

Small area estimation of general finite-population parameters based on grouped data

Yuki Kawakubo* and Genya Kobayashi*

Abstract

This paper proposes a new model-based approach to small area estimation of general finite-population parameters based on grouped data or frequency data, which is often available from sample surveys. Grouped data contains information on frequencies of some pre-specified groups in each area, for example the numbers of households in the income classes, and thus provides more detailed insight about small areas than area-level aggregated data. A direct application of the widely used small area methods, such as the Fay–Herriot model for area-level data and nested error regression model for unit-level data, is not appropriate since they are not designed for grouped data. The newly proposed method adopts the multinomial likelihood function for the grouped data. In order to connect the group probabilities of the multinomial likelihood and the auxiliary variables within the framework of small area estimation, we introduce the unobserved unit-level quantities of interest which follows the linear mixed model with the random intercepts and dispersions after some transformation. Then the probabilities that a unit belongs to the groups can be derived and are used to construct the likelihood function for the grouped data given the random effects. The unknown model parameters (hyperparameters) are estimated by a newly developed Monte Carlo EM algorithm using an efficient importance sampling. The empirical best predicts (empirical Bayes estimates) of small area parameters can be calculated by a simple Gibbs sampling algorithm. The numerical performance of the proposed method is illustrated based on the model-based and design-based simulations. In the application to the city level grouped income data of Japan, we complete the patchy maps of the Gini coefficient as well as mean income across the country.

Keywords: Grouped data; Latent variables; Mixed effects model; Monte Carlo; Small area estimation.

1 Introduction

Sample surveys are generally designed to estimate finite population parameters, such as total, mean, variance and quantiles. On the other hand, decision makers of both public and private agencies have become interested in such parameters for smaller subpopulation (small area) as well, created by cross classifying geographical and demographical variables, such as age, sex and race. However, direct survey estimators of small area parameters, sample mean, sample variance, sample quantiles and others, are often unstable and unreliable because the sample size for each area is too small mainly due to the budget constraint. In order to obtain more reliable estimators of small area parameters, the model-based approach which uses mixed effects models is becoming popular. The empirical best predictor or empirical Bayes estimator derived from

*Graduate School of Social Sciences, Chiba University, 1-33, Yayoi-cho, Inage-ku, Chiba, 263-8522, Japan, (E-mail: {kawakubo, gkobayashi}@chiba-u.jp)

mixed effects models, which is often called model based estimator, is more stable than the direct survey estimator because the model-based estimator borrows strength from other areas through the statistical model which connects across the areas with auxiliary variables from other data sources such as large-scale sample surveys and population census. Alternatively, the hierarchical Bayes approach to the model-based method has been also discussed in the literature. For the detail about small area estimation (SAE), see Datta and Ghosh (2012), Pfeiffermann (2013), Rao and Molina (2015) and others. There are two fundamental models for model-based SAE: the Fay–Herriot model for area-level aggregated data, which was first proposed to estimate the per capita income for small areas by Fay and Herriot (1979), and the nested error regression model for unit-level data (Battese et al., 1988). While only one population parameter, such as an areal mean, can be estimated at a time by using Fay–Herriot model, general finite population parameters can be estimated by using the nested error regression model and its extensions, proposed by Molina and Rao (2010), Guadarrama et al. (2018), Diallo and Rao (2018), Sugasawa and Kubokawa (2019) and others provided that a unit-level data is available. However, the Fay–Herriot model is more widely used in practice as the accessibility of unit-level data is limited in many cases.

Along with area-level aggregated measures of quantities of interest, as sample mean, sample surveys frequently report grouped data. Grouped data contains information on frequency distributions based on some predefined groups in each area and thus provides more insight about areas than an aggregated areal measure. The need to model for and to analyze a grouped data arises in many fields of statistical analysis and there exist theoretical developments regarding the grouped data analysis, see Heitjan (1989) and references therein. Especially in the analysis of income data, the individual households often are grouped into some predefined income classes (Chotikapanich, 2008). For example, Housing and Land Survey (HLS) conducted by Statistics Bureau of Japan in 2013 reports the numbers of households that fall into the five and nine income classes over 1265 municipalities. The grouped data literature, mainly from the view point of the income data analysis, predominantly focused on developing a more flexible underlying parametric or semiparametric form for a single nation, region or period. However, when we face the grouped data over multiple local areas as in the HLS data, the existing grouped data methods do not suffice. This is because the reported frequency distributions are based on the survey sampling, they are not reliable for areas with small sample sizes and thus call for a correction through an SAE method. It must be noted that none of the existing SAE methods can be used to reduce uncertainty in grouped data, because grouped data do not contain unit-level information that is required in the nested error regression model and an appropriate direct estimator that can be used in the Fay–Herriot model is difficult to define for many small area parameters. Therefore a new SAE method specifically designed for grouped data is required.

In this paper, we develop a new model-based SAE method which explicitly takes frequency distributions observed in grouped data into account and can estimate general finite population parameters including areal means. Since the frequency distribution in the grouped data counts the number of units that fall into each group, the multinomial likelihood function is adopted. We introduce the latent unit-level variables that represent the unit-level quantities of interest and that are supported within the range of each group. Then in order to connect the frequency distribution to the auxiliary variables within the SAE framework, these latent unit-level variables are assumed to follow a linear mixed model after some transformation. The linear mixed model adopts the random dispersion as well as random intercept, because the frequency distribution of each area provides the information on the scale of the distribution. While Jiang and Nguyen (2012) and Kubokawa et al. (2016) considered the heteroskedasticity in SAE, they did not consider the grouped data setting. Given the random effects, the probabilities that a unit

belongs to the groups can be derived and are used to construct the multinomial likelihood function for the grouped data. The unknown model parameters (hyperparameters) are estimated by maximizing the marginal likelihood which integrates out the random effects. Since the marginal likelihood cannot be evaluated analytically, we develop an EM algorithm (Dempster et al., 1977), where the E-step is carried out by Monte Carlo integration based on the sampling importance resampling (SIR) using an efficient importance sampling technique. After obtaining the estimates of hyperparameters, the empirical Bayes (EB) or equivalently empirical best predicts, of small area parameters, such as areal means and Gini coefficients, are easily calculated using the output from a simple Gibbs sampler, where the unobserved unit-level quantities are augmented as latent variables to simulate the finite population.

The rest of the paper is organized as follows. Section 2 describes the proposed model and methods for hyperparameter estimation and calculation of EB estimates. Section 3 presents the application of the proposed method to Japanese income dataset from HLS. The patchy maps of the areal mean income and Gini coefficient are completed using our method. In Section 4, the performance of the proposed model is examined through the model-based and design-based simulation studies. Finally, Section 5 concludes the paper with some discussion.

2 Proposed method

2.1 Model description

In each of m areas, we observe the grouped data that provides the frequency distribution over the mutually exclusive G groups divided by the known thresholds $0 = c_0 < c_1 < \dots < c_{G-1} < c_G = +\infty$. Let us denote the observed frequencies and sample size in the i th area by $\mathbf{y}_i = (y_{i1}, \dots, y_{iG})^\top$ for $i = 1, \dots, m$ and $n_i = \sum_{g=1}^G y_{ig}$, respectively, and thus y_{ig} counts the number of units that fall into the g th group in the i th area. Therefore, it can be regarded that \mathbf{y}_i follows the multinomial distribution. In order to model the group probabilities of the multinomial distribution that links the grouped data with the auxiliary variables and then to facilitate the small area parameter estimation (see Section 2.3), we introduce the positive latent variable $z_{ij} > 0$ for the j th unit in the i th area ($i = 1, \dots, m; j = 1, \dots, N_i$) that constitutes the population of the i th area and from which the units are sampled to construct the grouped data. We also let $\mathbf{z}_i = (z_{i1}, \dots, z_{iN_i})^\top$. Note that N_i is not the sample size but the population size and thus a finite population setting is considered. Without loss of generality, it is assumed that the first n_i values of z_{ij} 's are sampled. Then y_{ig} can be expressed as

$$y_{ig} = \sum_{j=1}^{n_i} I(c_{g-1} \leq z_{ij} < c_g), \quad (g = 1, \dots, G), \quad (1)$$

where $I(\cdot)$ is the indicator function. We take into account the variability of the frequency distribution by incorporating the sample size into our model.

In order to devise small area estimation for the grouped data, we assume that the latent z_{ij} after some transformation follows the linear mixed model:

$$\begin{aligned} h_\kappa(z_{ij}) &= \mathbf{x}_i^\top \boldsymbol{\beta} + b_i + \varepsilon_{ij}, \quad b_i \sim N(0, \tau^2), \\ \varepsilon_{ij} \mid \sigma_i^2 &\sim N(0, \sigma_i^2), \quad \sigma_i^2 \sim \text{IG}\left(\frac{\lambda}{2} + 1, \frac{\lambda \varphi_i}{2}\right), \quad \varphi_i = \exp(\mathbf{x}_i^\top \boldsymbol{\gamma}), \end{aligned} \quad (2)$$

or equivalently the following Bayesian model:

$$\begin{aligned} h_\kappa(z_{ij}) \mid \mu_i, \sigma_i^2 &\sim \text{N}(\mu_i, \sigma_i^2) \\ \mu_i &\sim \text{N}(\mathbf{x}_i^\top \boldsymbol{\beta}, \tau^2) \\ \sigma_i^2 &\sim \text{IG}\left(\frac{\lambda}{2} + 1, \frac{\lambda \varphi_i}{2}\right), \quad \varphi_i = \exp(\mathbf{x}_i^\top \boldsymbol{\gamma}), \end{aligned} \quad (3)$$

where $h_\kappa(\cdot)$ is an arbitrary parametric transformation with the parameter κ , \mathbf{x}_i is the area specific p -dimensional auxiliary variable vector, $\boldsymbol{\beta}$ is the unknown parameter vector of regression coefficients, b_i is the random area effect with the unknown variance parameter τ^2 and ε_{ij} is the error term with the area specific random variance σ_i^2 . It is further assumed that b_i 's and σ_i^2 's are mutually independent or equivalently μ_i 's and σ_i^2 's are mutually independent and that z_{ij} 's are conditionally independent given $\mathbf{b} = (b_1, \dots, b_m)^\top$ and $\boldsymbol{\sigma} = (\sigma_1^2, \dots, \sigma_m^2)^\top$. The mean of σ_i^2 is φ_i which is further modeled as $\varphi_i = \exp(\mathbf{x}_i^\top \boldsymbol{\gamma})$ using the auxiliary variables. While the model looks like a version of unit-level nested error regression model proposed in the small area estimation literature (Battese et al., 1988), there is a crucial difference that in the present setting we do not observe the unit-level \mathbf{z}_i 's but \mathbf{y}_i 's only. Also, the auxiliary variables \mathbf{x}_i are available only at the area-level.

Based on the statistical model (2) or (3), the conditional probability that z_{ij} falls in the g th group given b_i (or μ_i) and σ_i^2 is given by

$$\Pr(c_{g-1} \leq z_{ij} < c_g \mid b_i, \sigma_i^2) = \Phi\left\{\frac{h_\kappa(c_g) - \mu_i}{\sigma_i}\right\} - \Phi\left\{\frac{h_\kappa(c_{g-1}) - \mu_i}{\sigma_i}\right\}, \quad (4)$$

where $\mu_i = \mathbf{x}_i^\top \boldsymbol{\beta} + b_i$ and $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution.

Note that we model the unit-level variable z_{ij} , not the area-level variable like the Fay–Herriot model. However, the auxiliary variables are available only on the area-level. Hence, if the log transformation is used, the superpopulation of z_{ij} is the log-normal distribution with the same mean and variance within the same small area i , which is too restrictive. In this paper, a more flexible parametric transformation $h_\kappa(\cdot)$ is adopted to relax the restriction. Specifically, we use the Box–Cox transformation given by

$$h_\kappa(z) = \begin{cases} \frac{z^\kappa - 1}{\kappa}, & \kappa \neq 0, \\ \log(z), & \kappa = 0, \end{cases} \quad z > 0,$$

and $-1/\kappa < h_\kappa(z) < +\infty$ if $\kappa > 0$ and $-\infty < h_\kappa(z) < -1/\kappa$ if $\kappa < 0$.

Our goal is to estimate (predict) some characteristics of each area, such as the areal mean $\bar{z}_i = N_i^{-1} \sum_{j=1}^{N_i} z_{ij}$ and Gini coefficient defined as

$$\text{GINI}(\mathbf{z}_i) = \frac{1}{N_i} \left\{ N_i + 1 - \frac{2 \sum_{j=1}^{N_i} (N_i + 1 - j) z_{i(j)}}{N_i \bar{z}_i} \right\}, \quad (5)$$

where $\{z_{i(1)}, \dots, z_{i(N_i)}\}$ are sorted values of $\{z_{i1}, \dots, z_{iN_i}\}$ in non-decreasing order. To this end, we develop the empirical Bayes (EB) estimators of \bar{z}_i and $\text{GINI}(\mathbf{z}_i)$.

2.2 Hyperparameter estimation

The unknown model parameter vector is denoted by $\boldsymbol{\psi} = (\boldsymbol{\beta}^\top, \tau^2, \lambda, \kappa, \boldsymbol{\gamma}^\top)^\top$. If our model is seen as a Bayesian model (3), $\boldsymbol{\psi}$ is referred to as hyperparameters. Hereafter, $\boldsymbol{\psi}$ is referred to as the hyperparameters for the sake of clarity of terminology.

The hyperparameter $\boldsymbol{\psi}$ is estimated by maximizing the marginal likelihood:

$$L(\boldsymbol{\psi}; \mathbf{y}) = \prod_{i=1}^m \int f(\mathbf{y}_i | \mathbf{u}_i) \pi(\mathbf{u}_i) d\mathbf{u}_i, \quad (6)$$

where $\pi(\mathbf{u}_i)$ is the pdf of $\mathbf{u}_i = (b_i, \sigma_i^2)^\top \sim \text{N}(0, \tau^2) \times \text{IG}(\lambda/2 + 1, \lambda\varphi_i/2)$, and $f(\mathbf{y}_i | \mathbf{u}_i)$ is the conditional probability mass function (pmf) of \mathbf{y}_i given \mathbf{u}_i , which is given by the pmf of the multinomial distribution with n_i trials and the probabilities given by (4):

$$f(\mathbf{y}_i | \mathbf{u}_i) = \frac{n_i!}{y_{i1}! y_{i2}! \cdots y_{iG}!} \times \prod_{g=1}^G \left[\Phi \left\{ \frac{h_\kappa(c_g) - \mu_i}{\sigma_i} \right\} - \Phi \left\{ \frac{h_\kappa(c_{g-1}) - \mu_i}{\sigma} \right\} \right]^{y_{ig}}, \quad (7)$$

for $i = 1, \dots, m$. It is difficult to evaluate the marginal likelihood (6) analytically because of the integration with respect to \mathbf{u}_i . Thus we introduce the EM algorithm (Dempster et al., 1977) where the vector of random effects $\mathbf{u} = (\mathbf{u}_1^\top, \dots, \mathbf{u}_m^\top)^\top$ is regarded as the missing variable. The complete log-likelihood is given by

$$\log\{L^c(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})\} = \sum_{i=1}^m [\log\{f(\mathbf{y}_i | \mathbf{u}_i)\} + \log\{\pi(\mathbf{u}_i)\}].$$

In the k th iteration of the algorithm, the E-step calculates

$$Q(\boldsymbol{\psi} | \boldsymbol{\psi}^{(k-1)}) = E[\log\{L^c(\boldsymbol{\psi}; \mathbf{y}, \mathbf{u})\} | \mathbf{y}, \boldsymbol{\psi}^{(k-1)}],$$

where the expectation is taken with respect to the conditional distribution of \mathbf{u} given \mathbf{y} with the parameter value $\boldsymbol{\psi}^{(k-1)}$ from the $(k-1)$ th iteration. The M-step maximizes $Q(\boldsymbol{\psi} | \boldsymbol{\psi}^{(k-1)})$ with respect to $\boldsymbol{\psi}$. The maximizer, denoted by $\boldsymbol{\psi}^{(k)} = ((\boldsymbol{\beta}^{(k)})^\top, \tau^{2(k)}, \lambda^{(k)}, \kappa^{(k)}, (\boldsymbol{\gamma}^{(k)})^\top)^\top$, is obtained as

$$\begin{aligned} \tau^{2(k)} &= \frac{1}{m} E[\mathbf{b}^\top \mathbf{b} | \mathbf{y}, \boldsymbol{\psi}^{(k-1)}], \\ ((\boldsymbol{\beta}^{(k)})^\top, \kappa^{(k)})^\top &= \underset{(\boldsymbol{\beta}^\top, \kappa)^\top}{\operatorname{argmax}} E \left[\sum_{i=1}^m \log\{f(\mathbf{y}_i | \mathbf{u}_i)\} | \mathbf{y}, \boldsymbol{\psi}^{(k-1)} \right], \\ ((\boldsymbol{\gamma}^{(k)})^\top, \lambda^{(k)})^\top &= \underset{(\boldsymbol{\gamma}^\top, \lambda)^\top}{\operatorname{argmax}} E \left[\sum_{i=1}^m \log\{\pi(\sigma_i^2)\} | \mathbf{y}, \boldsymbol{\psi}^{(k-1)} \right]. \end{aligned}$$

Since it is difficult to evaluate the conditional expectation analytically in the E-step, we use the Monte Carlo integration based on the sampling importance resampling (SIR). Note that the conditional pdf of \mathbf{u} given \mathbf{y} is the product of the conditional pdfs of \mathbf{u}_i given \mathbf{y}_i :

$$\pi(\mathbf{u} | \mathbf{y}) = \prod_{i=1}^m \pi(\mathbf{u}_i | \mathbf{y}_i) \propto \prod_{i=1}^m f(\mathbf{y}_i | \mathbf{u}_i) \pi(\mathbf{u}_i),$$

where $\pi(\mathbf{u} | \mathbf{y})$ is the conditional pdf of \mathbf{u} given \mathbf{y} and $\pi(\mathbf{u}_i | \mathbf{y}_i)$ is the conditional pdf of \mathbf{u}_i given \mathbf{y}_i . Therefore, we apply the following SIR method independently for $i = 1, \dots, m$. Let

$q(\mathbf{u}_i \mid \mathbf{a}_i)$ denote the proposal density for \mathbf{u}_i where $\mathbf{a}_i \in \mathbb{R}^q$ is the parameter vector of the proposal distribution. In the SIR method, first a set of random numbers $\{\tilde{\mathbf{u}}_i^{(1)}, \dots, \tilde{\mathbf{u}}_i^{(S_1)}\}$ from $q(\mathbf{u}_i \mid \mathbf{a}_i)$ is generated. Then for each $\tilde{\mathbf{u}}_i^{(s)}$, the weight

$$\tilde{w}_{is} = \frac{f(\mathbf{y}_i \mid \tilde{\mathbf{u}}_i^{(s)})\pi(\tilde{\mathbf{u}}_i^{(s)})}{q(\tilde{\mathbf{u}}_i^{(s)} \mid \mathbf{a}_i)}, \quad s = 1, \dots, S_1,$$

is calculated. Finally, a set of samples of size S_2 , $\{\mathbf{u}_i^{(1)}, \dots, \mathbf{u}_i^{(S_2)}\}$, is drawn with replacement from $\{\tilde{\mathbf{u}}_i^{(1)}, \dots, \tilde{\mathbf{u}}_i^{(S_1)}\}$ based on the probability

$$\Pr(\mathbf{u}_i^{(r)} = \tilde{\mathbf{u}}_i^{(s)}) = \frac{\tilde{w}_{is}}{\sum_{s'=1}^{S_1} \tilde{w}_{is'}}, \quad s = 1, \dots, S_1, \quad r = 1, \dots, S_2.$$

For large S_1/S_2 , $\{\mathbf{u}_i^{(1)}, \dots, \mathbf{u}_i^{(S_2)}\}$ is approximately a set of independent random samples from $\pi(\mathbf{u}_i \mid \mathbf{y}_i)$. The expectations in the M-step are replaced with the Monte-Carlo estimates based on the SIR samples.

The performance of the SIR depends on the choice of the proposal distribution. It is ideal to employ a proposal distribution that well approximates the target distribution and we aim to achieve this by updating the value of \mathbf{a}_i through an iterative procedure proposed by Richard and Zhang (2007). Their efficient importance sampling (EIS) method determines the value $\hat{\mathbf{a}}_i$ such that it minimizes the Monte Carlo sampling variance of the importance weights with respect to the proposal distribution. In the current context, as shown by Richard and Zhang (2007), $\hat{\mathbf{a}}_i$ is determined through the following minimization problem

$$(\hat{c}_i, \hat{\mathbf{a}}_i^\top)^\top = \underset{(c_i, \mathbf{a}_i^\top)^\top}{\operatorname{argmin}} \int \{\log f(\mathbf{y}_i \mid \mathbf{u}_i) + \log \pi(\mathbf{u}_i) - c_i - \log g(\mathbf{u}_i \mid \mathbf{a}_i)\}^2 w_i(\mathbf{u}_i \mid \mathbf{a}_i) q(\mathbf{u}_i \mid \mathbf{a}_i) d\mathbf{u}_i, \quad (8)$$

where $g(\mathbf{u}_i \mid \mathbf{a}_i)$ is the kernel of the proposal density $q(\mathbf{u}_i \mid \mathbf{a}_i)$ such that $q(\mathbf{u}_i \mid \mathbf{a}_i) = g(\mathbf{u}_i \mid \mathbf{a}_i) / \int g(\mathbf{u}_i \mid \mathbf{a}_i) d\mathbf{u}_i$, c_i is a scalar that adjusts for the normalizing constants and $w_i(\mathbf{u}_i \mid \mathbf{a}_i) = f(\mathbf{y}_i \mid \mathbf{u}_i)\pi(\mathbf{u}_i)/q(\mathbf{u}_i \mid \mathbf{a}_i)$. The EIS method replaces (8) with a Monte Carlo approximation and proceeds by iteratively solving

$$(\hat{c}_i^{(t)}, \hat{\mathbf{a}}_i^{(t)\top})^\top = \underset{(c_i, \mathbf{a}_i^\top)^\top}{\operatorname{argmin}} \frac{1}{S_0} \sum_{s=1}^{S_0} \left\{ \log f(\mathbf{y}_i \mid \check{\mathbf{u}}_i^{(s)}) + \log \pi(\check{\mathbf{u}}_i^{(s)}) - c_i - \log g(\check{\mathbf{u}}_i^{(s)} \mid \mathbf{a}_i) \right\}^2 w_i(\check{\mathbf{u}}_i^{(s)} \mid \mathbf{a}_i^{(t-1)}), \quad (9)$$

where $(\hat{\mathbf{a}}_i^{(t)\top}, \hat{c}_i^{(t)})^\top$ denotes the value of $(\hat{\mathbf{a}}_i^\top, \hat{c}_i)^\top$ at the t th iteration of the EIS minimization and $\{\check{\mathbf{u}}_i^{(1)}, \dots, \check{\mathbf{u}}_i^{(S_0)}\}$ is the set of samples generated from $q(\mathbf{u}_i \mid \hat{\mathbf{a}}_i^{(t-1)})$ for $\check{\mathbf{u}}_i^{(s)} = (\check{b}_i^{(s)}, \check{\sigma}_i^{2(s)})^\top$. Richard and Zhang (2007) noted that S_0 does not have to be very large. In this paper, we employ $N(\theta_{i1}(\mathbf{a}_i), \theta_{i2}(\mathbf{a}_i)) \times \text{IG}(\theta_{i3}(\mathbf{a}_i), \theta_{i4}(\mathbf{a}_i))$ for $q(\mathbf{u}_i \mid \mathbf{a}_i)$ where $\mathbf{a}_i = (a_{i1}, a_{i2}, a_{i3}, a_{i4})^\top$ is the vector of natural parameters. Because the proposal distribution belongs to the exponential family where

$$\log g(\mathbf{u}_i^{(s)} \mid \mathbf{a}_i) = a_{i1}b_i + a_{i2}b_i^2 + a_{i3} \log(\sigma_i^2) + a_{i4} \frac{1}{\sigma_i^2},$$

for $a_{i1} = \theta_{i1}/\theta_{i2}$, $a_{i2} = -1/(2\theta_{i2})$, $a_{i3} = -(\theta_{i3} + 1)$ and $a_{i4} = -\theta_{i4}$, the solution for the EIS minimization (9) is given by the following generalized least squares (GLS) estimator

$$(\hat{c}_i^{(t)}, \hat{\mathbf{a}}_i^{(t)\top})^\top = (\mathbf{Z}_i^\top \mathbf{D}_i \mathbf{Z}_i)^{-1} \mathbf{Z}_i^\top \mathbf{D}_i \mathbf{f}_i \quad (10)$$

where $\mathbf{Z}_i = (\mathbf{1}_{S_0}, \check{\mathbf{b}}_i, \check{\mathbf{b}}_i^2, \mathbf{log}\check{\sigma}_i^2, \check{\sigma}_i^{-2})$, $\check{\mathbf{b}}_i, \check{\mathbf{b}}_i^2, \mathbf{log}\check{\sigma}_i^2, \check{\sigma}_i^{-2}$ and \mathbf{f}_i are $S_0 \times 1$ vectors with the s th elements given by $\check{b}_i^{(s)}, (\check{b}_i^{(s)})^2, \mathbf{log}(\check{\sigma}_i^{2(s)}), 1/\check{\sigma}_i^{2(s)}$ and $\mathbf{log} f(\mathbf{y}_i | \check{\mathbf{u}}_i^{(s)}) + \mathbf{log} \pi(\check{\mathbf{u}}_i^{(s)})$, respectively, and \mathbf{D}_i is the S_0 dimensional diagonal matrix with $w_i(\check{\mathbf{u}}_i^{(s)} | \hat{\mathbf{a}}_i^{(t-1)})$ on the s th diagonal position. In this paper, the EIS iteration is terminated when the relative change in $(\theta_{i1}(\mathbf{a}_i^{(t)}), \theta_{i2}(\mathbf{a}_i^{(t)}), \theta_{i3}(\mathbf{a}_i^{(t)}), \theta_{i4}(\mathbf{a}_i^{(t)}))^\top$ is below 10^{-3} . After the termination of the EIS iterations, the optimal parameters for the proposal distribution are obtained through $\hat{\theta}_{i1} = -\hat{a}_{1i}/(2\hat{a}_{2i})$, $\hat{\theta}_{i2} = -1/(2\hat{a}_{2i})$, $\hat{\theta}_{i3} = -\hat{a}_{3i} - 1$ and $\hat{\theta}_{i4} = -\hat{a}_{4i}$. See Richard and Zhang (2007) for more detailed implementation of the EIS method.

The initial values for the MCEM algorithm are determined as follows. Let us define $V_i = n_i^{-1} \sum_{g=1}^G \mathbf{log}(\bar{c}_g) \times y_{ig}$ where $\bar{c}_g = (c_{g-1} + c_g)/2$ for $g = 1, \dots, G-1$ and $\bar{c}_G = c_{G-1} + (c_{G-1} - c_{G-2})/2$, $\mathbf{V} = (V_1, \dots, V_m)^\top$ and $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_m)^\top$. Then, the initial value of $\boldsymbol{\beta}$ and τ^2 are determined as

$$\boldsymbol{\beta}^{(0)} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{V}, \quad \tau^{2(0)} = m^{-1} \|\mathbf{V} - \mathbf{X}\boldsymbol{\beta}^{(0)}\|^2.$$

The initial values of λ , κ and γ are determined by using the estimates based on the local model which modifies the model (2) as follows:

$$h_{\kappa_i}(z_{ij}) = \beta_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \mathbf{N}(0, \sigma_i^2), \quad (11)$$

where β_i, κ_i and σ_i^2 are the unknown parameters. Let $\hat{\beta}_i, \hat{\kappa}_i$ and $\hat{\sigma}_i^2$ denote the maximum likelihood estimates which independently maximizes the likelihood function for $i = 1, \dots, m$:

$$(\hat{\beta}_i, \hat{\kappa}_i, \hat{\sigma}_i^2)^\top = \underset{(\beta_i, \kappa_i, \sigma_i^2)^\top}{\operatorname{argmax}} \frac{n_i!}{y_{i1}! y_{i2}! \cdots y_{iG}!} \times \prod_{g=1}^G \left[\Phi \left\{ \frac{h_{\kappa}(c_g) - \beta_i}{\sigma_i} \right\} - \Phi \left\{ \frac{h_{\kappa}(c_{g-1}) - \beta_i}{\sigma} \right\} \right]^{y_{ig}}.$$

Then, the initial value of λ and κ are determined as

$$\lambda^{(0)} = 2 \times \{(\bar{\hat{\sigma}}^2)^2 / \widehat{V}(\hat{\sigma}^2) + 1\}, \quad \kappa^{(0)} = \bar{\hat{\kappa}},$$

where $\bar{\hat{\sigma}}^2$ and $\widehat{V}(\hat{\sigma}^2)$ are sample mean and variance of $\hat{\sigma}_i^2$'s over the areas and $\bar{\hat{\kappa}}$ is the sample mean of $\hat{\kappa}_i$'s. Furthermore, the initial value of γ is

$$\boldsymbol{\gamma}^{(0)} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\sigma},$$

where $\boldsymbol{\sigma} = (\hat{\sigma}_1^2, \dots, \hat{\sigma}_m^2)^\top$. This method generally provides reasonable initial values for the MCEM algorithm leading to a fast convergence. Although other initial values are also tried, the similar results are obtained with longer computing times.

To monitor the convergence of the MCEM algorithm, the criterion considered by Shi and Copas (2002) is used. In order to prevent premature termination of the algorithm due to the difference in the scale of the parameter values, the quantities $e_{k,(\boldsymbol{\beta})}$, $e_{k,(\tau^2)}$, $e_{k,(\kappa)}$, $e_{k,(\lambda)}$ and $e_{k,(\boldsymbol{\gamma})}$ is evaluated respectively for $\boldsymbol{\beta}$, τ^2 , κ , λ and $\boldsymbol{\gamma}$. In the case of $\boldsymbol{\beta}$, for example,

$$e_{k,(\boldsymbol{\beta})} = \frac{\|\tilde{\boldsymbol{\beta}}_1^{(k)} - \tilde{\boldsymbol{\beta}}_2^{(k)}\|}{\|\tilde{\boldsymbol{\beta}}_2^{(k)}\| + \delta}, \quad (12)$$

where $\tilde{\boldsymbol{\beta}}_1^{(k)} = H^{-1} \sum_{h=0}^{H-1} \boldsymbol{\beta}^{(k-h)}$, $\tilde{\boldsymbol{\beta}}_2^{(k)} = H^{-1} \sum_{h=0}^{H-1} \tilde{\boldsymbol{\beta}}^{(k-h-d)}$, and δ , H , and d are specified by the user. Then the EM algorithm is terminated in the k th iteration if

$$\max\{e_{k,(\boldsymbol{\beta})}, e_{k,(\tau^2)}, e_{k,(\kappa)}, e_{k,(\lambda)}, e_{k,(\boldsymbol{\gamma})}\} < \epsilon,$$

for some small value $\epsilon > 0$, and use $\tilde{\boldsymbol{\psi}}_1^{(k)} = (\tilde{\boldsymbol{\beta}}^{(k)\top}, \tilde{\tau}^{2(k)}, \tilde{\lambda}^{(k)}, \tilde{\kappa}^{(k)}, \tilde{\boldsymbol{\gamma}}^{(k)\top})^\top$ as the estimate of $\boldsymbol{\psi}$, which is denoted by $\hat{\boldsymbol{\psi}} = (\hat{\boldsymbol{\beta}}^\top, \hat{\tau}^2, \hat{\lambda}, \hat{\kappa}, \hat{\boldsymbol{\gamma}}^\top)^\top$ hereafter.

2.3 Calculation of empirical Bayes estimates

Here we propose the method to calculate EB estimates of some function of \mathbf{z}_i , which is denoted as $\zeta_i(\mathbf{z}_i)$ in general. The examples of $\zeta_i(\mathbf{z}_i)$ include the areal mean \bar{z}_i and Gini coefficients $\text{GINI}(\mathbf{z}_i)$ in (5). Under the quadratic loss, the Bayes estimator of $\zeta_i(\mathbf{z}_i)$ is its conditional expectation given the data, $E[\zeta_i(\mathbf{z}_i) \mid \mathbf{y}]$. Because of the independence over the areas, $E[\zeta_i(\mathbf{z}_i) \mid \mathbf{y}]$ is reduced to $E[\zeta_i(\mathbf{z}_i) \mid \mathbf{y}_i]$, which is denoted by

$$\xi_i(\boldsymbol{\psi}; \mathbf{y}_i) = E[\zeta_i(\mathbf{z}_i) \mid \mathbf{y}_i].$$

Because $\xi_i(\boldsymbol{\psi}; \mathbf{y}_i)$ is a function of the unknown parameter $\boldsymbol{\psi}$, we obtain the empirical Bayes (EB) estimator $\xi_i(\widehat{\boldsymbol{\psi}}; \mathbf{y}_i)$ by substituting $\widehat{\boldsymbol{\psi}}$ for $\boldsymbol{\psi}$ in the Bayes estimator. However, since it is impossible to evaluate the conditional expectation of $\zeta_i(\mathbf{z}_i)$ analytically, we calculate the EB estimates from the output of the following Gibbs sampler.

Let the random vector $\tilde{\mathbf{v}}_i = (v_{i1}, \dots, v_{in_i})^\top$ denote the sorted values of $\{h_{\hat{\kappa}}(z_{i1}), \dots, h_{\hat{\kappa}}(z_{in_i})\}$ in increasing order with size y_{i1}, \dots, y_{iG} and then the following relationship holds:

$$v_{ij} \leq v_{ik}, \quad \text{for all } j, k \text{ such that } j \leq \tilde{y}_{ig} < k, \text{ for all } g = 1, \dots, G,$$

where $\tilde{y}_{ig} = \sum_{g'=1}^g y_{ig'}$ for $g = 1, \dots, G$ and $n_i = \tilde{y}_{iG}$. For out-of-sample units, let $\check{\mathbf{v}}_i = (v_{i,n_i+1}, \dots, v_{iN_i})^\top = (h_{\hat{\kappa}}(z_{i,n_i+1}), \dots, h_{\hat{\kappa}}(z_{iN_i}))^\top$. Let $\mathbf{v}_i = (\tilde{\mathbf{v}}_i^\top, \check{\mathbf{v}}_i^\top)^\top = (v_{i1}, \dots, v_{iN_i})^\top$. To evaluate the conditional expectation of \mathbf{v}_i given \mathbf{y}_i , the sample from the joint conditional distribution of $\{\tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \mu_i, \sigma_i^2\}$ given \mathbf{y}_i is obtained by using the Gibbs sampling algorithm with the following full conditional distributions:

$$\begin{aligned} \mu_i \mid \tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \sigma_i^2, \mathbf{y}_i &\sim \text{N} \left(\frac{\sigma_i^2 \mathbf{x}_i^\top \hat{\boldsymbol{\beta}} + N_i \hat{\tau}^2 \bar{v}_i}{\sigma_i^2 + N_i \hat{\tau}^2}, \frac{\hat{\tau}^2 \sigma_i^2}{\sigma_i^2 + N_i \hat{\tau}^2} \right), \\ \mathbf{v}_{ij} \mid \mu_i, \tilde{\mathbf{v}}_i, \sigma_i^2, \mathbf{y}_i &\stackrel{\text{indep}}{\sim} \begin{cases} \text{TN}_{[h_{\hat{\kappa}}(c_0), h_{\hat{\kappa}}(c_1)]}(\mu_i, \sigma_i^2), & j = 1, \dots, \tilde{y}_{i1}, \\ \text{TN}_{[h_{\hat{\kappa}}(c_1), h_{\hat{\kappa}}(c_2)]}(\mu_i, \sigma_i^2), & j = \tilde{y}_{i1} + 1, \dots, \tilde{y}_{i2} \\ \vdots \\ \text{TN}_{[h_{\hat{\kappa}}(c_{G-1}), h_{\hat{\kappa}}(c_G)]}(\mu_i, \sigma_i^2), & j = \tilde{y}_{i,G-1} + 1, \dots, n_i, \end{cases} \quad (13) \\ \check{\mathbf{v}}_i \mid \mu_i, \tilde{\mathbf{v}}_i, \sigma_i^2, \mathbf{y}_i &\sim \text{N}_{N_i - n_i}(\mu_i \mathbf{1}_{N_i - n_i}, \sigma_i^2 \mathbf{1}_{N_i - n_i}), \\ \sigma_i^2 \mid \mu_i, \tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \mathbf{y}_i &\sim \text{IG} \left(\frac{N_i + \hat{\lambda}}{2} + 1, \frac{1}{2} \left\{ \hat{\lambda} \hat{\varphi}_i + \sum_{j=1}^{N_i} (v_{ij} - \mu_i)^2 \right\} \right), \end{aligned}$$

where $\bar{v}_i = N_i^{-1} \sum_{j=1}^{N_i} v_{ij}$ and $\text{TN}_{[a,b]}(\mu, \sigma^2)$ denotes the truncated normal distribution with the mean μ and variance σ^2 truncated to the interval $[a, b]$. The derivation of the full conditional distributions is given in Appendix A.1.

Let $\mathbf{v}_i^{(s)} = (v_{i1}^{(s)}, \dots, v_{iN_i}^{(s)})^\top$ be the s th output of \mathbf{v}_i from the Gibbs sampler ($s = 1, \dots, S_3$). Then the EB estimates $\xi_i(\widehat{\boldsymbol{\psi}}; \mathbf{y}_i)$ can be calculated as

$$\widehat{\xi_i(\boldsymbol{\psi}; \mathbf{y}_i)} = \frac{1}{S_3} \sum_{s=1}^{S_3} \zeta_i(h_{\hat{\kappa}}^{-1}(\mathbf{v}_i^{(s)})),$$

where $h_{\hat{\kappa}}^{-1}(\cdot)$ is the inverse Box–Cox transformation with parameter value $\hat{\kappa}$.

If the auxiliary variables \mathbf{x}_i 's are available for out-of-sample areas, $\zeta_i(\mathbf{z}_i)$ can be also predicted for an out-of-sample area $i = m + 1$ by $\xi_{m+1}(\widehat{\boldsymbol{\psi}})$ where $\xi_{m+1}(\boldsymbol{\psi}) = E[\zeta_{m+1}(\mathbf{z}_{m+1})]$, since \mathbf{y}

and z_{m+1} are mutually independent. This expectation can be calculated by the Monte Carlo integration that generates random numbers from the model (2) with the hyperparameters are fixed to their estimates.

3 Application to grouped income data of Japan

The proposed method is demonstrated by using the grouped income data obtained from Housing and Land Survey (HLS) of Japan in 2013. The data contains the number of households that fall in $G = 5$ and 9 income classes.¹ The income classes are defined in million Japanese Yen (M JPY) and the thresholds are given by $(c_1, c_2, c_3, c_4) = (3, 5, 7, 10)$ for $G = 5$ and $(c_1, c_2, c_3, c_4, c_5, c_6, c_7, c_8) = (1, 2, 3, 4, 5, 7, 10, 15)$ for $G = 9$. In this survey in 2013, 1265 out of 1899 municipalities in Japan were sampled. As a summary of the data, Figure 1 presents the proportions of the households in the in-sample-municipalities for each income class in the case of $G = 9$. The maps look incomplete because of the presence of the out-of-sample municipalities.

Using the proposed method, the EB estimates of the areal mean incomes and Gini coefficients are obtained. For the auxiliary variables, we use the total population denoted by P_i and working-age population denoted by WA_i obtained from Population Census (PC) of Japan in 2010 and set $\mathbf{x}_i = (1, \log P_i, \log WA_i)$ for the i th municipality. Since these auxiliary variables are also available for the out-of-sample municipalities of HLS, the model can be further utilised to complete the maps of the mean incomes and Gini coefficients.

¹ Only are the numbers of households in each income class adjusted for the population sizes accessible in the HLS data and the original sample sizes for the sampled municipalities of HLS are not published. How they are estimated for this analysis is described in Appendix A.2.

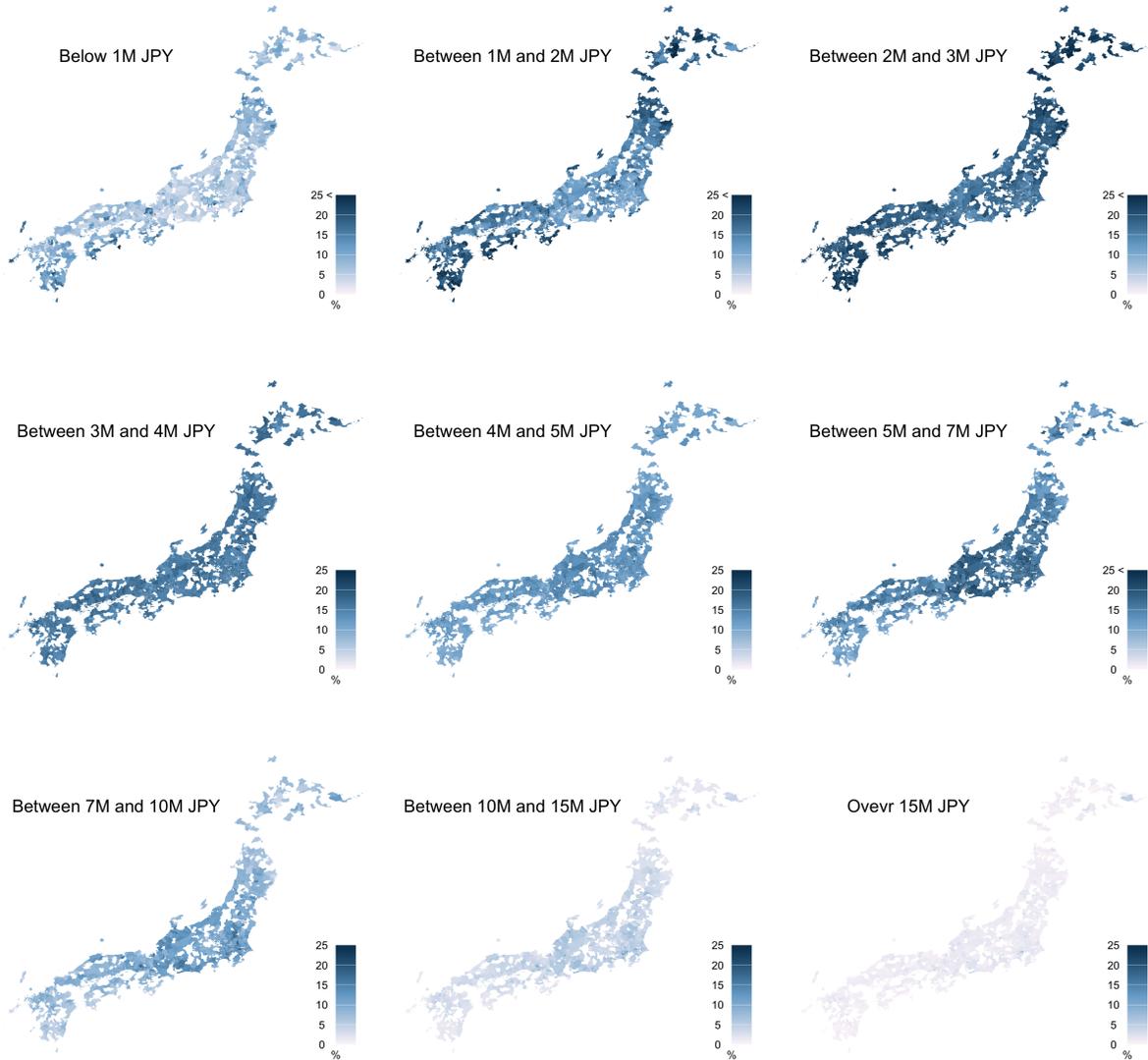


Figure 1: Proportions of households in in-sample-municipalities ($G = 9$)

To estimate the hyperparameters, we set $S_0 = 100$, $S_1 = 10000$, $S_2 = 500$, $H = 30$, $d = 5$, and $\delta = \epsilon = 0.001$ for the MCEM algorithm. The initial values are determined using the method described in Section 2.2. The convergence of the MCEM algorithm occurs relatively fast. We also tried other initial values obtained similar results. It is noted that the method in Section 2.2 took much shorter computing times. Figure 2 presents the 0.1, 0.5 and 0.9 quantiles of the effective sample size (ESS) divided by S_1 for the 1265 municipalities at each step of the MCEM algorithm. It is seen that the ESS is fairly high and stable over the EM iterations, especially for $G = 9$.

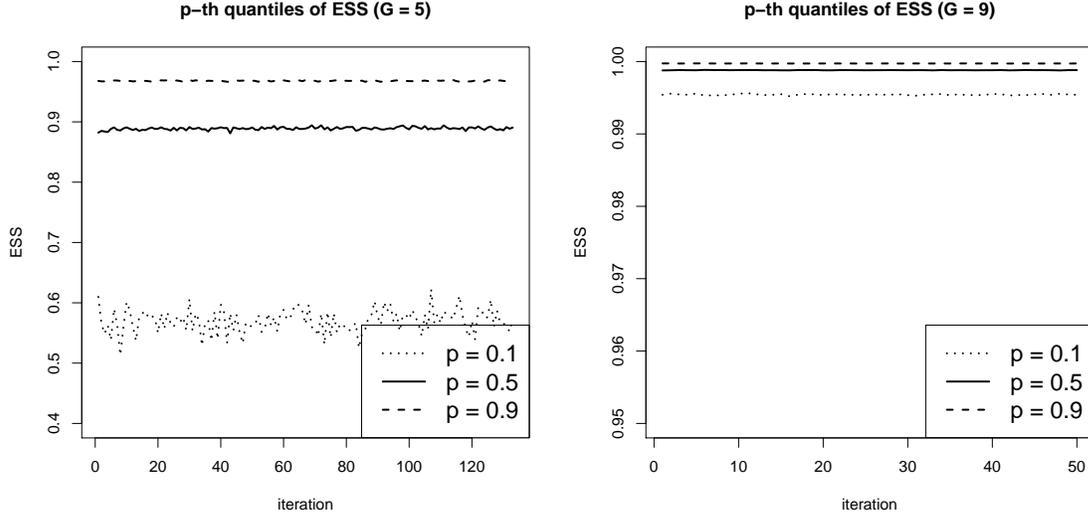


Figure 2: Quantiles of effective sample size (ESS)

The Bayes estimator of \bar{z}_i is denoted by $\xi_{1i}(\boldsymbol{\psi}; \mathbf{y}_i) = E(\bar{z}_i \mid \mathbf{y}_i)$ and that of $\text{GINI}(\mathbf{z}_i)$ is denoted by $\xi_{2i}(\boldsymbol{\psi}; \mathbf{y}_i) = E[\text{GINI}(\mathbf{z}_i) \mid \mathbf{y}_i]$. The EB estimates of \bar{z}_i and $\text{GINI}(\mathbf{z}_i)$ are calculated from the output of the Gibbs sampler (13) as

$$\widehat{\xi_{1i}}(\widehat{\boldsymbol{\psi}}; \mathbf{y}_i) = \frac{1}{S_3} \sum_{s=1}^{S_3} \left\{ \frac{1}{N_i} \sum_{j=1}^{N_i} h_{\hat{\kappa}}^{-1}(v_{ij}^{(s)}) \right\},$$

and

$$\widehat{\xi_{2i}}(\widehat{\boldsymbol{\psi}}; \mathbf{y}_i) = \frac{1}{S_3} \sum_{s=1}^{S_3} \frac{1}{N_i} \left\{ N_i + 1 - \frac{2 \sum_{j=1}^{N_i} (N_i + 1 - j) h_{\hat{\kappa}}^{-1}(v_{ij}^{(s)})}{\sum_{j=1}^{N_i} h_{\hat{\kappa}}^{-1}(v_{ij}^{(s)})} \right\},$$

where $\{v_{i(1)}^{(s)}, \dots, v_{i(N_i)}^{(s)}\}$ are sorted values of $\{v_{i1}^{(s)}, \dots, v_{iN_i}^{(s)}\}$ in non-decreasing order. In this analysis, we run the Gibbs sampler for $S_3 = 500$ iterations with the initial burn-in period of 50 iterations.

While it is generally difficult to define a reasonable direct estimator for these small area parameters from grouped data, for a comparison purpose, we may also think of the following “naive” estimator of the areal mean \bar{z}_i that uses the class midpoints given by

$$\widehat{\bar{z}}_i^{\text{naive}} = \frac{1}{n_i} \sum_{g=1}^G \bar{c}_g \times y_{ig} \quad (14)$$

where $\bar{c}_g = (c_{g-1} + c_g)/2$ for $g = 1, \dots, G-1$ and $\bar{c}_G = c_{G-1} + (c_{G-1} - c_{G-2})/2$. This estimator is naive particularly because the upper end \bar{c}_G has to be set and its choice is completely arbitrary. The choice of \bar{c}_G would have a huge impact on its performance. Note that the proposed approach has no arbitrariness with this respect as $c_G = \infty$ and (4) is well defined.

Figure 3 presents the estimates of the areal means based on the proposed method and naive method (14). By borrowing strength from the other municipalities through the statistical model (2), the proposed method can predict the income for the out-of-sample municipalities and provide the complete maps of the mean incomes and Gini coefficients. The boxplots of

Figure 4 compares the EB and naive estimates of the areal means for the sample areas. The figure indicates that the results for the naive estimates can vary between $G = 5$ and 9 resulting the lower mean incomes for some areas for $G = 5$ than for $G = 9$. This would be because the naive estimates cannot capture the behavior of the upper tail of the income distribution, which has an impact on the estimation of the mean income. In fact, we also considered the different values for \bar{c}_G for the naive estimates to demonstrate the impact. Figure 5 presents the boxplots of the naive estimates under the different values of \bar{c}_G for $G = 5$ and 9. The figure shows that the naive estimates exhibit severe sensitivity with respect to the setting of \bar{c}_G in the case of $G = 5$. While the sensitivity decreases for $G = 9$, the areal mean estimates for the high income areas still appear to increase with \bar{c}_G .

In order to assess the uncertainty of the estimators, we estimated the root mean squared error (RMSE) of the estimators for the sampled municipalities by using a parametric bootstrap method. Let $z_{ij}^{*(b)}$ ($i = 1, \dots, m$; $j = 1, \dots, N_i$) and $\{\mathbf{y}_1^{*(b)}, \dots, \mathbf{y}_m^{*(b)}\}$ denote the b th bootstrap sample ($b = 1, \dots, B$) generated from the models (1) and (2) with the hyperparameter fixed to the maximum likelihood estimate $\hat{\boldsymbol{\psi}}$. Then, the RMSE of the EB estimator of areal mean is estimated as

$$\widehat{\text{RMSE}}_i^{\text{EB}} = \sqrt{\frac{1}{B} \sum_{b=1}^B \left\{ \xi_{1i}(\hat{\boldsymbol{\psi}}; \mathbf{y}_i^{*(b)}) - \bar{z}_i^{*(b)} \right\}^2},$$

for a large B , where $\bar{z}_i^{*(b)} = N_i^{-1} \sum_{j=1}^{N_i} z_{ij}^{*(b)}$. For each b , we simply run the Gibbs sampler described in Section 2.3 to calculate the EB estimates given the estimate $\hat{\boldsymbol{\psi}}$ from the original data, not on the bootstrap samples. In the same way, the RMSE of the naive estimator is estimated as

$$\widehat{\text{RMSE}}_i^{\text{naive}} = \sqrt{\frac{1}{B} \sum_{b=1}^B \left\{ \hat{z}_i^{\text{naive}*(b)} - \bar{z}_i^{*(b)} \right\}^2},$$

where $\hat{z}_i^{\text{naive}*(b)} = n_i^{-1} \sum_{g=1}^G \bar{c}_g \times y_{ig}^{*(b)}$. Figure 6 presents the estimates of the RMSE of the EB estimators and naive estimators for the sampled areas. The naive estimators resulted in the large RMSE indicated by the darker shade of red in the case of $G = 5$. While the RMSE for the naive estimators improves as the number of income classes increases, the EB estimators resulted in the smaller RMSE. The figure also shows that the overall improvement in the RMSE of the EB estimators in the case of $G = 9$ over $G = 5$ is marginal compared to the naive estimators.

Finally, Figure 7 presents the EB estimates for the Gini coefficients for all municipalities and associated estimates of RMSE for the sampled municipalities. As in the case of the mean incomes, the proposed method can also predict the Gini coefficients for the out-of-sample municipalities to complete the map. The RMSE of the estimator of the Gini coefficient is estimated in the same way as that of the mean income by using the parametric bootstrap. The map for the case of $G = 9$ exhibits darker shades of blue than the map for $G = 5$ implying that the degree of inequality is greater across the country. This could be because that the data with $G = 9$ contains more information on the income distribution, especially on the upper tail of the distribution which can have an impact on the estimates. The figure also shows that the uncertainty regarding the Gini coefficients estimation decreases as the number of income classes in the data increases.

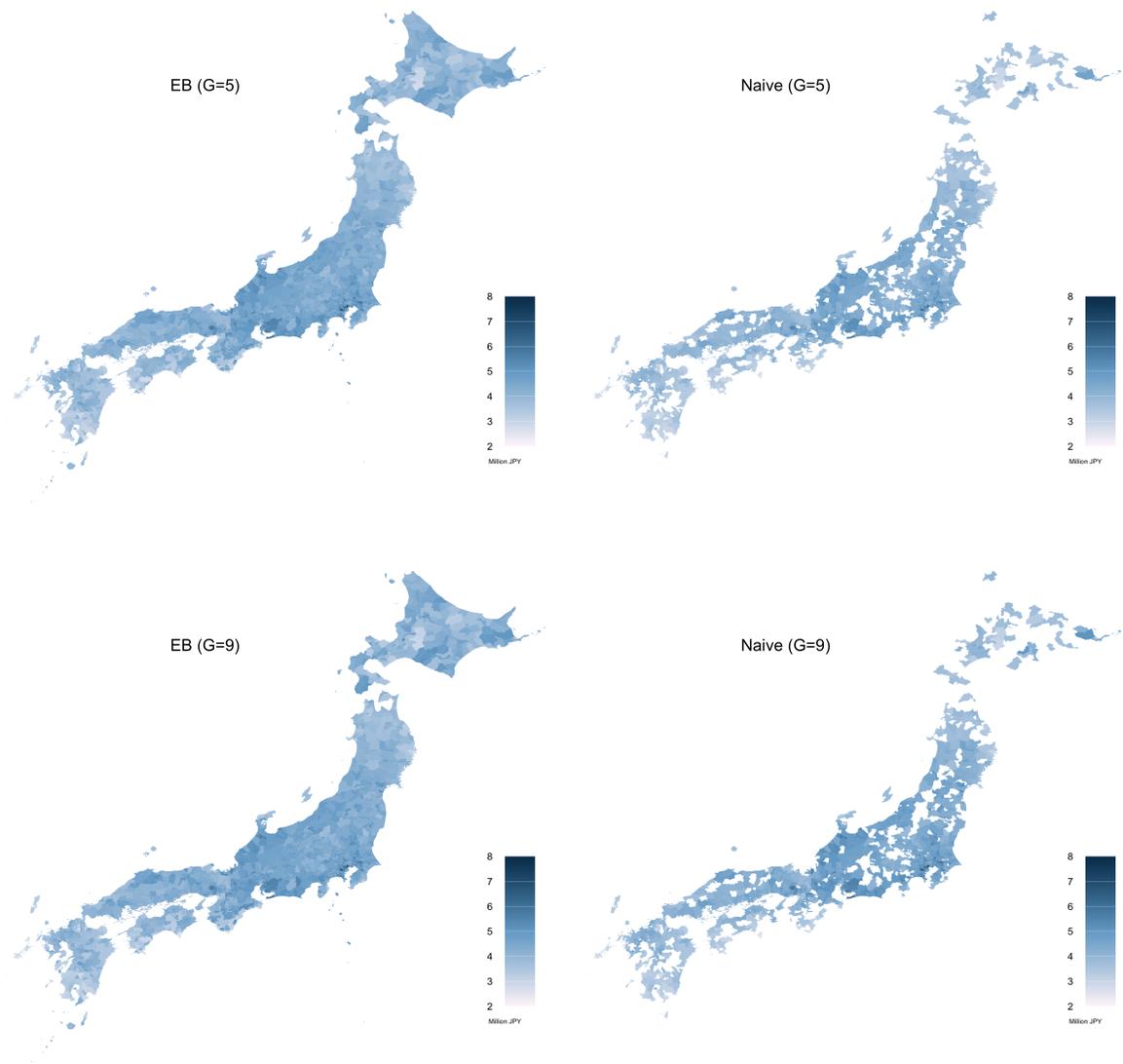


Figure 3: EB and naive estimates of areal means

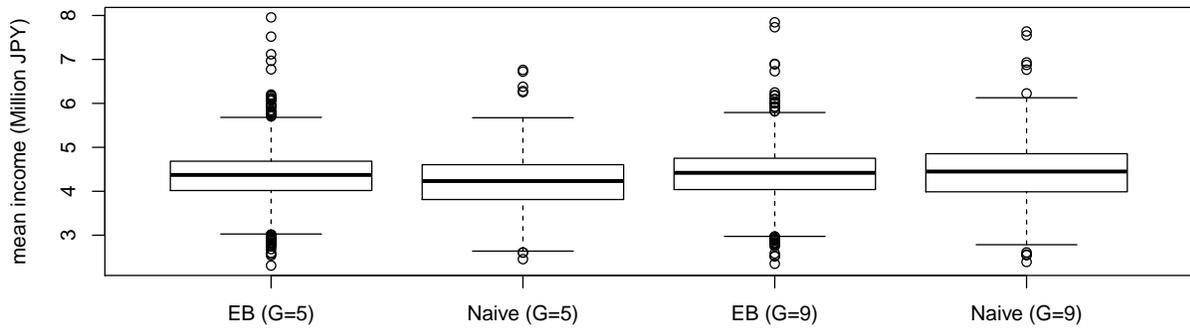


Figure 4: Boxplots of EB and naive estimates of areal means for the sampled areas

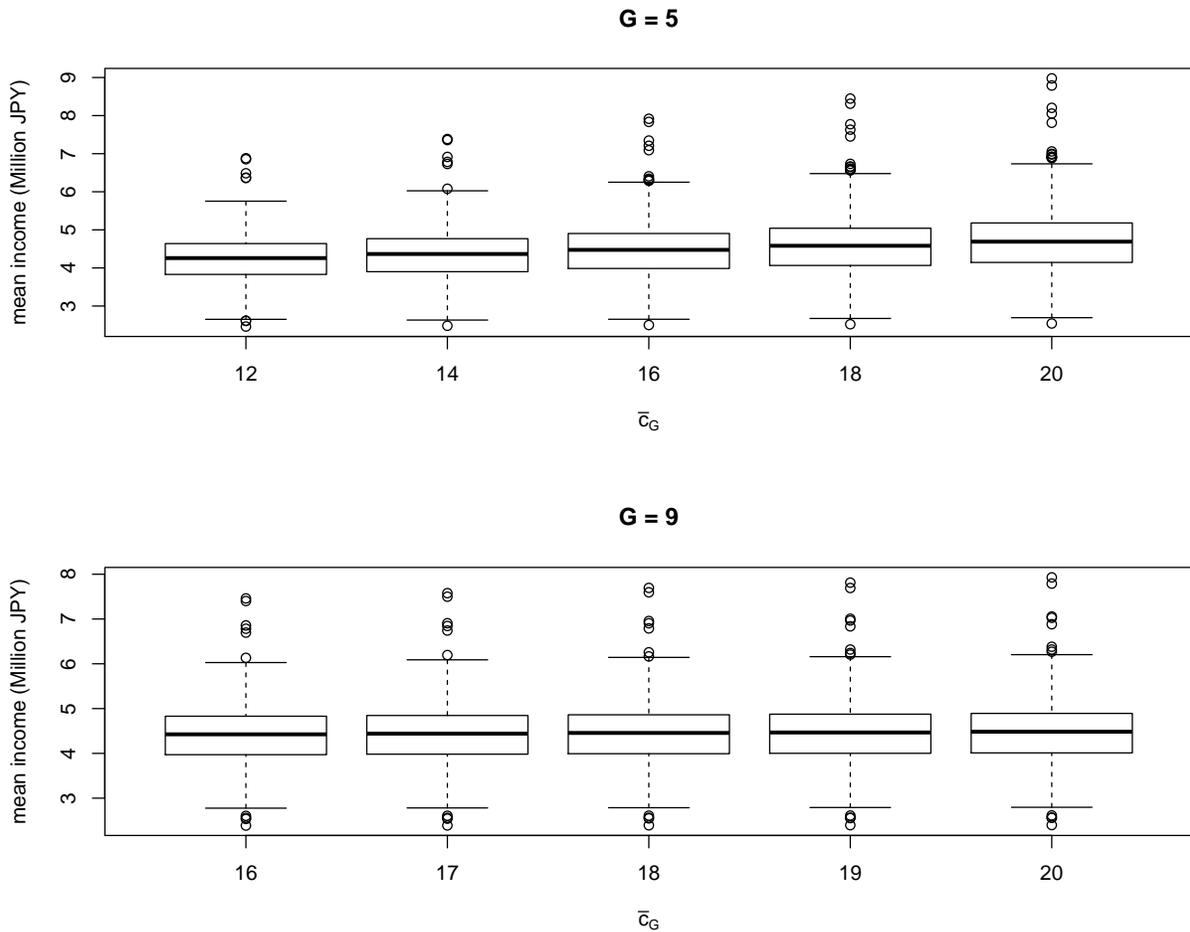


Figure 5: Boxplots of naive estimates of areal means under different values of \bar{c}_G

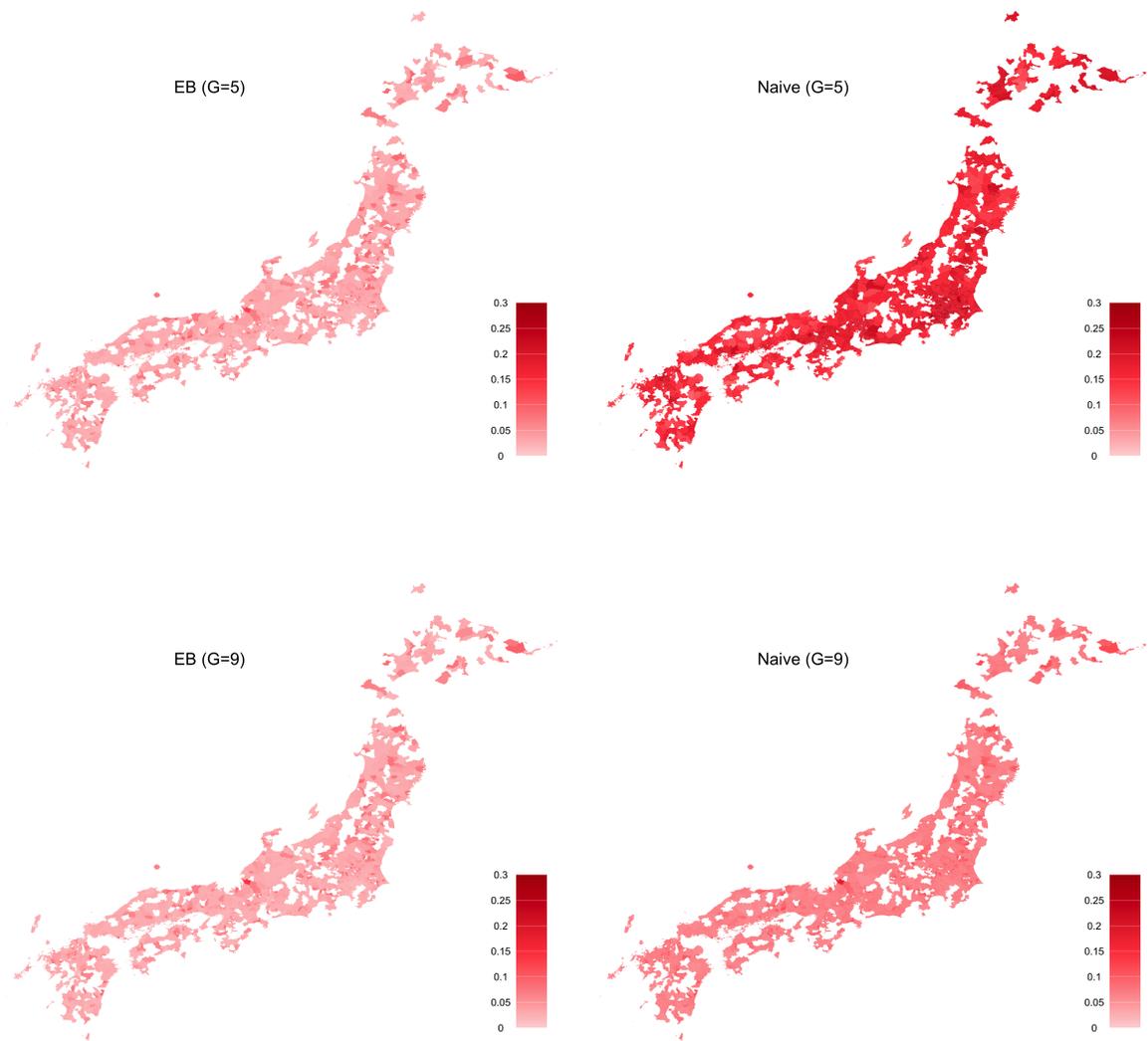


Figure 6: Estimates of RMSE of the naive estimators and EB estimators for areal means

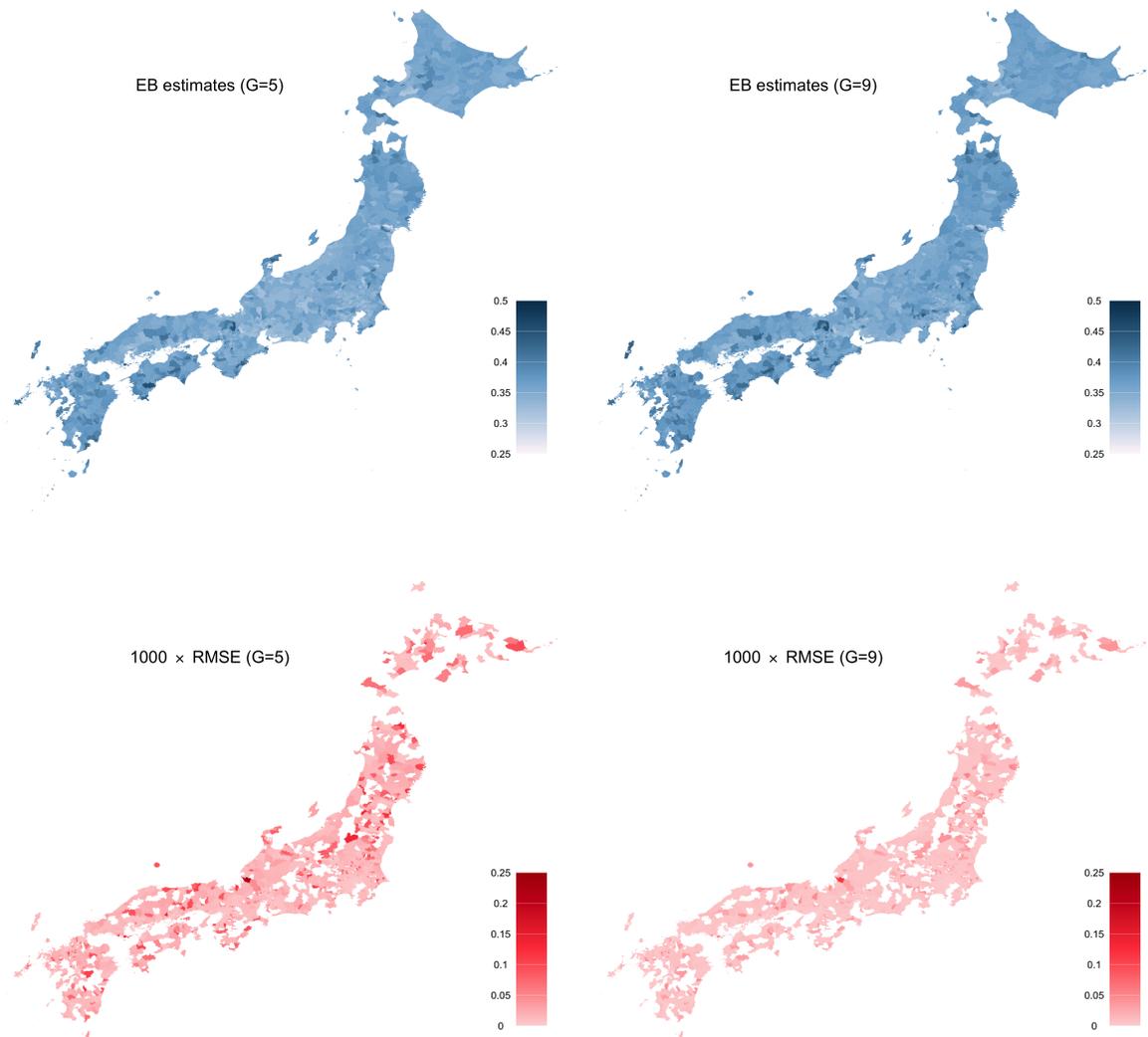


Figure 7: EB estimates and estimates of RMSE (multiplied by 1000) for Gini coefficients

4 Simulation Studies

4.1 Model-based simulation

In this section, the proposed approach is illustrated using the simulated data. The first simulation is a model-based simulation where (2) is the data generating process. The true parameter values are set to the estimates obtained in the real application in Section 3 and we use the same values of the auxiliary variables \boldsymbol{x}_i 's as the real data for the randomly chosen $m = 100$ areas out of the 1265 in-sample areas of HLS. Based on this setting, we generate $R = 100$ replications of z_{ij} 's with $N_i = 1000$ for all i and calculate the true mean \bar{z}_i and Gini coefficient $\text{GINI}(z_i)$. For each replication, we obtain a frequency distribution for each area from the simulated data $\{z_{i1}, \dots, z_{i,n_i}\}$. The two cases of the numbers of groups $G = 5$ and 9 with the same thresholds as HLS are considered. The sample sizes are set as $n_i = 10$ ($i = 1, \dots, 20$), $n_i = 50$ ($i =$

21, ..., 40), $n_i = 100$ ($i = 41, \dots, 60$), $n_i = 150$ ($i = 61, \dots, 80$), and $n_i = 200$ ($i = 81, \dots, 100$). The true parameter values and the auxiliary variables \mathbf{x}_i 's for $i = 1, \dots, m$ are fixed for all replications. The settings for the MCEM algorithm and the Gibbs sampler are the same as the real data analysis in Section 3.

In order to demonstrate the advantage of the present approach, the naive estimator of $\widehat{z}_i^{\text{naive}}$ in (14) is also considered again. The performance of the methods is compared by the simulated relative root MSE (RRMSE) over $R = 100$ replications of the data. The simulated RRMSE is calculated as

$$\text{RRMSE}(\widehat{z}_i) = \sqrt{\frac{1}{R} \sum_{r=1}^R \left(\frac{\widehat{z}_i^{(r)} - \bar{z}_i^{(r)}}{\bar{z}_i^{(r)}} \right)^2},$$

where $\widehat{z}_i^{(r)}$ is the EB or naive estimates and $\bar{z}_i^{(r)}$ is the true mean in the r th replication.

Figure 8 shows the result of the simulation. Noting that the horizontal axis represents the area index, the figure shows that the RRMSE decreases as the sample size increases both for the EB estimator and the naive estimator. In terms of RRMSE, the EB estimator improves on the naive estimator for all the areas. It is interesting to see that the improvement of the RRMSE is much larger for the areas with small sample sizes, especially for the areas with $n_i = 10$ and 50. This is because the EB estimator borrows strength from other areas even though the area sample size is small, while the naive estimator only uses the information of the target area. It is also observed that EB estimator for $G = 9$ resulted in better performance than for $G = 5$ for most of the areas. This is a natural result because the frequency distributions based on $G = 9$ contain more information of the distribution of the latent z_{ij} 's.

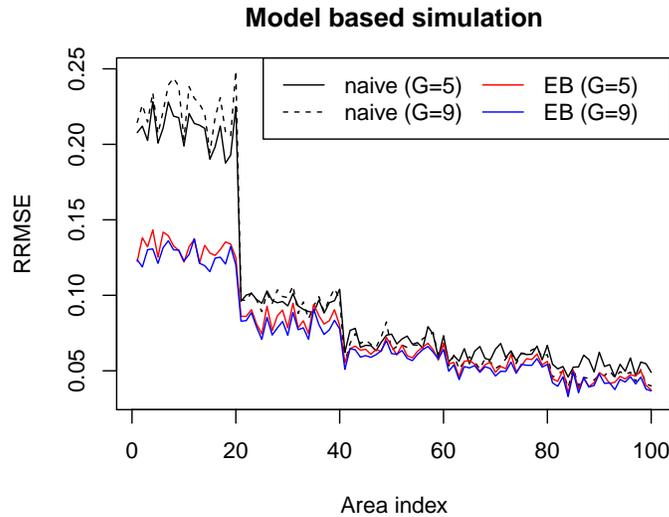


Figure 8: RRMSE of EB estimator and naive estimator for model based simulation

4.2 Design-based simulation

The second simulation is a design based simulation where (2) is not assumed to be the data generating process. For this simulation, the Spanish income dataset included in the R package `sae` developed by Molina and Marhuenda (2018).

This dataset contains the synthetic data on income and some related information of 17199 households including the province where the household is located and the gender of the head

of the household. There are 52 provinces in Spain and for each province the dataset is divided based on the gender of the head of the household. Therefore, this dataset consists of $m = 104$ small domains.

We generate the datasets for this design-based simulation study following the technique used by Chandra et al. (2012). First, a synthetic population is created for each domain by resampling with replacement from the original dataset and calculate the ‘true’ population mean for each dataset. Then 100 independent samples are obtained from the fixed synthetic populations based on the simple random sampling without replacement and form a frequency distribution for each domain.

As the auxiliary variables, we use $\mathbf{x}_i = (1, \text{NAT}_i, \text{WA}_i, \text{LABOR}_i)^\top$ where NAT_i is the proportion of the people holding Spanish nationality in the i th domain, WA_i is the proportion of the people who are in working age in the i th domain, and LABOR_i is the proportion of the people who are employed in the i th domain. For the transformation in (2), since the negative income observations are present for some households in this dataset, the following modified Box–Cox transformation is used:

$$h_\kappa(z) = \frac{(z - C)^\kappa - 1}{\kappa},$$

where C is equal to 0.1 less than the minimum income of the synthetic population. The same settings for the MCEM algorithm and Gibbs sampler as in the previous sections are used.

As in the previous sections, the performance of the proposed EB estimator and naive estimator is compared. Figure 9 shows the RRMSE for the EB and naive estimators. The figure shows that the EB estimator resulted in the better performance than the naive estimators in terms of RRMSE for most domains. In addition, the degree of improvement is larger in the case of $G = 5$, where the frequency distributions contain less information. Since this simulation setting does not assume a statistical model, we obtained an important implication that the proposed EB estimator performs well even when the statistical model is misspecified. This design based simulation can be seen as an empirical evidence to show the usefulness of our proposed method.

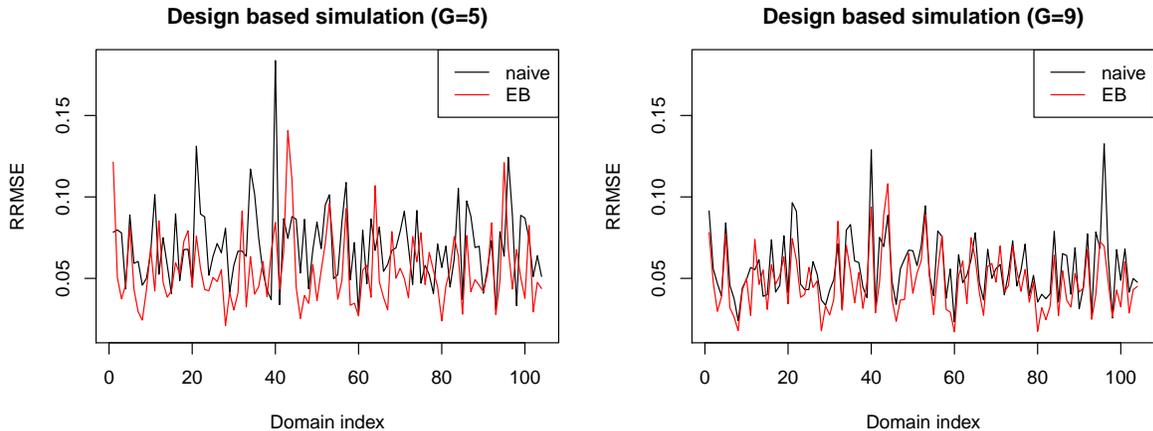


Figure 9: RRMSE of EB estimator and naive estimator based on design based simulation

5 Conclusion

We have proposed a new model-based small area estimation method for grouped data where only frequency distributions of the quantity of interest are observed at the area-level. In the proposed

model, the observed frequencies are linked with the area-level auxiliary variables through the unit-level latent variables which are modeled in a similar fashion to the nested error regression model. The model parameter is estimated easily by using the Monte Carlo EM algorithm based on the efficient importance sampling and the EB estimates of small area parameters are calculated by the output of the Gibbs sampler. From the application to the real data of Japan and simulation studies, we have shown that the proposed EB estimator performs better than the naive estimator.

Because our proposed model is in a general form, it can be applied to a wide variety of datasets. However, if we do focus on the income data, especially on the Gini coefficient or other poverty indicators, a probability distribution assumed by the small area model should provide good fit to the income distribution and provide a straightforward interpretation. The present model that assumes the normal distribution after a transformation may be limited in this sense. An extension of our model to the parametric income distribution is left for future studies.

Acknowledgments. This work is partially supported by JSPS KAKENHI (#19K13667, #18K12754). The computational results were obtained by using Ox version 6.21 (Doornik, 2007).

A Appendix

A.1 Derivation of the full conditional distributions (13)

Here the full conditional distributions of $\tilde{\mathbf{v}}_i$, $\check{\mathbf{v}}_i$, μ_i and σ_i^2 in (13) are derived. To avoid the notational complexity, we use the notation $p(\cdot)$ as the pdf or pmf for arbitrary random variable.

First, the joint conditional distribution of $\{\tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \mu_i, \sigma_i^2\}$ given \mathbf{y}_i , $p(\tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \mu_i, \sigma_i^2 \mid \mathbf{y}_i)$, is given by

$$p(\tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \mu_i, \sigma_i^2 \mid \mathbf{y}_i) = \frac{p(\mathbf{y}_i, \tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i \mid \mu_i, \sigma_i^2)p(\mu_i)p(\sigma_i^2)}{p(\mathbf{y}_i)}.$$

Thus it follows that

$$p(\tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \mu_i, \sigma_i^2 \mid \mathbf{y}_i) \propto p(\mathbf{y}_i, \tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i \mid \mu_i, \sigma_i^2)p(\mu_i)p(\sigma_i^2).$$

Note that $p(\mu_i) = \phi(\mu_i; \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, \hat{\tau}^2)$, where $\phi(\cdot; a, b)$ is the pdf of the normal distribution with the mean a and variance b and

$$p(\sigma_i^2) \propto (\sigma_i^2)^{-(\lambda/2+1)-1} \exp\left(-\frac{\hat{\lambda}\hat{\phi}_i}{2\sigma_i^2}\right).$$

Because out-of-sample $\check{\mathbf{v}}_i$ is independent of $\{\mathbf{y}_i, \tilde{\mathbf{v}}_i\}$ given $\{\mu_i, \sigma_i^2\}$, it follows that

$$\begin{aligned} p(\mathbf{y}_i, \tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i \mid \mu_i, \sigma_i^2) &= p(\mathbf{y}_i, \tilde{\mathbf{v}}_i \mid \mu_i, \sigma_i^2)p(\check{\mathbf{v}}_i \mid \mu_i, \sigma_i^2), \\ &= p(\mathbf{y}_i \mid \tilde{\mathbf{v}}_i, \mu_i, \sigma_i^2)p(\tilde{\mathbf{v}}_i \mid \mu_i, \sigma_i^2)p(\check{\mathbf{v}}_i \mid \mu_i, \sigma_i^2) \end{aligned}$$

where

$$p(\tilde{\mathbf{v}}_i \mid \mu_i, \sigma_i^2)p(\check{\mathbf{v}}_i \mid \mu_i, \sigma_i^2) = p(\mathbf{v}_i \mid \mu_i, \sigma_i^2) \propto \prod_{j=1}^{N_i} \phi(v_{ij}; \mu_i, \sigma_i^2),$$

for $\mathbf{v}_i = (\tilde{\mathbf{v}}_i^\top, \check{\mathbf{v}}_i^\top)^\top = (v_{i1}, \dots, v_{in_i}, v_{i,n_i+1}, \dots, v_{iN_i})^\top$. Furthermore, we can write the pmf of \mathbf{y}_i given $\{\tilde{\mathbf{v}}_i, \mu_i, \sigma_i^2\}$ as follows:

$$p(\mathbf{y}_i \mid \tilde{\mathbf{v}}_i, \mu_i, \sigma_i^2) = \left[\prod_{j=1}^{\tilde{y}_{i1}} I\{h_{\hat{\kappa}}(c_0) \leq v_{ij} < h_{\hat{\kappa}}(c_1)\} \right] \times \left[\prod_{j=\tilde{y}_{i1}+1}^{\tilde{y}_{i2}} I\{h_{\hat{\kappa}}(c_1) \leq v_{ij} < h_{\hat{\kappa}}(c_2)\} \right] \\ \times \cdots \times \left[\prod_{j=\tilde{y}_{i,G-1}+1}^{n_i} I\{h_{\hat{\kappa}}(c_{G-1}) \leq v_{ij} < h_{\hat{\kappa}}(c_G)\} \right],$$

where $I\{\cdot\}$ is the indicator function and $\tilde{y}_{ig} = \sum_{g'=1}^g y_{ig'}$, that is, $n_i = \sum_{g'=1}^G y_{ig'}$. Note that the value of $p(\mathbf{y}_i \mid \tilde{\mathbf{v}}_i, \mu_i, \sigma_i^2)$ only takes 1 or 0. Hence, the joint conditional distribution of $\{\tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \mu_i, \sigma_i^2\}$ given \mathbf{y}_i can be written as

$$p(\tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \mu_i, \sigma_i^2 \mid \mathbf{y}_i) \\ \propto \phi(\mu_i; \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, \hat{\tau}^2) \times (\sigma_i^2)^{(-\hat{\lambda}/2+1)-1} \exp\left(-\frac{\hat{\lambda}\hat{\varphi}_i}{2\sigma_i^2}\right) \\ \times \left[\prod_{j=1}^{\tilde{y}_{i1}} I\{h_{\hat{\kappa}}(c_0) \leq v_{ij} < h_{\hat{\kappa}}(c_1)\} \phi(v_{ij}; \mu_i, \sigma_i^2) \right] \times \left[\prod_{j=\tilde{y}_{i1}+1}^{\tilde{y}_{i2}} I\{h_{\hat{\kappa}}(c_1) \leq v_{ij} \leq h_{\hat{\kappa}}(c_2)\} \phi(v_{ij}; \mu_i, \sigma_i^2) \right] \\ \times \cdots \times \left[\prod_{j=\tilde{y}_{i,G-1}+1}^{n_i} I\{h_{\hat{\kappa}}(c_{G-1}) \leq v_{ij} < h_{\hat{\kappa}}(c_G)\} \phi(v_{ij}; \mu_i, \sigma_i^2) \right] \times \prod_{j=n_i+1}^{N_i} \phi(v_{ij}; \mu_i, \sigma_i^2).$$

Then, it follows that

$$p(\mu_i \mid \tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \sigma_i^2, \mathbf{y}_i) \propto \phi(\mu_i; \mathbf{x}_i^\top \hat{\boldsymbol{\beta}}, \hat{\tau}^2) \times \prod_{j=1}^{N_i} \phi(v_{ij}; \mu_i, \sigma_i^2) \\ p(\tilde{\mathbf{v}}_i \mid \mu_i, \check{\mathbf{v}}_i, \sigma_i^2, \mathbf{y}_i) \propto \left[\prod_{j=1}^{\tilde{y}_{i1}} I\{h_{\hat{\kappa}}(c_0) \leq v_{ij} \leq h_{\hat{\kappa}}(c_1)\} \phi(v_{ij}; \mu_i, \sigma_i^2) \right] \\ \times \left[\prod_{j=\tilde{y}_{i1}+1}^{\tilde{y}_{i2}} I\{h_{\hat{\kappa}}(c_1) \leq v_{ij} \leq h_{\hat{\kappa}}(c_2)\} \phi(v_{ij}; \mu_i, \sigma_i^2) \right] \\ \times \cdots \times \left[\prod_{j=\tilde{y}_{i,G-1}+1}^{n_i} I\{h_{\hat{\kappa}}(c_{G-1}) \leq v_{ij} < h_{\hat{\kappa}}(c_G)\} \phi(v_{ij}; \mu_i, \sigma_i^2) \right], \\ p(\check{\mathbf{v}}_i \mid \mu_i, \tilde{\mathbf{v}}_i, \sigma_i^2, \mathbf{y}_i) = \prod_{j=n_i+1}^{N_i} \phi(v_{ij}; \mu_i, \sigma_i^2), \\ p(\sigma_i^2 \mid \mu_i, \tilde{\mathbf{v}}_i, \check{\mathbf{v}}_i, \mathbf{y}_i) \propto (\sigma_i^2)^{(-\hat{\lambda}/2+1)-1} \exp\left(-\frac{\hat{\lambda}\hat{\varphi}_i}{2\sigma_i^2}\right) \prod_{j=1}^{N_i} \phi(v_{ij}; \mu_i, \sigma_i^2),$$

which leads to the full conditional distributions (13).

A.2 Appendix for the HLS data

HLS in 2013 was conducted based on the two stage stratified sampling. The first stage sampling strata corresponds to the sampling areas used in Population Census in 2010 and the second stage sampling strata consists of the households in the area. We have the information which areas are sampled in the first stage and the total number of the households in each area at the time when Population Census in 2010 was conducted. We also know which municipality the sampled areas in the first stage belong to. In the second stage, all households are sampled if the total number of the households in the area is less than 70, otherwise the number of sampled households is approximately 50. Combining these information, the sample size in each municipality is estimated.

References

- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An error-components model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28–36.
- Chandra, H., Salvati, N., Chambers, R. and Tzavidis, N. (2012). Small area estimation under spatial nonstationarity. *Computational Statistics and Data Analysis*, **56**, 2875–2888.
- Chotipakanich, D. (2008). *Modeling Income Distributions and Lorenz Curves*, Springer, New York.
- Datta, G. and Ghosh, M. (2012). Small area shrinkage estimation. *Statistical Science*, **27**, 95–114.
- Dempster, A.P., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, **39**, 1–38.
- Diallo, M.S. and Rao, J.N.K. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, **45**, 1092–1116.
- Doornik, J. (2007). *Ox: object oriented matrix programming*, Timberlake Consultants Press, London.
- Fay, R.E. and Herriot, R.A. (1979). Estimates of income for small places: An application of James–Stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.
- Guadarrama, M., Molina, I. and Rao, J.N.K. (2018). Small area estimation of general parameters under complex sampling designs. *Computational Statistics and Data Analysis*, **121**, 20–40.
- Heitjan, D. F. (1989). Inference from grouped continuous data: a review. *Statistical Science*, **4**, 164–179.
- Jaing, J. and Nguyen, T. (2012). Small area estimation via heteroscedastic nested-error regression. *The Canadian Journal of Statistics*, **40**, 588–603.
- Kubokawa, T., Sugasawa, S., Ghosh, M. and Chaudhuri, S. (2016). Prediction in heteroscedastic nested error regression models with random dispersions. *Statistica Sinica*, **26**, 465–492.
- Molina, I. and Marhuenda, Y. (2018). *sae*: Small Area Estimation. R package version 1.2.

- Molina, I. and Rao, J.N.K. (2010). Small area estimation of poverty indicators. *The Canadian Journal of Statistics*, **38**, 369–385.
- Rao, J.N.K. and Molina, I. (2015). *Small Area Estimation*, Wiley, New York.
- Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science*, **28**, 40–68.
- Richard, J.-F. and Zhang, W. (2007). Efficient high-dimensional importance sampling. *Journal of Econometrics*, **141**, 1385–1411.
- Shi, J.Q. and Copas, J. (2002). Publication bias and meta-analysis for 2×2 tables: an average Markov chain Monte Carlo EM algorithm. *Journal of the Royal Statistical Society Series B*, **64**, 221–236.
- Sugasawa, S. and Kubokawa, T. (2019). Adaptively transformed mixed-model prediction of general finite-population parameters. *Scandinavian Journal of Statistics*, to appear.