Journal of Survey Statistics and Methodology (2023) ***00****, 1–20*

CONJUGATE MODELING APPROACHES FOR SMALL AREA ESTIMATION WITH HETEROSCEDASTIC STRUCTURE

PAUL A. PARKER*
SCOTT H. HOLAN
RYAN JANICKI

Small area estimation (SAE) has become an important tool in official statistics, used to construct estimates of population quantities for domains with small sample sizes. Typical area-level models function as a type of heteroscedastic regression, where the variance for each domain is assumed to be known and plugged in following a design-based estimate. Recent work has considered hierarchical models for the variance, where the design-based estimates are used as an additional data point to model the latent true variance in each domain. These hierarchical models may incorporate covariate information but can be difficult to sample from in high-dimensional settings. Utilizing recent distribution theory, we explore a class of Bayesian hierarchical models for SAE that smooth both the design-based estimate of the mean and the variance. In addition, we develop a class of unit-level models for heteroscedastic Gaussian response data. Importantly, we incorporate both covariate information as well as spatial dependence, while retaining a conjugate model structure that allows for efficient sampling. We illustrate our methodology through an empirical simulation study as well as an application using data from the American Community Survey.

Paul A. Parker is with the Department of Statistics, University of California Santa Cruz, 1156 High St, Santa Cruz, CA 95064, USA and U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA. Scott H. Holan is with the Department of Statistics, University of Missouri, 146 Middlebush Hall, Columbia, MO 65211, USA and U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA. Ryan Janicki is with the U.S. Census Bureau, 4600 Silver Hill Road, Washington, DC 20233, USA

The authors would like to thank Associate Editor and two anonymous referees for comments that helped improve this paper. This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical issues are those of the authors and not those of the NSF or U.S. Census Bureau. This research was partially supported by the U.S. National Science Foundation (NSF) under NSF grants SES-1853096 and NCSE-2215168.

*Address correspondence to Paul A. Parker, Department of Statistics, University of California Santa Cruz, 1156 High St, Santa Cruz, CA 95064, USA; E-mail: paulparker@ucsc.edu.

https://doi.org/10.1093/jssam/smad002
Published by Oxford University Press on behalf of the American Association for Public Opinion Research 2023.

KEYWORDS: Gibbs sampling; ICAR; Mixed models; Multivariate log-Gamma; Spatial.

Statement of Significance

This article introduces new methodology for fitting Bayesian small area estimation models that consider various types of heteroscedastic structure. Current approaches do not yield conjugate full conditional distributions and are difficult to scale due to the challenge of sampling from the posterior distribution. Our proposed approach develops a conditionally conjugate model structure that allows for straightforward Gibbs sampling. This allows us to scale our approach by building richer structure into the model. Thus, we build on this by developing models with spatial dependence structure for both the mean and variance. Finally, we develop a model for unit-level data with heteroscedastic structure under informative sampling.

1. INTRODUCTION

Large-scale national surveys such as the American Community Survey (ACS) are typically designed to produce reliable estimates for a variety of demographic and household characteristics for large geographic regions. However, data users often need population estimates at smaller areas or domains than can be reliably provided by direct estimates, which use only domain-specific sample. By small area or domain, we mean any geographic region or demographic subpopulation for which the domain-specific sample size is small. These direct estimates, which use only domain-specific survey response data, may not be sufficiently precise for reliable inference due to small sample sizes and unreasonably high standard errors. To meet the demand for more granular estimates based on smaller sample sizes, model-based approaches, or small area estimation (SAE) models, are commonly used in place of direct estimates. Area-level models such as the popular Fay–Herriot (FH) model (Fay and Herriot 1979) can incorporate covariates or other dependencies to smooth the direct estimates and improve precision by "borrowing strength" from areas with large sample size. The FH model assumes that the sampling variance of the direct estimator is fixed and known, which is rarely the case in real survey data settings. In practice, the sampling variance is estimated from the survey data and then plugged into the model. Sampling variances tend to vary across geographic areas; thus, these area-level SAE models may be seen as a type of model for heteroscedastic data.

Recently, a variety of extensions to the FH model have been considered to address the issue of unknown sampling variances (You and Chapman 2006; You 2021). For example, Maiti et al. (2014) use a hierarchical modeling approach to jointly smooth both direct mean and variance estimates. Sugasawa et al. (2017) fit a similar model in a Bayesian setting. In both cases, covariates may be used to aid in the variance smoothing, but in the Bayesian case, Sugasawa et al. (2017) required the use of a Metropolis–Hastings step within the Gibbs sampler, due to a full-conditional distribution with unknown form.

Incorporating estimates of the sampling error variance of the direct estimator is a topic of significant interest in the course of producing model-based estimates for official statistics. In this direction, Bradley et al. (2016) propose a model to include both the direct estimates and an estimate of the sampling error variance in a Poisson area-level spatial change-of-support model for the ACS. The proposed model uses the variance estimates as another data source and exploits the Poisson equidispersion assumption (i.e., equal mean and variance) to condition on a common latent process.

The approach proposed here differs from Bradley et al. (2016) and exploits the distribution theory of Bradley et al. (2020). Specifically, we extend the approach proposed in Parker et al. (2021) for heteroscedastic data to the SAE setting for both area-level and unit-level models. Importantly, our approach yields models that are fully conjugate for both the mean and variance regression parameters and thus leads to extremely computationally efficient estimation (see sections 2 and 3 for additional detail).

Although SAE methods have a long and rich history, the literature on jointly modeling the mean (direct estimates) and variance (sampling error variance) is significantly more recent by comparison. For example, Savitsky and Gershunskaya (2022) propose a Bayesian nonparametric model that jointly models the mean and the variance in the context of the Consumer Expenditure Survey. Similarly, Polettini (2017) proposes a semiparametric Bayesian FH model that shrinks both the means and variance. In general, these approaches can be computationally expensive and, thus, there is a need for models that scale computationally to meet the high-dimensional demands that are faced by statistical agencies and subject-matter practitioners. The method proposed here meets this demand.

Outside of the SAE literature, there exists a substantial literature on joint models for the mean and covariance. For example, Pourahmadi (1999, 2011) and Chen and Dunson (2003), among others, model Cholesky-based factorizations of unstructured covariances. In general these methods are extremely useful; nevertheless, they are not immediately applicable to the SAE problems considered here. In particular, the SAE setting is comprised of a diagonal covariance (variance) structure with a model on the latent variances. A comprehensive review of Cholesky-based joint mean and covariance modeling can be found in Pourahmadi (2013) and the references therein.

There has also been significant research on modeling covariance structure using covariates, many of which arise in the spatio-temporal literature. For example, see Schmidt et al. (2011) and Gladish et al. (2014), among others. The main challenge that arises in this context is computational. In general, most of the proposed methods proceed using Bayesian methods and lead to non-conjugate updates, necessitating migration away from straightforward Gibbs sampling. One notable exception is proposed by Parker et al. (2021), which provides a starting point for the method proposed here.

This article also proposes a novel unit-level SAE model for heteroscedastic Gaussian survey data. The main complication of unit-level SAE modeling is accounting for the survey design in the model. When the survey design is noninformative, in the sense that the distribution of the sampled response values is the same as the distribution of the unsampled values, the effect of the survey design can largely be ignored. However, when the probability of selection in the survey is correlated with the response variable, the survey design is said to be informative, in which case the population distribution and the distribution for the sampled data will differ. Under informative sampling, the survey design must be carefully accounted for in the SAE model to avoid biased estimates. To this end, we also propose a heteroscedastic unit-level model under informative sampling using the pseudo-likelihood (Binder 1983; Skinner 1989; Savitsky and Toth 2016). For a comprehensive review of unit-level approaches to SAE, see Parker et al. (2019) and the references therein.

Although the methodology proposed here is extremely general and applies to a broad set of applications that are encountered by data users and official statistical agencies, our motivating example considers income estimation and is related to the Small Area Income and Poverty Estimates Program (SAIPE). In particular, SAIPE produces income estimates for all US states and counties (Bell et al. 2016) and in many cases, the estimates may be used in the administration of federal programs and the allocation of federal funds to local juridictions; for more details, see https://www.census.gov/programs-surveys/saipe/about.html. Therefore, using the ACS, we demonstrate our proposed approach by estimating the average income by PUMA for the state of California.

This article proceeds as follows. Section 2 provides background and introduces our heteroscedastic model for area-level data. Unit-level models are presented in section 3. To evaluate the effectiveness of our approach, an empirical simulation study is provided in section 4 and an analysis of ACS income data is presented in section 5. Discussion is provided in section 6.

2. AREA-LEVEL HETEROSCEDASTIC MODELS

2.1 Background

The FH model (Fay and Herriot 1979) is given by

$$\begin{aligned} y_i|\theta_i, \sigma_i^2 &\overset{\text{ind}}{\sim} \text{N}(\theta_i, \sigma_i^2),\ i = 1, \ldots, d \\ \theta_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + \nu_i \\ \nu_i &\overset{\text{ind}}{\sim} \text{N}(0, \sigma_\nu^2), \end{aligned}$$

where y_i is the direct estimate, σ_i^2 is the sampling variance of the direct estimate and is assumed known, and $i = 1, \ldots, d$ indexes the small areas of interest. Thus, the FH model may be seen as a type of heteroscedastic data model where the variance is known. Typically, the true value of σ_i^2 is unknown; thus, a design-based estimate, s_i^2, is plugged in instead.

To reflect the additional uncertainty attributed to s_i^2, a second data model may be used. For example, You and Chapman (2006) suggest the following model to address the issue of unknown sample variances:

$$\begin{aligned} y_i|\theta_i, \sigma_i^2 &\overset{\text{ind}}{\sim} \text{N}(\theta_i, \sigma_i^2),\ i = 1, \ldots, d \\ s_i^2|\sigma_i^2 &\overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{n_i - 1}{2}, \frac{n_i - 1}{2\sigma_i^2}\right),\ i = 1, \ldots, d \\ \theta_i &= \boldsymbol{x}_i'\boldsymbol{\beta} + \nu_i \\ \nu_i &\overset{\text{ind}}{\sim} \text{N}(0, \sigma_\nu^2) \\ \sigma_i^2 &\overset{\text{ind}}{\sim} \text{IG}(a_i, b_i), \end{aligned}$$

where n_i represents the sample size in area i and $\text{IG}(a, b)$ denotes an inverse gamma distribution with shape parameter a and scale parameter b. In principal, the data model for s_i^2 given here is only valid in the case of a simple random sample within area i. For complex sample designs, it may be important to give careful consideration to the degrees of freedom. Although estimation of the appropriate degrees of freedom is beyond the scope of this work, for more discussion on this matter, see Maples et al. (2009).

Sugasawa et al. (2017) provide a Bayesian extension of this approach that considers covariates in the variance model by letting $\sigma_i^2 \overset{\text{ind}}{\sim} \text{IG}\left(a_i, b_i \exp(\boldsymbol{x}_{2i}'\boldsymbol{\beta}_2)\right)$. Although the use of covariates here may improve small area estimates, Sugasawa et al. (2017) required the use of a Metropolis–Hastings sampler for $\boldsymbol{\beta}_2$, which can be extremely difficult to tune, especially in high dimensions.

2.2 Conjugate Priors for Heteroscedastic Models

The foundation of our modeling framework is the multivariate log-Gamma (MLG) distribution, introduced by Bradley et al. (2018, 2020). The MLG distribution was initially developed to model dependent data using a Poisson likelihood. The density for the MLG distribution is given as

$$f(\boldsymbol{y}) = \det(\boldsymbol{V}^{-1}) \left\{ \prod_{i=1}^{n} \frac{\kappa_i^{\alpha_i}}{\Gamma(\alpha_i)} \right\} \exp\left[\boldsymbol{\alpha}' \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) - \boldsymbol{\kappa}' \exp\left\{ \boldsymbol{V}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}) \right\} \right],$$

denoted by $\text{MLG}(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$. The length n vector $\boldsymbol{\mu}$ acts as a centrality parameter and the $n \times n$ matrix $\boldsymbol{V}$ controls the correlation structure. The length n vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\kappa}$ are shape and rate parameters, respectively. Sampling from $\boldsymbol{Y} \sim \text{MLG}(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$ is straightforward using the following steps:

(1) Generate a vector $\mathbf{g}$ as n independent Gamma random variables with shape α_i and rate κ_i, for $i = 1, \ldots, n$.
(2) Let $\mathbf{g}^* = \log(\mathbf{g})$.
(3) Let $\mathbf{Y} = \mathbf{V}\mathbf{g}^* + \boldsymbol{\mu}$.

In most cases, Bayesian inference using the MLG prior distribution also requires simulation from the conditional multivariate log-Gamma (cMLG) distribution. Letting $\mathbf{Y} \sim \text{MLG}(\boldsymbol{\mu}, \mathbf{V}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$, Bradley et al. (2018) show that $\mathbf{Y}$ can be partitioned into $(\mathbf{Y}_1', \mathbf{Y}_2')'$, where $\mathbf{Y}_1$ is r-dimensional and $\mathbf{Y}_2$ is $(n - r)$-dimensional. The matrix $\mathbf{V}^{-1}$ is also partitioned into $[\mathbf{H}\,\mathbf{B}]$, where $\mathbf{H}$ is an $n \times r$ matrix and $\mathbf{B}$ is an $n \times (n - r)$ matrix. Then,

$$\boldsymbol{Y_1} | \boldsymbol{Y_2} = \boldsymbol{d}, \boldsymbol{\mu}^*, \boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa} \sim \text{cMLG}(\boldsymbol{\mu}^*, \boldsymbol{H}, \boldsymbol{\alpha}, \boldsymbol{\kappa}),$$

with density

$$M \exp\left\{ \boldsymbol{\alpha}' \boldsymbol{H} \boldsymbol{Y_1} - \boldsymbol{\kappa}' \exp(\boldsymbol{H} \boldsymbol{Y_1} - \boldsymbol{\mu}^*) \right\},$$

where $\boldsymbol{\mu}^* = \mathbf{V}^{-1}\boldsymbol{\mu} - \mathbf{B}\mathbf{d}$, and M is a normalizing constant. Bradley et al. (2018) show that it is also straightforward to sample a draw from the cMLG distribution using the linear transformation $(\mathbf{H}'\mathbf{H})^{-1}\mathbf{H}'\mathbf{Y}$, where $\mathbf{Y}$ is first sampled from $\text{MLG}(\boldsymbol{\mu}, \mathbf{I}, \boldsymbol{\alpha}, \boldsymbol{\kappa})$.

Another key result given by Bradley et al. (2018) is that $\text{MLG}(\mathbf{c}, \alpha^{1/2}\mathbf{V}, \alpha\mathbf{1}, \alpha\mathbf{1})$ converges in distribution to a multivariate normal distribution with mean $\mathbf{c}$ and covariance matrix $\mathbf{V}\mathbf{V}'$ as the value of α approaches infinity. This allows for the computational benefits of MLG priors while maintaining effectively the same shape and structure as a Gaussian prior.

Although the original use of the MLG distribution was as a prior in high-dimensional Poisson regression, recently, Parker et al. (2021) found that the MLG distribution acts as a conjugate prior for variance regression when

using a negative log link function. This insight is the basis for our proposed approach.

2.3 Proposed Area-Level Approach

To account for the additional uncertainty around s_i^2 in the context of area-level SAE, we construct a joint model for the direct point estimate and variance. Critically, this model relies on the MLG distribution as a conjugate prior for the variance regression parameters. Our heteroscedastic area-level model (HALM) is given as

$$
\begin{aligned}
y_i|\theta_i, \sigma_i^2 &\overset{\text{ind}}{\sim} \text{N}(\theta_i, \sigma_i^2),\ i = 1, \ldots, d \\
s_i^2|\sigma_i^2 &\overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{n_i - 1}{2}, \frac{n_i - 1}{2\sigma_i^2}\right),\ i = 1, \ldots, d \\
\theta_i &= \boldsymbol{x}_i'\boldsymbol{\beta}_1 + \eta_{1i} \\
-\log(\sigma_i^2) &= \boldsymbol{x}_i'\boldsymbol{\beta}_2 + \eta_{2i} \\
\boldsymbol{\eta}_1|\sigma_{\eta_1}^2 &\sim \text{N}(\mathbf{0}, \sigma_{\eta_1}^2\mathbf{I}) \\
\boldsymbol{\eta}_2|\sigma_{\eta 2}^2 &\sim \text{MLG}(\mathbf{0}, \alpha^{1/2}\sigma_{\eta_2}\mathbf{I}, \alpha\mathbf{1}, \alpha\mathbf{1}) \\
\boldsymbol{\beta}_1 &\sim \text{N}(\mathbf{0}, \sigma_{\beta_1}^2\mathbf{I}) \\
\boldsymbol{\beta}_2 &\sim \text{MLG}(\mathbf{0}, \alpha^{1/2}\sigma_{\beta_2}\mathbf{I}, \alpha\mathbf{1}, \alpha\mathbf{1}) \\
\sigma_{\eta_1}^2 &\sim \text{IG}(a, b) \\
\sigma_{\eta_2} &\sim \text{N}^+(0, c).
\end{aligned}
$$

Here, y_i represents the direct estimate of an unknown population quantity, θ_i, while s_i^2 represents the design-based variance around this estimate. Note that this model requires $n_i > 1$, $i = 1, \ldots, d$. The unknown population quantity is written as a linear combination of the length p vector of covariates, $\boldsymbol{x}_i$, as well as an area-level random effect, η_{1i}. The unknown variance, σ_i^2, is modeled using the negative log link function as a linear combination of $\boldsymbol{x}_i$ as well as an additional area-level random effect, η_{2i}. To establish conjugate full-conditional distributions, β_2 takes on an MLG prior distribution that is asymptotically equivalent to a $\text{N}(\mathbf{0}, \sigma_\beta^2\mathbf{I})$ distribution. Similarly, conditional on $\sigma_{\eta_2}^2$, $\boldsymbol{\eta}_2$ takes an MLG prior distribution that is asymptotically equivalent to a $\text{N}(\mathbf{0}, \sigma_{\eta_2}^2\mathbf{I})$ distribution. Finally, we place a conjugate inverse Gamma prior distribution on $\sigma_{\eta_1}^2$ as well as a half-normal prior on σ_{η_2}. We note that this prior for σ_{η_2} is not conjugate and, thus, requires a Metropolis–Hastings step. However, this is only for a single parameter and we have found that there is very little effect on the mixing of the MCMC due to this. We use a random-walk Metropolis–Hastings step with a Normal distribution truncated below at zero for a proposal

distribution, although, depending on the setting, it may be helpful to consider other proposals. The model is completed by the specification of $\alpha, \sigma^2_{\beta_1}, \sigma^2_{\beta_2}, a, b, c > 0$. In practice, we work with relatively diffuse priors by using $\sigma^2_{\beta_1} = \sigma^2_{\beta_2} = 1{,}000$, $a = b = 0.5$, and $c = 5$. The value for α should be sufficiently large to invoke the asymptotic equivalence to the multivariate normal distribution. Similar to Bradley et al. (2018), we have found $\alpha = 1{,}000$ to be sufficient for our purposes, although in some other cases this may be data dependent. The full conditional distributions for this model are given in appendix A in the supplementary data online.

Often within SAE, more precise estimates can be generated through the use of spatial dependence modeling (e.g., see Marhuenda et al. 2013; Porter et al. 2015). This motivates the need for spatially correlated prior structures for the mean and variance models. To this end, we develop a spatial variant of the HALM, termed the spatial heteroscedastic area-level model (SHALM). This model is similar to HALM, with the replacement of the prior structure for $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$,

$$\begin{aligned}\boldsymbol{\eta}_1|\sigma^2_{\eta_1} &\sim \mathrm{N}\left(\mathbf{0}, \sigma^2_{\eta_1}(\boldsymbol{D} - \boldsymbol{W})^{-1}\right)\\ \boldsymbol{\eta}_2|\sigma^2_{\eta 2} &\sim \mathrm{MLG}(\mathbf{0}, \alpha^{1/2}\sigma_{\eta_2}(\boldsymbol{D} - \boldsymbol{W})^{-1/2}, \alpha\mathbf{1}, \alpha\mathbf{1}).\end{aligned}$$

Here, the $d \times d$ matrix $\boldsymbol{W}$ is an area-level adjacency matrix, with entry $W_{ij} = 1$ if areas i and j share a border and $W_{ij} = 0$ otherwise. By convention, an area is not considered a neighbor of itself, resulting in a zero value for all diagonal elements. The $d \times d$ matrix $\boldsymbol{D}$ is a diagonal matrix, where the i^{th} entry corresponds to the number of neighbors shared by area i, or equivalently, the sum of the ith row of the matrix $\boldsymbol{W}$. This prior for $\boldsymbol{\eta}_1$ is known as the intrinsic conditional autoregressive (ICAR) prior (Besag et al. 1991). Note that the matrices $\boldsymbol{D}$ and $\boldsymbol{W}$ are computed prior to model fitting and do not involve any unknown parameters. Similarly, the prior for $\boldsymbol{\eta}_2$ is asymptotically equivalent to an ICAR prior.

3. UNIT-LEVEL HETEROSCEDASTIC MODELS

An increasingly common alternative to area-level models for SAE is that of unit-level modeling. Unit-level models opt to model the survey data directly rather than the design-based estimates as in the area-level case. For example, the basic unit-level model (BULM) was introduced by Battese et al. (1988) and is written as

$$\begin{aligned}y_{ij} &\overset{ind}{\sim} \mathrm{N}(\mu_{ij}, \sigma^2), j \in \mathcal{S}\\ \mu_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \eta_i\\ \eta_i &\overset{iid}{\sim} \mathrm{N}(0, \sigma^2_\eta).\end{aligned}$$

Here, y_{ij} is the response for unit j in the sample, $\mathcal{S}$, residing in area i, while $\boldsymbol{x}_{ij}$ is a vector of unit-level covariates. It is important to note that the covariates used in the unit-level model must be known for both sampled and non-sampled individuals. The area-level random effects, η_i, allow for dependence among respondents within the same area. For this model, as well as other unit-level modeling approaches, the model may be fit using the observed sample data and predictions can be made for the entire non-sampled population using $\hat{\mu}_{ij} = \boldsymbol{x}'_{ij}\hat{\boldsymbol{\beta}} + \hat{\eta}_i$. In essence, this results in a synthetic population that may be aggregated as necessary to produce area-level estimates at the desired spatial resolution. Finally, in a Bayesian setting, this may be done for each sample of the parameters from the posterior distribution.

One major limitation of the BULM is that it assumes the survey design to be ignorable. Many surveys result in an informative sampling scheme in which there is a relationship between the response of interest and the unit probabilities of selection. Let $\mathcal{U}_i$ be an enumeration of the finite population in area i, and let $\mathcal{S}_i \subset \mathcal{U}_i$ be the survey sample from area i selected according to a known sampling scheme with inclusion probabilities $P(j \in \mathcal{S}_i) = \pi_{ij}$. Define the survey weights as $w_{ij} = 1/\pi_{ij}$. Informative survey designs occur when survey inclusion indicators are correlated with survey response variables, even after conditioning on observable covariates and design variables. In these situations, use of a model that does not consider the survey design may result in large biases. One popular solution to this problem is the use of an exponentially weighted pseudo-likelihood (Binder 1983; Skinner 1989). More recently, Savitsky and Toth (2016) popularized the use of a pseudo-likelihood in general Bayesian model settings. This results in a pseudo-posterior distribution that is proportional to the product of the pseudo-likelihood and the prior distribution,

$$\hat{\pi}(\boldsymbol{\theta}|\mathbf{y}, \tilde{\mathbf{w}}) \propto \left\{ \prod_{j \in \mathcal{S}} f(y_{ij}|\boldsymbol{\theta})^{\tilde{w}_{ij}} \right\} \pi(\boldsymbol{\theta}).$$

In this case, $\tilde{w}_{ij}$ represents the survey weights after scaling to sum to the sample size. For example, a Bayesian pseudo-likelihood alternative to the BULM may be written as

$$\begin{aligned} \boldsymbol{y}|\boldsymbol{\mu}, \sigma^2 &\propto \prod_{j \in \mathcal{S}} \mathrm{N}(y_{ij}|\mu_{ij}, \sigma^2)^{\tilde{w}_{ij}} \\ \mu_{ij} &= \boldsymbol{x}'_{ij}\boldsymbol{\beta} + \eta_i \\ \eta_i &\overset{iid}{\sim} \mathrm{N}(0, \sigma^2_\eta). \end{aligned} \tag{1}$$

Although there are many other approaches to account for an informative sample design, our focus here will be strictly on the Bayesian pseudo-likelihood. For an overview of alternative approaches, see Parker et al. (2019).

Another limitation of the BULM is the assumption of constant variance across survey units. In practice, the dispersion of a particular response of interest may vary along with certain covariates or by geographic region.

3.1 Proposed Unit-Level Approach

To address limitations of the BULM, we propose a unit-level model that accounts for a possibly informative sample design while also relaxing the constant variance assumption. This approach is termed the heteroscedastic unit-level model (HULM) and is written as

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\sigma}^2 &\propto \prod_{j\in\mathcal{S}} \mathrm{N}(y_{ij}|\mu_{ij},\sigma_{ij}^2)^{\tilde{w}_{ij}} \\
\mu_{ij} &= \boldsymbol{x}_{ij}'\boldsymbol{\beta}_1 + \eta_{1i} \\
-\log(\sigma_{ij}^2) &= \boldsymbol{x}_{ij}'\boldsymbol{\beta}_2 + \eta_{2i} \\
\boldsymbol{\eta}_1|\sigma_{\eta_1}^2 &\sim \mathrm{N}(\mathbf{0},\sigma_{\eta_1}^2) \\
\boldsymbol{\eta}_2|\sigma_{\eta 2}^2 &\sim \mathrm{MLG}(\mathbf{0},\alpha^{1/2}\sigma_{\eta_2}\mathbf{I},\alpha\mathbf{1},\alpha\mathbf{1}) \\
\boldsymbol{\beta}_1 &\sim \mathrm{N}(\mathbf{0},\sigma_{\beta_1}^2) \\
\boldsymbol{\beta}_2 &\sim \mathrm{MLG}(\mathbf{0},\alpha^{1/2}\sigma_{\beta_2}\mathbf{I},\alpha\mathbf{1},\alpha\mathbf{1}) \\
\sigma_{\eta_1}^2 &\sim \mathrm{IG}(a,b) \\
\sigma_{\eta_2} &\sim \mathrm{N}^+(0,c).
\end{aligned}$$

This approach uses a Gaussian pseudo-likelihood with individual mean and variance to model the data. The individual means, μ_{ij}, are written as a linear combination of the length p covariate vector, $\boldsymbol{x}_{ij}$, as well as an area-level random effect, η_{1i}. Using a negative log link function, the individual variances, σ_{ij}^2, are also modeled as a linear combination of $\boldsymbol{x}_{ij}$ and an additional area-level random effect, η_{2i}.

The model structure for $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$ is identical to the HALM. In particular, $\boldsymbol{\eta}_1$ and $\boldsymbol{\eta}_2$ are modeled hierarchically with unknown variance parameters while both $\boldsymbol{\eta}_2$ and $\boldsymbol{\beta}_2$ take MLG prior distributions to allow for computational feasibility. Although not directly explored here, it is straightforward to extend this approach to consider spatially correlated random effects similar to the SHALM (e.g., see Sun et al. 2022). The full conditional distributions for this model are given in appendix B in the supplementary data online.

4. EMPIRICAL SIMULATION STUDY

In many cases, the decision whether to use an area-level versus a unit-level model will depend on whether an analyst has access to unit-level microdata.

In addition, in situations where many areas contain little or no data, an area-level approach may be infeasible due to the lack of appropriate design-based estimates. Here, we aim to compare a variety of both area-level and unit-level models by devising an appropriate simulation study. However, we note that, in practice, it is possible that some subset of the models explored here may not be appropriate.

To construct our simulation study, we require a population of individuals that we may sample from. Rather than using a synthetic population drawn from some parametric distribution, we instead take an existing survey dataset and treat it as our population to preserve many of the characteristics associated with the real data. In particular, we use the public-use microdata sample (PUMS) from the ACS. We restrict our scope to the 1-year PUMS data with positive income for the state of California only, as it is often of interest to estimate the mean income among the population that actually earns an income. This empirical population contains roughly 179,000 individuals. Each individual is associated with a geographic region known as the public-use microdata area (PUMA). The state of California contains 265 PUMAs. Ultimately our goal is to estimate average income of positive income earners by PUMA using a sample from this population.

We consider two different approaches for subsampling from the empirical population. First, we take a stratified random sample by PUMA with a simple random sample without replacement of five observations per PUMA. Second, we take a probability proportional to size sample using the Poisson method (Brewer et al. 1984) with an expected total sample size of 1,000. For the second sample design, we construct the size variables as $\exp(2 + 0.3 \times w_{ij} + 0.3 \times \tilde{y}_{ij})$, where w_{ij} is the original scaled survey weight and $\tilde{y}_{ij}$ is the income for unit j in area i, after scaling to have mean zero and variance one. The use of income in the size variable enforces an informative design. For both sampling approaches, we repeat the sampling and estimation procedure $K = 100$ times. Horvitz–Thompson direct estimates of mean income are calculated for each PUMA.

We consider two different unit-level models for this study. First, we present the Bayesian pseudo-likelihood alternative of the BULM (PL-BULM) given in (1). We also compare to the proposed HULM. We note that exploratory analysis indicated that the spread of income did not vary by PUMA, so for the HULM, we constrain $\boldsymbol{\eta}_2 = \mathbf{0}$. For both unit-level approaches, we model income after taking a log transformation, and we use age, sex, and race as covariates. We note that at the unit level, Gaussian models for income are a starting point, but further work is necessary in this area. For the PUMS data in particular, there is some rounding and top-coding that occurs as a disclosure avoidance mechanism. The unit-level models explored here are adequate for characterization of the first moment of income, but a more complex model, such as a censored or inflated model, seems necessary to adequately model the full distributional uncertainty.

We also consider four area-level models. First, we fit the basic FH model. Second, we consider the model used by Sugasawa et al. (2017) that shrinks both the design-based mean and variance estimates (STK). Lastly, we fit both the proposed HALM and SHALM. All area-level models are fit after log transforming the design-based estimates of income and using the delta method for variance estimates. Log population size was used as a covariate. All models were fit using MCMC with 3,000 iterations, discarding the first 1,000 iterations as burn-in. Convergence was assessed via traceplots of the sample Markov chains, with no lack of convergence detected.

We are primarily interested in two forms of assessment for these models. First, we examine the root mean squared error (RMSE) of our point estimates,

$$\sqrt{\sum_{k=1}^{K}\frac{(\hat{\theta}_k-\theta)^2}{K}}.$$

Here, θ represents the true population quantity of interest while $\hat{\theta}_k$ represents an estimate for sample dataset k. RMSE has the desirable property of being composed of both a bias and variance term. We also consider the interval score (Gneiting and Raftery 2007) for our 95 percent credible interval estimates,

$$\frac{1}{K}\sum_{k=1}^{K}\left\{(u_k-\ell_k)\ +\ \frac{2}{\alpha}(\ell_k-\theta)I(\theta<\ell_k)\ +\ \frac{2}{\alpha}(\theta-u_k)I(\theta>u_k)\right\},$$

where $\alpha=0.05$, u_k is the upper bound of the interval and ℓ_k is the lower bound of the interval for sample dataset k. For the interval score, a lower score is desirable. Thus, narrow intervals are rewarded, but a penalty is incurred if the interval misses the true value. Along with these, we report the absolute bias,

$$\left|\frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k-\theta\right|$$

and the coverage rate,

$$\frac{1}{K}\sum_{k=1}^{K}I(\ell_k<\theta<u_k).$$

Results for the stratified sampling design and the probability proportional to size design are summarized in tables 1 and 2, respectively. All results are averaged across PUMAs. RMSE is presented relative to the direct estimator, where a value less than 1 indicates a reduction in RMSE relative to the direct estimator. For the stratified sampling design, all models were able to reduce the RMSE relative to the direct estimator with the exception of the PL-BULM.

Table 1. Empirical Simulation Results for Stratified Random Sampling by PUMA using the 2018 1-Year American Community Survey Public-Use Microdata Sample

Estimator	Rel. RMSE	Abs. bias ($\times 10^3$)	Cov. rate	Int. score ($\times 10^4$)
PL-BULM	1.080 (1.08)	20.299 (9.46)	0.368 (0.37)	30.570 (25.52)
HULM	0.687 (0.79)	11.356 (7.37)	0.596 (0.37)	14.720 (17.14)
FH	0.694 (0.11)	5.126 (4.80)	0.894 (0.04)	8.790 (3.55)
HALM	0.640 (0.18)	7.677 (7.60)	0.956 (0.09)	6.695 (4.13)
SHALM	**0.561** (0.17)	6.759 (6.34)	0.933 (0.12)	**6.031** (4.82)
STK	0.636 (0.15)	6.958 (6.82)	0.952 (0.07)	6.648 (3.92)

NOTE.—All results are averaged across PUMAs (standard deviation across PUMAs is shown in parentheses). RMSE is presented relative to the direct estimator. For RMSE and Interval Score, the best performing estimator is given in bold.

Table 2. Empirical Simulation Results for Probability Proportional to Size Sampling using the 2018 1-Year American Community Survey Public-Use Microdata Sample

Estimator	Rel. RMSE	Abs. bias ($\times 10^3$)	Cov. rate	Int. score ($\times 10^4$)
PL-BULM	0.766 (0.71)	19.607 (9.33)	0.392 (0.33)	30.846 (23.48)
HULM	0.646 (0.65)	16.509 (8.48)	0.461 (0.35)	23.574 (20.17)
FH	0.492 (0.18)	6.815 (4.27)	0.955 (0.04)	8.855 (4.33)
HALM	0.442 (0.26)	9.765 (6.95)	0.958 (0.09)	6.800 (2.76)
SHALM	**0.401** (0.21)	8.481 (6.13)	0.913 (0.14)	**6.267** (4.15)
STK	0.429 (0.24)	9.195 (6.73)	0.958 (0.09)	6.468 (2.92)

NOTE.—All results are averaged across PUMAs (standard deviation across PUMAs is shown in parentheses). RMSE is presented relative to the direct estimator. For RMSE and Interval Score, the best performing estimator is given in bold.

Thus, the proposed HULM was able to offer substantial improvement over a model that assumes constant variance. For the HULM, the interval estimates were also improved relative to the PL-BULM, although as discussed previously, further model development here is desirable. In terms of area-level approaches, the FH model performed worst in terms of both RMSE and interval score. Therefore, shrinkage of both the mean and variance appears to be important to improve the quality of generated estimates. The HALM and STK model resulted in quite similar RMSE and interval scores. Finally, the SHALM was able to leverage spatial dependence resulting in the lowest RMSE and interval scores across all models. Results were similar for the probability proportional to size design, indicating robustness to the assumption of a simple random sample within each area used in the variance shrinkage model.

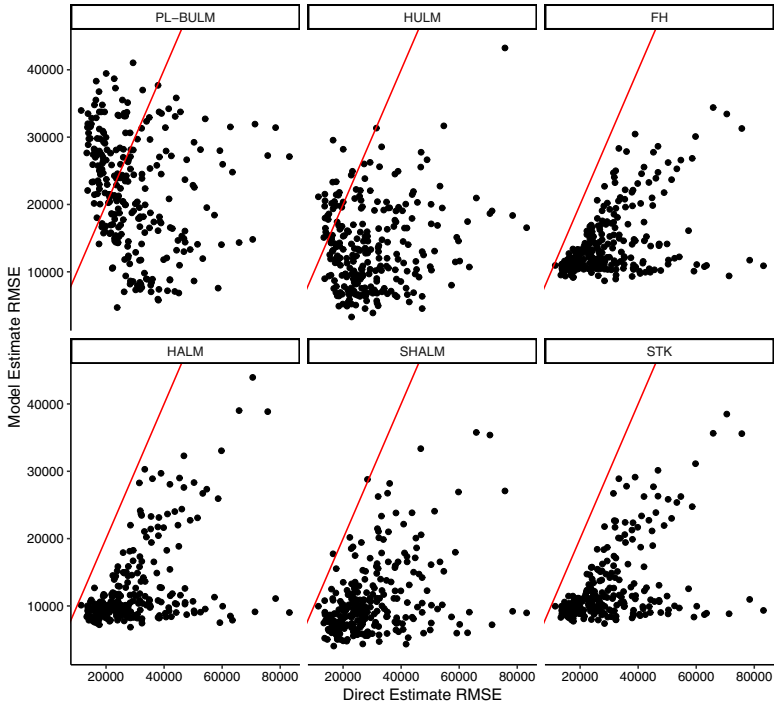

Figure 1. Empirical Simulation Results of Direct Versus Model-Based RMSE by PUMA for Stratified Random Sampling using the 2018 1-Year American Community Survey Public-Use Microdata Sample.

We also compare the RMSE by PUMA between the direct estimates and each model-based estimate. Figure 1 presents these results for the stratified sample design. With the exception of the PL-BULM, most to all PUMAs experience a reduction in RMSE relative to the direct estimator, as indicated by points that fall below the one-to-one line. Among the area-level models, there is a general downward shift in points for the STK, HALM, and STK models compared to the FH model. This indicates a general reduction in RMSE for most areas when compared to the FH model. Similarly, the SHALM appears to have a general downward shift when compared to the STK and HALM methods. Similar results are presented for the probability proportional to size design in figure 2, for which similar patterns hold.

Taken collectively, this simulation illustrates that heteroscedastic modeling techniques may be used to improve the quality of small area estimates. At the area level, these techniques may be used to simultaneously shrink both the design-based means and variances. At the unit level, our framework allows for respondent-specific variances when modeling continuous data. In both cases, our framework has the potential to improve the precision of associated small

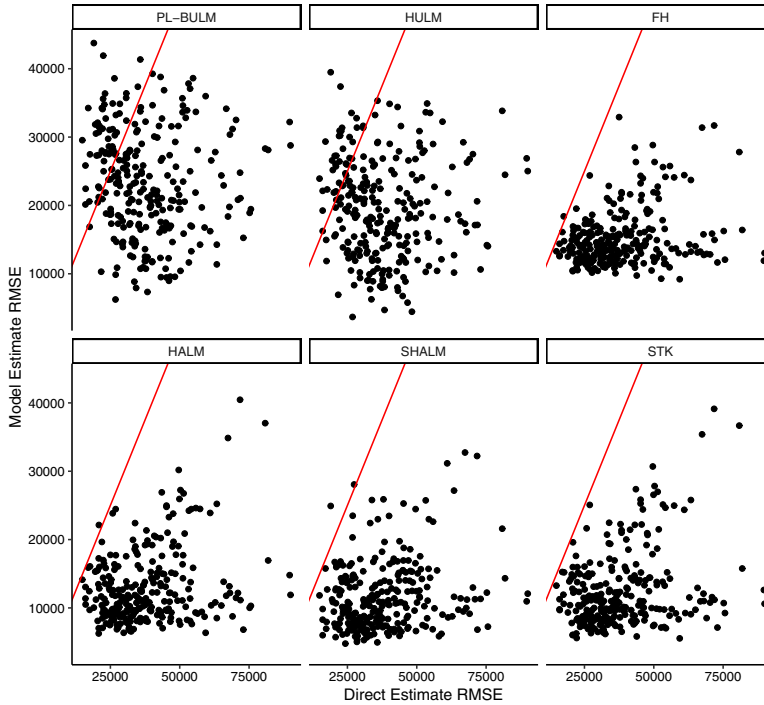

Figure 2. Empirical Simulation Results of Direct Versus Model-Based RMSE by PUMA for Probability Proportional to Size Sampling using the 2018 1-Year American Community Survey Public-Use Microdata Sample.

area estimates relative to approaches that do not have flexible models for the variance.

5. ANALYSIS OF ACS INCOME DATA

One important application of SAE techniques is the estimation of mean or median income for various geographies. For example, the SAIPE produces income estimates for all US states and counties (Bell et al. 2016). In many cases, the estimates produced by SAIPE or similar programs may be used to allocate critical federal aid. Thus, improving the quality of model-based estimates for various outcomes such as income constitutes an important research problem. To this end, we demonstrate an application of our proposed SHALM approach by estimating the average income of positive income earners by PUMA for the state of California using the 2018 1-year ACS PUMS sample. PUMA sample sizes ranged from 265 to 1,297. The SHALM is fit analogously to section 4, with the exception that the entire PUMS dataset was used rather

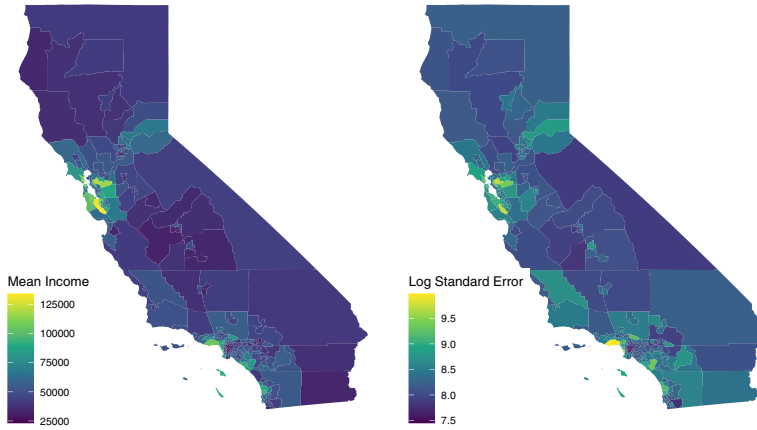

Figure 3. Model-Based Estimates of Mean Income by Public-Use Microdata Area along with Associated Log-Transformed Standard Errors. Estimates are constructed using the 2018 1-year American Community Survey Public-Use Microdata Sample.

than a subsample. For reference, the model took roughly 6.5 min to fit using a standard 2.3-GHz Intel Core i9 processor.

The model-based estimates of mean income by PUMA are shown in figure 3, along with their standard errors. The estimates are generally as expected with higher incomes around city centers with high cost of living, such as the San Francisco Bay area and Los Angeles, and lower incomes in more rural parts of the state. Uncertainty is higher in areas that had lower sample sizes, but also in some areas with very high estimated income. For these counties, there was considerably more spread in the observed incomes, which contributes to the uncertainty around the estimated mean.

We also compare the model-based estimates to the direct estimates in figure 4. Here, we see that the two estimates generally agree, with points falling close to the one-to-one line. However, we also see that the model-based approach tends to result in slightly higher estimates for areas with low average income and slightly lower estimates for areas that exhibited high average income.

Finally, we compare the ratio of the SHALM-based estimates and the direct estimates to the PUMA sample sizes in figure 5. We expect that the model-based and direct estimates would be more similar for larger sample sizes. This is generally the case, as there is more variability around the one-to-one line for smaller sample sizes than for large sample sizes.

6. DISCUSSION

Small area/domain estimation is an area of wide-spread interest both for data users and official statistical agencies. Consequently, there has been significant

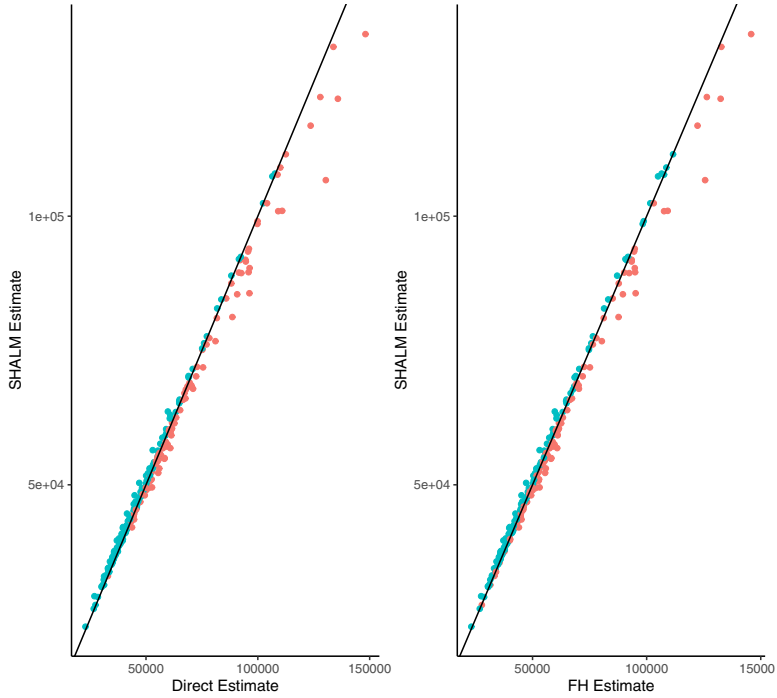

Figure 4. Direct Versus SHALM-Based Estimates of Mean Income by Public-Use Microdata Area as well as FH Model-Based versus SHALM-Based Estimates. Estimates are constructed using the 2018 1-year American Community Survey Public-Use Microdata Sample.

research on both area-level and unit-level models with the goal of improving the precision of target estimates. Nevertheless, model-based approaches typically focus on modeling the mean, with a few notable exceptions. As illustrated here, in the presence of heteroscedasticity, simultaneously modeling both the mean and variance can achieve estimates with both reduced MSE and superior frequentist coverage properties.

By extending Bradley et al. (2020) and Parker et al. (2021), our approach to simultaneously modeling the mean and variance produces fully conjugate updates for hard-to-estimate parameters and is, therefore, extremely computationally efficient. Importantly, our approach is extremely flexible and allows for the incorporation of spatial dependence and covariates in the portion of the model for the variance.

Although our main focus is on area-level modeling, we also introduce unit-level models that account for informative sampling through the use of the pseudo-likelihood. These models can be extremely effective in situations where

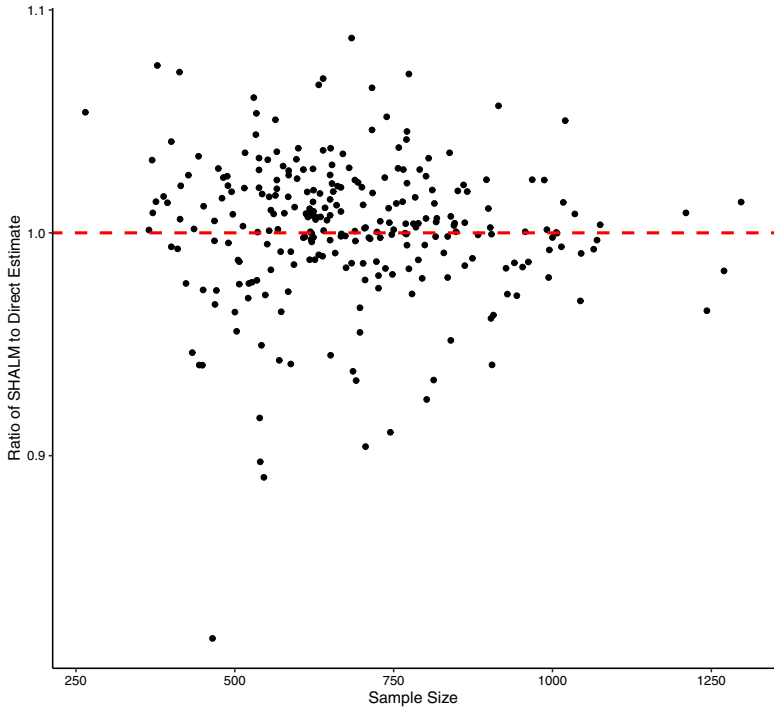

Figure 5. Sample Size Versus Ratio of SHALM-Based Estimates to Direct Estimates. Estimates are constructed using the 2018 1-year American Community Survey Public-Use Microdata Sample.

tabulations are desired for custom-geographies and/or situations when internal aggregation consistency is desired.

Our motivating example focused on modeling income and showcased the gains achieved from using our proposed approach. Nevertheless, for this example, our area-level models outperformed the models that were estimated at the unit level. One reason for this is that the Gaussian assumption at the unit level may not be an optimal starting point for this application due to data issues that arise from disclosure avoidance mechanisms. In this direction, there are several avenues for future work, including extensions to non-Gaussian data or Gaussian mixtures for both unit- and area-level models. In addition, future work also includes the extension to multivariate applications and applications where data integration is advantageous.

Supplementary Materials

Supplementary materials are available online at academic.oup.com/jssam.

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.

Bell, W. R., Basel, W. W., and Maples, J. J. (2016), "An overview of the U. S. Census Bureau's Small Area Income and Poverty Estimates Program," in *Analysis of Poverty Data by Small Area Estimation*, ed. M. Pratesi, New York: Wiley, pp. , 349–378.

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, 1–20.

Binder, D. A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, 51, 279–292.

Bradley, J. R., Holan, S. H., and Wikle, C. K. (2020), "Bayesian Hierarchical Models with Conjugate Full-Conditional Distributions for Dependent Data from the Natural Exponential Family," *Journal of the American Statistical Association*, 115, 2037–2052.

———. (2018), "Computationally Efficient Multivariate Spatio-Temporal Models for High-Dimensional Count-Valued Data (with Discussion)," *Bayesian Analysis*, 13, 253–310.

Bradley, J. R., Wikle, C. K., and Holan, S. H. (2016), "Bayesian Spatial Change of Support for Count-Valued Survey Data with Application to the American Community Survey," *Journal of the American Statistical Association*, 111, 472–487.

Brewer, K., Early, L., and Hanif, M. (1984), "Poisson, Modified Poisson and Collocated Sampling," *Journal of Statistical Planning and Inference*, 10, 15–30.

Chen, Z., and Dunson, D. B. (2003), "Random Effects Selection in Linear Mixed Models," *Biometrics*, 59, 762–769.

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.

Gladish, D. W., Wikle, C., and Holan, S. (2014), "Covariate-Based Cepstral Parameterizations for Time-Varying Spatial Error Covariances," *Environmetrics*, 25, 69–83.

Gneiting, T., and Raftery, A. E. (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378. 477

Maiti, T., Ren, H., and Sinha, S. (2014), "Prediction Error of Small Area Predictors Shrinking Both Means and Variances," *Scandinavian Journal of Statistics*, 41, 775–790. 3

Maples, J., Bell, W., and Huang, E. T. (2009), "Small Area Variance Modeling with Application to County Poverty Estimates from the American Community Survey," in Proceedings of the Section on Survey Research Methods, Alexandria, VA: American Statistical Association, pp. 5056–5067.

Marhuenda, Y., Molina, I., and Morales, D. (2013), "Small Area Estimation with Spatio-Temporal Fay–Herriot Models," *Computational Statistics & Data Analysis*, 58, 308–325.

Parker, P. A., Holan, S. H., and Wills, S. A. (2021), "A General Bayesian Model for Heteroskedastic Data with Fully Conjugate Full-Conditional Distributions," *Journal of Statistical Computation and Simulation*, 91, 3207–3227.

Parker, P. A., Janicki, R., and Holan, S. H. (2019), "Unit Level Modeling of Survey Data for Small Area Estimation under Informative Sampling: A Comprehensive Overview with Extensions," arXiv:1908.10488.

Polettini, S. (2017), "A Generalised Semiparametric Bayesian Fay–Herriot Model for Small Area Estimation Shrinking Both Means and Variances," *Bayesian Analysis*, 12, 729–752.

Porter, A. T., Wikle, C. K., and Holan, S. H. (2015), "Small Area Estimation via Multivariate Fay–Herriot Models with Latent Spatial Dependence," *Australian & New Zealand Journal of Statistics*, 57, 15–29.

Pourahmadi, M. (1999), "Joint Mean-Covariance Models with Applications to Longitudinal Data: Unconstrained Parameterisation," *Biometrika*, 86, 677–690.

———. (2011), "Covariance Estimation: The GLM and Regularization Perspectives," *Statistical Science*, 26, 3, 369–387.

———. (2013), *High-Dimensional Covariance Estimation: With High-Dimensional Data* (Vol. 882), Hoboken, NJ: John Wiley & Sons.

Savitsky, T. D., and Gershunskaya, J. (2022), "Bayesian Nonparametric Joint Model for Domain Point Estimates and Variances under Biased Observed Variances," *Journal of Survey Statistics and Methodology*. DOI: 10.1093/jssam/smac003.

Savitsky, T. D., and Toth, D. (2016), "Bayesian Estimation under Informative Sampling," *Electronic Journal of Statistics*, 10, 1677–1708.

Schmidt, A. M., Guttorp, P., and O'Hagan, A. (2011), "Considering Covariates in the Covariance Structure of Spatial Processes," *Environmetrics*, 22, 487–500.

Skinner, C. J. (1989), "Domain Means, Regression and Multivariate Analysis," in *Analysis of Complex Surveys*, eds. C. J. Skinner, D. Holt, and T. M. F. Smith, Chichester: Wiley, pp. 80–84, Chapter 2.

Sugasawa, S., Tamae, H., and Kubokawa, T. (2017), "Bayesian Estimators for Small Area Models Shrinking Both Means and Variances," *Scandinavian Journal of Statistics*, 44, 150–167.

Sun, A., Parker, P. A., and Holan, S. H. (2022), "Analysis of Household Pulse Survey Public-Use Microdata via Unit-Level Models for Informative Sampling," *Stats*, 5, 139–153.

You, Y. (2021), "Small area Estimation Using Fay-Herriot Area Level Model with Sampling Variance Smoothing and Modeling," *Survey Methodology*, 47, 2, 361–371.

You, Y., and Chapman, B. (2006), "Small Area Estimation Using Area Level Models and Estimated Sampling Variances," *Survey Methodology*, 32, 97.