LFS releases monthly unemployment rate estimates for large areas such as the nation and provinces as well as local areas (small areas) such as Census Metropolitan Areas (CMAs, *i.e.*, cities with population more than 100,000) and other Urban Centres (UCs) across Canada. Although national and provincial estimates get the most media attention, sub-provincial estimates of the unemployment rate are also very important. They are used by the Employment Insurance (EI) program to determine the rules used to administer the program. In addition, the unemployment rates for CMAs and UCs receive close scrutiny at local levels. However, many local areas do not have large enough samples to produce adequate direct estimates, since the LFS is designed to produce adequate or reliable estimates at the national level and provincial level. The estimated coefficient of variation (CV) level for the nation is about 2% and 4% to 7% for provinces. However, the CVs for CMAs and UCs range from about 7% to 50%. Some UCs have CVs even larger than 50%. The direct LFS estimates for some local areas are not reliable with very large CVs due to the small sample sizes for those areas. Therefore, alternative estimators, in particular, model-based estimators, are considered to improve the direct LFS estimates for small areas. The objective in this paper is to obtain a reliable model-based estimator that is an improvement over the direct LFS estimator in terms of small and stable CVs.

In general, direct survey estimators, based only on the domain-specific sample data, are typically used to estimate parameters for large domains such as the nation and provinces. 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 census counts and 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; see Rao (2003). The model-based estimators are indirect estimators in the sense that these estimators are obtained by using small

1. Yong You, Household Survey Methods Division, Statistics Canada, Ottawa, Ontario, K1A 0T6. E-mail: yong.you@statcan.ca.

area models, direct estimates and related auxiliary variables. The model-based estimators are obtained to improve the direct design-based estimators in terms of precision and reliability, that is, smaller CVs. Traditional small area estimators borrow strength either from similar small areas or from the same area across time, but not both. In recent years, several approaches to borrowing strength simultaneously across both space and time have been developed. Estimators based on the approach developed by Rao and Yu (1994), Ghosh, Nangia and Kim (1996), Datta, Lahiri, Maiti and Lu (1999), Datta, Lahiri and Maiti (2002) and You, Rao and Gambino (2000, 2003), successfully exploit the two dimensions simultaneously to produce improved estimates with desirable properties for small areas. In particular, You *et al.* (2000, 2003) studied the model-based estimation of unemployment rates for local sub-provincial areas such as CMAs and Census Agglomerations (CAs) across Canada. They obtained efficient model-based estimators and adequate model fit for the LFS unemployment rate estimation. However, the model proposed by You *et al.* (2000, 2003) has some limitations. In this paper, we discuss these limitations and propose a new integrated model for the LFS unemployment rate estimation under hierarchical Bayes (HB) framework. The idea is to model the parameters of interest and the sampling variances together as suggested in You *et al.* (2003) and You and Chapman (2006). We will apply the proposed model to the 2005 LFS data and obtain the model-based unemployment rate estimates. Comparison of the HB estimates with the direct LFS estimates and model fit analysis will also be provided.

This paper is organized as follows. In Section 2, we present and discuss various small area models proposed in the literature for the unemployment rate estimation. In Section 3, we discuss the problem of smoothing and modeling the sampling covariance matrix. In Section 4, an integrated non-linear mixed effects model is proposed in a hierarchical Bayes framework, and the use of Gibbs sampling to generate samples from the joint posterior distribution is described. In Sections 5, we apply the proposed model to LFS data and obtain the HB estimates for small area unemployment rates. Model analysis and evaluation are also provided. And finally in Section 6 we offer some concluding remarks and future work directions.

2. Small area models

2.1 Cross-sectional model

Cross-sectional or area level models are used to produce reliable model-based estimates by combining area level auxiliary information and direct area level estimates. A basic area level model is the well-known Fay-Herriot model (Fay and Herriot 1979). This model has two components: (1) a sampling model for the direct survey estimates, and (2) a linking model that relates the small area parameters to area level auxiliary variables through a linear regression model. For the LFS monthly unemployment rate estimation, let θ_{it} denote the true unemployment rate for the i^{th} CMA/UC at a particular time (month) t, where $i = 1, ..., m$, where m is the number of CMA/UCs, and let y_{it} denote the direct LFS estimate of θ_{it}. Then the sampling model for y_{it} can be expressed as

$$y_{it} = \theta_{it} + e_{it}, \quad i = 1, ..., m, \tag{1}$$

where e_{it} is the sampling error associated with the direct estimator y_{it}. The sampling error is assumed to be normally distributed as $e_{it} \sim N(0, \sigma_{it}^2)$ where σ_{it}^2 is the sampling variance. The linking model for the true unemployment rate θ_{it} may be written as

$$\theta_{it} = x_{it}'\beta + v_i, \quad i = 1, ..., m, \tag{2}$$

where x_{it} is the auxiliary variable and v_i is area-specific random effect. For each time point (each month), we can use the Fay-Herriot model for the monthly direct estimates. The Fay-Herriot model combines cross-sectional information but does not borrow strength over the past time periods.

2.2 Cross-sectional and time series model

Because of the LFS sample design and rotation pattern, there is substantial sample overlap over six month time periods within each area. As a result, for a particular area i, the correlation between the sampling errors e_{it} and e_{is} $(t \neq s)$ need to be taken into account. You *et al.* (2000, 2003) proposed a cross-sectional and time series model for the LFS unemployment rate estimates. You *et al.* (2000, 2003) only used previous six months of data to predict the current month rate since the LFS sample rotation is based on a six month cycle. Each month, one sixth of the LFS sample is replaced. Thus after six months, the correlation between estimates is weak (see Section 2.1 for the lag correlation coefficients). Let $y_i = (y_{i1}, ..., y_{iT})'$, $\theta_i = (\theta_{i1}, ..., \theta_{iT})'$, and $e_i = (e_{i1}, ..., e_{iT})'$, where $T = 6$ here. By assuming that e_i follows a multivariate normal distribution with mean vector 0 and sampling covariance matrix $\sum_i$, we have

$$y_i \sim N\left(\theta_i, \sum\nolimits_i\right), \quad i = 1, ..., m.$$

Thus y_i is assumed to be design-unbiased for θ_i. The sampling covariance matrix $\sum_i$ is unknown in the model. Direct estimates of the sampling covariance matrices are available. It is customary to assume a known sampling variance in area level model-based small area estimation (Rao 2003). For example, the traditional Fay-Herriot model assumes the sampling variance known in the model. Usually

a smoothed estimator of the sampling variance is used. However, recent development on modeling the sampling variance provides an alternative approach to handle the problem of sampling variance; for example, see Wang and Fuller (2003), You and Dick (2004) and You and Chapman (2006). For the unemployment rate estimation, details on smoothing and modeling the sampling variances are given in Section 3.

To borrow strength across regions and time periods, and following You *et al.* (2000, 2003) we can model the true unemployment rate θ_{it} by a linear regression model with random effects through auxiliary variables x_{it}, that is,

$$\theta_{it} = x_{it}'\beta + v_i + u_{it}, \quad i = 1, \ldots, m, \; t = 1, \ldots, T, \qquad (3)$$

where v_i is a area random effect assumed to be $N(0, \sigma_v^2)$ and u_{it} is a random time and area component. We can further assume that u_{it} follows a random walk process over time period $t = 1, \ldots, T$, that is,

$$u_{it} = u_{i,t-1} + \varepsilon_{it}, \qquad (4)$$

where $\varepsilon_{it} \sim N(0, \sigma_\varepsilon^2)$. Then $\text{cov}(u_{it}, u_{is}) = \min(t, s)\, \sigma_\varepsilon^2$. The regression vector β and the variance components σ_v^2 and σ_ε^2 are unknown in the model and need to be estimated. Combining the model (1), (3) and (4), we obtain a linear mixed model with time components as

$$y_{it} = x_{it}'\beta + v_i + u_{it} + e_{it}, \quad i = 1, \ldots, m, \; t = 1, \ldots, T. \qquad (5)$$

You *et al.* (2003) showed that the cross-sectional and time series model (5) is better than the Fay-Herriot model in terms of smoothing the direct estimates and CV reduction over the direct estimates for the LFS unemployment rate estimation.

We have used a random walk model for u_{it}. Rao and Yu (1994) used a stationary autoregressive model for u_{it}. You *et al.* (2003) showed that the random walk model on u_{it} had provided better model fit to the unemployment rate estimation than the autoregressive AR(1) model. Datta *et al.* (1999) also used a random walk model to estimate the US unemployment rates at the state level.

2.3 Log-linear linking model

However, a limitation of the model (3) is that the linking model for the parameter of interest, the true unemployment rate θ_{it}, is a linear model with normal random effects. Since θ_{it} is a positive number between 0 and 1, and it is close to 0, the linear linking model with normal random effects may lead to negative estimates for θ_{it} for some small areas. To avoid this problem, You, Chen and Gambino (2002) proposed a log-linear linking model for θ_{it} as follows:

$$\log(\theta_{it}) = x_{it}'\beta + v_i + u_{it}, \quad i = 1, \ldots, m, t = 1, \ldots, T. \qquad (6)$$

You and Rao (2002) also studied the log-linear linking model for the Fay-Herriot model as the unmatched sampling and linking models with application in the Canadian census undercoverage estimation. The results of You and Rao (2002) and You *et al.* (2002) have shown that the log-linear linking model performs very well in the small area estimation problems. In this paper, we therefore will use the log-linear linking model (6) for the true unemployment rate θ_{it}.

3. Sampling variance

In general, we can obtain direct sampling variance estimates from survey data. However, these direct estimates are unstable if sample sizes are small. In area level models of small area estimation, the sampling variances are usually assumed to be known (*e.g.*, Fay and Herriot 1979; Datta *et al.* 1999; You and Rao 2002). If the sampling variances are assumed to be known in the model, then reliable (smoothed) estimates of sampling variances are constructed using other auxiliary data and models usually through generalized variance functions (*e.g.*, Dick 1995; Datta *et al.* 1999). In this paper alternatively, we model sampling variance covariance matrix using the direct estimates in a specific way such that we do not need to assume the sampling variances and covariances are known in the model. Thus we simplify the problem of smoothing unknown sampling variance and integrate the sampling variance modeling part into the whole model.

3.1 Smoothing sampling covariance matrix

You *et al.* (2000, 2003) used two steps to smooth the sampling covariance matrix. The first step is to obtain a smoothed or common CV for each CMA/UC by computing the average CVs for each CMA/UC over a certain time period, denoted as $\overline{\text{CV}}_i$, where $i = 1, 2, \ldots, m$. The second step is to obtain the average lag correlation coefficients over time and all CMA/UCs, denoted as $\bar{\rho}_{|t-s|}$ for the time lag $|t - s|$. This step involves intensive computation. We have used three years (1999 to 2001) of LFS data to compute the smoothed correlation coefficients. We treat the smoothed values over both time and space as the true values in the model. The one-month lag (lag-1) correlation coefficient is obtained as $\bar{\rho}_1 = 0.48$, lag-2 correlation coefficient is $\bar{\rho}_2 = 0.31$, lag-3 is $\bar{\rho}_3 = 0.21$, lag-4 is $\bar{\rho}_4 = 0.16$, lag-5 is $\bar{\rho}_5 = 0.11$ and $\bar{\rho}_6 = 0.1$. After lag 6, the lag correlation coefficient is less than 0.1. The lag correlation coefficients decrease as the lag increases. This is consistent with the rotation pattern of the LFS design. Figure 1 shows the smoothed lag correlation coefficients for the LFS unemployment rate estimates.

By using these smoothed CVs and lag correlation coefficients, a smoothed covariance matrix $\hat{\Sigma}_i$ can be obtained with diagonal elements $\hat{\sigma}_{it}^2 = (\overline{\mathrm{CV}}_i)^2 y_{it}^{2\,i}$ and off-diagonal elements $\hat{\sigma}_{its} = \bar{\rho}_{|t-s|}\hat{\sigma}_{it}\hat{\sigma}_{is}$. The smoothed $\hat{\Sigma}_i$ is then treated as known in the model. The study of You *et al.* (2000, 2003) suggests that using the smoothed $\hat{\Sigma}_i$ in the model can significantly improve the estimates in terms of CV reduction compared to the HB estimates obtained using the direct survey estimates of Σ_i in the model. For more details of the result, see You *et al.* (2003).

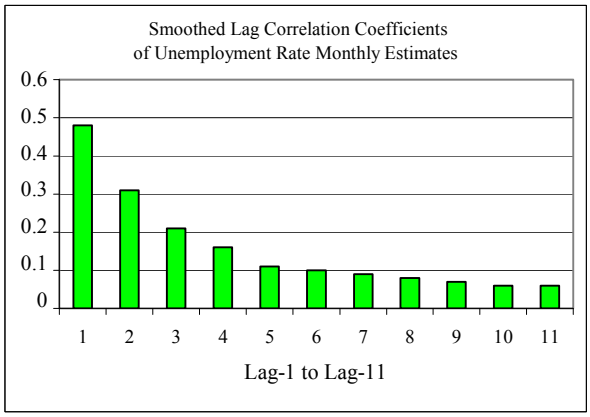

Figure 1 Smoothed unemployment rate lag correlation coefficients

3.2 Equal CV modeling approach

The main problem of the method of You *et al.* (2000, 2003) is that the smoothed sampling covariance matrices depend on the direct survey estimates y_{it}, whereas the y_{it}'s are not reliable for some small regions. Note that the true sampling variance can be written as $\sigma_{it}^2 = \theta_{it}^2(\mathrm{CV}_{it})^2$. Based on the assumption of common CV over time for a given area, You *et al.* (2003) suggested in their concluding remarks to use estimates of the form $\tilde{\sigma}_{it}^2 = \theta_{it}^2(\overline{\mathrm{CV}}_i)^2$ and $\tilde{\sigma}_{its} = \bar{\rho}_{|t-s|}(\tilde{\sigma}_{it}\tilde{\sigma}_{is})$ for the smoothed variances and covariances respectively. Then the new smoothed sampling covariance matrix $\tilde{\Sigma}_i$ has diagonal elements $\tilde{\sigma}_{it}^2$ and off-diagonal elements $\tilde{\sigma}_{its}$. However, under this method, the sampling covariance matrix $\tilde{\Sigma}_i$ becomes unknown in the model, since $\tilde{\sigma}_{it}^2$ and $\tilde{\sigma}_{its}$ depend on the unknown parameters θ_{it}, whereas θ_{it} is related to a linking model. The advantage of this method is that the model structure of the sampling covariance matrix is clearly specified. This method is better than the smoothing method in the sense that the sampling covariance is clearly specified and not treated as known.

3.3 Equal design effects modeling approach

An alternative modeling approach is based on the assumption of common design effects as suggested in Singh, You and Mantel (2005) and Singh, Folsom and Vaish (2005) to smooth the sampling variance σ_{it}^2. The design effect (deff) for the ith area at time t may be approximately written as

$$\mathrm{deff}_{it} = \frac{\sigma_{it}^2}{\theta_{it}(1-\theta_{it})/n_{it}},$$

where n_{it} is the corresponding sample size. Then the sampling variance σ_{it}^2 can be written as $\sigma_{it}^2 = \theta_{it}(1-\theta_{it})\cdot\mathrm{deff}_{it}/n_{it}$. Let $\tau_{it} = \mathrm{deff}_{it}/n_{it} = \sigma_{it}^2/(\theta_{it}(1-\theta_{it}))$. Then we can estimate τ_{it} using the direct estimates of θ_{it} and σ_{it}^2 as $\hat{\tau}_{it} = \hat{\sigma}_{it}^2/(y_{it}(1-y_{it}))$. For each area, based on the assumption of a common deff and a common sample size over time, we can obtain a smoothed average factor $\bar{\tau}_i$ as $\bar{\tau}_i = \sum_{t=1}^{T}\hat{\tau}_{it}/T$. Then a smoothed sampling variance can be obtained as $\tilde{\sigma}_{it}^2 = \theta_{it}(1-\theta_{it})\cdot\bar{\tau}_i$, which again depends on θ_{it} as well. The sampling covariance is still in the form of $\tilde{\sigma}_{its} = \bar{\rho}_{|t-s|}(\tilde{\sigma}_{it}\tilde{\sigma}_{is})$, as in You *et al.* (2003). Note that $\bar{\tau}_i$ is a moving average of $\hat{\tau}_{it}$ over the time period T in the model. In practice, however, alternatively one may use a longer time series data to obtain more stable estimate of $\bar{\tau}_i$ for each area if necessary. In this paper, we will use the common design effects model for unemployment rate estimation based on the smoothed moving average factor $\bar{\tau}_i$ as we borrow information from the past time period T.

4. Hierarchical Bayes inference

In this section, we propose an integrated cross-sectional and time series log-linear model for the unemployment rate estimation. We apply the hierarchical Bayes approach to the model. Estimates of posterior means and posterior variances are obtained by using the Gibbs sampling method.

4.1 Integrated hierarchical Bayes model

We now propose the integrated cross-sectional and time series log-linear model in a hierarchical Bayes framework as follows:

- Conditional on $\theta_i = (\theta_{i1}, \ldots, \theta_{iT})'$, $[y_i \mid \theta_i] \sim \text{ind } N(\theta_i, \sum_i(\theta_i))$;
- Conditional on β, u_{it} and σ_v^2, $[\log(\theta_{it}) \mid \beta, u_{it}, \sigma_v^2] \sim \text{ind } N(x_{it}'\beta + u_{it}, \sigma_v^2)$;
- Conditional on $u_{i,t-1}$ and σ_ε^2, $[u_{it} \mid u_{i,t-1}, \sigma_\varepsilon^2] \sim \text{ind } N(u_{i,t-1}, \sigma_\varepsilon^2)$;
- $\sum_i(\theta_i)$ depends on θ_i with diagonal elements $\tilde{\sigma}_{it}^2 = \theta_{it}(1-\theta_{it})\cdot\bar{\tau}_i$ and off-diagonal elements $\tilde{\sigma}_{its} = \bar{\rho}_{|t-s|}(\tilde{\sigma}_{it}\tilde{\sigma}_{is})$.
- Marginally β, σ_v^2 and σ_ε^2 are mutually independent with priors given as $\beta \propto 1$, $\sigma_v^2 \sim \mathrm{IG}(a_1, b_1)$, and $\sigma_\varepsilon^2 \sim \mathrm{IG}(a_2, b_2)$, where IG denotes an inverse gamma

distribution and a_1, b_1, a_2, b_2 are known positive constants and usually set to be very small to reflect our vague knowledge about σ_v^2 and σ_ε^2.

Remarks:

1. The proposed HB model has used a log-linear linking model for the small area parameter of interest θ_{it} as suggested in You *et al.* (2002) and You and Rao (2002).
2. The sampling covariance matrix $\sum_i$ is unknown in the model, and it is specified as a function of unknown small area parameter θ_i as suggested in You and Rao (2002) and You *et al.* (2003).
3. We have used the assumption of common design effects for small areas as suggested in Singh, You and Mantel (2005).
4. The proposed HB model overcomes the limitations of the model of You *et al.* (2000, 2003) in terms of log-linear modeling and specification of unknown sampling covariance matrix modeling. In particular, we model the unknown sampling covariance matrix through the small area parameters θ_i using smoothed estimates of design effects for each areas.

We are interested in estimating the true unemployment rate θ_{it}, and in particular, the current unemployment rate θ_{iT}. In the HB analysis, θ_{iT} is estimated by its posterior mean $E(\theta_{iT} \mid y)$ and the uncertainty associated with the estimator is measured by the posterior variance $V(\theta_{iT} \mid y)$. We use the Gibbs sampling method (Gelfand and Smith 1990; Gelman and Rubin 1992) to obtain the posterior mean and the posterior variance of θ_{iT}.

4.2 Gibbs sampling inference

The Gibbs sampling method is an iterative Markov chain Monte Carlo sampling method to simulate samples from a joint distribution of random variables by sampling from low dimensional densities to make inference about the joint and marginal distributions (Gelfand and Smith 1990). The most prominent application is for inference within a Bayesian framework. In Bayesian inference one is interested in the posterior distribution of the parameters. Assume that $y_i \mid \theta$ has conditional density $f(y_i \mid \theta)$ for $i = 1, ..., n$ and that the prior information about $\theta = (\theta_1, ..., \theta_k)'$ is summarized by a prior density $\pi(\theta)$. Let $\pi(\theta \mid y)$ denote the posterior density of θ given the data $y = (y_1, ..., y_n)'$. It may be difficult to sample from $\pi(\theta \mid y)$ directly in practice due to the high dimensional integration with respect to θ. However, one can use the Gibbs sampler to construct a Markov chain $\{\theta^{(g)} = (\theta_1^{(g)}, ..., \theta_k^{(g)})'\}$ with $\pi(\theta \mid y)$ as the limiting distribution. For illustration, let $\theta = (\theta_1, \theta_2)'$. Starting with an initial set of values $\theta^{(0)} = (\theta_1^{(0)}, \theta_2^{(0)})'$, we generate $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$ by sampling $\theta_1^{(g)}$ from $\pi(\theta_1 \mid \theta_2^{(g-1)}, y)$ and $\theta_2^{(g)}$ from $\pi(\theta_2 \mid \theta_1^{(g-1)}, y)$. Under certain regularity conditions, $\theta^{(g)} = (\theta_1^{(g)}, \theta_2^{(g)})'$ converges in distribution to $\pi(\theta \mid y)$ as $g \to \infty$. Marginal inference about $\pi(\theta_i \mid y)$ can be based on the marginal samples $\{\theta_i^{(g+k)}; k = 1, 2, ...\}$ for large g.

For the proposed integrated HB model, to obtain the posterior estimation of unemployment rate, we implement the Gibbs sampling method by generating samples from the full conditional distributions of the parameters β, σ_v^2 and σ_ε^2, u_{it} and θ_i. These full conditional distributions are given in the Appendix. The distributions of β, σ_v^2 and σ_ε^2, u_{it} are standard normal or inverse gamma distributions that can be easily sampled. However, the conditional distribution of θ_i does not have a closed form. We use the Metropolis-Hastings algorithm within the Gibbs sampler (Chib and Greenberg 1995) to update θ_i. Following You *et al.* (2002) and You and Rao (2002), the full conditional distribution of θ_i in the Gibbs sampler can be written as

$$\theta_i \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u \propto h(\theta_i) f(\theta_i),$$

where

$$h(\theta_i) = \left|\sum_i(\theta_i)\right|^{-1} \exp\{-\frac{1}{2}(y_i - \theta_i)'\sum_i^{-1}(y_i - \theta_i)\}$$

and

$$f(\theta_i) = \exp\left\{-\frac{1}{2\sigma_v^2}(\log(\theta_i) - \mathrm{x}_i'\beta - u_i)'(\log(\theta_i) - \mathrm{x}_i'\beta - u_i)\right\} \cdot \left(\prod_{t=1}^{T}\frac{1}{\theta_{it}}\right).$$

To update θ_i, we proceed as follows:

1. For $t = 1, ..., T$, draw $\theta_{it}^{(k+1)} \sim \log N(x_{it}'\beta^{(k+1)} + u_{it}^{(k+1)}, \sigma_v^{2(k+1)})$, then we have $\theta_i^{(k+1)} = (\theta_{i1}^{(k+1)}, ..., \theta_{iT}^{(k+1)})'$.
2. Compute the rejection probability

$$\alpha(\theta_i^{(k)}, \theta_i^{(k+1)}) = \min\{\frac{h(\theta_i^{(k+1)})}{h(\theta_i^{(k)})}, 1\}.$$

3. Generate $\lambda \sim \text{Uniform}(0,1)$, if $\lambda < \alpha(\theta_i^{(k)}, \theta_i^{(k+1)})$, then accept $\theta_i^{(k+1)}$; otherwise reject $\theta_i^{(k+1)}$ and set $\theta_i^{(k+1)} = \theta_i^{(k)}$.

To implement Gibbs sampling, we follow the recommendation of Gelman and Rubin (1992) and independently run $L (L > 2)$ parallel chains, each of length $2d$. The first d iterations of each chain are deleted. The convergence monitoring is based on the potential scale reduction factor as suggested in Gelman and Rubin (1992) and adopted by You *et al.* (2003) for estimating θ_{iT}. Details are given in You *et al.* (2003). Estimates of the posterior mean $E(\theta_{iT} \mid y)$ and the posterior variance $V(\theta_{iT} \mid y)$ are obtained based on the samples generated from the Gibbs sampler.

5. Application to LFS data

5.1 Estimation

We use the 2005 January to June LFS unemployment rate estimates, y_{it}, in our data analysis. In addition to the direct estimates y_{it} and the sampling covariance matrices used in the small area models, auxiliary administrative variables are needed in the models. For the unemployment rate estimation, local area employment insurance (EI) monthly beneficiary rate is used as auxiliary data x_{it} in the model. The beneficiary rate is calculated as the ratio of the number of persons applying EI benefit over the number of persons in the labour force. There are 72 CMA/UCs across Canada. One UC (Miramichi) does not have the EI data. So we consider $m = 71$ CMA/UCs in the model. Within each area, we consider six consecutive monthly estimates y_{it} from January 2005 to June 2005, so that $T = 6$. For the January to June 2005 data, the overall average (over 71 CMA/UCs and 6 months) unemployment rate is 0.076, and the overall average EI beneficiary rate is 0.059. For the proposed small area model, the parameter of interest θ_{iT} is the true unemployment rate for area i in June 2005, where $i = 1, \ldots, 71$. To implement the Gibbs sampler, we have used 10 parallel runs, each of length 2000. The first 1,000 iterations are deleted as "burn-in" periods. The hyper-parameters for variance components in the model are set to be 0.0001 to reflect the vague knowledge about σ_v^2 and σ_ε^2.

We now present the posterior estimates of the unemployment rates under the proposed integrated HB model given in section 4.1 using the Gibbs sampling method. Figure 2 displays the LFS direct estimates and the HB model-based estimates of the June 2005 unemployment rates for the 71 CMA/UCs across Canada. The 71 CMA/UCs appear in the order of population size with the smallest UC (Dawson Creek, BC) on the left and the largest CMA (Toronto, ON) on the right. For the point estimates, the HB estimates leads to moderate smoothing of the direct LFS estimates. For the CMAs with large population sizes and therefore large sample sizes, the direct estimates and the HB estimates are very close to each other as expected, particularly for Toronto, Montreal and Vancouver; for smaller UCs, the direct and HB estimates differ substantially for some regions.

Figure 3 displays the CVs of the estimates. The CV of the HB estimate is taken as the ratio of the square root of the posterior variance and the posterior mean. It is clear from Figure 3 that the direct estimates have very large CVs, particularly for the UCs, the CVs are very large and unstable. The HB estimates have very small and stable CVs compared to the direct estimates. The efficiency gain of the HB estimates is obvious, particularly for the UCs with smaller population sizes. More precisely, we computed the percent CV reduction for the HB estimators based on the data of June 2005. The percent CV reduction is computed as the difference of the direct CV and HB CV relative to the direct CV. The average CV reduction for UCs is 63% and the CV reduction for CMAs is 35%. As expected, the proposed model has achieved a large CV reduction over the direct estimates, particularly for smaller UCs with smaller sample sizes.

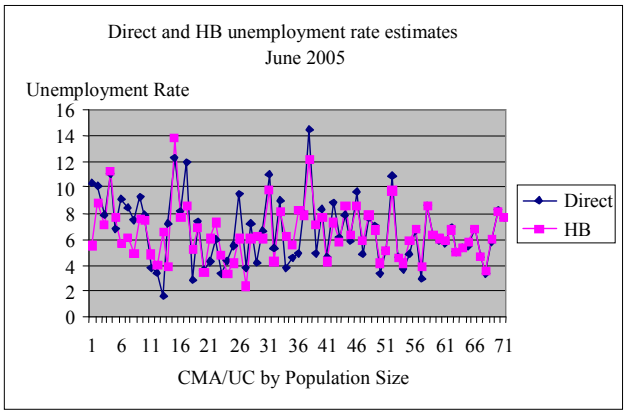

Figure 2 Comparison of direct and HB estimates

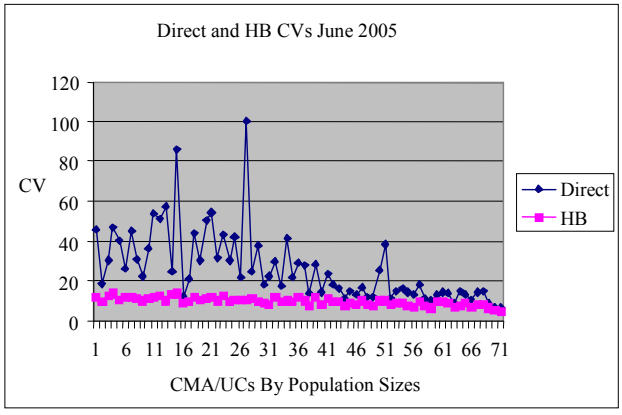

Figure 3 Comparison of direct and HB CVs

5.2 Model fit using posterior predictive distribution

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

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

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 $D(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ compared to the posterior predictive distribution of $D(y_{\text{rep}}, \theta \mid y_{\text{obs}})$ with any significant difference indicates a model failure. Meng (1994) and Gelman, Carlin, Stern and Rubin (1995) proposed the posterior predictive p-value as

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

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 $D(y_{\text{obs}}, \theta \mid y_{\text{obs}})$ should be near the central part of the histogram of the $D(y_{\text{rep}}, \theta \mid y_{\text{obs}})$ 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 $f(\theta \mid y_{\text{obs}})$, and let y^*_{rep} represent a draw from $f(y_{\text{rep}} \mid \theta^*)$. Then marginally y^*_{rep} is a sample from the posterior predictive distribution $f(y_{\text{rep}} \mid y_{\text{obs}})$. 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 $D(y^*_{\text{rep}}, \theta^*)$ and $D(y_{\text{obs}}, \theta^*)$. Then the p-value is estimated by the proportion of times $D(y^*_{\text{rep}}, \theta^*)$ exceeds $D(y_{\text{obs}}, \theta^*)$.

For the proposed HB model, the discrepancy measure used for overall fit is given by $d(y, \theta) = \sum_{i=1}^{m} (y_i - \theta_i)' \sum_i^{-1} (y_i - \theta_i)$. This measure has been used by Datta *et al.* (1999) and You *et al.* (2003). We computed the p-value by combining the simulated θ^* and y^* from all 10 parallel runs. We obtained an estimated average p-value about 0.38. 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. 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 2002; 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.

To compare the proposed model with the model of You *et al.* (2003), we computed the divergence measure of Laud and Ibrahim (1995) based on the posterior predictive distribution. The expected divergence measure of Laud and Ibrahim (1995) is given by $d(y^*, y_{\text{obs}}) = E(k^{-1} \| y^* - y_{\text{obs}} \|^2 \mid y_{\text{obs}})$, where k is the dimension of y_{obs} and y^* is a sample from the posterior predictive distribution $f(y \mid y_{\text{obs}})$. Between two models, we prefer a model that yields a smaller value of this measure. As in Datta, Day and Maiti (1998) and You *et al.* (2003), we approximated the divergence measure $d(y^*, y_{\text{obs}})$ by using the simulated samples from the posterior predictive distribution. Using the Gibbs sampling multiple outputs, we obtained a divergence measure in the range of 8 to 9 for the proposed model, and about 12 to 14 for the model of You *et al.* (2003). Thus the divergence measure suggests a better fit of the proposed integrated HB model for the LFS unemployment rate estimation.

5.3 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 LFS estimates and the HB model-based estimates. The regression method is suggested by Brown, Chamber, Heady and Heasman (2001). If the model-based estimates are close to the true unemployment rates, then the direct LFS estimators should behave like random variables whose expected values correspond to the values of the model-based estimates. We plot the model-based HB estimates as X and the direct LFS estimates as Y, and see 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 June 2005 data, let Y be the direct unemployment rate estimates, and X be the model-based HB estimates. We obtain the estimated α value as 1.0207 with standard error 0.0281. Figure 4 shows a scatter plot with the fitted regression line.

The regression result shows no significant difference from $Y = X$. Therefore, we conclude that the model-based estimates derived from the proposed model are consistent with the direct LFS estimates with no extra possible bias included. The result may also indicate no evidence of any bias due to possible model misspecification.

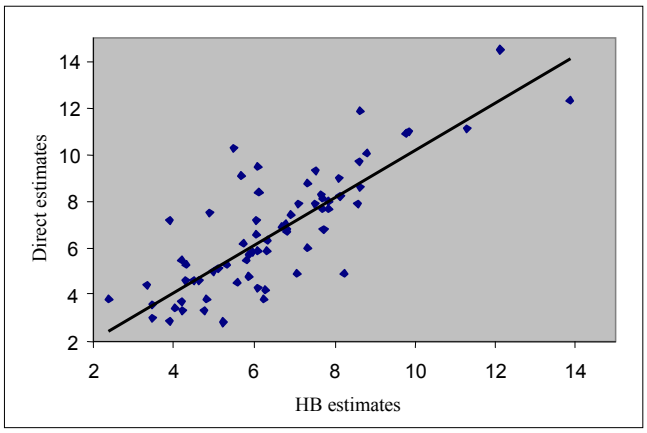

Figure 4 Scatter plot with regression line

6. Concluding remarks and future work

In this paper we have reviewed some small area models including the Fay-Herriot model and the cross-sectional and time series model of You *et al.* (2003). In view of the previous work, we have proposed an integrated non-linear cross-sectional and time series model to obtain model-based estimates of unemployment rates for CMA/UCs across Canada using the LFS data. The proposed model overcomes the limitations of the previous work. In particular, we can model the sampling variance as a function of the small area mean by assuming either a common CV for a given area or a common deff for a given area. Our data analysis has shown that the proposed model fits the data quite well. The hierarchical Bayes estimates, based on the model, improve the direct survey estimates significantly in terms of CV reduction, especially for UCs with small population sizes.

We plan to use alternative modeling approach for the sampling variance. Recently You and Dick (2004) and You and Chapman (2006) has used the HB approach to model the sampling variance directly without specifying the form of the sampling variance under the frame of the Fay-Herriot model. The model automatically takes into account the variability of estimating the sampling variances. In particular, You and Dick (2004) applied the model to the census undercoverage estimation problem and obtained efficient HB census undercoverage estimates for small domains across Canada. It will be interesting to adopt the same idea to the cross-sectional and time series model and compare the results with the current work. The purpose of comparison is to establish a reliable and easy-to-implement model for the LFS model-based unemployment rate estimation for small areas.

We plan to produce the model-based estimates for a relative long time period, for example, 24 months from 2004 to 2005. We will compare the 24 months model-based estimates with the 24 months direct estimates, particularly for the large CMAs to study the smoothing effects of the proposed model. The model-based estimates should follow the pattern of direct LFS estimates for large CMAs, which indicates that the smoothing effects on time series effects are reasonable. The purpose is to verify the robustness of the proposed model-based estimates over time.

Appendix

In the following, we present the full conditional distributions for the Gibbs sampler under the proposed HB model. Let $Y = (Y_1', ..., Y_m')'$, $X = (X_1', ..., X_m')'$, $\theta = (\theta_1', ..., \theta_m')'$, and $u = (u_1', ..., u_m')'$, with $Y_i' = (y_{i1}, ..., y_{iT})$, $X_i' = (x_{i1}, ..., x_{iT})$, $\theta_i' = (\theta_{i1}, ..., \theta_{iT})$, and $u_i' = (u_{i1}, ..., u_{iT})$, we obtain the full conditional distributions as follows:

- $\beta \mid Y, \sigma_v^2, \sigma_\varepsilon^2, u, \theta \sim$
  $N((X'X)^{-1}X'(\log(\theta) - u), \sigma_v^2(X'X)^{-1})$;

- $\sigma_v^2 \mid Y, \beta, \sigma_\varepsilon^2, u, \theta \sim$
  $\mathrm{IG}\left(\left(a_1 + mT/2, b_1 + \sum_{i=1}^{m}\sum_{t=1}^{T}(\log(\theta_{it}) - x_{it}'\beta - u_{it})^2\right)/2\right)$;

- $\sigma_\varepsilon^2 \mid Y, \beta, \sigma_v^2, u, \theta \sim$
  $\mathrm{IG}\left(\left(a_2 + m(T-1)/2, b_2 + \sum_{i=1}^{m}\sum_{t=2}^{T}(u_{it} - u_{i,t-1})^2\right)/2\right)$;

- For $i = 1, ..., m$,
  $u_{i1} \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i2}, \theta \sim$
  $$N\left(\left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\log(\theta_{i1}) - x_{i1}'\beta}{\sigma_v^2} + \frac{u_{i2}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\right);$$

- For $i = 1, ..., m$, and $2 \le t \le T-1$,
  $u_{it} \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,t-1}, u_{i,t+1}, \theta \sim$
  $$N\left(\left(\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\log(\theta_{it}) - x_{it}'\beta}{\sigma_v^2} + \frac{u_{i,t-1} + u_{i,t+1}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{2}{\sigma_\varepsilon^2}\right)^{-1}\right);$$

- For $i = 1, ..., m$,
  $u_{iT} \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u_{i,T-1}, \theta \sim$
  $$N\left(\left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\left(\frac{\log(\theta_{iT}) - x_{iT}'\beta}{\sigma_v^2} + \frac{u_{i,T-1}}{\sigma_\varepsilon^2}\right), \left(\frac{1}{\sigma_v^2} + \frac{1}{\sigma_\varepsilon^2}\right)^{-1}\right);$$

- For $i = 1, ..., m$,
  $\theta_i \mid Y, \beta, \sigma_v^2, \sigma_\varepsilon^2, u \propto$
  $$\left|\sum\nolimits_i\right|^{-1/2}\exp\left\{-\frac{1}{2}(y_i - \theta_i)'\sum\nolimits_i^{-1}(y_i - \theta_i)\right\}$$
  $$\times \exp\left\{-\frac{1}{2\sigma_v^2}\sum_{t=1}^{T}(\log(\theta_{it}) - x_{it}'\beta - u_{it})^2\right\}\left(\prod_{t=1}^{T}\frac{1}{\theta_{it}}\right).$$

Acknowledgements

The author would like to thank the Editor, the Associate Editor and one referee for their comments and suggestions. This work was partially supported by Statistics Canada Methodology Branch Research Block Fund.

References

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

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.

Chib, S., and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49, 327-335.

Datta, G.S., Day, B. and Maiti, T. (1998). Multivariate Bayesian small area estimation: An application to survey and satellite data. *Sankhyā*, 60, 344-362.

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.

Datta, G.S., Lahiri, P. and Maiti, T. (2002). Empirical Bayes estimation of median income of four-person families by state using time series and cross-sectional data. *Journal of Statistical Planning and Inference*, 102, 83-97.

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

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.

Gambino, J.G., Singh, M.P., Dufour, J., Kennedy, B. and Lindeyer, J. (1998). *Methodology of the Canadian Labour Force Survey*, Statistics Canada, Catalogue No. 71-526.

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., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995). *Bayesian Data Analysis.* London: Chapman and Hall.

Gelman, A., and Rubin, D.B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Ghosh, M., Nangia, N. and Kim, D.H. (1996). Estimation of median income of four-person families: A Bayesian time series approach. *Journal of the American Statistical Association*, 91, 1423-1431.

Laud, P., and Ibrahim, J. (1995). Predictive model selection. *Journal of Royal Statistical Society*, Series B, 57, 247-262.

Meng, X.L. (1994). Posterior predictive *p* value. *The Annals of Statistics*, 22, 1142-1160.

Rao, J.N.K., and Yu, M. (1994). Small area estimation by combining time series and cross-sectional data. *The Canadian Journal of Statistics*, 22, 511-528.

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

Singh, A.C., Folsom, R.E., Jr. and Vaish, A.K. (2005). Small area modeling for survey data with smoothed error covariance structure via generalized design effects. *Federal Committee on Statistical methods Conference proceedings*, *Washington*, *D.C.*, *www.fcsm.gov*.

Singh, A., You, Y. and Mantel, H. (2005). Use of generalized design effects for variance function modeling in small area estimation from survey data. Presentation at the 2005 Statistical Society of Canada Annual Meeting, Regina, SK.

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.

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

You. Y., Chen, E. and Gambino, G. (2002). Nonlinear mixed effects cross-sectional and time series models for unemployment rate estimation. *2002 Proceedings of the American Statistical Association, Section on Government Statistics*, Alexandria, VA: American Statistical Association. 3883-3888.

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*, Alexandria, VA: American Statistical Association, 1836-1840.

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

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 25-32.