

HIERARCHICAL BAYES SMALL AREA ESTIMATION FOR THE CANADIAN COMMUNITY HEALTH SURVEY

Qian M. Zhou¹ and Yong You²

ABSTRACT

Area level models such as Fay-Herriot models (Fay and Herriot, 1979) have been widely used to obtain reliable model-based estimators in small area estimation. However, in the model, two strong assumptions are made. One is that the sampling error variances are customarily assumed to be known, and the other is that the area-specific random effects are assumed to be independent and identically distributed. In this paper, we propose four full hierarchical Bayes (HB) models which relax these two strong assumptions by constructing Gaussian conditional autoregressive (CAR) models on the area-specific effects to induce spatial correlation, and/or assuming the sampling variances unknown. Through analysis of the survey data from Cycle 1.1 of Canadian Community Health Survey (CCHS), we make comparison among the HB model-based estimates and direct design-based estimates for the rate of asthma for the 20 health regions in BC province. Our results have shown that the model-based estimates perform better than the direct estimates. In addition, the proposed area-level CAR models have smaller CVs than the Fay-Herriot model which imposes independent area-specific random effects. Moreover, larger number of neighbours offers more efficient information in CAR models, leading to greater CV reduction over the Fay-Herriot model.

KEY WORDS: Area-specific effects, Fay-Herriot models, Gaussian conditional autoregressive, Hierarchical Bayes.

RÉSUMÉ

Niveau de la zone tels que les modèles de Fay-Herriot modèles (Fay et Herriot, 1979) ont été largement utilisées pour obtenir des modèles à base d'estimateurs dans l'estimation petite zone. Toutefois, dans le modèle, deux hypothèses fortes sont prises. La première est que l'erreur d'échantillonnage écarts sont habituellement supposé être connu, et l'autre est que la zone des effets aléatoires sont supposés être indépendants et identiquement distribués. Dans cet article, nous proposons quatre hiérarchique de Bayes (HB) des modèles qui se détendre ces deux hypothèses par la construction de Gauss autorégressive conditionnelle (CAR) modèles sur la zone des effets d'induire la corrélation spatiale, et / ou en supposant que l'échantillonnage des écarts inconnu. Par l'analyse des données de l'enquête du cycle 1.1 de la Communauté canadienne de l'Enquête sur la santé (ESCC), nous faisons la comparaison entre le modèle DP-fondé des estimations et directe conception fondée sur des estimations pour le taux de l'asthme pour les 20 régions sanitaires dans la province de la Colombie-Britannique. Nos résultats ont montré que le modèle basé sur les estimations de meilleurs résultats que les estimations directes. En outre, le projet de zone niveau des modèles de voiture sont plus petits que les CV Fay-Herriot modèle indépendant qui impose domaine des effets aléatoires. En outre, plus grand nombre de voisins plus d'information efficace dans des modèles de voiture, conduisant à une plus grande réduction de CV au cours de la Fay-Herriot modèle.

MOTS CLÉS: Autorégressive conditionnelle gaussienne; Fay-Herriot modèles; hiérarchique de Bayes; la zone des effets spécifiques.

1. INTRODUCTION

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. In the first year of the survey cycle, a general population health survey of a large sample (130,000 persons) is 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, Prince Edward Island, Quebec and others. More details of the design are provided in Béland (2002). In this paper, we focus on the cycle 1.1 of the CCHS, which collected data in totally 136 health regions in the 10 provinces and

¹ Qian M. Zhou, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, 200 University Avenue West, Waterloo, Ontario, Canada, N2L 3G1, q2zhou@math.uwaterloo.ca

² Yong You, Household Survey Method Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6, yongyou@statcan.ca

three territories. We are interested in estimating the disease rate for the 20 health regions in BC province, from the data collected in Cycle 1.1.

Direct estimates, based only on the domain-specific sample data, usually provide reliable estimates of the parameter of interest for large areas such as provinces and nations. However, due to cost or other reasons, it is seldom possible to have a large enough overall sample size to support adequate direct estimates for all the smaller areas of interest. In particular, for small areas such as local health regions and age-sex domains, direct estimators are likely to yield large standard errors. For example, for the 20 health regions in BC province, the estimated coefficients of variation (CVs) of the rate of asthma range from 10% to 21%. Therefore, it is necessary to use indirect estimates that borrow strength by using values of the variable of interest from related areas, thus increasing the “effective” sample sizes. These values are brought into the estimation process through an explicit model that provides a link to related areas through the use of supplementary information such as census counts or administrative record; see Rao (2003). There are two broad classifications for these models: area level models and unit level models. Area level models are based on area direct survey estimators and unit level models are based on individual observations in areas. This paper we focus on area level models.

Area level models such as Fay-Herriot models (Fay and Herriot, 1979) have been widely used to obtain reliable model-based estimators for small areas. However, in Fay-Herriot models, several strong assumptions have been made. For example, the sampling variances are assumed as known, but this assumption is rarely possible in practice. You and Chapman (2006) considered the situation where the sampling variances are unknown and estimated individually by direct estimators. Another strong assumption is that Fay-Herriot models assume area-specific random effects in the linking model independent and identically distributed, but in some applications prior knowledge may indicate that geographically close areas tend to have similar values of the variable of interest, i.e. there exists locally spatially structured variation. Thus, it may be more realistic to construct spatial models on the area-specific effects to induce the correlation among them.

The objective of this paper is to obtain reliable model-based estimate for the disease rate at health regions level within provinces. In section 2, we propose several area level models which relax the strong assumptions described above based on the basic Fay – Herriot model by incorporating spatial structure on the area-specific effects, and/or assuming the sampling variances unknown. In section 3, we obtain the Hierarchical Bayes estimators of the parameter of interest and the posterior variances through the Gibbs sampling method. In section 4, through the data analysis of Cycle 1.1 of CCHS, we compare the performance of the direct design-based estimates with the model-based estimates, and moreover, compare the proposed models with the basic Fay-Herriot model to investigate the effects of incorporating spatial structure on the area-specific effects. Finally in section 5, we offer some conclusions and discussions.

2. SMALL AREA ESTIMATION MODELS

2.1 Fay-Herriot Model

Let θ_i , the area parameter of interest, denote the underlying rate of certain disease for the i th area or health region, where $i = 1, \dots, m$, and m is the total number of areas. A basic area level model assumes that the θ_i is related to area-specific auxiliary data $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ through a linear model

$$\theta_i = \mathbf{x}_i' \boldsymbol{\beta} + v_i, \quad i = 1, \dots, m \quad (1)$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)'$ is the $p \times 1$ vector of regression coefficients, and the v_i 's are area-specific random effects assumed to be independent and identically distributed (i.i.d.) with $E(v_i) = 0$ and $\text{var}(v_i) = \sigma_v^2$. The assumption of normality may also be included. This model is referred to as a linking model for θ_i .

The basic area level model also assumes that a direct survey estimator y_i (usually design-unbiased) of the parameter of interest θ_i is available whenever the area sample size $n_i > 1$. It is customary to assume that

$$y_i = \theta_i + e_i, \quad i = 1, \dots, m \quad (2)$$

where the e_i 's are the sampling error associated with the direct estimator y_i . We also assume that the e_i 's are independent normal random variables with mean $E(e_i | \theta_i) = 0$ and sampling variance $\text{var}(e_i | \theta_i) = \sigma_e^2$. The model (2) is

referred to as a sampling model for the direct survey estimator y_i . Combining these two components (1) and (2) leads to the well-known area level linear mixed model, Fay-Herriot model (Fay & Herriot, 1979)

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + v_i + e_i, \quad i = 1, \dots, m \quad (3)$$

In the basic Fay-Herriot model (3), the sampling variance σ_i^2 are usually assumed as known, which is a very strong assumption, but it is impractical in many cases. Generally, we can use direct sampling variance estimates from the survey data. However, these direct estimates are unstable if sample sizes are small. Therefore, in practice, a smoothed estimator of σ_i^2 is used in the model and treated as known. In the paper by You (2008), equal design effects modeling approach was applied to obtain a smooth estimator of sampling variances. The design effect for the i th area may be approximately written as $deff_i = s_i^2 / s_{ri}^2$, for $i = 1, \dots, m$, where s_i^2 is the unbiased direct estimate of sampling variance based on the complex sampling design, and s_{ri}^2 is the estimate of sampling variance based on the assumption of simple random sampling design. For each area, based on the assumption of a common design effect suggested in You (2008) and Singh, You and Mantel (2005), a smoothed factor $deff$ can be obtained by $deff = \sum_{i=1}^m deff_i / m$. Then a smoothed sampling variance estimate $\tilde{\sigma}_i^2$ can be obtained by $\tilde{\sigma}_i^2 = s_{ri}^2 \cdot deff$.

Instead of plugging in the smoothed estimates of sampling variances in the model, alternatively we can model the sampling variance directly. In the paper by Wang and Fuller (2003) and You and Chapman (2006), they assume the sampling variance σ_i^2 unknown and estimate σ_i^2 by an unbiased direct estimator s_i^2 , which are independent of the direct survey estimator y_i . They also assume that $d_i s_i^2 \sim \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$, and n_i is the sample size for the i th area. You and Chapman (2006) considered the full HB approach with the Gibbs sampling method which automatically takes into account the extra uncertainty associated with the estimation of σ_i^2 . In this paper, we consider both of the smoothing and modeling approaches for the sampling variances.

2.2 Spatial Linking Model

Another strong assumption made in the basic Fay – Herriot model (3) is that the area-specific random effects v_i are i.i.d. normal variables capturing geographically unstructured heterogeneity among areas. To incorporate spatially-correlated effects in the model, Gaussian Markov random fields (MRF) models are the most commonly used when “neighboring” areas can be defined. In the class of MRF model, the conditional distribution of area-specific random effect v_i in area i , given the values of v_j ’s in all other areas $j \neq i$, depends only on the values of the neighboring areas. Thus in this model, area-specific random effects have a locally dependent prior probability structure, their joint distribution is determined (up to a normalizing constant) by these conditional distribution. Leroux et. al. (1999) and MacNab (2000, 2003) proposed the following spatial model on the area-specific random effects $\mathbf{v} = (v_1, \dots, v_m)'$:

$$\mathbf{v} \sim \text{MVN}(\mathbf{0}, \Sigma(\sigma_v^2, \lambda)) \quad (4)$$

$$\Sigma(\sigma_v^2, \lambda) = \sigma_v^2 \mathbf{D}^{-1}, \quad \mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I} \quad (5)$$

where σ_v^2 is a spatial dispersion parameter and λ is a spatial autocorrelation parameter, $0 \leq \lambda \leq 1$; \mathbf{I} is an identity matrix of dimension m ; \mathbf{R} , commonly known as the neighborhood matrix, is an m by m square matrix. Its i th diagonal element equal to the number of neighbors of the area i , given by $w_{i+} = \sum_{j=1}^m w_{ij}$, where $w_{ij} = 1$ if i and j are adjacent areas; $w_{ij} = 0$, otherwise. Its off-diagonal elements in each row equal to -1 if the corresponding two areas are neighbors and 0 otherwise. The spatial model (4) - (5) results in the following conditional distribution of v_i :

$$v_i | v_{-i} \sim \text{N} \left(\frac{\lambda}{1 - \lambda + \lambda w_{i+}} \sum_{j \neq i} w_{ij} v_j, \frac{\sigma_v^2}{1 - \lambda + \lambda w_{i+}} \right), \quad i = 1, \dots, m.$$

Such models are also known as Gaussian conditional autoregressive (CAR) model. The CAR model (4) - (5) becomes the intrinsic autoregressive model (Besag et al., 1991), given by

$$v_i | v_{-i} \sim N\left(\frac{\sum_j w_{ij} v_j}{w_{i+}}, \frac{\sigma_u^2}{w_{i+}}\right),$$

if $\lambda = 1$. On the other hand, if $\lambda = 0$, the CAR model (4) - (5) reduces to the independent linking model (1) which assumes independence on the area-specific effects v_i . It is necessary to point out that the conditional mean and variances of $v_i | v_{-i}$ are weighted sums of the corresponding “global smoothing” moments from the basic Fay-Herriot model and “local smoothing” moments from the intrinsic autoregressive model:

$$\begin{aligned} E(v_i | v_{-i}) &= \frac{1 - \lambda}{1 - \lambda + \lambda w_{i+}} \times 0 + \frac{\lambda w_{i+}}{1 - \lambda + \lambda w_{i+}} \times \left(\sum_{j \neq i} w_{ij} v_j / w_{i+} \right) \\ \text{Var}(v_i | v_{-i}) &= \frac{1 - \lambda}{1 - \lambda + \lambda w_{i+}} \times \sigma_v^2 + \frac{\lambda w_{i+}}{1 - \lambda + \lambda w_{i+}} \times (\sigma_v^2 / w_{i+}) \end{aligned}$$

Thus model (4)-(5) is a balance between the independent linking model and the intrinsic CAR model. The spatial correlation parameter λ measures the extent of the spatial effects for “local smoothing” of the neighboring areas. The modeling structure (5) captures both the unstructured heterogeneity among areas and the spatial correlation effects of the neighboring area.

3. HIERARCHICAL BAYES INFERENCE

In order to estimate θ_i , the parameter of interest, we apply a hierarchical Bayes (HB) approach using the Gibbs sampling method. Compared to other approaches such as EBLUP and empirical Bayes (EB), HB approach is straightforward and the inference for θ_i are “exact” unlike the EB or EBLUP. Moreover, HB approach can deal with complex small area models using Monte Carlo Markov Chain (MCMC) method, which overcomes the computational difficulties of multi-dimensional integrations of posterior quantities to a large extent.

Let $\mathbf{y} = (y_1, \dots, y_m)'$, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_m)'$, and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)'$. We first construct two HB models without and with spatial structure under the assumption that the sampling variance σ_i^2 are assumed known and replaced by the smoothed estimate $\tilde{\sigma}_i^2$.

Model 1: Fay-Herriot model

- $y_i | \theta_i \sim N(\theta_i, \sigma_i^2 = \tilde{\sigma}_i^2)$, for $i = 1, \dots, m$;
- $\theta_i | \beta, \sigma_v^2 \sim N(\mathbf{x}_i' \beta, \sigma_v^2)$, for $i = 1, \dots, m$;
- Priors for the parameters (β, σ_v^2) : $\pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$, where a_0, b_0 are chosen to be very small known constants to reflect vague knowledge on σ_v^2 . N stands for normal distribution and IG for inverse gamma distribution.

Model 2: Proposed area level CAR model

- $\mathbf{y} | \boldsymbol{\theta} \sim \text{MVN}(\boldsymbol{\theta}, \mathbf{E})$, where \mathbf{E} is a diagonal matrix with the i th diagonal element $\sigma_i^2 = \tilde{\sigma}_i^2$;
- $\boldsymbol{\theta} | \beta, \sigma_v^2 \sim \text{MVN}(\mathbf{X}\beta, \sigma_v^2 \mathbf{D}^{-1})$, where $\mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I}$, with \mathbf{I} , an identity matrix of dimension m , and \mathbf{R} , the neighborhood matrix;
- Priors for the parameters $(\beta, \lambda, \sigma_v^2)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \text{Uniform}(0, 1)$, where $0 \leq \lambda \leq 1$; $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$, where a_0, b_0 are chosen to be very small known constants. MVN stands for the multivariate normal distribution.

We also consider two HB models with the sampling variance σ_i^2 unknown and modeled by the direct unbiased estimator s_i^2 .

Model 3: Fay-Herriot model with unknown sampling variances (You and Chapman, 2006)

- $y_i | \theta_i, \sigma_i^2 \sim N(\theta_i, \sigma_i^2)$, for $i = 1, \dots, m$;
- $d_i s_i^2 | \sigma_i^2 \overset{ind}{\sim} \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$, for $i = 1, \dots, m$;
- $\theta_i | \beta, \sigma_v^2 \sim N(\mathbf{x}'_i \beta, \sigma_v^2)$, for $i = 1, \dots, m$;
- Priors for the parameters $(\beta, \sigma_v^2, \sigma_i^2, i = 1, \dots, m)$: $\pi(\beta) \propto 1$; $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$, $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i)$ for $i = 1, \dots, m$, where a_i, b_i ($0 \leq i \leq m$) are chosen to be very small known constants to reflect vague knowledge on σ_i^2 and σ_v^2 .

Model 4: Proposed area level CAR model with unknown sampling variances

- $\mathbf{y} | \boldsymbol{\theta}, \sigma_1^2, \dots, \sigma_m^2 \sim \text{MVN}(\boldsymbol{\theta}, \mathbf{E})$, where \mathbf{E} is a diagonal matrix with the i th diagonal element σ_i^2 ;
- $d_i s_i^2 | \sigma_i^2 \overset{ind}{\sim} \sigma_i^2 \chi_{d_i}^2$, where $d_i = n_i - 1$, for $i = 1, \dots, m$;
- $\boldsymbol{\theta} | \beta, \sigma_v^2 \sim \text{MVN}(\mathbf{X}\beta, \sigma_v^2 \mathbf{D}^{-1})$, where $\mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I}$;
- Priors for the parameters $(\beta, \lambda, \sigma_v^2, \sigma_i^2, i = 1, \dots, m)$: $\pi(\beta) \propto 1$; $\pi(\lambda) \sim \text{Uniform}(0, 1)$, where $0 \leq \lambda \leq 1$; $\pi(\sigma_v^2) \sim \text{IG}(a_0, b_0)$; $\pi(\sigma_i^2) \sim \text{IG}(a_i, b_i)$ for $i = 1, \dots, m$, where a_i, b_i ($0 \leq i \leq m$) are chosen to be very small known constants.

In the HB approach, we use the posterior mean $E(\theta_i | \mathbf{y})$ as a point estimate of θ_i and the posterior variance $\text{Var}(\theta_i | \mathbf{y})$ as a measure of variability. We implement Gibbs sampling method (Gelfand and Smith, 1990) by drawing samples of the parameters from the joint posterior.

For Model 1, the full conditional distributions of $(\theta_1, \dots, \theta_m, \beta, \sigma_v^2)$ for the Gibbs sampler are:

- $[\theta_i | y_i, \beta, \sigma_v^2] \sim N[\gamma_i y_i + (1 - \gamma_i) \mathbf{x}'_i \beta, \tilde{\sigma}_i^2 \gamma_i]$, where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \tilde{\sigma}_i^2)$, for $i = 1, \dots, m$;
- $[\beta | \boldsymbol{\theta}, \sigma_v^2] \sim N\left[\left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \left(\sum_{i=1}^m \mathbf{x}_i \theta_i\right), \sigma_v^2 \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}'_i\right)^{-1}\right]$;
- $[\sigma_v^2 | \boldsymbol{\theta}, \beta] \sim \text{IG}\left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2} \sum_{i=1}^m (\theta_i - \mathbf{x}'_i \beta)^2\right]$.

For Model 2, the full conditional distributions of $(\boldsymbol{\theta}, \beta, \lambda, \sigma_v^2)$ for the Gibbs sampler are:

- $[\boldsymbol{\theta} | \mathbf{y}, \beta, \lambda, \sigma_v^2] \sim \text{MVN}(\boldsymbol{\Lambda} \mathbf{y} + (\mathbf{I} - \boldsymbol{\Lambda}) \mathbf{X} \beta, \boldsymbol{\Lambda} \mathbf{E})$, where $\boldsymbol{\Lambda} = (\mathbf{E}^{-1} + \mathbf{D} / \sigma_v^2)^{-1} \mathbf{E}^{-1}$ with $\mathbf{E} = \text{diag}\{\tilde{\sigma}_1^2, \dots, \tilde{\sigma}_m^2\}$ and $\mathbf{D} = \lambda \mathbf{R} + (1 - \lambda) \mathbf{I}$;
- $[\beta | \boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \text{MVN}\left[(\mathbf{X}' \mathbf{D} \mathbf{X})^{-1} \mathbf{X}' \mathbf{D} \boldsymbol{\theta}, \sigma_v^2 (\mathbf{X}' \mathbf{D} \mathbf{X})^{-1}\right]$;
- $[\lambda | \boldsymbol{\theta}, \beta, \sigma_v^2] \propto |[\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}]^{-1}|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma_v^2} (\boldsymbol{\theta} - \mathbf{X} \beta)' [\lambda \mathbf{R} + (1 - \lambda) \mathbf{I}] (\boldsymbol{\theta} - \mathbf{X} \beta)\right\}$;
- $[\sigma_v^2 | \boldsymbol{\theta}, \beta, \lambda] \sim \text{IG}\left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2} (\boldsymbol{\theta} - \mathbf{X} \beta)' \mathbf{D} (\boldsymbol{\theta} - \mathbf{X} \beta)\right]$.

For Model 3, the full conditional distributions of $(\theta_1, \dots, \theta_m, \beta, \sigma_v^2, \sigma_1^2, \dots, \sigma_m^2)$ for the Gibbs sampler are:

- $[\theta_i | y_i, \beta, \sigma_i^2, \sigma_v^2] \sim N[\gamma_i y_i + (1 - \gamma_i) \mathbf{x}'_i \beta, \sigma_i^2 \gamma_i]$, where $\gamma_i = \sigma_v^2 / (\sigma_v^2 + \sigma_i^2)$, for $i = 1, \dots, m$;
- $[\beta | \boldsymbol{\theta}, \sigma_v^2] \propto N\left[\left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}'_i\right)^{-1} \left(\sum_{i=1}^m \mathbf{x}_i \theta_i\right), \sigma_v^2 \left(\sum_{i=1}^m \mathbf{x}_i \mathbf{x}'_i\right)^{-1}\right]$;
- $[\sigma_i^2 | y_i, \theta_i] \sim \text{IG}\left(a_i + \frac{d_i + 1}{2}, b_i + \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2}\right)$, where $d_i = n_i - 1$, for $i = 1, \dots, m$;

- $[\sigma_v^2 | \boldsymbol{\theta}, \boldsymbol{\beta}] \sim \text{IG}\left[a_0 + \frac{1}{2}m, b_0 + \frac{1}{2}\sum_{i=1}^m (\theta_i - \mathbf{x}'_i \boldsymbol{\beta})^2\right]$.

For Model 4, the full conditional distributions of $(\boldsymbol{\theta}, \boldsymbol{\beta}, \lambda, \sigma_v^2, \sigma_1^2, \dots, \sigma_m^2)$ for the Gibbs sampler are:

- $[\boldsymbol{\theta} | \mathbf{y}, \boldsymbol{\beta}, \lambda, \sigma_v^2, \sigma_1^2, \dots, \sigma_m^2] \sim \text{MVN}(\boldsymbol{\Lambda}\mathbf{y} + (\mathbf{I} - \boldsymbol{\Lambda})\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Lambda}\mathbf{E})$, where $\boldsymbol{\Lambda} = (\mathbf{E}^{-1} + \mathbf{D}/\sigma^2)^{-1}\mathbf{E}^{-1}$ with $\mathbf{E} = \text{diag}\{\sigma_1^2, \dots, \sigma_m^2\}$ and $\mathbf{D} = \lambda\mathbf{R} + (1 - \lambda)\mathbf{I}$;
- $[\boldsymbol{\beta} | \boldsymbol{\theta}, \lambda, \sigma_v^2] \sim \text{MVN}\left[(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\theta}, \sigma_v^2(\mathbf{X}'\mathbf{D}\mathbf{X})^{-1}\right]$;
- $[\lambda | \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2] \propto |[\lambda\mathbf{R} + (1 - \lambda)\mathbf{I}]^{-1}|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})'[\lambda\mathbf{R} + (1 - \lambda)\mathbf{I}](\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})\right\}$
- $[\sigma_i^2 | y_i, \theta_i] \sim \text{IG}\left(a_i + \frac{d_i + 1}{2}, b_i + \frac{(y_i - \theta_i)^2 + d_i s_i^2}{2}\right)$, where $d_i = n_i - 1$, for $i = 1, \dots, m$;
- $[\sigma_v^2 | \boldsymbol{\theta}, \boldsymbol{\beta}, \lambda] \sim \text{IG}\left[a_0 + \frac{m}{2}, b_0 + \frac{1}{2}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})' \mathbf{D}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})\right]$.

The distributions of $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, σ_i^2 and σ_v^2 are standard multivariate normal or inverse gamma distributions that can be easily sampled. However, the conditional distribution of λ does not have a closed form. We use the Metropolis-Hastings algorithm within the Gibbs sampler (Chip and Greenberg, 1995) to update λ . The full conditional distribution of λ in the Gibbs sampler can be written as $[\lambda | \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma_v^2] \propto h(\lambda)f(\lambda)$, where $f(\lambda)$ is a density function of the uniform distribution $\text{Uniform}(0,1)$ given as $f(\lambda) \propto 1$, where $0 \leq \lambda \leq 1$, and $h(\lambda)$ is a function given by

$$h(\lambda) \propto |[\lambda\mathbf{R} + (1 - \lambda)\mathbf{I}]^{-1}|^{-\frac{1}{2}} \times \exp\left\{-\frac{1}{2\sigma_v^2}(\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})'[\lambda\mathbf{R} + (1 - \lambda)\mathbf{I}](\boldsymbol{\theta} - \mathbf{X}\boldsymbol{\beta})\right\}.$$

We use $f(\lambda)$ as the ‘‘candidate’’ generating density function in the Metropolis-Hastings updating step. To implement the Gibbs sampling, we use $L=5$ parallel runs each with a ‘‘burn-in’’ length of $B=2000$ and Gibbs sampling size of $G=5000$. For Model 2 and Model 4, in order to reduce the autocorrelation which results from the accept-rejection algorithm in the run, we take every 5th iteration after the ‘‘burn-in’’ period. Therefore, for Model 1 and Model 3, we have $n=5000$ samples for each run, and for Model 2 and Model 4, $n=1000$ samples for each run.

In the following section, we apply the proposed four HB models in Section 3 to estimate disease rates for health regions using CCHS Cycle 1.1 data.

4. DATA ANALYSIS

The health region-level survey of Cycle 1.1 consists of common content to meet basic health data requirements on an ongoing basis. The questionnaire contains the information about several chronic health conditions, including food allergies, asthma, arthritis or rheumatism, diabetes and etc. In this paper, we are interested in estimating the rate of asthma in the 20 health regions of BC province. Based on the map, for each health region, we define the corresponding adjacent health regions. It is necessary to point out that the two health regions Capital and Vancouver are not adjacent since they are separated by the ocean. However, due to the intensive connection between these two areas on transportation, economics, tourism and other aspects, it is reasonable to define that they are neighbors as well.

From the survey data of Cycle 1.1, we obtained eight variables for each health region to estimate the rate of asthma as follows: (1) sample size, (2) direct estimate of the number of persons who have asthma, (3) total population size, (4) number of persons who have asthma as one of the symptoms of the chronic disease, (5) number of persons who have asthma as the main symptom of the chronic disease, (6) number of persons who have diabetes as one of the symptoms of the chronic disease, (7) number of persons who have diabetes as the main symptom of the chronic disease, and (8) number of visits to hospitals. For each health region, the direct estimate y_i of the rate of asthma θ_i is obtained as the ratio of number of people having asthma over the corresponding population size, for $i = 1, \dots, m$. The six variables 3, 4, 5, 6, 7, and 8 are used as the area-specific auxiliary data $\mathbf{x}_i = (x_{i1}, \dots, x_{i6})'$.

In the literatures of disease mapping (e.g., Mollié, 1996; Maiti, 1998; MacNab 2003), Poisson or Binomial distribution is usually assumed in the sampling model for the direct estimator y_i . However, in our application, the direct estimator y_i is obtained based on the complex sampling design conducted in the survey. Thus, it is more reasonable to assume normal approximation on the direct estimator y_i .

Firstly we present the HB estimates of the rate of asthma under the Model 1 and 2 in which the sampling variances σ_i^2 are assumed as known. We use the smoothed estimate $\tilde{\sigma}_i^2$ obtained by the smoothing technique in You (2008) described in Section 2. Figure 1 displays the direct estimates and the HB model-based estimates from Model 1 and Model 2 for the 20 health regions in BC province. The health regions appear in the order of sample size with the largest (South Fraser Valley) on the left and the smallest (Peace Liard) on the right. Model 1 (Fay-Herriot model) and Model 2 (CAR model) give the similar point estimates, and both the model-based estimates lead to moderate smooth estimates compared to the direct estimates. Moreover, the direct estimates and two HB estimates of the disease rate are very close for some health regions with large sample sizes, but for some areas with smaller sample sizes, they differ to some extent.

Figure 1: Comparison of direct and HB model-based estimates under the Fay-Herriot model 1 and CAR model 2

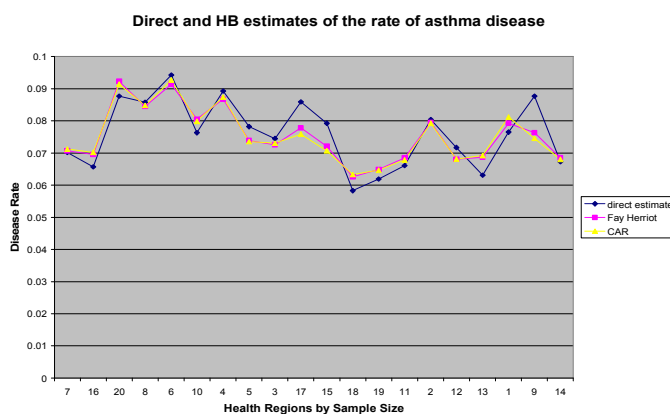


Figure 2 presents the CVs of the direct and two HB model-based estimates with the health regions ordered by the sample sizes from the largest to smallest. The CVs of HB estimates are obtained by dividing the squared root of the posterior variance by the posterior mean. As expected, the CVs of the direct estimates show a clear tendency of increasing as the sample size decreases, which demonstrates the unreliability of direct estimates in the areas with small sample sizes. However, the two model-based estimates give smoother CVs. Moreover, the two HB model-based estimates exhibit a great improvement over the direct design-based estimates in terms of precision and reliability, that is, smaller CVs. The average CV reduction of the HB estimates under Model 1 (Fay-Herriot model) is about 22.7% with the range of 7.8% to 40.5%, and the average reduction of the CVs for the HB estimates under Model 2 (CAR model) is 27.8% with the range of 12.5% to 52.1%, compared to the direct estimates.

Figure 2: Comparison of direct and HB CVs under the Fay-Herriot model 1 and CAR model 2 with the health regions sorted by the sample size

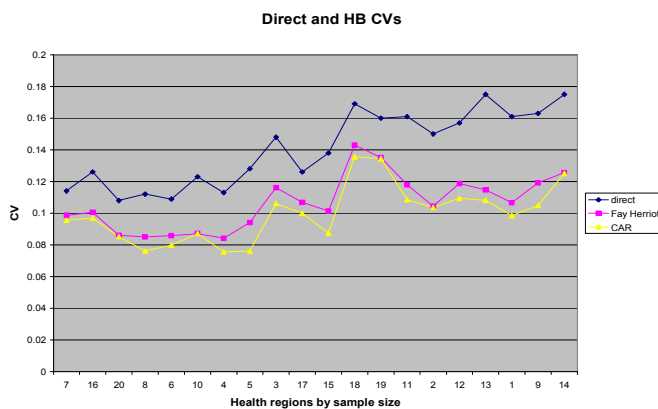


Figure 3 also displays the CVs of the direct and HB estimates, while the health regions are sorted by the number of neighboring regions from largest to smallest in order to investigate the effects of incorporating the spatial structure in the model. It shows that the HB estimates from the CAR Model 2 has smaller CVs than the estimates from the basic Fay-Herriot model. In addition, the improvement of the CAR model over the Fay-Herriot model is stronger in the areas with more neighbors, but these two models give very close CVs in the regions with less adjacent areas. Table 1 lists the reduction of the CVs under the CAR model over the Fay-Herriot model across the health regions with the same number of neighbors. The results in Table 1 present the CV reduction of the CAR model for both the cases of known and unknown sampling variances. For example, for known σ_i^2 (smoothed $\tilde{\sigma}_i^2$), for areas with only 2 neighbors, the average CV reduction of CAR model over the Fay-Herriot model is only around 0.9%, whereas for areas with 7 neighbors, the average CV reduction for CAR model is as high as around 20%. The numerical results also confirm the clear trend of more CV reduction under the CAR model over the Fay-Herriot model as the number of neighbors increases. Thus, more neighboring areas provide more information in the spatial structure to improve the precision and reliability of the HB estimates.

Figure 3: Comparison of direct and HB CVs under the Fay-Herriot model 1 and CAR model 2 with the health regions sorted by the number of neighbors

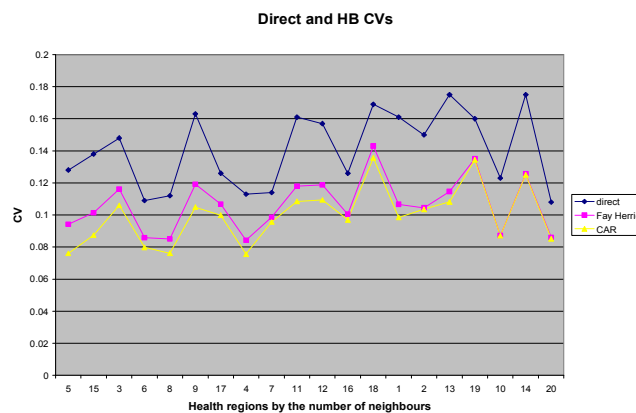


Table 1: The average CV reduction of CAR model over the Fay Herriot model

The number of neighbours	The average CV reduction	
	σ_i^2 known	σ_i^2 unknown
7	19.2%	20.7%
6	13.7%	11.0%
5	8.9%	8.7%
4	6.3%	6.0%
3	3.7%	3.5%
2	0.9%	1.8%

We also obtained point estimates and CVs under Model 3 and Model 4. The comparison of the results shows a great agreement with the comparison among the direct estimates and two HB model-based estimates under the Model 1 and 2.

4. CONCLUSION AND DISCUSSION

In this paper we have studied the well-known Fay-Herriot model in which two strong assumptions are made. One is that the sampling variances σ_i^2 are assumed as known. You (2008) used the smoothed variances $\tilde{\sigma}_i^2$ obtained by the equal design effects modeling approach, and You and Chapman (2006) instead modeled the sampling variances σ_i^2 directly by the unbiased estimator s_i^2 . The other assumption is that the area-specific random effects are assumed independent and identically distributed. Gaussian CAR model was proposed in the literature for disease mapping to incorporate spatially-correlated effects. According to the previous work, we propose four HB models which relax these two strong assumptions

to investigate the effect of including the geographically structured distribution in the model where the sampling variances σ_i^2 are replaced by the smoothed estimate $\tilde{\sigma}_i^2$ or modeled by the direct estimator s_i^2 .

In the data analysis which aims at estimating the rate of asthma for the 20 health regions in BC provinces, the model-based estimates achieve a great improvement over the direct estimates in terms of moderately smoothed point estimates and much smaller CVs. In addition, we find that whenever the sampling variances are assumed as known or unknown, the proposed area level CAR models have smaller CVs than the Fay-Herriot model which imposes independent area-specific random effects. Moreover, the CV reduction of CAR model over the Fay-Herriot model is greater for the areas with more neighbors.

One possible limitation of our proposed model is that the linking model for the disease rate θ_i is a linear model with normal random effects. Since θ_i takes value between 0 and 1, and it is close to 0 for some rare disease, the linear linking model with normal random effects may lead to negative estimates for θ_i for some small areas in practice if the sampling variances vary substantially. You and Rao (2002) proposed a log-linear linking model for the Fay-Herriot model as the unmatched sampling and linking models as follows:

$$y_i = \theta_i + e_i, \quad i = 1, \dots, m$$
$$\log(\theta_i) = \mathbf{x}_i' \boldsymbol{\beta} + v_i, \quad i = 1, \dots, m$$

In the future work, the proposed CAR models can be extended to the unmatched sampling and linking models with the sampling variance known or unknown. We will also plan to evaluate the estimation effects of different spatial models (e.g., Best, Richardson and Thomson, 2005) as well as the effects of spatial structures. We also plan to study different methods to test the overall fit of the proposed models, and also to assess model fit at the individual area level. For data analysis, we will produce model-based health status estimates based on the proposed models for health regions across Canada and evaluate the possibility of extending the model-based approach to lower level estimates such as age-sex domains within health regions. One way to do this is to extend the area level models to unit level model with spatial correlation structure between area level variations.

REFERENCES

- Béland, Y. (2002). Canadian Community Health Survey Methodological overview. Health report, Statistics Canada, Catalogue no. 82-003-XPE, Vol. 13, No. 3, ISSN 0840-6529.
- Besag, J., York, J. and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43, 1-59 (with discussion).
- 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.
- Chip, S. and Greenberg, E. (1995). Understanding the Metropolis-Hastings algorithm. *The American Statistician*, 49, 327-335.
- 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.
- Gelfand, A. E. and Smith, A. F. M. (1991). Gibbs sampling for marginal posterior expectations. *Communications In Statistics – Theory and Methods*, 20, 1747-1766.
- Leroux, B. G., Lei, X., Breslow, N. (1999). Estimation of disease rates in small areas : a new mixed model for spatial dependence. In *Statistical Models in Epidemiology, the Environment and Clinical Trials*, Halloran ME, Berry D (eds). Springer-Verlag: New York, 135-178.
- MacNab, Y. C., Dean, C. B. (2000). *Parametric bootstrap and penalized quaslikelihood inference in conditional autoregressive models*. *Statistics in Medicine*, 19(17/18), 2421-2436.

- MacNab, Y. C. (2003). Hierarchical Bayesian spatial modeling of small-area rates of non-rare disease. *Statistics in Medicine*, 22, 1761-1773.
- Maiti, T. (1998). Hierarchical Bayes estimation of mortality rates for disease mapping. *Journal of Statistical Planning and Inference*, 69, 339-348.
- Mollié, A. (1996). Bayesian mapping of disease. In Gilks, W.R., Richardson, S., Spiegelhalter, D.J., editors. In *Markov Chain Monte Carlo in Practice*. London: Chapman and Hall, 359-379.
- Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley & Sons, New York.
- 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.
- 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. (2006). Model-based small area unemployment rate estimation for the Canadian Labour Force Survey. Methodology Branch working paper, HSMD-2006-004E, Statistics Canada.
- 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*, 20, 3-15.