SSC Annual Meeting, June 2009
Proceedings of the Survey Methods Section

AREA LEVEL MODELING APPROACHES TO SMALL AREA ESTIMATION WITH APPLICATIONS

Yong You[1]

ABSTRACT

In this paper we discuss the Fay-Herriot model and extended area level models including unmatched models, time series models and spatial correlation models. We also discuss sampling variance smoothing and modeling methods as well as benchmarking small area estimates using EBLUP and HB approaches. Then we will apply the hierarchical Bayes approach with Gibbs sampling method to various area level models and discuss the use of small area models in different surveys for small area estimation at Statistics Canada. Comparison of model-based and direct estimates will be presented. Finally we will give conclusion and some suggestions.

KEY WORDS: Hierarchical Bayes; Sampling variance; Spatial models; Unmatched models.

RÉSUMÉ

Cet article étudie le modèle Fay-Herriot et des modèles régionaux plus étendus tels que les modèles sans correspondance, de séries chronologiques et de corrélation spatiale. Nous discutons du nivellement de la variance de l'échantillon et des méthodes de modélisation, ainsi de l'étalonnage des estimations sur petits domaines par le biais de la MPLSBE et du HB. Nous utilisons l'approche hiérarchique bayésienne et la méthode d'échantillonnage de Gibbs avec divers modèles régionaux dans le cadre de diverses enquêtes de Statistiques Canada. Enfin, nous comparons l'estimation par modèle et l'estimation directe et présentons des conclusions et des recommandations.

MOTS CLÉS: Approche hiérarchique bayésienne; modèles non appariés; modèles spatiaux; variance d'échantillonnage.

1. INTRODUCTION

Model-based estimates have been widely used in practice to provide reliable indirect estimates for small areas in recent years. In general, small area models are classified into two groups: unit level models and area level models. Unit level models are generally based on observation units from surveys and auxiliary variables associated with each observations, whereas area level models are based on direct survey estimates aggregated from the unit level data and area level auxiliary variables. Therefore area level models generally have the ability to protect confidentiality of microdata. Another advantage of area level modeling is that it takes into account the survey design through the use of the direct survey estimates and related design-based variance estimates. Various area level models have been proposed to improve the precision of the direct survey estimates. Among the area level models, the Fay-Herriot model (Fay and Herriot, 1979) is a basic area level model. The Fay-Herriot model has a sampling model for the direct survey estimates and a linking model for the small area parameters of interest. The sampling model assumes that there exists a direct survey estimator y_i, which is usually design unbiased, for the small area parameter θ_i such that $y_i = \theta_i + e_i$, $i = 1,...,m$, where the e_i is the sampling error and m is the number of small areas. It is customary to assume that e_i's are independently normal random variables $e_i \sim N(0, \sigma_i^2)$. The linking model for θ_i is given as $\theta_i = x_i'\beta + v_i$, where $x_i = (x_{i1},...,x_{ip})'$ is a vector of auxiliary variables, $\beta = (\beta_1,...,\beta_p)'$ is a p × 1 vector of regression coefficients, and the v_i's are area-specific random effects assumed to be *iid* with mean 0 and model variance σ_v^2. The assumption of normality is generally also included, even though it is more difficult to justify this assumption. The model variance is unknown and needs to be estimated from the data. The area level random effects v_i capture the unstructured heterogeneity among areas that are not explained by the sampling variances. For the Fay-Herriot model, the sampling variance is usually assumed to be known. In practice,

[1] Yong You, Household Survey Methods Division, Statistics Canada, Ottawa, K1A 0T6, Canada, yongyou@statcan.gc.ca.

smoothed estimators of the sampling variances are used in the Fay-Herriot model and then treated as known. The smoothing of sampling variance estimates usually makes the use of generalized variance function (GVF) method (e.g., Dick, 1995; Wolter, 2007). The assumption of linear linking model and known sampling variance in the sampling model are two major limitations of the Fay-Herriot model. In recent years various extended models have been proposed in practice for many applications. We will discuss some extended area level models in Section 2 including unmatched models, time series models and spatial correlation models. Given the complex small area models proposed in applications, HB approach with the Gibbs sampling method has been widely used to overcome the computation difficulties and obtain the posterior estimates for small area parameters. One advantage of the HB approach is that it is relatively straightforward and the inferences about the small area parameters are exact unlike the EBLUP approach, even in the case where the area-specific sample sizes are very small, which is usually a typical small area estimation problem. The HB approach will automatically take into account the uncertainties associated with unknown parameters in the model. By using the Gibbs sampling method for HB inference, posterior predictive model checking becomes popular in practice. In this paper, we will discuss the use of Bayesian method for model checking and evaluation as well.

The paper is organized as follows: Section 2 presents some extended area level modes. In Section 3, we discuss the problem of sampling variance smoothing and modeling for area level models. Section 4 presents benchmarking methods for model-based small area estimates. Then we discuss Bayesian model checking and model fit analysis in Section 5. In Section 6, we present three important applications of area level models to different surveys of Statistics Canada. And finally we offer some conclusion remarks in Section 7 for small area estimation using area level models.

2. AREA LEVEL MODELS

In this section, we briefly describe three important extended area levels for small area estimation.

2.1 Unmatched models

You and Rao (2002) studied unmatched small area models with a general linking function for small area parameters θ_i as $g(\theta_i) = x_i'\beta + v_i$, where $g(\cdot)$ is a suitable non-linear function to relate θ_i, auxiliary variables x_i, and random effects v_i. This model is particularly useful in modeling proportions by using log or logit linking model for θ_i. For example, a log-linear linking model was used in You and Rao (2002) for census undercoverage estimation and in You (2008) for unemployment rate estimation. Mohadjer et al. (2007) and Liu, Lahiri and Kalton (2008) studied logit linking models in their applications.

2.2 Time series models

Time series model is an important extension of the Fay-Herriot model to borrow strength across regions and over time periods simultaneously to achieve considerable efficiency gains, particularly for repeated surveys such as the monthly Canadian Labour Force Survey (LFS) and US Current Population Survey (CPS). You, Rao and Gambino (2003) has shown that the cross-sectional and time series model is more efficient than the Fay-Herriot model and improves the direct estimation considerably in terms of CV reduction. For more details on the models and applications, see, for example, Rao and Yu (1994), Datta, Lahiri, Maiti and Lu (1999), You, Rao and Gambino (2003), and You (2008) for modeling and smoothing the covariance matrices.

2.3 Spatial correlation models

Spatial correlation random effects can also be included in the linking model to account for spatial dependence among areas. The spatial models are commonly used in health data analysis and disease mapping. In practice, various spatial models have been proposed in the literature; see Best, Richardson and Thomson (2005) for a good review and discussion on various spatial random effects models and applications. For example, a spatial correlation effect u_i can be added to the linking model of small area parameters. Zhou and You (2007) adopted the spatial linking model studied by MacNab (2003) and combined it with the sampling model of Fay-Herriot for health status small area estimation using HB approach. For other discussion and applications of spatial models in small area estimation, see Rao (2003).

3. SAMPLING VARIANCE SMOOTHING AND MODELING

3.1 Sampling Variance Smoothing

The Fay-Herriot model assumes that the sampling variances σ_i^2 are known in the model. This is a very strong assumption in practice. Usually a smoothed estimator of σ_i^2 is used in the model and then treated as known. The method of generalized variance function (GVF) is commonly used to obtain the smoothed estimates of the sampling variances (Wolter, 2007). In this method, a regression model is constructed for the direct estimates of sampling variances by using some external auxiliary variables, and the final fitted values are obtained as the smoothed estimates (e.g., Dick, 1995; Bizier, You, Veilleux and Grondin, 2009). The GVF method requires a separate model for the direct estimates of the sampling variances. The GVF method is widely used in small area estimation to obtain smoothed or fitted variance estimates. Alternatively, we may use design effects to smooth the sampling variances based on the assumption of common design effects across regions. Singh, Folsom and Vaish (2005) suggested the use of generalized design effects in the smoothing procedure for the sampling covariance matrices. You (2008a) also have used the model of common design effects to smooth the sampling variances over time. Design effects have been studied in survey data analysis for various purposes, and particularly useful in variance estimation (Park and Lee, 2004; Gambino, 2009). For proportions, let $\hat{p}_{iw}$ be the direct design-based estimate for a proportion p_i and $\hat{V}_i$ the corresponding direct sampling variance under complex design for the i-th small area. Then the estimated design effect (deff) can be approximately computed as (Gambino, 2009): $deff_i = \dfrac{\hat{V}_i}{\hat{p}_{iw}(1-\hat{p}_{iw})/n_i + \hat{V}_i/n_i}$, or the sampling variance $\hat{V}_i$ can be written as $\hat{V}_i = deff_i \cdot \dfrac{\hat{p}_{iw}(1-\hat{p}_{iw})}{n_i} \cdot (1-\dfrac{deff_i}{n_i})^{-1}$, where n_i is the sample size of the i-th small area. If the sample size n_i is large, then the term $(1-deff_i/n_i)^{-1}$ may be negligible. Then $\hat{V}_i$ reduces to $\hat{V}_i = deff_i \cdot \hat{p}_{iw}(1-\hat{p}_{iw})/n_i$. We can, for example, compute the design effects for all small areas, and then compute the common (smoothed) deff as: $\overline{deff} = \sum_{i=1}^{m} deff_i / m$. Then, smoothed sampling variances for proportion or rate can be obtained as: $\tilde{V}_i = \overline{deff} \cdot \dfrac{\hat{p}_{iw}(1-\hat{p}_{iw})}{n_i} \cdot (1-\dfrac{\overline{deff}}{n_i})^{-1}$. For total or count direct estimates, if we know the corresponding population size N_i, then we can transform the total estimate to proportion estimate, and the corresponding sampling variance is $\hat{V}_i(rate) = \hat{V}_i(total)/N_i^2$. We can then apply the smoothing method for proportions to get smoothed variance estimates $\tilde{V}_i(rate)$, and the smoothed sampling variance for total can be simply obtained as follows: $\tilde{V}_i(total) = \tilde{V}_i(rate) \cdot N_i^2$.

3.2 Sampling Variance Modeling

Alternative to the smoothing approach, we can model the sampling variance directly using various models. For example, suppose that the sampling variances σ_i^2 are unknown in the Fay-Herriot model and are estimated directly by unbiased estimators s_i^2. The estimators s_i^2 are independent of the direct survey estimators y_i of θ_i. Then a sampling model for the sampling variances can be constructed as $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$ and n_i is the sample size for the *i*-th area (Rivest and Vandal, 2002; Wang and Fuller, 2003; You and Chapman, 2006). A log-linear regression model may also be considered for σ_i^2 as $\log(\sigma_i^2) = x_i'\gamma + \xi_i$. We can also model the sampling variances indirectly in the sense that the sampling variances can be expressed as a function of small area parameters or design effects. For example, You and Rao (2002) proposed a log-linear unmatched model with the sampling variance written as σ_i^2 as $\sigma_i^2 = \theta_i^2 \cdot (cv)_i^2$, where θ_i is the unknown small area mean of interest and cv_i is the smoothed coefficient of variation (CV). In applications, the smoothed CV can be obtained over time series data for a given local area, as shown in the small area unemployment rate estimation of You, Rao and Gambino (2003). If the parameter of interest is proportion or rate, we can model the sampling variance using smoothed design effects as $\sigma_i^2 = \theta_i(1-\theta_i) \cdot deff / n_i$, where $deff$ is a known and smoothed design effect for a given area and n_i is the small area sample size (e.g., You, 2008a; Liu, Lahiri and Kalton, 2008), particularly You (2008a) used smoothed design effects over time to model the sampling variance. Indirect modeling of sampling variance is actually a combination of smoothing and modeling methods.

4. BENCHMARKING SMALL AREA ESTIMATES

Benchmarking the small area estimates is a desired property for practical users. In this section, we briefly describe two commonly used benchmarking methods for model-based small area estimates.

4.1 Hierarchical Bayes Approach

You, Rao and Dick (2004) proposed an approach to benchmark the HB estimators such that the Benchmarked HB (BHB) estimates add up to the sum of the direct estimates. To measure the variability of the BHB estimators, posterior mean squared error (PMSE), similar to the posterior variance associated with the HB estimators, was proposed. 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 (You, Rao and Dick, 2004). Let $\hat{\theta}_i^{HB} = E(\theta_i \mid y)$ denote the HB estimator of the small area parameter θ_i and $V(\theta_i \mid y)$ the posterior variance of θ_i. Let $\hat{\theta}_i^{BHB}$ denote the BHB estimator of θ_i such that $\hat{\theta}_i^{BHB}$ is a function of the HB estimators $\hat{\theta}_i^{HB}$, i.e., $\hat{\theta}_i^{BHB} = h(\hat{\theta}_1^{HB},...,\hat{\theta}_m^{HB})$ for some known function $h(\cdot)$, and satisfies the benchmark property: $\sum_{i=1}^m \hat{\theta}_i^{BHB} = \sum_{i=1}^m y_i$. Measure of the variability associated with the BHB estimator $\hat{\theta}_i^{BHB}$ is given by the PMSE as $\text{PMSE}(\hat{\theta}_i^{BHB}) = E\left[(\hat{\theta}_i^{BHB} - \theta_i)^2 \mid y\right] = (\hat{\theta}_i^{BHB} - \hat{\theta}_i^{HB})^2 + V(\theta_i \mid y)$. The PMSE is readily obtained from the posterior variance and the estimators $\hat{\theta}_i^{HB}$ and $\hat{\theta}_i^{BHB}$. The BHB estimators can be constructed using various methods such as the ranking ratio method. The advantage of the BHB estimator and the PMSE is that the benchmarking procedure is well-defined and very easy to compute. The method can be applied to any small area models under the HB inference.

4.2 EBLUP Approach

Wang, Fuller and Qu (2008) proposed a simple method to benchmark the EBLUP estimators under the Fay-Herriot model. Assuming the sampling variances σ_i^2 are known, one can construct an augmented model by using the sampling variances as one more auxiliary variable in the linking model, that is, let $z_i = (x_i, \sigma_i^2) = (x_{i1},...,x_{ip}, \sigma_i^2)'$, and the linking model for the small area parameter θ_i is changed to $\theta_i = z_i'\beta + v_i$. Then it can be shown that the EBLUP estimator under this augmented Fay-Herriot model has a self-benchmarking property such that $\sum_{i=1}^m \hat{\theta}_i^{EBLUP} = \sum_{i=1}^m y_i$, as long as the regression parameter β is estimated using the generalized least squares method. The MSE can be estimated using the usual EBLUP formulas. If σ_i^2 is unknown, then smoothed estimates of σ_i^2 can also be used in this method. This augmented model method is very easy to use in practice, and more importantly, the corresponding MSE estimator is also available. But this method can only be applied to the Fay-Herriot model, unlike the BHB method of You, Rao and Dick (2004), which can be applied to various area level models under the HB inference.

5. MODEL CHECKING AND EVALUATION

For hierarchical Bayes model comparison, deviance information criterion (DIC) proposed by Spiegelhalter, Best, Carlin and van der Linde (2002) is commonly used in recent years to compare non-nested and mixed effects Bayesian models. The DIC is based on the deviance of the model $D(\theta)$, which is equal to minus twice the log-likelihood of the model, and DIC is usually computed as $DIC = D(\hat{\theta}) + 2p_D$, where $D(\hat{\theta})$ is the deviance of the model evaluated at the posterior mean of the model parameters, which summarizes the goodness of fit of the model, and p_D is the effective number of parameters, which captures the complexity of the model. p_D is defined as $p_D = \overline{D}(\theta) - D(\hat{\theta})$, and $\overline{D}(\theta)$ is the posterior mean of the deviance of the model. Thus the DIC is defined as the summation of goodness of fit of the model and the model complexity. Smaller values of DIC indicate a better model fit. Computation of DIC is relatively straightforward provided that the deviance $D(\theta)$ is available in closed form, and p_D may be calculated after the Gibbs sampling run by taking the sample mean of the simulated values of $D(\theta)$ minus the plug-in estimate of the deviance $D(\hat{\theta})$.

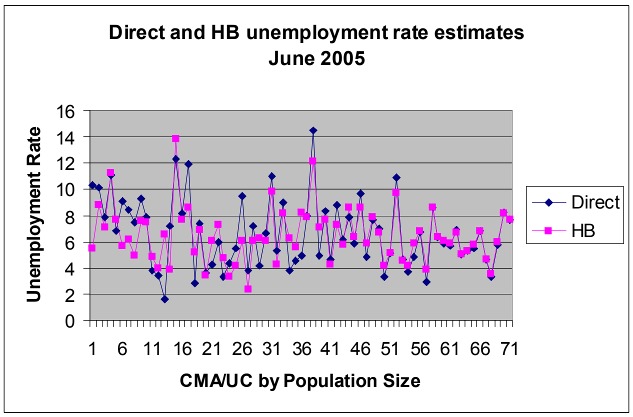
For the overall model fit analysis, the method of posterior predictive distribution is commonly used. Let y_{rep} denote the replicated observation under the model. The posterior predictive distribution of y_{rep} given the observed data y_{obs} is defined as $f(y_{rep} \mid y_{obs}) = \int f(y_{rep} \mid \theta) f(\theta \mid y_{obs}) d\theta$. In this approach, a test statistic $T(y,\theta)$ that depends on the data y and possibly the parameter θ can be defined and the observed value $T(y_{obs},\theta \mid y_{obs})$ compared to the posterior predictive distribution of $T(y_{rep},\theta \mid y_{obs})$ with any significant difference indicates a model failure. Lack of fit of the data with respect to the posterior predictive distribution can be measured by p-value of the test quantity (Gelman, Meng and Stern, 1996). The posterior predictive p-value is defined as $p = P(T(y_{rep},\theta) \geq T(y_{obs},\theta) \mid y_{obs})$. If the given model adequately fits the observed data, then $T(y_{obs},\theta \mid y_{obs})$ should be near the central part of the histogram of the $T(y_{rep},\theta \mid y_{obs})$ 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 model checking has been criticized for being conservative due to the double use of the observed data; see, for example, 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. To carry out the posterior predictive model checking, we need to specify a test quantity $T(y,\theta)$. You (2008b) studied several test quantities in posterior predictive model checking through a simulation study and proposed a test quantity given as $T(y,\theta) = |\max(y_i) - \text{mean}(\theta_i)| - |\min(y_i) - \text{mean}(\theta_i)|$ for small area models. It is shown in You (2008b) that the proposed test quantity $T(y,\theta)$ is sensitive to the choice of distribution of random effects and different mean functions under the Fay-Herriot model. A similar test quantity is also suggested in Gelman, Carlin, Stern and Rubin (2004) for posterior predictive model checking. To access model fit at the individual observation level, we can also compute the individual predictive probability values p_i^* as $p_i^* = P(y_{i(rep)} < y_{i(obs)} \mid y_{obs})$; see, for example, Gelfand (1996) and Daniels and Gatsonis (1999). These individual predictive probabilities provide information on the degree of consistent overestimation or underestimation of the observed data.

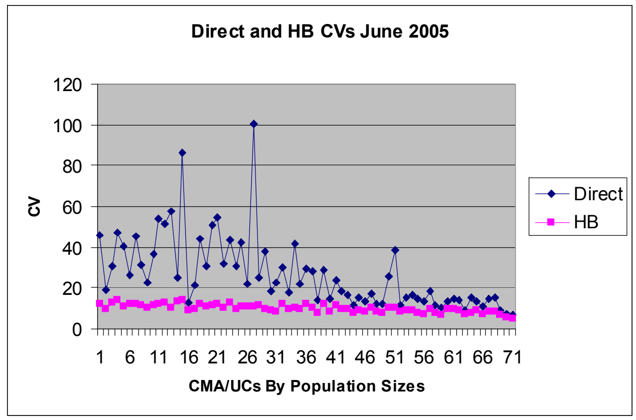
To evaluate the model-based estimates under the proposed models with respect to the direct survey estimates, Brown, Chambers, Heady and Heasman (2001) proposed four diagnostic methods, namely, bias diagnostic, goodness of fit diagnostic, coverage diagnostic and calibration diagnostic. These methods are relatively easy to implement and useful in model evaluation. For example, for bias diagnostic, we can consider a simple method of regression analysis for the direct estimates and the model-based estimates. You (2008a) also used the regression analysis method for model bias diagnostics. If the model-based estimates are close to the true values of the small area disease rate, then the direct survey estimates, which are assumed to be unbiased for the true disease rates, should behave like random variables whose expected values correspond to the values of the model-based estimates. That means the model-based estimates should be unbiased predictors of the direct estimates. In terms of regression analysis, we basically fit the regression model $Y = \alpha X$ to the data and estimate the coefficient α, and to see how close the regression line is to $Y = X$.

6. APPLICATIONS

6.1 LFS Unemployment Rate Estimation

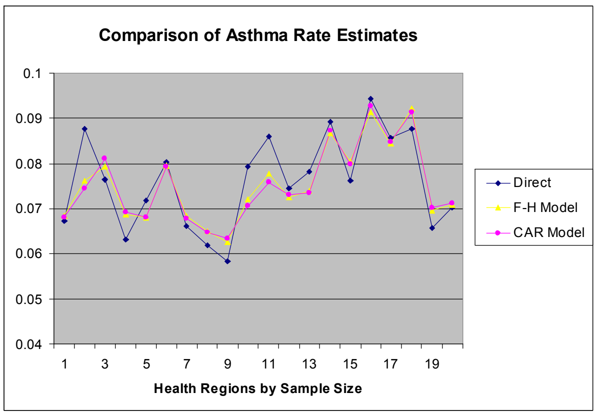
The Canadian Labour Force Survey (LFS) produces monthly estimates of the unemployment rate at national and provincial levels. The LFS also releases unemployment estimates for sub-provincial areas such as Census Metropolitan Areas (CMAs) and Urban Centers (UCs). However, for some sub-provincial areas, 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. Various models have been used in practice to obtain model-based estimates for unemployment rate estimation (e.g., Datta, Lahiri, Maiti and Lu, 1999; You, Rao and Gambino, 2003). In particular, You (2008a) proposed an integrated non-linear mixed effects model under the hierarchical Bayes 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. The model of You (2008a) is a log-linear area level model with area random effects and random walk time components to model the true unemployment rate for the i-th area at time point t. The model has several components as follows: $[y_i \mid \theta_i] \sim N(\theta_i, \Sigma_i(\theta_i))$, $\theta_i = (\theta_{i1},...,\theta_{iT})'$ where θ_{it} is the unemployment rate, and it is modeled as $[\log(\theta_{it}) \mid \beta, u_{it}, \sigma_v^2] \sim N(x_{it}'\beta + u_{it}, \sigma_v^2)$, and the time component u_{it} is modeled as $[u_{it} \mid u_{i,t-1}, \sigma_\varepsilon^2] \sim N(u_{i,t-1}, \sigma_\varepsilon^2)$, the variance covariance matrix $\Sigma_i(\theta_i)$ depends on the parameters θ_i through smoothed design effects across areas; for details, see You (2008a). 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. As in You (2008a), we consider m=71 CMA/UCs in the model and within each area, we consider six consecutive monthly estimates y_{it} from January 2005 to June 2005, so that $T = 6$. The parameter of interest θ_{iT} is the true unemployment rate for area i in June 2005, where i = 1, …, 71. To implement the Gibbs sampler, we have used 10 parallel runs, each of length 2000. As shown in You (2008a), the proposed integrated model performs very well and leads to reliable model-based estimates for local areas in terms of CV reduction, model fit analysis and comparison of the model-based estimates. The following two figures show the comparison of point estimates and CV estimates. It is clear that model-based method provides much more stable CVs than the direct estimates; for more detailed discussion on the results, see You (2008a).

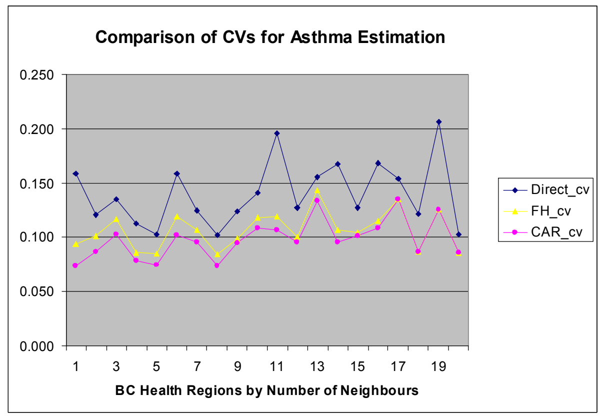




6.2 CCHS Disease Rate Estimation

The Canadian Community Health Survey (CCHS) is a federal survey conducted by Statistics Canada. The primary objective of CCHS is to provide timely and reliable estimates of health determinants, health status and health system utilization across Canada. It is a cross-sectional survey which operates on a two-year collection cycle. The first year of the survey cycle "x.1" targets individuals aged 12 or older who are living in private dwellings, and it is a general population health survey of a large sample (130,000 persons) designed to provide reliable estimates at the health region, provincial and national levels. The second year of the survey cycle "x.2" has a smaller sample (30,000 persons) allocated based on provincial sample buy-ins and is designed to provide provincial and national level results on specific focused health topics. Although national and provincial estimates are very important, there is an increasing demand for health data at lower levels of geography voiced by a number of provinces including British Columbia (BC), Prince Edward Island (PEI), Quebec and others. Cycle "x.1" of the CCHS collected data corresponding to 136 health regions in the 10 provinces and three territories. It primarily used two sampling frame. The first one, used as the primary frame, was based on the area frame designed for the Canadian Labour Force Survey, and within the area frame, a multistage stratified cluster design was used to sample dwellings. The second frame consists of a list of telephone numbers. Random digit dialing methodology is used in some of the health regions for cost reasons. In this paper, we use a small data set from Cycle 1.1 as an example to demonstrate the analysis. We are interested in estimating the disease rate for local health regions within provinces. Zhou and You (2007) proposed a hierarchical Bayes spatial model to obtain reliable disease rate estimates at health region levels for the Canadian Community Health Survey (CCHS). The Zhou-You model is given in as follows: $\mathbf{y} \mid \boldsymbol{\theta}, \sigma_1^2, \cdots, \sigma_m^2 \sim \text{MVN}(\boldsymbol{\theta}, \mathbf{E})$, where $\mathbf{E}$ is a diagonal matrix with the i-th diagonal element equal to the sampling variance σ_i^2; $\boldsymbol{\theta} \mid \beta, \sigma_v^2 \sim \text{MVN}(\mathbf{X}\beta, \sigma_v^2\mathbf{D}^{-1})$, where $\mathbf{D} = \lambda\mathbf{R} + (1-\lambda)\mathbf{I}$, and σ_v^2 is a spatial dispersion parameter and λ is a spatial autocorrelation parameter, $0 \le \lambda \le 1$; $\mathbf{I}$ is an identity matrix of dimension m; **R,** commonly known as the neighbourhood matrix. The sampling variance σ_i^2 can be treated as known or unknown in the model. The linking is a conditional autoregressive model which includes both local smoothing of neighbouring areas and global smoothing. The proposed model captures both the unstructured heterogeneity among areas and the spatial correlation effects of the neighbouring area; for detail discussions, see Zhou and You (2007). Applying the Zhou-You model to the CCHS cycle 1.1 data of asthma estimates for the 20 health regions of BC province, we compared the direct estimates with the model-based estimates from the Fay-Herriot model and the proposed spatial correlation model. The following two figures show the comparison of the point estimates and CV estimates. Both the Fay-Herriot model and the spatial model have produced similar point estimates and have achieved efficient CV reduction over the direct estimates. More importantly, the proposed spatial model is more efficient than the Fay-Herriot model as it borrows more information using spatial correlation structure. It is clear that more CV reduction under the spatial model over the Fay-Herriot model as the number of neighbours increases. Thus, more neighbouring areas can provide more information in the spatial structure to improve the precision and reliability of the model-based estimates.





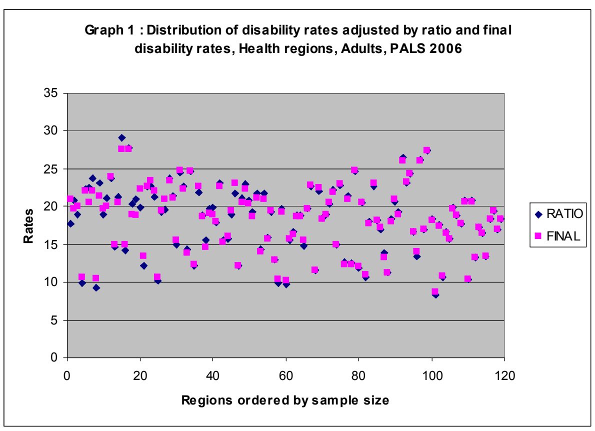
6.3 PALS Disability Estimation

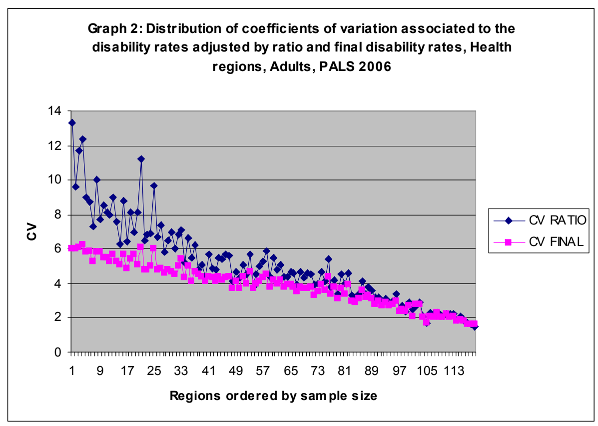
The Participation and Activity Limitation Survey (PALS) is a post-census survey that collects information about persons with disabilities whose everyday activities are limited because of a health-related condition or problem. The PALS produces estimates on various disability statistics for adults and children at the national, provincial and sub-provincial levels. The 2006 PALS uses 2006 Census as a sampling frame to identify its population. The 2006 Census questionnaire includes two general questions on activity limitations. The PALS sample is selected from people who respond “yes” to at least one of the general activity limitation questions in the Census. The PALS interview begins with the census activity limitation filter questions identical to the two Census general questions on activity limitations followed by a series of detailed screening questions on disability. A person is considered as disabled if he or she answers at least one “yes” to the Census filter question on activity limitation of daily living and at least one “yes” to the PALS screening questions. The 2006 PALS sample size is 48,000, consisting of approximately 39,000 adults and 9,000 children. The sample was selected using a two-phase stratified design where at the first phase, a Census questionnaire was distributed to approximately one out of five persons, and at the second phase, a sample of persons who answered “yes” to the Census filter questions is selected according to a stratified systematic sample. The strata are created in order to allow good estimates by provinces and territories. Interviews are conducted by telephone with the interviewers using a computer assisted collection methodology. Two questionnaires are used, one for adults aged 15 and over and one for children under the age of 15. The interviews for the children’s questionnaire are conducted with the parent or guardian of the child.

The small area estimation of disability data includes the estimation of disability counts and rates for adults and children at two sub-provincial areas across Canada. The first small area set consists of census metropolitan areas (CMAs) and census agglomerations (CAs) for which the PALS has respondents. There are 114 CMAs and CAs across Canada. At the CMA/CA level, the sample sizes for adults range between 12 and 1751, and for children, between 3 and 448. The second small area set consists of health regions (HRs) as defined by the provincial departments of health in 2007. There are 119 health regions across Canada. At the health region level, the sample sizes available for adults range between 16 and 1653, and for children, between 3 and 358.

The model-based small area estimation procedure for PASL disability data includes the following five steps: (1) ratio adjustment of direct estimates, (2) sampling variance modeling, (3) selection of auxiliary variables, (4) log-linear unmatched models for disability rates estimation, and (5) benchmarking the model-based estimates. The ratio adjustment is used to ensure that the adjusted direct estimates of disability counts are consistent with (less than or equal to) the census counts of persons who answer "yes" to the census filter questions. The log-linear unmatched model (You and Rao, 2002) is then applied to the ratio adjusted direct estimates of disability rates at CMA/CA or HR level for adults and children, separately. Let $\hat{p}_i^R$ denote the ratio adjusted direct estimate of disability rate for the i-th area. The log-linear unmatched model for the true disability rate p_i is given as $\hat{p}_i^R = p_i + e_i$, $i = 1,..., m$, and the linking model is a log-linear regression model: $\log(p_i) = x_i'\beta + \nu_i$, $i = 1,..., m$, where the e_i's are sampling errors associated with the direct estimate $\hat{p}_i^R$ and the ν_i's are random effects related to the linking model for $\log(p_i)$. The variance of e_i, namely, the sampling variance, is smoothed using a generalized variance function model give as $\tilde{V}(\hat{p}_i^R) = \tilde{V}(e_i) = \exp\{\beta_0 + \beta_1 \log(M_i) + \beta_2 \log(N_i)\}$, where M_i represents the number of people having responded "yes" to at least one of the Census filter questions, and N_i represents the total population of the i-th area. Hierarchical Bayes (HB) inference with Gibbs sampling method is used to obtain the HB disability rate estimates $\hat{p}_i^{HB}$. Then these HB estimates are benchmarked to get the final estimates $\hat{p}_i^{FINAL}$ such that the final benchmarked HB estimates add up to the direct estimates of provincial totals, using the method of You, Rao and Dick (2004) discussed in Section 4.

The following two figures show the comparison of point estimates and CVs of ratio adjusted direct estimates and the final benchmarked HB estimates. For the point estimates, it is clear that when sample sizes are large, the direct estimates and final model-based estimates are close to each other as expected, as more weights are put on the direct estimates in the model-based estimates when the sample sizes are large. For CVs, both direct CV and model CV decrease as the sample increases. Also, when sample size is large, the efficiency gain of the model is small as both direct CV and model CV are close to each other. When sample size is small, the efficiency gain is big as expected. For more details on the PALS small area estimates and methods, see Bizier, You, Veilleux and Grondin (2009).





7. CONCLUSIONS

We have discussed the well-known Fay-Herriot model and various extended area level models for small area estimation including unmatched models, time series models and spatial models. These area level models can be applied to various survey data to improve the direct survey estimates for small areas. We also discussed the problem of sampling variance smoothing and modeling approaches for area level models. Benchmarking is a desired property for practical users to obtain more confidence on model-based estimates. We have reviewed some methods under both the HB and EBLUP approaches. We also outlined some model checking and evaluation methods for area level models. We then provided three examples of area level modeling approaches for small area data from different surveys including the Labour Force Survey, the Canadian Community Health Survey and the Participation and Activity Limitation Survey. For example, we considered a log linear unmatched time series model with sampling covariance modeling to improve the direct unemployment rate estimation for sub-provincial areas across Canada using the LFS data (You, 2008a). For the CCHS, we have studied a spatial correlation model for disease rate estimation with sampling variance smoothing and modeling for sub-provincial health regions (Zhou and You, 2007). Particularly Zhou and You (2007) showed that the proposed spatial model can improve the direct estimate and performs better than the Fay-Herriot model as the number of neighbour areas increases.

REFERENCES

Bell, W.R. (1995). Bayesian sampling error modeling with application. Proceedings of the Survey Methods Section, Statistical Society of Canada Annual Meeting, 19-26.

Bell, W.R., and Huang, E.T. (2006). Dealing with influential observations and outliers in small area estimation. XXIII International Symposium on Methodological Issues, Ottawa, Canada.

Best, N., Richardson S. and Thomson, A. (2005). A comparison of Bayesian spatial models for disease mapping. *Statistical Methods in Medical Research*, 14, 35-39.

Bizier, V, You, Y., Veilleux, L. and Grondin, C. (2009). Model-based approach to small area estimation of disability count and rate using data from the 2006 Participation and Activity Limitation Survey. Methodology Branch Working Paper, SSMD-2009-003E/F, Statistics Canada.

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.

Daniels, M.J. and Gatsonis, C. (1999). Hierarchical generalized linear models in the analysis of variations in health care utilization. *Journal of the American Statistical Association*, 94, 29-42.

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

Fay, R. E. and Herriot, R. A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data, *Journal of the American Statistical Association*, 74, 268-277.

Gambino, J. (2009). Design effect caveats. *The American Statistician*, 63, 141-146.

Gelfand, A.E. (1996). Model determination using sampling-based methods. In Markov Monte Carlo in Practice (W.R. Gilks, S. Richardson and D.J. Spiegelhalter, eds.), Chapman & Hall, London, pp. 145-161.

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

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*, 2nd Edition. Chapman & Hall/CRC.

Liu, B., Lahiri, P. and Kalton, G. (2008). Hierarchical Bayes modeling of survey-weighted small area proportions. Unpublished manuscript.

MacNab, Y.C. (2003). Hierarchical Bayesian spatial modeling of small area rates of non-rare disease. *Statistics in Medicine, 22*, 1761-1773.

Mohadjer, L., Rao, J.N.K., Liu, B., Krenzke, T. and VanDekerckhove, W. (2007). Hierarchical Bayes small area estimates of adult literacy using unmatched sampling and linking models. *2007 Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Mukhopadhyay, P. and Maiti, T. (2004). Two stage non-parametric approach for small area estimation. *Proceedings of the American Statistical Association, Section on Survey Research Methods [CD-ROM]*, 4058-4065, Alexandria, VA: American Statistical Association.

Park, I. and Lee, H. (2004). Design effects for the weighted mean and total estimators under complex survey sampling. *Survey Methodology*, 30, 183-193.

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

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

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.

Singh, A.V., Folsom, R.E. 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.

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

Spiegelhalter, D.J., Best, N., Carlin, B.P. and van de Linde, A. (2002). Bayesain measures of model complexity and fit. *Journal of Royal Statistical Society, B*, 64, 583-639.

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. (2008). Small area estimation under a restriction. *Survey Methodology*, 34, 29-36.

Wolter, K.M. (2007). *Introduction to Variance Estimation*. Second Edition, Springer.

You, Y. (2008a). An integrated modeling approach to unemployment rate estimation for sub-provincial areas of Canada. *Survey Methodology*, 34, 19-27.

You, Y. (2008b). Small area estimation using area level models with model checking and applications. *Statistical Society of Canada 2008 Proceedings of the Survey Methods Section*.

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 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.

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.

Zhou, Q.M. and You, Y. (2007). Hierarchical Bayes small area estimation for the Canadian Community Health Survey. Methodology Branch Working Paper, HSMD-2007-007E, Statistics Canada.