SMALL AREA ESTIMATION USING AREA LEVEL MODELS WITH MODEL CHECKING AND APPLICATIONS

Yong You[1]

ABSTRACT

In this paper, we first discuss various area level models for small area estimation including a basic area level model and some extended area level models. We also discuss some sampling variance modeling and smoothing methods. Secondly, we will discuss the problem of model checking using posterior predictive distributions for area level models. We propose a new test quantity and compare it with other several test quantities for posterior predictive model checking through a simulation study. Finally, we give some conclusion and application areas in some Statistics Canada surveys.

KEY WORDS: Hierarchical Bayes; Posterior predictive p-value; Sampling variance; Spatial models; Unmatched models.

RÉSUMÉ

Dans ce document, nous examinons en premier lieu divers modèles au niveau de la région pour l'estimation de petits domaines, dont un modèle de base au niveau de la région et certains modèles élargis. Nous considérons également quelques méthodes de modélisation de la variance d'échantillonnage et de lissage. En deuxième lieu, nous discutons du problème de la vérification du modèle par l'emploi de distributions prédictives a posteriori pour les modèles au niveau de la région. Nous proposons une nouvelle statistique de test et nous la comparons à plusieurs autres statistiques de test utilisées pour la vérification de modèles avec distributions prédictives a posteriori dans une étude de simulation. Finalement, nous présentons des conclusions et nous proposons des secteurs d'application dans certaines enquêtes de Statistique Canada.

MOTS CLÉS: approche hiérarchique bayésienne; valeur p prédictive a posteriori; variance d'échantillonnage; modèles spatiaux; modèles non appariés.

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

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

the sampling variances. For the Fay-Herriot model, the sampling variance is usually assumed to be known. In practice, 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). Recently Singh, Folsom and Vaish (2005) suggested the use of generalized design effects in the smoothing procedure for the sampling covariance matrices. You (2008) have used the model of common design effects to smooth the sampling variances. 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. Another limitation of the Fay-Herriot model is that it is a cross-sectional model only. In many applications, temporal and spatial correlation effects can be employed in the model to improve the effectiveness of model-based estimates over the direct survey estimates. We will discuss some other extended models including time series and spatial models in Section 2 as well.

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 the posterior predictive distribution for model fit analysis, and particularly, we will discuss the use of posterior predictive p-values for overall model fit analysis and compare several test quantities through a simulation study in Section 3. Finally, in Section 4, we will present some applications of area level modeling and offer some conclusion remarks.

2. EXTENSIONS OF FAY-HERRIOT MODEL

2.1 Sampling variance modeling

Suppose that the sampling variances σ_i^2 are unknown in the 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. Rivest and Vandal (2002) and Wang and Fuller (2003) considered EBLUP approach and obtained extra term added to the mean squared error (MSE) estimator to account for the variability associated with the estimation of the sampling variances. You and Chapman (2006) considered the same problem using a HB approach. You and Chapman (2006) also have shown that the posterior estimates and the corresponding CVs are stable and not sensitive to the choice of parameters in the vague proper priors. For multivariate cases, a Wishart distribution can be used to model the sampling variance covariance matrix and to make inference about the parameters in the sampling error model; for example, see Bell (1995).

We can also model the sampling variance indirectly in the sense that the sampling variance 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 design effect for a given area and n_i is the small area sample size (e.g., You, 2008; Liu, Lahiri and Kalton, 2008), particularly You (2008) used smoothed design effects over time to model the sampling variance. Indirect modeling of sampling variance is actually a combination of smoothing and modeling.

2.2 Some extended models

The small area parameters θ_i are not directly observable, it can be more difficult to specify a suitable distribution for θ_i. In the Fay-Herriot model, v_i is usually assumed to have a normal distribution in practice as an approximation, but it is more difficult to justify. To evaluate the effects of model checking and sensitivity to the linking model specification, we consider alternative t-distribution and exponential distribution for v_i, and accordingly for θ_i, we consider linking models

$\theta_i \sim t_\lambda(x_i'\beta, \sigma_v^2)$ and $\theta_i \sim \exp(x_i'\beta)$. The t-distribution has longer tails and the exponential distribution is not symmetric and thus might represent a serious failure of the assumption of the normal linking model in the Fay-Herriot model. Bell and Huang (2006) also considered t distributions for the random effects in small area estimation. In Section 3, we consider the t and exponential distributions in a simulation study and use the Fay-Herriot model for model checking to evaluate the performance of the posterior predictive p-values. The linear linking model for θ_i in the Fay-Herriot model may not be appropriate for all applications in practice. Mukhopadhyay and Maiti (2004) studied the case of a non-linear mean function for θ_i as $\theta_i = m(x_i) + v_i$, where $m(\cdot)$ is a smooth mean function which defines the true relation between auxiliary variables x_i and small area parameters θ_i. In this paper, we will use the non-linear mean functions in the simulation study for the posterior predictive model checking of the Fay-Herirot model.

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.

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.

Spatial 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 more discussion and applications of spatial models in small area estimation, see Rao (2003).

3. POSTERIOR PREDICTIVE MODEL CHECKING

To evaluate the overall fit of a Bayesian model, we study the use of posterior predictive p-value (e.g., see Gelman, Carlin, Stern and Rubin, 1995) defined as $p = P(T(y_{rep}, \theta) \geq T(y_{obs}, \theta) \mid y_{obs})$, where y_{rep} denote the replicated observation under the model and $T(y, \theta)$ be a test statistic that depends on the data y and possibly the parameter θ. If a model fits the observed data, then $T(y_{obs}, \theta \mid y_{obs})$ should be near the central part of the histogram of the $T(y_{rep}, \theta \mid y_{obs})$. Consequently, the posterior predictive p-value is expected to be near 0.5. Extreme p-values (near 0 or 1) suggest poor fit. To carry out the posterior predictive model checking, we need to specify a test quantity $T(y, \theta)$. Because a model can fail to reflect the process that generated the data in any number of ways, we should compute the posterior predictive p-values for a variety of test quantities in order to evaluate any possible model failure. In our study we consider several test quantities as follows: $T_1 = \sum_{i=1}^{m}(y_i - \theta_i)^2 / \text{var}(y_i \mid \theta)$, a general goodness-of-fit test statistic that resembles the classical χ^2 goodness-of-fit measure; $T_2 = \max(y_i)$, $T_3 = \min(y_i)$, $T_4 = \text{mean}(y_i)$, and $T_5 = \text{var}(y_i)$. For the area level models, we propose a new test quantity $T_6 = |\max(y_i) - \text{mean}(\theta_i)| - |\min(y_i) - \text{mean}(\theta_i)|$, which should be sensitive to asymmetry of the distribution. In the next section we will compare the performance of these six test quantities in the posterior predictive model checking through simulation studies. The posterior predictive p-value model checking has been criticized for being conservative due to the double use of the observed data. However, 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.

Simulation study on random effects: We use simulation to assess the performance of the posterior predictive model checking for the specification of random effects and mean functions. For random effects model checking, we use the milk data used in You and Chapman (2006) as study case. There are 43 areas. From the milk data, we get estimates of regression parameters $\beta' = (\beta_0, \beta_1, \beta_2, \beta_3) = (0.967, 1.097, 1.195, 0.726)$. We use these β estimates as true values in the simulation study. The x variables are indicator variables with $x_i'\beta = \beta_j$ if $i \in j$ -th major area. The 4 major areas are 1-7, 8-14, 15-25 and 26-43, as used in You and Chapman (2006). We then generate the small area mean θ_i from $\theta_i \sim \exp(x_i'\beta)$, $\theta_i \sim t(\upsilon, x_i'\beta, \sigma_v^2)$ and $\theta_i \sim N(x_i'\beta, \sigma_v^2)$, where the degrees of freedom for the t distribution are $\upsilon = 2$, 10, 20 and 30, and the variance component σ_v^2 is 0.0126, which is the method of moments estimate of σ_v^2 from the milk data. After we have generated θ_i, we can generate y_i from $y_i \sim N(\theta_i, s_i^2)$, $i = 1,...,43$, where s_i^2 are the sampling variances given in the milk data and treated as known. For each setup, we have generated 1000 data sets. We fit the Fay-Herriot model to each generated data set and apply the HB approach to obtain the posterior predictive p-values. The Gibbs sampling procedure is applied to each generated data set with a "burn-in" period of 500 and Gibbs sampling size of 1000. Table 1 presents the average p-value for each test quantity based on the 1000 generated data sets.

Table 1. Posterior predictive p-values for six test quantities when the Fay-Herirot model used to analyze the simulated data with different random effects.

	$\theta_i \sim \exp$	$\theta_i \sim t_2$	$\theta_i \sim t_{10}$	$\theta_i \sim t_{20}$	$\theta_i \sim t_{30}$	$\theta_i \sim N$
P-value1	0.496	0.498	0.477	0.456	0.448	0.412
P-value2	0.399	0.403	0.335	0.370	0.405	0.500
P-value3	0.186	0.207	0.347	0.401	0.432	0.539
P-value4	0.499	0.498	0.500	0.498	0.498	0.499
P-value5	0.505	0.509	0.513	0.507	0.506	0.507
P-value6	0.227	0.244	0.286	0.344	0.387	0.518

First we note that when the Fay-Herirot model is used in the correct data set, that is, θ_i is generated from the normal distribution $\theta_i \sim N(x_i'\beta, \sigma_v^2)$, the p-values in the last column are all close to 0.5 except p-value1, which has slightly lower value 0.412. The p-value4 and p-value5 are always close to 0.5 for all the populations, which indicate the test quantities $T_4 = \text{mean}(y_i)$, and the sample variance $T_5 = \text{var}(y_i)$ are not good in posterior predictive model checking. When θ_i is generated from the exponential distribution, p-value3 is 0.186 and p-value6 is 0.227, and these two values are very small compared to 0.5. When θ_i is generated from t_2, p-value3 and p-value6 are also small with p-value3 = 0.207 and p-value6 = 0.244. We note that as the degrees of freedom increase, the values of p-value3 and p-value6 also increase to close to 0.5. This should be expected since the t-distribution goes to normal as the the degrees of freedom goes to infinity. One interesting result is about p-value1. P-value1 is close to 0.5 for all the populations except that the normal population has a slightly lower value. From table 1, we conclude that the test quantity $T_3 = \min(y_i)$ and the proposed test quantity $T_6 = |\max(y_i) - \text{mean}(\theta_i)| - |\min(y_i) - \text{mean}(\theta_i)|$ are sensitive in posterior predictive model checking for the specification of random effects, particularly for the exponential and t distribution.

Simulation study on mean functions: Following Mukhopadhyay and Maiti (2004), we consider the following three mean functions in the linking model for small area parameters: (1) Linear function: $f_1(x) = 50 + 2x$; (2) Cubic function: $f_2(x) = 0.01 + 0.2x - 0.05x^3$; (3) Exponential function: $f_3(x) = \exp(0.5x)$. We consider m = 99 areas, and x_i is generated from uniform distribution U(0,5). For each mean function $f_k(x)$, we generate the small area mean θ_i from normal distribution $N(f_k(x_i), \sigma_v^2)$, for i = 1, …, 99, where σ_v^2 is chosen as 0.25. Then we generate y_i from normal distribution $N(\theta_i, \sigma_e^2)$, where the sampling variance σ_e^2 is taken as 0.1 for areas from 1 to 33, 0.25 for areas from 34 to 66, and 0.5 for areas from 67 to 99. We have generated 500 simulated data sets. We then fit the Fay-Herriot model to the

simulated data sets and perform model checking. Table 3 presents the average p-value for each test quantity based on the 500 generated data sets for the three different mean functions of small area means.

Table 3. Posterior predictive p-values for six test quantities when the Fay-Herirot model used to analyze the simulated data with different mean functions

	Linear function	Cubic function	Exponential function
P-value1	0.481	0.487	0.492
P-value2	0.519	0.766	0.331
P-value3	0.498	0.635	0.148
P-value4	0.493	0.497	0.498
P-value5	0.497	0.502	0.496
P-value6	0.512	0.773	0.151

As expected, when the mean function is the linear function, that is, the Fay-Herriot model is the correct model for the data, all posterior predictive p-values are close to 0.5. When the mean functions are cubic and exponential functions, the Fay-Herriot model is no longer a correct model. The values of p-value1, p-value4 and p-value5 are still close to 0.5, which indicates that the test quantities fail in the model checking. When the mean function is cubic, p-value2 and p-value6 are 0.766 and 0.773, respectively. When the mean function is exponential, p-value3 and p-value6 are 0.148 and 0.151, respectively, which are far from 0.5. The results have show that the test quantities $T_2 = \max(y_i)$, $T_3 = \min(y_i)$ and the proposed test quantity $T_6 = |\max(y_i) - \text{mean}(\theta_i)| - |\min(y_i) - \text{mean}(\theta_i)|$ are useful in model checking. From the above results, we conclude that the proposed test quantity $T_6 = |\max(y_i) - \text{mean}(\theta_i)| - |\min(y_i) - \text{mean}(\theta_i)|$ is effective in model checking in all the cases and thus the best among all the test quantities.

4. CONCLUDING REMARKS

We have discussed the well-known Fay-Herriot model and various extended area level models for small area estimation. In particular, we discussed the problem of sampling variance smoothing and modeling approaches for area level models. We also discussed some important extensions including unmatched models, time series models and spatial models. All these modeling approaches can be applied to various survey data to improve small area direct survey estimates. At Statistics Canada, we have studied and applied various area level modeling methods to many surveys including the Labour Force Survey (LFS), the Canadian Community Health Survey (CCHS), the Reverse Record Check for census undercoverage and the Participation and Activity Limitation Survey (PALS). 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, 2008). 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. Model checking is important in model-based small area estimation. In this paper we studied several test quantities for model checking using the posterior predictive p-value. We proposed a test quantity depending on the direct survey estimates and small area parameters for area level model checking. A simulation study based on the Fay-Herriot model has shown that the proposed test quantity performs very well in the check of overall model fit. The proposed test quantity can be used together along with other test quantities for overall model fit analysis for small area estimation.

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.

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

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.

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.

Gelman, A., Carling, J.B., Stern, H.S. and Rubin, D.B. (1995) Bayesian Data Analysis. New York: Chapman & Hall.

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.

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.

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

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 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.