Journal of the Royal Statistical Society Series A: Statistics in Society, 2025, **00**, 1–17
https://doi.org/10.1093/jrsssa/qnaf063

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS
Journal of the Statistics Society
Series A
Statistics in Society

Original Article

Spatially selected and dependent random effects for small area estimation with application to rent burden

Sho Kawano, Paul A. Parker and Zehang Richard Li

Department of Statistics, University of California - Santa Cruz, Santa Cruz, CA 95064, USA

Address for correspondence: Sho Kawano, Department of Statistics, University of California - Santa Cruz, 1156 High Street, Santa Cruz, CA 95064, USA. Email: shkawano@ucsc.edu

Abstract

Area-level models for small area estimation typically rely on areal random effects to shrink design-based direct estimates towards a model-based predictor. Incorporating the spatial dependence of the random effects into these models can further improve the estimates when there are not enough covariates to fully account for the spatial dependence of the areal means. A number of recent works have investigated models that include random effects for only a subset of areas, in order to improve the precision of estimates. However, such models do not readily handle spatial dependence. In this paper, we introduce a model that accounts for spatial dependence in both the random effects as well as the latent process that selects the effects. We show how this model can significantly improve predictive accuracy via an empirical simulation study based on data from the American Community Survey, and illustrate its properties via an application to estimate county-level median rent burden.

Keywords: American Community Survey, Bayesian hierarchical model, rent burden, shrinkage prior, spike-and-slab

1 Introduction

Small area estimation (SAE) has become an integral tool in official statistics to produce estimates for quantities of interest, such as means and totals, for subpopulations with small sample size. The term *small area* usually refers to geographic subregions but can also be demographic groups or a combination of the two. SAE models play a vital role in the effective implementation of government policies. One of the many policy-relevant estimates of interest is rent burden, defined as the share of household income spent on rent.

The cost of housing is a very pressing issue in the United States. According to 2017–2021 ACS 5-year estimates, over 19 million US renter households were paying more than 30% of income on rent, making them *cost burdened* (Cromwell, 2022). More recently in 2022, 22.5 million households, half of all US renters, were estimated to be cost burdened, based on the ACS 2022 1-year estimates (Harvard Joint Center for Housing Studies, 2024; U.S. Census Bureau, 2023). Furthermore, 12.1 million households are estimated to be *severely cost burdened*, paying more than half of their incomes on rent. These severe rent burdens heavily affect households with lower incomes that do not have sufficient funds to cover basic necessities after paying rent (Desmond, 2018; U.S. Bureau of Labor Statistics, 2024).

US housing policy is highly decentralized. Although the central (or 'federal') government administers subsidies to housing consumers, most subsidies are directed at homeowners rather than renters, despite renters typically falling into lower-income categories (Crump & Schuetz, 2021).

Received: April 18, 2024. **Revised:** May 12, 2025. **Accepted:** May 13, 2025

In other key areas of housing policy, especially in land use planning and development permitting, the federal government has quite limited powers (Quigley, 2008). Most of the consequential policy debates affecting renters today are therefore happening at the level of local and state governments. Thus, providing accurate and granular estimates at the local and state level is critical to good policy making.

Direct survey estimators often fail to produce reliable estimates at the level of a small area (e.g. counties or census tracts). This is due to the fact that transnational surveys like the American Community Survey (ACS) are usually designed for accuracy at a high level of aggregation. Thus, model-based approaches may become necessary for reliable SAE without the development of new surveys. In this paper, we primarily focus on what is known as area-level modelling, where aggregated direct estimates at discrete areas of interest are modelled. This approach contrasts with unit-level SAE models, where individual survey responses are directly modelled. For a review of unit-level models, see Parker, Janicki, et al. (2023).

The foundational area-level model was introduced by Fay and Herriot (1979). Their model utilizes a regression-based estimate with independent, normally distributed random effects that have and a common variance. The estimates from such a model can be viewed as a weighted combination of the direct survey estimates and the regression-based estimates, where the weights are constructed according to the ratio of the survey variance and the variance of the random effects. If the survey variance is relatively large, then the Fay–Herriot (FH) model puts more weight on the regression-based estimate and vice versa. This allows for more accurate estimation in areas with small sample size by *borrowing strength* from areas with larger samples. The FH model has become one of the most widely used tools for SAE. Extensions that generalize the FH model to new types of data and/or relax modelling assumptions have grown substantially over the years.

One convenient assumption the FH model makes is that the survey variances are fixed and known, when in reality they are estimated via the sample design and then plugged into the model. You (2016) and Sugasawa et al. (2017) introduce Bayesian models that use a weighted estimate similar to the FH approach to model the survey variance in addition to the small area mean, with the latter allowing the use of covariates. These models were compared in You (2021). Parker, Holan, et al. (2023) further extend the approach by proposing a model that yields conjugate full conditional distributions to further allow for spatial modelling of the survey variances.

There is also a growing body of literature dedicated to proposing alternate structures for the random effects used within the FH model. For instance, instead of assuming independence, one might consider a spatial structure to the random effects. This allows for models that account for spatial dependence of the quantities of interest, which is commonly observed in many SAE applications. A common distributional choice is the conditional autoregressive (CAR) structure. Examples of previous work using the CAR structure include Zhou and You (2008) and Porter et al. (2015), the latter providing an extension for multivariate data. Another common choice is the simultaneous autoregressive (SAR) structure. For examples, see Singh et al. (2005), Petrucci and Salvati (2006), and Schmid and Münnich (2014). Chung and Datta (2022) provides a Bayesian overview of CAR, SAR, and other spatial models. The authors also demonstrate how the Bayesian approach extends naturally to settings where direct estimates are not available for some areas. The quickly growing spatial SAE literature suggests that incorporating spatial dependence can substantially improve the precision of model-based estimates.

Another area of exploration has been to address the normality and common variance assumptions of the random effects. These assumptions can cause the FH model to be less robust in situations where large random effect values are needed for relatively few areas. Many recent papers have proposed extensions that move away from one or both of these simplifying assumptions. An overview of this topic is given by Jiang and Rao (2020). One approach is to assume heavier-tailed distributions for the random effects, which was explored by Datta and Lahiri (1995) and Fabrizi and Trivisano (2010), among others. Chakraborty et al. (2016) use a mixture of two normal distributions, one with a small variance and the other with a large variance, to relax the common variance assumption. Finally, Janicki et al. (2022) take a novel approach by using a Bayesian nonparametric (BNP) model to estimate counts for multi-way contingency tables (e.g. race and age groups by county). Using a BNP prior to model both the fixed and random effects is an effective way to model contingency tables that exhibit heterogeneity both in the coefficient-response relationships as well as spatial patterns.

A more recent (and somewhat related) avenue of research was spurred by a number of works that have questioned whether all of the random effects are necessary in the presence of appropriate covariates. The first round of inquiry was done via a hypothesis test by Datta et al. (2011). But the test could only test whether all or none of the random effects are necessary and the presence of large effects for even one area could result in a small p-value. This has led to development of models that use *shrinkage* priors, pushing the unnecessary random effects towards zero. More specifically, Datta and Mandal (2015) proposed an SAE model with a discrete-normal spike-and-slab prior that shrinks the unnecessary random effects to zero, selecting which random effects are necessary. A continuous analogue to the Datta–Mandal (DM) model were introduced by Tang et al. (2018) with a multivariate extension by Ghosh et al. (2022), using the class of global–local shrinkage priors for the random effects. In the global–local approach, the variance of each random effect is a product of a global shrinkage parameter and a local parameter that serves as an area-specific adjustment.

Shrinkage guided by the spatial structure of the areas of interest may further improve the accuracy of model-based estimates. Intuitively, if a random effect is not necessary in a given region, it may be likely that it is also not necessary in a neighbouring area. On the other hand, if a random effects are necessary for two neighbouring areas, the value of the random effects may be similar as well. Therefore, spatial dependence can be built in on two different levels: in the latent selection process that selects/shrinks the random effects as well as in the magnitude of the random effects that are selected.

Building in spatial dependence for the random effects themselves is somewhat straightforward within the global–local framework, as demonstrated by Tang and Ghosh (2023), who used a CAR prior for the random effects. However, building in spatial dependence for the shrinkage process is complicated by the distributional requirements of the global–local prior. For instance, using a spatial horseshoe prior on the random effects would require the use of a Gaussian copula on the local parameters (Reich & Staicu, 2021). Given these complications, a modelling approach that uses a spike-and-slab prior may provide a more natural starting point.

In this paper, we introduce a model that incorporates spatial dependence on both the selection process and the random effects themselves, extending the work of Datta and Mandal (2015). The remainder of this paper is organized as follows. In Section 2, we first discuss area-level modelling, introducing the FH model in addition to the relevant spatial and shrinkage extensions. In Section 3, we introduce our proposed model and discuss prior specification. In Section 4, we describe posterior inference for our proposed model, including the full conditional distributions of each parameter. In Section 5, we conduct an empirical simulation study that shows how the proposed model can significantly improve the accuracy of both point and interval estimates of median rent burden, using ACS county-level data for North Carolina. In Section 6, we use the proposed model to estimate the median rent burden for all counties in the South Atlantic Census Division. Finally, we conclude with a discussion in Section 7.

2 Area-level modelling

Consider a survey in a study region partitioned into n small areas. Let the small area means $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ be the quantities of interest. For each area $i = 1, \ldots, n$, a survey is used to provide design-based direct estimates of θ_i, denoted as y_i, with survey variances d_i, which are assumed to be known. Let N_i represent the population size in area i and y_{ij} denote the value of some variable of interest for the jth population unit in area i, where $j \in \{1, \ldots, N_i\}$. Under the design-based perspective, the small area means are treated as fixed but unknown quantities (i.e. $\theta_i = \frac{1}{N_i}\sum_{j=1}^{N_i} y_{ij}$). The commonly used Horvitz and Thompson (1952) estimator incorporates design information through inverse-probability weighting,

$$y_i = \frac{1}{N_i}\sum_{j\in\mathcal{S}_i}\frac{y_{ij}}{\pi_{ij}},$$

where $\mathcal{S}_i$ denotes the collection of sample indices for area i and π_{ij} represents the probability that the jth unit in area i is included in the sample. Direct estimators like the Horvitz–Thompson have important properties such as being design unbiased.

However, for areas where the sample size is small, direct estimators can have unreasonably high standard errors (SE), which may necessitate the use of a model. A standard area-level model can be written as

$$[\boldsymbol{y} \mid \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}) \quad \text{and} \quad \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$, the $n \times n$ covariance matrix is $\boldsymbol{D} = \text{diag}\{d_1, \ldots, d_n\}$, $\boldsymbol{X}$ is a $n \times j$ full-rank covariate matrix, and $\boldsymbol{\beta}$ is the corresponding coefficient vector (a.k.a. fixed effects). Finally, the length n vector $\boldsymbol{u}$ represents the random effects. The random effects $\boldsymbol{u}$ capture the variability in θ_i that cannot be fully explained by the covariates in $\boldsymbol{X}$. A great deal of research in SAE is focused on the distributional choice for $\boldsymbol{u}$.

The random effects for the foundational FH model (Fay & Herriot, 1979) are assumed to be independent and identically distributed (IID) normal random variables with common variance:

$$[\boldsymbol{u} \mid \sigma^2] \sim N_n(0, \sigma^2 \boldsymbol{I}).$$

In a Bayesian setting, the FH model is often used with the improper prior $\pi(\beta, \sigma^2) \propto 1$. Despite its widespread acceptance, there are some shortcomings with this model, as previously mentioned. These limitations are generally driven by the assumption that the random effects are IID and normally distributed. Assuming that all random effects are normally distributed with a common variance σ^2 causes issues with robustness. The independence is also a questionable assumption because spatial dependence of the random effects are often observed in practice.

2.1 Incorporating spatial dependence in the random effects

Instead of assuming independence among random effects, one could model some dependence structure in the random effects

$$[\boldsymbol{u} \mid \sigma^2, \rho] \sim N_n(0, \sigma^2 \mathcal{Q}^{-1}(\rho)),$$

where σ^2 represents the random effect variance and ρ denotes additional parameters needed for $\mathcal{Q}$, the precision matrix. A common choice for $\mathcal{Q}$ is the CAR structure (Besag, 1974). The CAR precision matrix has the form $\mathcal{Q} = \text{diag}\{A_{ii}\}_{i=1}^n - \rho \boldsymbol{A}$ where $\rho \in (-1, 1)$ is the spatial correlation parameter and $\boldsymbol{A}$ is a $n \times n$ spatial adjacency matrix whose i, kth element a_{ik} is equal to one if areas i and k are seen as spatially adjacent (e.g. share a border), and equal to zero otherwise. Positive (negative) values of ρ indicate positive (negative) spatial correlation with stronger spatial correlation indicated by values close to ± 1, although it is common practice to restrict $\rho \geq 0$. Setting $\rho = 1$ results in the degenerate Intrinsic CAR (ICAR) prior.

Existing literature indicates that the CAR model may lead to misleading results when there is no spatial correlation actually present in the data (Leroux et al., 2000; Wakefield, 2007). A flexible alternative to the CAR prior is a Besag–York–Mollié (BYM) prior (Besag et al., 1991) that includes a separate spatial and nonspatial random effect components,

$$\boldsymbol{u} = \boldsymbol{v}_1 + \boldsymbol{v}_2 \text{ where } [\boldsymbol{v}_1] \sim N_n(0, \sigma_1^2 \boldsymbol{I}) \quad \text{and} \quad [\boldsymbol{v}_2] \sim N_n(0, \sigma_2^2 \mathcal{Q}^-).$$

Note that sum-to-zero constraints are placed on $\boldsymbol{v}_1, \boldsymbol{v}_2$ for identifiability and that $\mathcal{Q}^-$ is the generalized inverse of the ICAR precision matrix. Under this prior, the random effects are a sum of two components: IID normal effects $\boldsymbol{v}_1$ and the ICAR spatial effects $\boldsymbol{v}_2$. When there is zero spatial correlation present, the posterior values of the random effects would be dominated by $\boldsymbol{v}_1$, making the estimates similar to an IID FH model. Conversely, if the heterogeneity is very spatially dependent on the posterior values of $\boldsymbol{v}_2$ would be larger than $\boldsymbol{v}_1$. There are also many instances where having both the IID effects in addition to the ICAR spatial effects helps best capture the heterogeneity. The overall balance of the spatial vs. nonspatial components can be determined by comparing the posterior distribution of the random effect variances σ_1^2, σ_2^2.

In order to aid in the prior specification of σ_2^2, we found it useful to scale the ICAR precision matrix by dividing it by the geometric mean of the diagonal elements, as recommended by Sørbye and Rue (2014). We denote the *scaled* ICAR precision matrix by $\boldsymbol{Q}$. Scaling the ICAR precision matrix allows the parameter σ_2^2 to reflect the typical variance, allowing the prior of the IID variance σ_1^2 to be roughly comparable to that of σ_2^2 (Freni-Sterrantino et al., 2018; Riebler et al., 2016).

2.2 Incorporating random effect shrinkage

Datta and Mandal (2015) introduced the first SAE model that uses Bayesian variable selection methodology for the random effects. Specifically, they use a spike-and-slab prior that assumes a discrete-normal mixture for the random effects

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{\delta} \odot \boldsymbol{v}, \\ [v_i \mid \delta_i = 1, \sigma^2] &\overset{\text{ind}}{\sim} N(0, \sigma^2), \\ [v_i \mid \delta_i = 0] &= 0, \\ [\delta_i \mid p] &\overset{\text{iid}}{\sim} \text{Bernoulli}(p), \\ [p] &\sim \text{Beta}(a, b). \end{aligned}$$

To complete the model, they use an inverse-gamma prior on σ^2 and $\pi(\beta) \propto 1$. Note that we use $\odot$ to represent the operator for an element-wise vector product. In the DM model, a given random effect, v_i, for area i is only included if the corresponding selection indicator, δ_i, is equal to 1. This approach relaxes the common variance assumption because σ^2 only applies to random effects for areas that are selected (i.e. $\delta_i = 1$). Meanwhile, the effects that are not selected are degenerate with zero variance. Note that in the extreme case if random effects for all areas are selected (i.e. $\delta_i = 1$ for all $i = 1, \ldots, n$), then the DM is equivalent to the independent FH model.

The mechanism behind the shrinkage of the random effects for the DM model can be understood from the posterior selection probability, $\check{p}_i$, for a given area i,

$$\check{p}_i = \frac{p \cdot \phi\left(y_i \mid \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i,\, d_i\right)}{p \cdot \phi\left(y_i \mid \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i,\, d_i\right) + (1-p) \cdot \phi\left(y_i \mid \boldsymbol{x}_i^\top \boldsymbol{\beta},\, d_i\right)}, \tag{1}$$

where $\phi(z \mid \mu, \kappa^2)$ is the normal density function with mean μ and variance κ^2, evaluated at the point z. The value of $\check{p}_i$ can be interpreted as the posterior probability that a random effect is necessary in the area i. The global level of shrinkage across all areas is influenced by the parameter p. If the posterior distribution of p is close to one, it indicates that most areas need random effects and vice versa. Area-specific selection probabilities also depend on the normal likelihood evaluations both with and without the random effect. For example, even if p has mass towards zero, if the fit without the random effect is very poor for a given area i, then $\phi(y_i \mid \boldsymbol{x}_i^\top \boldsymbol{\beta} + v_i, d_i)$ will be much greater than $\phi(y_i \mid \boldsymbol{x}_i^\top \boldsymbol{\beta}, d_i)$, which will be reflected in the selection probability.

3 The spatially selected and dependent random effects model

We propose a model that accounts for spatial dependence on both the selection process and the random effects themselves. The data model is given by

$$[\boldsymbol{y} \mid \boldsymbol{\theta}] \sim N_n(\boldsymbol{\theta}, \boldsymbol{D}) \quad \text{and} \quad \boldsymbol{\theta} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u},$$

where n again is the number of small areas. Again, $\boldsymbol{y}$ is the vector of direct estimates, $\boldsymbol{D}$ is the $n \times n$ covariance matrix, $\boldsymbol{X}$ is a $n \times j$ full-rank covariate matrix, and $\boldsymbol{\beta}$ is the corresponding coefficient

vector. Similar to Datta and Mandal (2015), we consider a spike-and-slab prior for the random effects, although we include spatial structure,

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{\delta} \odot (\boldsymbol{v}_1 + \boldsymbol{v}_2), \\ [\boldsymbol{v}_1] &\sim N_n(0, \sigma_1^2 \boldsymbol{I}), \\ [\boldsymbol{v}_2] &\sim N_n(0, \sigma_2^2 \boldsymbol{Q}^-), \\ [\delta_i \mid p_i] &\overset{\text{ind}}{\sim} \text{Bernoulli}(p_i). \end{aligned}$$

Note that for identifiability purposes, we place a sum-to-zero constraint on $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, while $\boldsymbol{Q}^-$ is the generalized inverse of the ICAR precision matrix, which is scaled by the geometric mean of the diagonal elements, as suggested by Sørbye and Rue (2014). Again, we use $\odot$ to represent the operator for an element-wise vector product. Also note that the selection indicators themselves are assumed independent, but each has a separate selection probability parameter, p_i. We further model dependence within the selection process through the prior on these selection probabilities:

$$\begin{aligned} \text{logit}(\boldsymbol{p}) &= \boldsymbol{\psi}_1 + \boldsymbol{\psi}_2, \\ [\boldsymbol{\psi}_1 \mid s_1^2] &\sim N_n(0, s_1^2 \boldsymbol{I}), \\ [\boldsymbol{\psi}_2 \mid s_2^2] &\sim N_n(0, s_2^2 \boldsymbol{Q}^-), \end{aligned}$$

where, again, $\boldsymbol{Q}^-$ is the generalized inverse of the scaled ICAR precision matrix. We used the logit link to map probability to the entire real line in order to model spatial dependence with a multivariate normal distribution. This allows for the use of the BYM structure on the logit effects that form the selection probabilities, mirroring what was done for the random effects. Moreover, using the logit link allows the use of Pólya-Gamma (PG) data augmentation (Polson et al., 2013) which has computational advantages.

This model can capture spatial and nonspatial heterogeneity in both the random effects and the selection process. Thus, we refer to this model as the spatially selected and dependent (SSD) random effects model. Using the BYM structure twice allows for greater flexibility that avoids having to predetermine a spatial or nonspatial structure to the random effects and selection. The overall balance of the spatial vs. nonspatial components can be determined by comparing the posterior distribution of σ_1^2, σ_2^2 for the random effects and s_1^2, s_2^2 for the selection process.

3.1 Prior specification

To simplify prior specification for the SSD model, we recommend fitting on scaled data ($\boldsymbol{y}$, $\boldsymbol{D}$), similar to what is done in other shrinkage methods such as the Bayesian Lasso (Park & Casella, 2008). Specifically, we recommend scaling such that the direct estimates, $\boldsymbol{y}$, have zero mean and unit variance. That is, if $\bar{y}$ is the sample mean and s_y is the sample standard deviation, then each direct estimate is scaled by $(y_i - \bar{y})/s_y$ for $i = 1, \ldots, n$. Also, the survey variances are correspondingly scaled d_i/s_y^2 to form the diagonal matrix $\boldsymbol{D}$.

In the SSD model, priors for three sets of parameters need to be specified: the regression coefficients $\boldsymbol{\beta}$, the random effect variances σ_1^2, σ_2^2, and the logit variances s_1^2, s_2^2. The choice of prior on the random effect variances σ_1^2, σ_2^2 require some care, as they can influence the selection probabilities of the random effects, which may affect the estimates. In order for the model to be able to distinguish between the degenerate *spike* at zero and the *slab* normal distribution, a prior needs to avoid variance values that are too small, while still being sufficiently diffuse. The inverse-gamma distributions have a gap near the origin, which can be adjusted to provide the necessary qualities. Thus, we place inverse-gamma priors on the random effect variances with $[\sigma_1^2] \sim \text{IG}(c, d)$, and $[\sigma_2^2] \sim \text{IG}(c, d)$. Note that $X \sim \text{IG}(c, d)$ refers to an inverse-gamma distribution with the density: $f(x \mid c, d) \propto (1/x)^{c+1} exp(-d/x)$, where $c, d > 0$ are the shape and scale parameters. We conducted a detailed sensitivity analysis on multiple datasets and found that priors that avoids values near zero and are sufficiently diffuse, such as IG(5, 5), work well for scaled data. The result of the sensitivity analysis is included in the online supplementary materials.

For the regression coefficients, we specify the proper prior $[\boldsymbol{\beta}] \sim N_j(\mathbf{0}, k^2\boldsymbol{I})$ where the hyperparameter k^2 is sufficiently large as to be noninformative. This could depend on the scale of the data, but for the suggested standardized data, we recommend $k^2 = 100^2$. The prior choice of s_1^2, s_2^2, the logit variances, can affect convergence of the MCMC chains but they have minimal impact on the estimates themselves, especially given that these parameters are far down the model hierarchy. We let $s_1^2 \sim \text{IG}(r, q)$ and $s_2^2 \sim \text{IG}(r, q)$ for fixed hyperparameters r, q, where we recommend $r = 5$ and $q = 10$.

3.2 Relationships to related models

Note that each of the models discussed in Section 2 is a limiting case of the SSD model introduced here. The BYM is an SSD model where all of the random effects are selected (i.e. $\delta_i = 1$ for all $i = 1, \ldots, n$). The DM is a limiting case of the SSD model where the spatial component, both in the random effects and the selection process is zero (i.e. if σ_2^2 and s_2^2 are degenerate at zero). The independent FH model is the limiting case of the SSD model if both are true: all of the random effects are IID only (i.e. σ_2^2 and s_2^2 are degenerate at zero) and the random effects are all selected (i.e. $\delta_i = 1$ for all $i = 1, \ldots, n$).

Also, note that the posterior selection probability for the SSD model is

$$\tilde{p}_i = \frac{p_i \cdot \phi\left(y_i \mid \boldsymbol{x}_i^\top\boldsymbol{\beta} + v_{1i} + v_{2i},\, d_i\right)}{p_i \cdot \phi\left(y_i \mid \boldsymbol{x}_i^\top\boldsymbol{\beta} + v_{1i} + v_{2i},\, d_i\right) + (1 - p_i) \cdot \phi\left(y_i \mid \boldsymbol{x}_i^\top\boldsymbol{\beta},\, d_i\right)}, \tag{2}$$

where y_i are the scaled direct estimates, d_i are the scaled survey variances, and ϕ is the normal density function as previously described. Note that the structure of the posterior selection probability here is essentially the same as that of the DM model (1), except that the random effect has the form $v_{1i} + v_{2i}$ and there is distinct p_i for each area, instead of shared, global p parameter. Just like the DM model, area-specific selection probabilities also depend on the normal likelihood evaluations both with and without the random effect. Thus, areas with larger magnitude random effects will lead to more definitive conclusions about whether an effect is necessary during posterior inference. But unlike the DM, there will be some spatial smoothing resulting from the BYM structure on the effects and the selection probabilities [i.e. $\text{logit}(p_i) = \psi_{1i} + \psi_{2i}$]. Please see Figure 4 in the online supplementary materials for a visual explanation.

4 Posterior inference

Let $\boldsymbol{\Omega}$ denote the set of parameters in the SSD model. Then, the full posterior distribution can be written, up to a constant of proportionality, as

$$\begin{aligned}
\pi(\boldsymbol{\Omega} \mid \boldsymbol{y}, \boldsymbol{X}, \boldsymbol{D}) \propto\ & \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\delta} \odot [\boldsymbol{v}_1 + \boldsymbol{v}_2])^\top \boldsymbol{D}^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\delta} \odot [\boldsymbol{v}_1 + \boldsymbol{v}_2])\right\} \\
& \times \left(\frac{1}{\sigma_1^2}\right)^{n/2}\left(\frac{1}{\sigma_2^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma_1^2}\boldsymbol{v}_1{}^\top\boldsymbol{v}_1 - \frac{1}{2\sigma_2^2}\boldsymbol{v}_2{}^\top\boldsymbol{Q}\boldsymbol{v}_2\right\} \\
& \times \prod_{i=1}^{n} p_i^{\delta_i}(1 - p_i)^{1-\delta_i} \\
& \times \left(\frac{1}{s_1^2}\right)^{n/2}\left(\frac{1}{s_2^2}\right)^{n/2} \exp\left\{-\frac{1}{2s_1^2}\boldsymbol{\psi}_1{}^\top\boldsymbol{\psi}_1 - \frac{1}{2s_2^2}\boldsymbol{\psi}_2{}^\top\boldsymbol{Q}\boldsymbol{\psi}_2\right\} \\
& \times \pi(\beta)\pi(\sigma_1^2)\pi(\sigma_2^2)\pi(s_1^2)\pi(s_2^2).
\end{aligned}$$

This model and the chosen priors leads to full conditional distributions all from standard parametric families, when PG data augmentation is applied (Polson et al., 2013). Then, Gibbs sampling can be used to sample from the posterior distribution. We will denote $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{D}$ as *data* for ease of notation. Please see Section 4 in the online supplementary materials for the derivation of the full conditional distributions.

Full conditional distributions of $\boldsymbol{v_1}$, $\boldsymbol{v_2}$, and $\boldsymbol{\beta}$. The fixed and random effects $\boldsymbol{\beta}$ and $\boldsymbol{v_1}$, $\boldsymbol{v_2}$ can be sampled in a block. In order to do so, we set $\boldsymbol{\gamma} = (\boldsymbol{\beta}^\top, \boldsymbol{v}_1^\top, \boldsymbol{v}_2^\top)^\top$ and $\boldsymbol{Z} = (\boldsymbol{X}\ \Delta\ \Delta)$ where $\Delta = \text{diag}\{\delta_i\}_{i=1}^n$ and $\boldsymbol{Z\gamma} = \boldsymbol{X\beta} + \boldsymbol{\delta} \odot (\boldsymbol{v_1} + \boldsymbol{v_2})$. Provided that there exists i such that $\delta_i \neq 0$, the full conditional of $\boldsymbol{\gamma}$ is then

$$[\boldsymbol{\gamma} \mid \sigma_1^2, \sigma_2^2, \boldsymbol{\delta}, \text{data}] \sim N_{j+2n}\left(\boldsymbol{P}_\gamma^{-1}\boldsymbol{D}^{-1}\boldsymbol{Z}^\top\boldsymbol{y}, \boldsymbol{P}_\gamma^{-1}\right),$$

where

$$\boldsymbol{P}_\gamma = \boldsymbol{Z}^\top\boldsymbol{D}^{-1}\boldsymbol{Z} + \boldsymbol{\Lambda}_\gamma \quad \text{and} \quad \boldsymbol{\Lambda}_\gamma = \begin{pmatrix} \boldsymbol{I}_j/k^2 & 0 & 0 \\ 0 & \boldsymbol{I}_n/\sigma_1^2 & 0 \\ 0 & 0 & \boldsymbol{Q}/\sigma_2^2 \end{pmatrix}.$$

Note that we have n areas, j covariates, and $\boldsymbol{I}_m$ denotes an identity matrix of rank m and $\boldsymbol{Q}$ again is the scaled ICAR precision matrix. If $\delta_i = 0$ for all $i = 1, \ldots, n$, then $\boldsymbol{v}_1 = \boldsymbol{v}_2 = 0$ and the posterior distribution of $\boldsymbol{\beta}$ is:

$$[\boldsymbol{\beta} \mid \sigma_1^2, \sigma_2^2, \textit{data}] \sim N_j\left(\boldsymbol{P}_\beta^{-1}\boldsymbol{D}^{-1}\boldsymbol{X}^\top\boldsymbol{y}, \boldsymbol{P}_\beta^{-1}\right) \quad \text{where } \boldsymbol{P}_\beta = \boldsymbol{X}^\top\boldsymbol{D}^{-1}\boldsymbol{X} + \boldsymbol{I}_j/k^2.$$

Full conditional distribution of $\boldsymbol{\delta}$. For the SSD model, the posterior selection probability $\tilde{p}_i$ for area i, is given by equation 2 in Section 3.2. The full conditional distribution of $\boldsymbol{\delta}$ is then

$$[\delta_i \mid p_i, v_{1i}, v_{2i}, \boldsymbol{\beta}, \text{data}] \overset{\text{ind}}{\sim} \text{Bern}(\tilde{p}_i) \quad \text{for } i = 1, \ldots, n.$$

Full conditional distributions of $\boldsymbol{\psi_1}$, $\boldsymbol{\psi_2}$, and latent variables $\boldsymbol{w}$. The posterior conditional distribution of $\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$ can be sampled with PG data augmentation (Polson et al., 2013). A random variable with a PG distribution with parameters $b > 0$ and $c \in \mathbb{R}$ is denoted as $X \sim \text{PG}(b, c)$. For more on the PG random variable, including its formal definition, we refer you to Polson et al. (2013). We have provided a detailed explanation of PG data augmentation in the online supplementary materials (Section 4).

We use the PG data augmentation strategy to sample the full conditional of $\boldsymbol{\psi}_1$, $\boldsymbol{\psi}_2$ in a block. We set the blocked logit effects $\Psi = (\boldsymbol{\psi}_1^\top,\ \boldsymbol{\psi}_2^\top)^\top$ and $\boldsymbol{H} = (\boldsymbol{I}_n\ \boldsymbol{I}_n)$. We introduce PG latent variables $[w_i] \sim \text{PG}(1, 0)$ and set $\boldsymbol{W} = \text{diag}\{w_i\}_{i=1}^n$. We also set $\boldsymbol{\kappa} = (\delta_1 - 1/2, \ldots, \delta_n - 1/2)^\top$.

Conditional on the latent variables w_i and the variance parameters (s_1^2, s_2^2), the blocked logit effects, can be sampled from

$$[\Psi \mid s_1^2, s_2^2, \boldsymbol{W}] \sim N_{2n}(\boldsymbol{P}_\Psi^{-1}\boldsymbol{H}^\top\boldsymbol{\kappa}, \boldsymbol{P}_\Psi^{-1}),$$

where

$$\boldsymbol{P}_\Psi = \boldsymbol{H}^\top\boldsymbol{W}\boldsymbol{H} + \boldsymbol{\Lambda}_\Psi \quad \text{and} \quad \boldsymbol{\Lambda}_\Psi = \begin{pmatrix} \boldsymbol{I}_n/s_1^2 & 0 \\ 0 & \boldsymbol{Q}/s_2^2 \end{pmatrix}.$$

The latent variables used for the PG data augmentation are also conjugate. Conditional on the logit effects $\boldsymbol{\Psi}$, w_i can be sampled from

$$[w_i \mid \Psi] \overset{\text{ind}}{\sim} \text{PG}(1, h_i^\top\Psi)$$

for $i = 1, \ldots, n$, where h_i is the ith row of $\boldsymbol{H} = (\boldsymbol{I}_n\ \boldsymbol{I}_n)$ and $h_i^\top\Psi = \psi_{1i} + \psi_{2i}$.

Full conditional distributions of σ_1^2, σ_2^2 and s_1^2, s_2^2. The full conditional for the two sets of variance parameters are

$$[\sigma_1^2 \mid \boldsymbol{v}_1] \sim \text{IG}(n/2 + c, \boldsymbol{v}_1^\top \boldsymbol{v}_1/2 + d) \quad \text{and} \quad [\sigma_2^2 \mid \boldsymbol{v}_2] \sim \text{IG}(n/2 + c, \boldsymbol{v}_2^\top \boldsymbol{Q} \boldsymbol{v}_2/2 + d),$$
$$[s_1^2 \mid \boldsymbol{\psi}_1] \sim \text{IG}(n/2 + r, \boldsymbol{\psi}_1^\top \boldsymbol{\psi}_1/2 + q) \quad \text{and} \quad [s_2^2 \mid \boldsymbol{\psi}_2] \sim \text{IG}(n/2 + r, \boldsymbol{\psi}_2^\top \boldsymbol{Q} \boldsymbol{\psi}_2/2 + q).$$

Note that these sets of variance parameters, (σ_1^2, σ_2^2) and (s_1^2, s_2^2), are not necessarily on the same scale.

5 Empirical simulation study

5.1 Simulation description

In constructing our simulation study, we aim to generate data that behave similarly to what might be observed in practice. Rather than generating data synthetically from a model, we take direct estimates from an existing survey dataset and add noise based on the reported survey variances to generate data. We took this approach to preserve many of the characteristics associated with the real data when creating synthetic datasets. The setup is similar to what is done in Bradley et al. (2015, 2018), and Janicki et al. (2022).

Specifically, the data for this simulation study was based on the publicly available ACS 5-year estimates from 2015 to 2019 (U.S. Census Bureau, 2020). The code used to conduct this simulation study was written in R (R Core Team, 2023) using publicly available packages. For details, please see the Data Availability section.

Let z_i be the observed direct estimate of median rent as a percentage of household income (median rent burden) for the ith county in North Carolina. Note that there are $n = 100$ counties in total. Also, let d_i be the reported sampling variance associated with z_i. In this simulation study, we treat the $\boldsymbol{z} = (z_1, \ldots, z_n)^\top$ as the *truth*, or the unobserved true small area means of interest. Figure 1 shows the spatial distribution of the $\boldsymbol{z}$ and a scatterplot of $\sqrt{d_i}$ and z_i. The map Figure 1(a) provides evidence for the spatial correlation present throughout the state, including a cluster of counties in the eastern portion of state with the highest rent burden. There are also multiple instances of low and high-rent burden counties being adjacent to each other (known as negative spatial correlation). This includes Gates and Perquimans counties, near the northeastern corner of the state, that represent the highest and lowest values in the state. The scatterplot Figure 1(b) shows that the counties with highest rent burdens also have some of the largest survey variances. All of these factors will test the robustness of the various estimators.

In this simulation study, we run $G = 300$ simulations, each with a different synthetic dataset that is created by perturbing $\boldsymbol{z}$ with noise distribution that uses the reported survey variances. Specifically, we generate

$$y_i^{(g)} \overset{\text{ind}}{\sim} N(\theta_i = log(z_i), d_i/z_i^2)$$

for $i = 1, \ldots, n$ and $g = 1, \ldots, 100$, where i is the index for a given county and g is the index for the simulation. As the rent burden can be skewed, with a lower bound at zero, we take a log transformation to better meet the Gaussian assumption. Note that since the sampling variance for the log transformed rent burden is not known, we use the delta method to estimate this quantity. Note that although we use the log transformation to generate the data, the predictive performance is all assessed on the original scale [assessed on the estimates of z_i, not $\log(z_i)$].

In each simulation, we use the synthetic dataset $\boldsymbol{y}^{(g)} = \{y_1^{(g)}, \ldots, y_n^{(g)}\}$ and $\boldsymbol{D} = \text{diag}\{d_i/z_i^2\}_{i=1}^{n=100}$ to predict $\boldsymbol{z}$. We compare the predictive performance of the SSD model with four other methods: three model-based methods and the direct survey estimate using this synthetic dataset. The three other model-based estimates we compared were the independent FH model, the DM spike-and-slab model, and a FH model with BYM spatial random effects. As a reminder, BYM models spatial dependence but does not select random effects. The DM model performs selection but does not incorporate spatial dependence for the random effects or the selection process like the SSD.

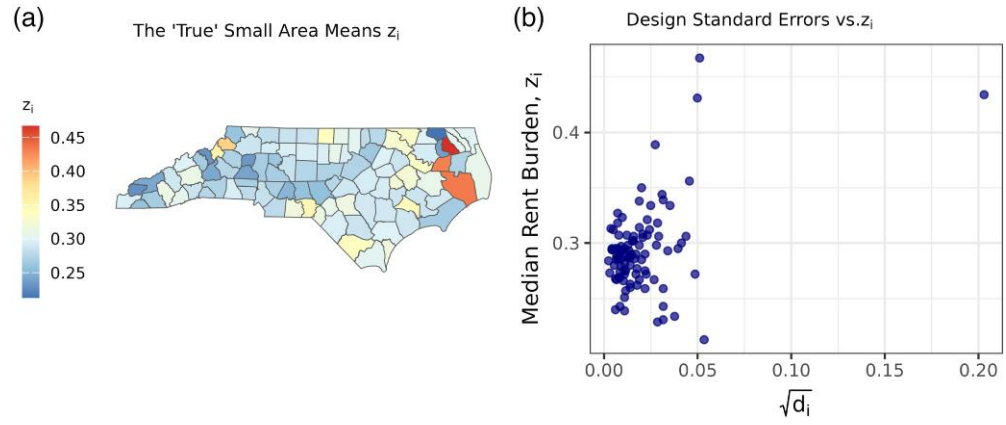

Figure 1. (a) A map of the direct estimates z_i of rent burden, which we treat as the true small area means for our simulation study. (b) A scatterplot of the survey standard errors $\sqrt{d_i}$ and z_i.

The model-based estimates used $\{\boldsymbol{y}^{(g)}, \boldsymbol{D}\}$ in addition to covariates $\boldsymbol{X}$ to fit the model. The matrix $\boldsymbol{X}$ contained ACS estimates for rates of college graduates, residents receiving public assistance income or Supplemental Nutrition Assistance Program in the past 12 months, residents who own a car, the poverty rate, racial makeup of residents (White, Black, Native, Asian), and the rate of Hispanic residents.

The model-based approaches were all fit using Markov Chain Monte Carlo (MCMC). Each model used the same MCMC sample size ($S = 2{,}000$) with varying amounts of burn-in; the non-spatial models used 9,000 iterations while the spatial models used 2,000 iterations for burn-in. A randomized set of simulations was checked and lack of convergence was not detected based on a thorough visual inspection trace plots for all parameters and all models.

The standard improper prior $\pi(\boldsymbol{\beta}, \sigma^2_{FH}) \propto 1$ was used for the FH. The DM used an improper prior $\pi(\boldsymbol{\beta}) \propto 1$, in addition to an Empirical-Bayes like prior $\sigma^2_{DM} \sim \text{IG}(3, 2\bar{d})$, where $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$, was recommended by the authors (Datta & Mandal, 2015) and acts as a scaling mechanism. The BYM model used $\pi(\boldsymbol{\beta}) \propto 1$ and an noninformative IG(c, c) prior for σ^2_1, σ^2_2 with $c = 5 \times 10^{-5}$. Finally, the SSD model which, again, was fit after scaling the data, used a noninformative $\boldsymbol{\beta} \sim N_j(\mathbf{0}, k^2 I_j)$ with $k^2 = 100^2$, an IG(5, 5) prior for σ^2_1, σ^2_2, and IG(5, 10) for the logit-variance parameters s^2_1, s^2_2. Note that both the BYM and SSD models scaled the ICAR precision matrix. All of the models used noninformative priors for the fixed effects. Note that the random effect variance priors are not easily comparable across models. This is due to some models having one random effect term while the spatial models have two, as well as the SSD model using scaled data.

5.2 Assessment

We compare the predictive capability of the various methods for $G = 300$ simulations. For a given county i and dataset g, let $\hat{z}_i^{(g)}$ denote the point estimate (posterior mean for the model-based estimates) of rent burden and $\{\hat{l}_i^{(g)}, \hat{h}_i^{(g)}\}$ be the lower and upper endpoints of the 90% credible interval ($\alpha = 0.1$). We compare the performances of the various methods using mean squared error (MSE), coverage rate, and interval score, averaged across the $G = 300$ simulations. We also compare the absolute bias of the estimators produced by each method. The coverage rate and interval scores are not computed for the direct estimates as they do not produce credible intervals. Note that the predictive performance is all assessed on the original scale [assessed on the estimates of z_i, not $\log(z_i)$].

The average of MSE serves as an overall measure of performance, balancing both bias and variance of the estimator. It is given by

$$\text{Avg MSE} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{n}\sum_{i=1}^{n}(\hat{z}_i^{(g)} - z_i)^2.$$

Table 1. A comparison of estimates from various methods in our empirical simulation study, which used ACS median rent burden data from North Carolina

Estimator	MSE	90% Cr. Int. coverage rate	Interval score for 90% Cr. Int.	Absolute bias
Direct estimate	11.1×10^{-4}	–	–	0.0018
BYM	7.0×10^{-4}	0.828	0.1124	0.0110
FH	6.9×10^{-4}	0.833	0.1093	0.0108
DM	6.8×10^{-4}	0.789	0.0980	0.0110
SSD model	5.4×10^{-4}	0.894	0.0769	0.0087

Note. The simulation study contained 300 simulation iterations. The Direct Estimate along with four model-based estimates were compared: independent FH model, the DM model, FH model with BYM effects, and the proposed SSD model. For the models, posterior means were used as point estimates and 90% Credible Intervals (Cr. Int.) were used for the interval estimates. Bold values indicate the best performing model for a given metric. ACS = American Community Survey; MSE = mean squared error; FH = Fay–Herriot; DM = Datta–Mandal; BYM = Besag–York–Mollié; SSD = spatially selected and dependent.

The average Coverage Rate indicates how well the 90% credible intervals perform in terms of capturing the true small area means across areas and simulations. This is given by

$$\text{Avg Coverage Rate} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{I}\{\hat{l}_i^{(g)} < z_i\} \cdot \boldsymbol{I}\{z_i < \hat{u}_i^{(g)}\}.$$

The interval score is a more comprehensive way to assess interval estimates, as discussed by Gneiting and Raftery (2007). The interval score penalizes for length of the interval and missed coverage on both the upper and lower endpoints. The average interval score for the simulation study is given by

$$\begin{aligned}\text{Avg Interval Score} = \frac{1}{G}\sum_{g=1}^{G}\frac{1}{n}\sum_{i=1}^{n} (\hat{u}_i^{(g)} - \hat{l}_i^{(g)}) + \frac{2}{\alpha}(\hat{l}_i^{(g)} - z_i)\boldsymbol{I}\{\hat{l}_i^{(g)} > z_i\} \\ + \frac{2}{\alpha}(z_i - \hat{u}_i^{(g)})\boldsymbol{I}\{z_i > \hat{u}_i^{(g)}\},\end{aligned}$$

where α corresponds to the $(1-\alpha)\times 100\%$ credible interval ($\alpha = 0.1$ in this case).

Finally, we also compare the average absolute bias for each estimator, given by

$$\text{Avg Absolute Bias} = \frac{1}{n}\sum_{i=1}^{n}\left| z_i - \frac{1}{G}\sum_{g=1}^{G}\hat{z}_i^{(g)}\right|.$$

5.3 Results

Table 1 summarizes the results from the simulation study. Comparing the average MSE, the SSD model outperforms the direct estimate by 52%, the BYM by 23%, the FH by 22%, and the DM by 21%. Note that the two best-performing methods used shrinkage priors for the random effects. Figure 2 compares the log MSE of the estimates of the SSD model against those from the other methods for the 300 simulations. We see that the SSD model consistently outperforms the other methods across simulations and can be used to greatly improve MSE of the small area estimates. Moreover, among the model-based approaches, the SSD model achieves the lowest bias by a 19% margin compared to the FH (the model with the second-lowest bias).

In terms of coverage rate, the SSD model is the only model whose 90% credible intervals actually contains the truth, z, about 90% of the time with a coverage rate of 89.4%. The BYM, FH, and DM all exhibit varying degrees of undercoverage. This superior coverage is achieved without

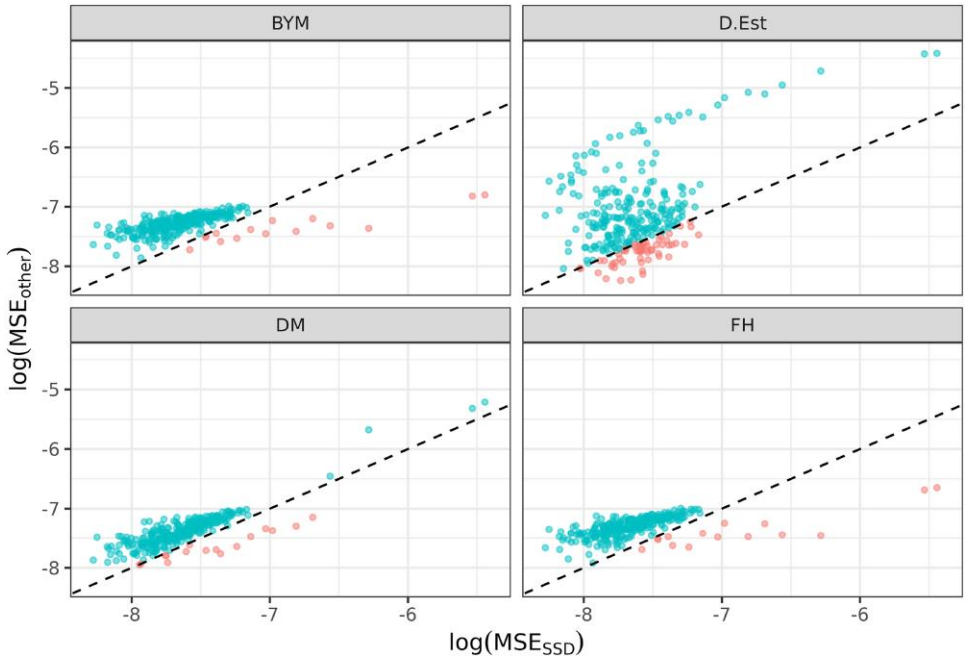

Figure 2. Comparison of the Log mean squared error (MSE) of the estimates from the spatially selected and dependent (SSD) model against alternative methods for each of the 300 simulations. Points above the dotted line indicate simulations where the SSD model had lower MSE while the points below the dotted line correspond to simulations where the SSD model had higher MSE.

excessively large intervals, as indicated by the interval score; the SSD model has a lower interval score than the DM by 22% (the model with the second lowest score). See Section 1 in the online supplementary materials for more details.

6 Estimates of median rent burden for the South Atlantic Census Division

In this section, we fit the SSD model in order to estimate median rent burden in the United States South Atlantic Census Division, using 2015–2019 5-year ACS data (U.S. Census Bureau, 2020). The South Atlantic Census Division comprises of eight states plus the District of Columbia. There are $n = 588$ counties in total for this Census Division. We consider the same covariates used in Section 5, in addition to adding state-level fixed effects. Also, we used the same priors from the simulation study.

The MCMC algorithm for the SSD model was run for 4,000 iterations with the first 1,500 discarded as burn-in. No lack of convergence was detected. For North Carolina datasets with $n = 100$ areas used for the simulation study, all of the candidate models take less than a minute to fit. For the South Atlantic Census Division with $n = 588$ areas, the runtimes for the spatial models are significantly higher: CAR takes ≈48 min, BYM takes ≈37 min, and SSD takes ≈66 min on a 2021 M1 Macbook Pro. The high computational burden in high dimensions is a problem with spatial models in general. That said, we do not think this degree of computational burden is a real obstacle to the practical use of these models. All of the code used to conduct this analysis was written in R (R Core Team, 2023). For details, see the Data Availability section.

The maps in the first row of Figure 3 compare the estimates from our proposed SSD model with the direct estimates. We see that both estimators exhibit the same overall spatial pattern, but there is a smoothing effect through the use of the model-based estimates, as expected. The impact of the smoothing is noticeable near the boundary between West Virginia and Virginia, in central South Carolina, as well as eastern North Carolina. The maps in the second row of Figure 3 compare the log SE of the direct estimates with the log SE (posterior standard deviations) from the SSD model. We can see a reduction in the log SE from the SSD throughout the division. The reduction is

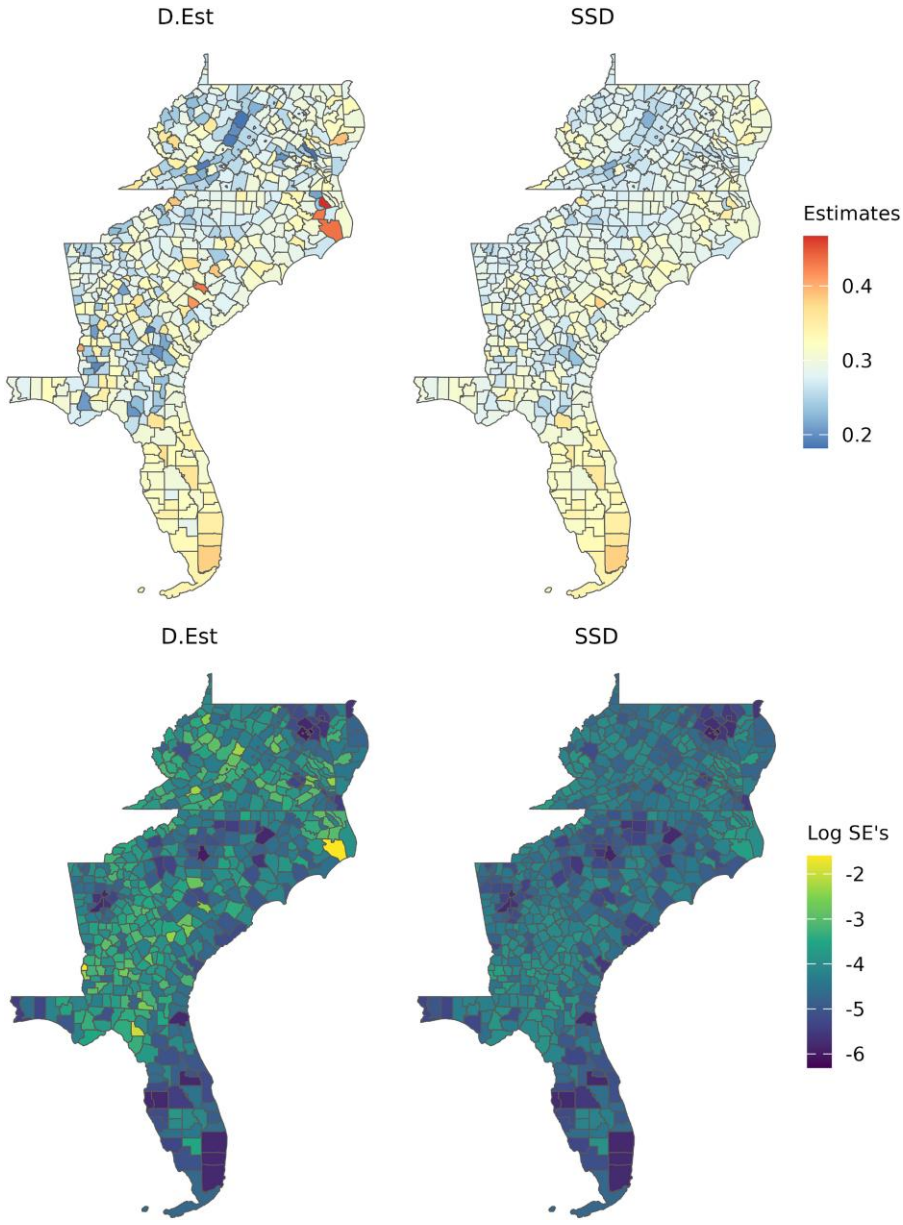

Figure 3. Comparison of the estimates and standard errors for county-level median rent burden using the direct estimates (D.Est) and the spatially selected and dependent (SSD) model. The study region is the South Atlantic Census Division, which consists of 588 counties from D.C. and eight states.

especially noticeable in West Virginia, Western and Eastern ends of North Carolina, Georgia, and Northern Florida.

We also fit the DM model to estimate median rent burden for comparison. A Monte-Carlo Simulation of Geary's C test with 1,000 iterations was performed on the posterior means of the random effects from the DM model. The resulting p-value was below 0.001. This indicates that there is a strong spatial dependence in the random effects of the DM, despite the model's assumption of independence and the inclusion of state-level fixed effects. More details on the random effects in this analysis, including a comparison of random effects from different models can be found in Section 2 of the online supplementary materials.

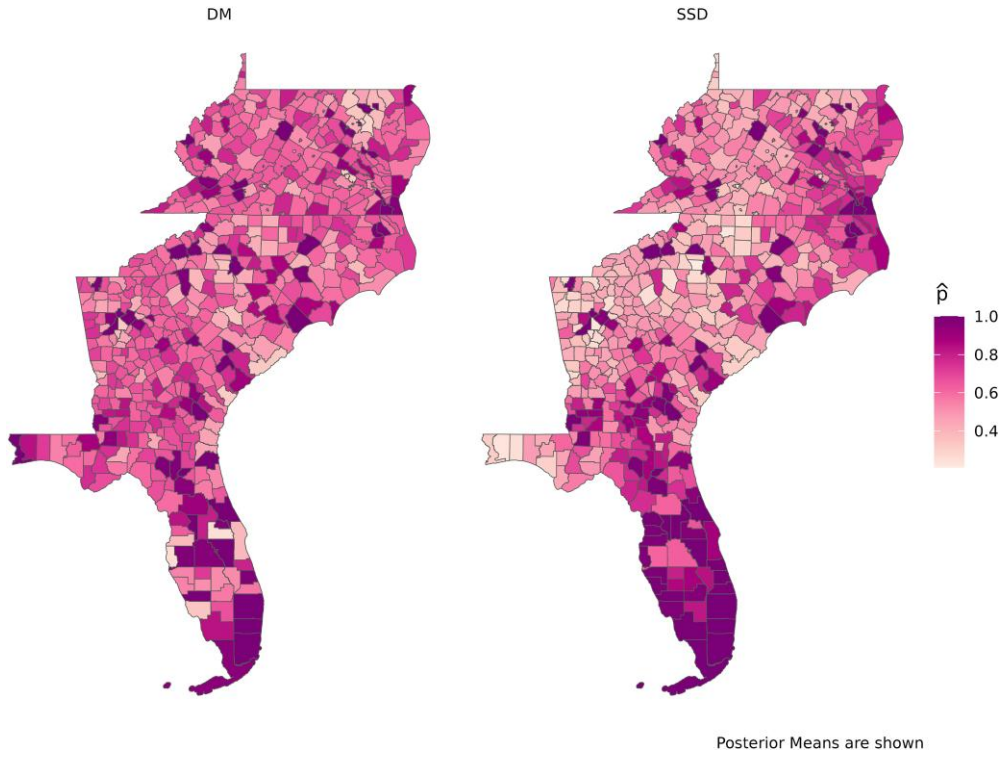

Figure 4. Comparison of the posterior mean of the random effect selection probabilities between the Datta–Mandal (DM) (left) and the spatially selected and dependent (SSD) model (right). If the selection probability for a given county is high, the data indicates the need for inclusion of a random effect for that county.

Finally, Figure 4 shows the posterior means of the random effect selection probabilities of the DM and SSD models. We can see in Figure 4 that the selection probabilities of the DM model also have a strong spatial pattern, again, despite the inclusion of state-level fixed effects. Most counties in central and southern Florida, for example, have selection probabilities close to 1. The spatial pattern is shown by the selection probabilities in the DM model is directly modelled in the SSD approach. Correspondingly, we see that the SSD selection probabilities exhibit a smoother spatial pattern. This analysis, as well as the empirical simulation study, suggests that incorporating spatial dependence in both the random effects and the shrinkage/selection probabilities may be useful for improving estimates in certain settings.

7 Discussion

In this work, we develop a new SAE model combining both shrinkage of random effects and spatial modelling. The proposed Bayesian model incorporates spatial dependence through priors on both the random effects and the selection probabilities of the spike-and-slab prior, extending the approach by Datta and Mandal (2015). Inference of our model can be carried out with a computationally efficient Gibbs sampler by using PG data augmentation (Polson et al., 2013).

Using data from the American Community Survey, we conducted an empirical simulation study and data analysis, both centred around estimation of median rent burden, a policy-relevant statistic. In the simulation study, we showed that the new model can produce far more accurate point and interval estimates, compared to standard approaches (direct survey estimates and the independent FH) and approaches that use shrinkage or spatial priors alone (DM and the BYM). We also demonstrated the ability of the SSD model to reduce uncertainty in the data analysis compared to the direct estimate. Both the simulation study and the data analysis illustrate the benefits of incorporating spatial dependence for both the random effects and the latent selection process.

High computational burden is an issue for many spatial models. The Gibbs sampler for fitting our proposed model requires two inversions of the covariance matrix for every MCMC iteration: one for the blocked fixed and random effects and the other for the blocked logit parameters (see Section 4). These matrix inversions can make fitting the model time-intensive for applications with a large number of small areas. Future work involves development of a more efficient sampler or use of approximations to help speed up the computation for massive datasets.

The simulation study and data analysis in this article both focused on estimating county-level median rent burden using ACS data. However, the SSD model is of general interest and applicable to many other Small Area Estimation scenarios where the covariates explain the variable of interest well in many areas and the area-level random effects are expected to exhibit spatial shrinkage and dependence. The random effect priors used for the SSD model could also be adapted in a straightforward manner to spatial models in other domains such as disease mapping.

Conflicts of interest: None declared.

Funding

This research was partially supported by the U.S. National Science Foundation (NSF) under NSF Grant NCSE-2215169. 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 author and not those of the NSF.

Data availability

The code needed to download the data and reproduce the results for this paper are available at https://github.com/sho-kawano/ssd_paper_code and is written in R v4.2.3 (R Core Team, 2023). All of the data used is publicly available. The R package `tidycensus` v1.4.4 (Walker & Herman, 2024) was used to download ACS data. The TIGER shapefiles for the counties were downloaded using `tigris` v2.0.3 (Walker, 2023).

Supplementary material

Supplementary material is available online at *Journal of the Royal Statistical Society: Series A*.

References

Besag J. (1974, January). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, *36*(2), 192–225. https://doi.org/10.1111/j.2517-6161.1974.tb00999.x

Besag J., York J., & Mollié A. (1991, March). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, *43*(1), 1–20. https://doi.org/10.1007/BF00116466

Bradley J. R., Holan S. H., & Wikle C. K. (2015, December). Multivariate spatio-temporal models for high-dimensional areal data with application to longitudinal employer-household dynamics. *The Annals of Applied Statistics*, *9*(4), 1761–1791. https://doi.org/10.1214/15-AOAS862

Bradley J. R., Holan S. H., & Wikle C. K. (2018). Computationally efficient multivariate spatio-temporal models for high-dimensional count-valued data (with discussion). *Bayesian Analysis*, *13*(1), 253–310. https://doi.org/10.1214/17-BA1069

Chakraborty A., Datta G. S., & Mandal A. (2016). A two-component normal mixture alternative to the Fay–Herriot model. *Statistics in Transition New Series*, *17*(1), 67–90. https://doi.org/10.59170/stattrans-2016-004

Chung H. C., & Datta G. S. (2022, December) Bayesian spatial models for estimating means of sampled and non-sampled small areas. *Survey Methodology*, Statistics Canada (Catalogue No. 12-001-X, Vol. 48, No. 2.). http://www.statcan.gc.ca/pub/12-001-x/2022002/article/00012-eng.htm

Cromwell M. (2022, December). Renters more likely than homeowners to spend more than 30% of income on housing in almost all counties. U.S. Census Bureau Website. https://www.census.gov/library/stories/2022/12/housing-costs-burden.html

Crump S., & Schuetz J. (2021, April). U.S. rental housing markets are diverse, decentralized, and financially stressed. Web Articles, The Brookings Institution. https://www.brookings.edu/articles/us-rental-housing-markets/

Datta G. S., Hall P., & Mandal A. (2011, March). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, *106*(493), 362–374. https://doi.org/10.1198/jasa.2011.tm10036

Datta G. S., & Lahiri P. (1995, August). Robust hierarchical Bayes estimation of small area characteristics in the presence of covariates and outliers. *Journal of Multivariate Analysis*, *54*(2), 310–328. https://doi.org/10.1006/jmva.1995.1059

Datta G. S., & Mandal A. (2015, October). Small area estimation with uncertain random effects. *Journal of the American Statistical Association*, *110*(512), 1735–1744. https://doi.org/10.1080/01621459.2015.1016526

Desmond M. (2018, January). Heavy is the house: Rent burden among the American urban poor. *International Journal of Urban and Regional Research*, *42*(1), 160–170. https://doi.org/10.1111/ijur.2018.42.issue-1

Fabrizi E., & Trivisano C. (2010, February). Robust linear mixed models for small area estimation. *Journal of Statistical Planning and Inference*, *140*(2), 433–443. https://doi.org/10.1016/j.jspi.2009.07.022

Fay R. E., & Herriot R. A. (1979, June). Estimates of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, *74*(366a), 269–277. https://doi.org/10.1080/01621459.1979.10482505

Freni-Sterrantino A., Ventrucci M., & Rue H. (2018). A note on intrinsic conditional autoregressive models for disconnected graphs. *Spatial and Spatio-temporal Epidemiology*, *26*, 25–34. https://doi.org/10.1016/j.sste.2018.04.002

Ghosh T., Ghosh M., Maples J. J., & Tang X. (2022, July). Multivariate global–local priors for small area estimation. *Stats*, *5*(3), 673–688. https://doi.org/10.3390/stats5030040

Gneiting T., & Raftery A. E. (2007, March). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, *102*(477), 359–378. https://doi.org/10.1198/016214506000001437

Harvard Joint Center for Housing Studies (2024). *America's rental housing 2024* (Technical report). Harvard Kennedy School. https://www.jchs.harvard.edu/sites/default/files/reports/files/Harvard_JCHS_Americas_Rental_Housing_2024.pdf

Horvitz D. G., & Thompson D. J. (1952, December). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, *47*(260), 663–685. https://doi.org/10.1080/01621459.1952.10483446

Janicki R., Raim A. M., Holan S. H., & Maples J. J. (2022, March). Bayesian nonparametric multivariate spatial mixture mixed effects models with application to American community survey special tabulations. *The Annals of Applied Statistics*, *16*(1), 144–168. https://doi.org/10.1214/21-AOAS1494

Jiang J., & Rao J. S. (2020, March). Robust small area estimation: An overview. *Annual Review of Statistics and Its Application*, *7*(1), 337–360. https://doi.org/10.1146/statistics.2020.7.issue-1

Leroux B. G., Lei X., & Breslow N. (2000). Estimation of disease rates in small areas: A new mixed model for spatial dependence. In M. E. Halloran, & D. Berry (Eds.), *Statistical models in epidemiology, the environment, and clinical trials* (pp. 179–191). Springer. ISBN 978-1-4612-1284-3. https://doi.org/10.1007/978-1-4612-1284-3_4

Park T., & Casella G. (2008, June). The Bayesian Lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Parker P. A., Holan S. H., & Janicki R. (2023, February). Conjugate modeling approaches for small area estimation with heteroscedastic structure. *Journal of Survey Statistics and Methodology*, *12*(4), 1061–1080. https://doi.org/10.1093/jssam/smad002

Parker P. A., Janicki R., & Holan S. H. (2023, September). A comprehensive overview of unit-level modeling of survey data for small area estimation under informative sampling. *Journal of Survey Statistics and Methodology*, *11*(4), 829–857. https://doi.org/10.1093/jssam/smad020

Petrucci A., & Salvati N. (2006, June). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological, and Environmental Statistics*, *11*(2), 169. https://doi.org/10.1198/108571106X110531

Polson N. G., Scott J. G., & Windle J. (2013, December). Bayesian inference for logistic models using Pólya-Gamma latent variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349. https://doi.org/10.1080/01621459.2013.829001

Porter A. T., Wikle C. K., & Holan S. H. (2015). Small area estimation via multivariate Fay–Herriot models with latent spatial dependence. *Australian & New Zealand Journal of Statistics*, *57*(1), 15–29. https://doi.org/10.1111/anzs.2015.57.issue-1

Quigley J. M. (2008, May). Housing policy in the United States. *UC Berkeley: Berkeley program on housing and urban policy* (NO. W06-001B). https://escholarship.org/uc/item/89p9r7w9

R Core Team (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Reich B. J., & Staicu A.-M. (2021, December). Bayesian variable selection in spatial regression models. In *Handbook of Bayesian variable selection* (1st ed., pp. 251–270). Chapman and Hall/CRC. ISBN 978-1-00-308901-8. https://doi.org/10.1201/9781003089018-11

Riebler A., Sørbye S. H., Simpson D., & Rue H. (2016, August). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, *25*(4), 1145–1165. https://doi.org/10.1177/0962280216660421

Schmid T., & Münnich R. T. (2014, August). Spatial robust small area estimation. *Statistical Papers*, *55*(3), 653–670. https://doi.org/10.1007/s00362-013-0517-y

Singh B. B., Shukla G. K., & Kundu D. (2005). Spatio-temporal models in small area estimation. *Survey Methodology*, *31*(12), 183–195. https://www150.statcan.gc.ca/n1/pub/12-001-x/2005002/article/9053-eng.pdf

Sørbye S. H., & Rue H. (2014, May). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, *8*, 39–51. https://doi.org/10.1016/j.spasta.2013.06.004

Sugasawa S., Tamae H., & Kubokawa T. (2017, March). Bayesian estimators for small area models shrinking both means and variances. *Scandinavian Journal of Statistics*, *44*(1), 150–167. https://doi.org/10.1111/sjos.v44.1

Tang X., & Ghosh M. (2023). Global–local priors for spatial small area estimation. *Calcutta Statistical Association Bulletin*, *75*(2), 141–154. https://doi.org/10.1177/00080683231186378

Tang X., Ghosh M., Ha N. S., & Sedransk J. (2018, October). Modeling random effects using global–local shrinkage priors in small area estimation. *Journal of the American Statistical Association*, *113*(524), 1476–1489. https://doi.org/10.1080/01621459.2017.1419135

U.S. Bureau of Labor Statistics (2024). Consumer price index for all Urban consumers: Rent of primary residence in U.S. city average. FRED, Federal Reserve Bank of St. Louis. https://fred.stlouisfed.org/series/CUUR0000SEHA

U.S. Census Bureau (2020, December). American community survey 2015–2019 5-year estimates. Accessed via tidycensus package.

U.S. Census Bureau (2023, September). American community survey 2022 1-year estimates. https://www.census.gov/programs-surveys/acs/data.html

Wakefield J. (2007, April). Disease mapping and spatial regression with count data. *Biostatistics (Oxford, England)*, *8*(2), 158–183. https://doi.org/10.1093/biostatistics/kxl008

Walker K. (2023). *Tigris: Load census TIGER/line shapefiles*. https://CRAN.R-project.org/package=tigris

Walker K., & Herman M. (2024). *Tidycensus: Load US census boundary and attribute data as 'tidyverse' and 'sf'-ready data frames*. https://walker-data.com/tidycensus/

You Y. (2016). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models with applications. *Methodology branch working paper, ICCSMD-2016-03-E, Statistics Canada, Ottawa, Canada*.

You Y. (2021, December). Small area estimation using Fay–Herriot area level model with sampling variance smoothing and modeling. *Survey Methodology*, *47*(2), 361–370.

Zhou Q. M., & You Y. (2008, May). Hierarchical Bayes small area estimation for the Canadian community health survey. *Survey Methodology*, *37*(1), 25–37.