Spatial Two-Component Mixture Model for State-Level Cash Rental Rates

Lu Chen* Balgobin Nandram†

Abstract

The United States Department of Agriculture's National Agricultural Statistics Service conducts the Cash Rents Survey to provide the basis for state and county estimates of the cash rent paid for each agricultural land-use category. The use of small area models in estimation process has gained increased attention by statistical agencies. They can "borrow strength" from related areas across space and/or time or through auxiliary information and these models can provide "indirect" but reliable estimates for small areas while also increasing the precision. However, some of the realized sample sizes are too small to support reliable direct estimates, and there are outliers. In addition, quantities of interests for geographically contiguous small areas in Cash Rents Survey display a spatial pattern. Therefore, we propose a hierarchical Bayesian area-level two-component mixture model with spatial random effects to account for outliers and spatial correlation. Moreover, the model incorporates two years of data, and a discounting factor for the first year provides a not-too-tight prior for the hyperparameters. That is, it avoids correlations between the two years of data by using a power prior that partially discounts past data. We assess the effectiveness of the spatial model based on a case study from 2021 and 2022 Cash Rents Survey. The results show superior performance of the proposed model over the direct estimates.

Key Words: Block Gibbs sampler, Grid method, Mixture model, Power prior, Outliers, Small area estimation, Spatial model

*National Institute of Statistical Sciences and USDA National Agricultural Statistics Service, 1400 Independence Avenue, SW, Room 6412 B, Washington, DC 20250-2054. Email: lchen@niss.org.

†Worcester Polytechnic Institute and USDA National Agricultural Statistics Service, Department of Mathematical Sciences, Stratton Hall, 100 Institute Road, Worcester, MA 01609. E-mail: balnan@wpi.edu

1. Introduction

The United States Department of Agriculture's (USDA's) National Agricultural Statistics Service (NASS), one of thirteen US Federal Statistical Agencies, provides official statistics of both state and county-level cash rental rates for irrigated cropland, non-irrigated cropland, pasture land, and cropland. The cash rents estimates provide basic information needed by farmers to make decisions for both short-term and long-term agricultural production planning. These estimates may be used by individual producers in planning for their agricultural operation or by Agricultural Extension Services or university staff in developing operating budgets for agricultural operations in their locale. However, some of the realized sample sizes at the state level may be too small to support reliable direct estimates.

Traditionally, the NASS Agricultural Statistics Board has relied on survey indications combined with historic information to produce official statistics using the survey estimates as a foundation informed by auxiliary information, including historical data, reliable administrative data, and other non-survey data. Although external reviews have consistently found that NASS estimates are the gold standard for the agricultural industry, the process lacked transparency and reproducibility and did not lead to valid measures of uncertainty. In 2021, a small area model for county-level cash rental rates was developed based on annual data collection and implemented in 2021 (Chen and Nandram (2022), Young and Chen (2022)) to improve the transparency and reproducibility. In what follows we discuss the research conducted on the state-level models. Small area models are investigated since it would give alternatives to state estimates ahead of county estimates, and the county-level model with state effects can be used to follow up. The objective for the state-level model is to get a good state model that can be used early in the season to provide more accurate benchmarks for county-level estimates.

USDA's NASS has explored different model-based approaches for cash rental rate estimates and traditionally focused on county-level models. In a frequentist framework, Berg et al. (2014) propose a univariate area-level model that involves fitting two sets of survey-based direct estimates. They assume that the variances for the two years are the same. The model can be characterized as an extension of the Fay-Herriot model (Fay and Herriot (1979)). Erciulescu et al. (2019) propose a HB bivariate unit-level model under the normality assumptions. The model is flexible to allow the variances to differ between the two time-points. However, it is computationally intensive to fit the models in production. In 2021, Chen and Nandram (2022) developed a hierarchical Bayesian (HB) two-component mixture model with power prior at the county level

to accommodate outliers and the model was implemented in production. The model addresses the issues of the assumption of constant variances for two consecutive years and the computational intensity by proposing a general HB model that puts the two years of data together and avoids the two-year correlations by using a power prior.

Recent studies and papers related to NASS small area estimation research on crops county estimates, farm labor program, and county-level cash rental rates program have shown that the HB small area models can incorporate auxiliary sources of data with survey estimates to improve the precision and increase the accuracy of related NASS official estimates. The National Academies of Sciences, Engineering, and Medicine Committee on National Statistics (CNSTAT) published a report (2018) to discuss several needs and requirements for NASS county-level crop estimates that were illuminated during the activities of the CNSTAT panel. NASS has taken important steps towards realizing the vision and recommendation in the report, as indicated in the recent published CNSTAT report, Chapter 8 (2023). Because the interest of NASS programs is in constructing summaries for different levels of geography (county, state, regional, and U.S. levels), the Bayesian approach to model fitting and estimation is preferable. Nandram et al. (2022) and Chen et al. (2022b) proposed and implemented HB subarea-level models with inequality constraints to produce county-level estimates that satisfy important relationships between the estimates and administrative data, along with the associated measures of uncertainty. These models for all crops with federally mandated reporting requirements were implemented for the 2020 crop year. Measures of uncertainty were published with the resulting official statistics. Chen et al. (2022a) discussed several HB subarea-level models in support of estimates of interest in the Farm Labor Survey. The resulting framework provided a complete set of coherent estimates for all required geographic levels and the modeling process was incorporated into the official Farm Labor publication for the first time in 2020. Moreover, as mentioned before, the two-component mixture model for county-level cash rental rates was developed based on annual data collection and implemented in 2021 (Chen and Nandram, 2022). Young and Chen (2022) summarize the current NASS small area models in production.

Our study addresses the spatial correlation among the state-level cash rental rates, especially using an conditional autoregressive (CAR) hierarchical component to account for spatial association. The use of CAR specifications for modeling areal data was first introduced by Besag (1974). Bayesian methods have been the dominating paradigm for spatial models including a CAR component (Sun et al., 1999; He and Song, 2000; Hodges et al., 2003; Banerjee et al., 2014). Based on the data exploration analysis, we found

significant spatial correlation among the estimates. Therefore, we assign the CAR prior for the spatial random effects in the hierarchical model. In addition, similar to the county-level model, the state-level model also addresses the issues of the assumption of constant variances for two consecutive years and the computational intensity by proposing a general HB model that puts the two years of data together and avoids the two-year correlations by using a power prior (Chen and Ibrahim, 2000) that partly discounts past data. In addition, outliers issues are taken into account to increase robustness for the standard HB area-level model with the normality assumptions. One way to accommodate outliers is to use a two-component mixture model (e.g., Gershunskaya and Lahiri, 2017). They use an Empirical Bayes approach assisted by the expectation-maximization (EM) algorithm. Chakraborty et al. (2019) and Goyal et al. (2021) have provided a full Bayesian approach for the unit-level nested error regression model. By using a two-component mixture of normal distributions, this model accommodates populations where a small portion of unit-level errors come from a secondary distribution with a larger variance than the primary distribution. In practice, because unit-level models generally require substantially more computational time, area-level models are more applicable for the production of official statistics that are published on tight timelines. Therefore, we focus on the area-level two-component mixture model.

The paper is structured as follows. Section 2 describes the survey procedures and background. The proposed HB spatial two-component mixture model with technical details is presented in Section 3. In Section 4, a case study using 2021 and 2022 cash rent data illustrates the performance of the model. Section 5 provides a summary and some discussion of possible future research for cash rental rates and small area estimation more generally.

2. Data Sources and Requirements

The Cash Rents Survey is conducted on an annual basis. The survey obtains acres rented and cash rental rates from a statistically representative sample of farmers and ranchers in the United States, excluding Alaska. This survey provides the basis for estimates of the current year's state-level and county-level cash rents paid for irrigated cropland, non-irrigated cropland, and permanent pastureland. From 1950 to 1974, a list survey of real estate appraisers was used to estimate state-level cash rents. Beginning in 1974, producers provided information about their rental agreements by responding to questions on the June Area Survey. In the 2008 farm bill, NASS was mandated to provide mean rental rates for all counties (not just states) with at least

20,000 acres of crop land.

The target population for the cash rents estimate program is all farms and ranches with $1,000 or more in agricultural sales (or potential sales) who rent land from others on a cash rent basis. The Cash Rent Survey sample is selected from a list frame of farm and ranch operators maintained by NASS. NASS is constantly seeking qualifying farming operations from outside sources to be added to the list. A profile, known as control data, of each operation is maintained, which indicates what the farm has historically produced and a general indication of size. This information allows NASS to define sampling populations that are specific to each survey and employ advanced and more efficient sample designs. Samples for the Cash Rents Survey are drawn with a county-level stratified design to produce state and county-level estimates. Large operations in each county are stratified into the census strata, where all are included in the sample. The national sample size for the Cash Rents Survey is approximately 260,000. The sample is stratified by state and county within state to produce state and county-level estimates. Data collection occurs from late February until the end of June. Variances for cash rental rate estimates are constructed using a second-order Taylor series expansion for the ratio.

From the Cash Rents Survey, county, state, and national rental rates (dollars/acre) for each land-use category (irrigated, non-irrigated, and pasture) are published (see Figure 1 for the 2022 state-level published cash rental rate estimates for cropland, combined land type with irrigated land and non-irrigated land). Although total value of cash rents and acres rented on a cash basis are computed, these values have not been published. Historically, the state-level direct survey estimates were reviewed and, if deemed appropriate, adjusted by NASS staff or the Agricultural Statistics Board. The primary reason for adjustment was a large difference between the previous year's published cash rental rate and the current year's survey estimate due to small sample sizes. The adjusted estimate was restricted to being between, or on, the current survey estimate and the previous year's published estimate so that the direction of change was honored. If the number of responses within a state was substantial, then the survey estimate received the greatest weight. If only few responses were received, then the previous year's published value and the estimates from surrounding counties or the agricultural statistics district were given more weight. Any model to replace the expert opinion needs to follow the guidelines used in the review process.

In 2021, the two-component mixture model was implemented for the county-level cash rental rate production process. It accounts for the previous-year estimates and also accommodates the outliers, providing both key indicators for the county-level official statistics and the corresponding measures of uncertainties.

However, the state-level estimates are published before the county-level ones, and they are based on the review of the Agricultural Statistics Board. The state-level models are also under research to provide more transparent and reproducible methods. This paper describe the process and the model for the state-level model. It would provide alternatives to state estimates ahead of county estimates.

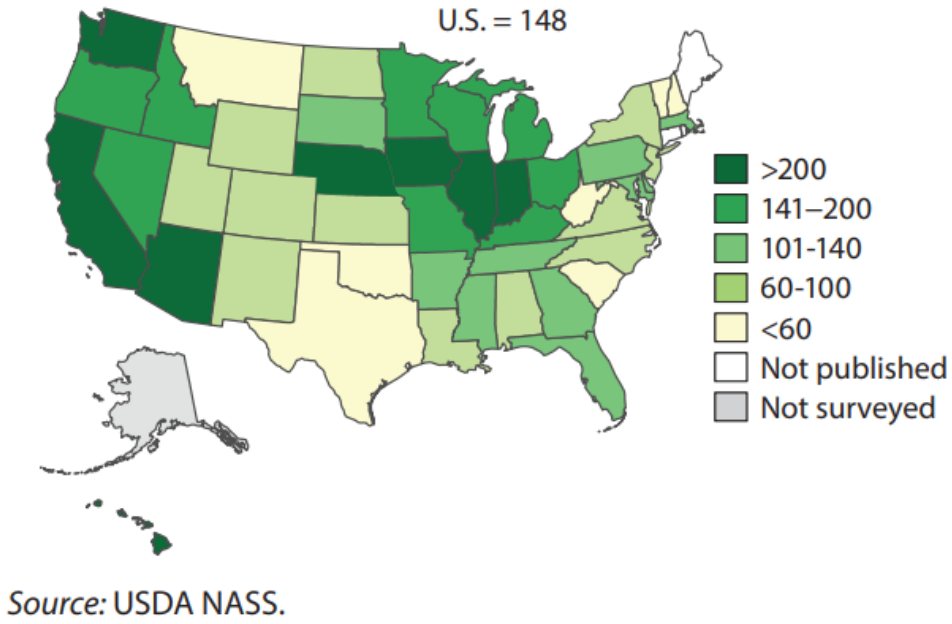

Figure 1: State-level published cash rental rate estimates for cropland

3. Model

In this section, we describe a HB spatial two-component mixture model that accounts for spatial association and outliers and uses the power prior to control how much of the past data is used. Bayesian spatial models are commonly used when the data are not independent and have spatial patterns. The spatial random effects are often described by a CAR distribution, and a CAR prior is adopted to this distribution. For the outlier operations, the two-component mixture model is applied (Chen and Nandram, 2022). In addition, since the research interest is only in the second year (i.e., the year of the survey), we can use the data in the prior year(s) to obtain a power prior (Ibrahim and Chen, 2000; Ibrahim, et al., 2015) for the parameters of the current year. In this way, the correlation between survey data from the two consecutive years can be avoided, which reduces the computational time to a manageable scale.

3.1 A Hierarchical Bayesian Spatial Two-Component Mixture Model

We consider a model for cash rents data at the state level for both years. The model penalizes last year's data and also incorporates outliers in the mixture model with a penalty. In addition, neighboring states should be similar, and therefore a spatial analysis may be useful. For all states, we consider the incidence matrix W, with diagonal elements being zeros and off-diagonal elements being ones if two states are in the same neighborhood and zero otherwise. This is symmetric matrix with real eigenvalues. Let $r_i = \sum_{j=1}^{\ell} w_{ij}, i = 1, \ldots, \ell$ and R denote the diagonal matrix, $R = \text{Diag}(r_i, i = 1, \ldots, \ell)$. The spatial matrix we considered is $(R - \gamma W)^{-1}$, the CAR model, where γ is the so-called spatial correlation and $\lambda_1^{-1} < \gamma < \lambda_\ell^{-1}$ for matrix $(R - \gamma W)^{-1}$ to be positive definite, where $\lambda_1, ..., \lambda_\ell$ are the eigenvalues of W. Note that λ_1 is negative and λ_ℓ is positive, and so γ is in a limited sub-interval of $(-1, 1)$.

For our dataset of 49 states including three covariates, the survey estimates, and the corresponding survey standard errors, we found significant spatial correlation. We estimated the spatial correlation using Moran's I (Moran, 1950). We obtained $0.553, 0.524, 0.451$ for the three covariates; and 0.811 for the survey estimates; and 0.293 for the survey variances, with the p-values in all cases being at most 0.009; for the standard errors, the spatial correlation is 0.528. This confirms that there is significant spatial correlation.

For modeling, the subscript 1 will denote year 1 (e.g., previous year), and the subscript 2 will denote year 2 (e.g., the current year). Let $\hat{\theta}_{ts}$ denote the direct estimates and $\hat{\sigma}_{ts}^2$ denote the corresponding sampling variances, where $t = 1, 2$ indicating years and $s = 1, \ldots, S$ indicating states. The spatial two-component mixture model is,

$$\hat{\theta}_{1s} \mid z_{1s} = 0 \sim \text{Normal}(\underset{\sim}{x}_{1s}'\underset{\sim}{\beta} + \phi_{1s}, \rho\frac{\hat{\sigma}_{1s}^2}{a}), \;\; \hat{\theta}_{1s} \mid z_{1s} = 1 \sim \text{Normal}(\underset{\sim}{x}_{1s}'\underset{\sim}{\beta} + \phi_{1s}, \frac{\hat{\sigma}_{1s}^2}{a})$$

and

$$\hat{\theta}_{2s} \mid z_{2s} = 0 \sim \text{Normal}(\underset{\sim}{x}_{2s}'\underset{\sim}{\beta} + \phi_{2s}, \rho\hat{\sigma}_{2s}^2), \;\; \hat{\theta}_{2s} \mid z_{2s} = 1 \sim \text{Normal}(\underset{\sim}{x}_{2s}'\underset{\sim}{\beta} + \phi_{2s}, \hat{\sigma}_{2s}^2).$$

We list the priors in the following. First, the random effect $\underset{\sim}{\phi}_t$ is assigned the CAR prior,

$$\underset{\sim}{\phi}_t \mid \Omega \overset{ind}{\sim} \text{Normal}(\underset{\sim}{0}, \delta^2(R - \gamma W)^{-1}), t = 1, 2,$$

where W is the adjacency matrix as described above and R is the diagonal matrix.

Second, the latent variables are z_{ts}. It is computationally advantageous to augment the joint posterior

density for mixture models using the latent binary variables. Define z_{ts}, where $z_{ts} = 1$ if a state is an outlier and $z_{ts} = 0$ otherwise. The priors are

$$z_{ts} \sim \text{Bernoulli}(p_t), p_t \sim \text{Uniform}(0, \frac{1}{2}), t = 1, 2, s = 1, \ldots, S.$$

Third, for parameter (δ^2, γ), the prior is $\pi(\delta^2, \gamma) \propto 1/\delta^2, \lambda_1^{-1} < \gamma < \lambda_\ell^{-1}, \delta^2 > 0$.

Fourth, an empirical diffuse prior is assigned to the coefficients $\boldsymbol{\beta}$, that is, a multivariate normal prior distribution with fixed and known mean and variance and covariance matrix $\boldsymbol{\beta} \sim MN(\hat{\boldsymbol{\beta}}, 1000\hat{\Sigma}_{\hat{\boldsymbol{\beta}}})$. Here, $\hat{\boldsymbol{\beta}}$ are the least squares estimates of $\boldsymbol{\beta}$ obtained from fitting a simple linear regression model of the county-level survey estimates on the auxiliary data $\boldsymbol{x}_{ij}$ and $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}}$ is the estimated covariance matrix of $\hat{\boldsymbol{\beta}}$.

3.2 Computation

We can write these assumptions compactly as follows. Let

$$d_{10s}^2 = \rho\frac{\hat{\sigma}_{1s}^2}{a}, \;\; d_{11s}^2 = \frac{\hat{\sigma}_{1s}^2}{a}, \;\; d_{20s}^2 = \rho\hat{\sigma}_{2s}^2, \;\; d_{21s}^2 = \hat{\sigma}_{2s}^2, \;\; s = 1, \ldots, S.$$

The model can be written as

$$\hat{\theta}_{ts} \mid z_{ts} = u \sim \text{Normal}(\underset{\sim}{x}_{ts}'\underset{\sim}{\beta} + \phi_{ts}, d_{tus}^2), t = 1, 2, u = 0, 1, s = 1, \ldots, S. \tag{1}$$

The assumption (1) represents the entire likelihood function. Let $\Omega = \{a, \rho, \gamma, \delta^2\}$ denote the set of all hyper-parameters. Then given Ω, the $\hat{\theta}_{ts}$ are independent with

$$f(\hat{\theta}_{ts} \mid z_{ts}, \Omega) = \frac{1}{(2\pi d_{t0s}^2)^{(1-z_{ts})/2}} \frac{1}{(2\pi d_{t1s}^2)^{z_{ts}/2}} \exp\{-(\frac{1 - z_{ts}}{2d_{t0s}^2} + \frac{z_{ts}}{2d_{t1s}^2})(\hat{\theta}_{ts} - \underset{\sim}{x}_{ts}'\underset{\sim}{\beta} - \phi_{ts})^2\}$$

$$= \frac{1}{(2\pi d_{t0s}^2)^{(1-z_{ts})/2}} \frac{1}{(2\pi d_{t1s}^2)^{z_{ts}/2}} \frac{1}{(\frac{1-z_{ts}}{d_{t0s}^2} + \frac{z_{ts}}{d_{t1s}^2})^{1/2}} \text{Normal}_{\phi_{ts}} \left\{ \hat{\theta}_{ts} - \underset{\sim}{x}_{ts}'\underset{\sim}{\beta}, \frac{1}{\frac{1-z_{ts}}{d_{t0s}^2} + \frac{z_{ts}}{d_{t1s}^2}} \right\}.$$

We can run an ordinary Gibbs sampler with all the conditions. We have encountered difficulties with $\underset{\sim}{\beta}$ mixing slowly; a possible cause is its connection to $\underset{\sim}{\phi}$. So we draw $\underset{\sim}{\beta}$ and $\underset{\sim}{\phi}$ together, and this is done using the decomposition,

$$\pi(\underset{\sim}{\beta}, \underset{\sim}{\phi} \mid \Omega, \underset{\sim}{\hat{\theta}}) = \pi(\underset{\sim}{\beta} \mid \Omega, \underset{\sim}{\hat{\theta}})\pi(\underset{\sim}{\phi} \mid \underset{\sim}{\beta}, \Omega, \underset{\sim}{\hat{\theta}}).$$

To obtain the CPD of $\underset{\sim}{\phi}_t$, we define

$$\underset{\sim}{\mu}_t = (\mu_{ts}), \mu_{ts} = \hat{\theta}_{ts} - \underset{\sim}{x}'_{ts}\underset{\sim}{\beta} - \phi_{ts}, s = 1, \ldots, S,$$

$$D_t = \text{Diag}\{\frac{1}{\frac{1-z_{ts}}{d^2_{t0s}} + \frac{z_{ts}}{d^2_{t1s}}}, s = 1, \ldots, S\}, t = 1, 2.$$

Then,

$$\underset{\sim}{\phi}_t \mid \Omega, \underset{\sim}{y} \sim \text{Normal}(\underset{\sim}{\mu}_t, D_t), t = 1, 2,$$

and this is combined with the prior,

$$\underset{\sim}{\phi}_t \mid \Omega \overset{ind}{\sim} \text{Normal}(\underset{\sim}{0}, \delta^2(R - \gamma W)^{-1}), t = 1, 2,$$

to get the CPD of $\underset{\sim}{\phi}_t$,

$$\underset{\sim}{\phi}_t \mid \Omega, \underset{\sim}{y} \overset{ind}{\sim} \text{Normal}\left\{[D_t^{-1} + \frac{1}{\delta^2}(R - \gamma W)]^{-1} D_t^{-1}\underset{\sim}{\mu}_t, [D_t^{-1} + \frac{1}{\delta^2}(R - \gamma W)]^{-1}\right\}, t = 1, 2.$$

Therefore, each $\underset{\sim}{\phi}_t$ can be drawn independently from a multivariate normal density.

The integrated CPD of $\underset{\sim}{\beta}$ has two parts, with and without the prior on $\underset{\sim}{\beta}$. We first integrate out $\underset{\sim}{\phi}_t, t = 1, 2$, and then we combine this with the prior on $\underset{\sim}{\beta}$; sampling of τ is another CPD, not needed in this discussion (done in a previous section). We have the following,

$$\hat{\theta}_{ts} \mid z_{ts} = u \overset{ind}{\sim} \text{Normal}(\underset{\sim}{x}'_{ts}\underset{\sim}{\beta} + \phi_{ts}, d^2_{tus}), t = 1, 2, u = 0, 1, s = 1, \ldots, S,$$

which we write more compactly as

$$\underset{\sim}{\hat{\theta}}_t \mid \underset{\sim}{\beta}, \underset{\sim}{\phi}_t \overset{ind}{\sim} \text{Normal}(X_t\underset{\sim}{\beta} + \underset{\sim}{\phi}_t, D_t), t = 1, 2, \tag{2}$$

and for ease of presentation, we drop all obvious conditioning variables. Recall that our prior on $\underset{\sim}{\phi}_t$ is

$$\underset{\sim}{\phi}_t \overset{ind}{\sim} \text{Normal}\{\underset{\sim}{0}, \delta^2(R - \gamma W)^{-1}\}, t = 1, 2. \tag{3}$$

Then, integrating out $\underset{\sim}{\phi}_t$ from (2) and (3), we have

$$\hat{\underset{\sim}{\theta}}_t \mid \underset{\sim}{\beta} \overset{ind}{\sim} \text{Normal}\{X_t\underset{\sim}{\beta}, \delta^2(R-\gamma W)^{-1} + D_t\}, t = 1, 2. \tag{4}$$

Recall that the conditioning is actually on all variables.

First, we consider the integrated CPD of $\underset{\sim}{\beta}$ with the prior, $\pi(\underset{\sim}{\beta}) = 1$. Letting

$$\hat{\Sigma} = \left[\sum_{t=1}^{2} X_t'\{\delta^2(R-\gamma W)^{-1} + D_t\}^{-1}X_t\right]^{-1}, \ \hat{\underset{\sim}{\beta}} = \hat{\Sigma}\left[\sum_{t=1}^{2} X_t'\{\delta^2(R-\gamma W)^{-1} + D_t\}^{-1}\hat{\underset{\sim}{\theta}}_t\right],$$

we have

$$\underset{\sim}{\beta} \mid \hat{\underset{\sim}{\theta}} \sim \text{Normal}(\hat{\underset{\sim}{\beta}}, \hat{\Sigma}). \tag{5}$$

Second, recall that originally we assume a hierarchical prior for $\underset{\sim}{\beta}$,

$$\underset{\sim}{\beta} \mid \tau \sim \text{Normal}(\underset{\sim}{\beta}, \tau\Sigma_o), \tag{6}$$

where $\underset{\sim}{\beta}$ and $\Sigma_o = 1000\Sigma_{\hat{\beta}}$ are specified as before. It is now easy to combine (5) and (6) to finally get

$$\underset{\sim}{\beta} \mid \tau, \hat{\underset{\sim}{\theta}} \sim \text{Normal}(\Lambda\hat{\underset{\sim}{\beta}} + (I-\Lambda)\underset{\sim}{\beta}_o, (I-\Lambda)\Sigma_o),$$

where $\Lambda = \{\hat{\Sigma}^{-1} + \Sigma_o^{-1}\}^{-1}\hat{\Sigma}^{-1} = \Sigma_o(\hat{\Sigma} + \Sigma_o)^{-1}$, and only one matrix inversion is needed.

Therefore, we can draw $(\underset{\sim}{\beta}, \underset{\sim}{\phi})$ from their joint CPD, and this is more efficient than drawing $\underset{\sim}{\beta}$ and $\underset{\sim}{\phi}$ from their respective CPDs. Once $\underset{\sim}{\phi}_2$ and $\underset{\sim}{\beta}$ are drawn, one can estimate $\underset{\sim}{\theta}_2 = X_2\underset{\sim}{\beta} + \underset{\sim}{\phi}_2$.

It is remarkable by simply drawing (δ^2, γ) simultaneously, we are able to improve the Gibbs sampler enormously.

We note here that the joint CPD of (δ^2, γ) comes from $\underset{\sim}{\phi}_t \mid \delta^2, \gamma \overset{ind}{\sim} \text{Normal}(\underset{\sim}{0}, \delta^2(R-\gamma W)^{-1}), t = 1, 2, \pi(\delta^2, \gamma) \propto 1/\delta^2, 0 < \gamma < 1, \delta^2 > 0$. Then, the joint CPD of (δ^2, γ) is

$$\pi(\delta^2, \gamma \mid \underset{\sim}{\phi}) \propto \frac{|R-\gamma W|}{(\delta^2)^{S+1}} e^{-\frac{1}{2\delta^2}\sum_{t=1}^{2}\underset{\sim}{\phi}_t'(R-\gamma W)\underset{\sim}{\phi}_t}, \delta^2 > 0, \ \lambda_1^{-1} < \gamma < \lambda_\ell^{-1}.$$

Therefore,

$$\delta^2 \mid \gamma, \underset{\sim}{\phi} \sim \text{InvGam}\{\frac{2S}{2}, \frac{\sum_{t=1}^{2}\underset{\sim}{\phi}_t'(R-\gamma W)\underset{\sim}{\phi}_t}{2}\},$$

and integrating out δ^2,

$$\pi(\gamma \mid \underset{\sim}{\phi}) \propto \frac{|R - \gamma W|}{\{\sum_{t=1}^{2} \underset{\sim}{\phi}_t'(R - \gamma W)\underset{\sim}{\phi}_t\}^S},\ \lambda_1^{-1} < \gamma < \lambda_\ell^{-1}.$$

Then, $\pi(\delta^2, \gamma \mid \underset{\sim}{\phi}) = \pi(\delta^2, \gamma \mid \underset{\sim}{\phi})\pi(\gamma \mid \underset{\sim}{\phi})$ (i.e., draw γ from $\pi(\gamma \mid \underset{\sim}{\phi})$ and δ^2 from $\pi(\delta^2, \gamma \mid \underset{\sim}{\phi})$).

4. Case Study

In this section, 2021 and 2022 cash rental rates data are selected as the case study. We fit the model in Section 3 for each land type (non-irrigated land, irrigated land and pasture land) for 49 states in the US.

The covariates $\underset{\sim}{x}_{ts}$ are the known auxiliary information used in the model and include an intercept, the corresponding previous year state-level official estimates, the number of positive responses, and the National Commodity Crop Productivity Indices (NCCPIs). NCCPIs, which measure the quality of the soil for growing non-irrigated crops in climate conditions best suited for various crops, are available at the county-level and state-level in the US. They are also correlated to the crop yield. Other covariates are the percentages of farmland by land types in the 2017 Census of Agriculture and the state-level population densities from Census 2020.

In Section 4.1, we discuss the model fit, Bayesian diagnostics, and the computation time related to the model. In Section 4.2, we show nationwide comparisons among model, survey, and published estimates.

4.1 Model Fit and Estimation

The spatial two-component mixture model introduced in Section 3 is a useful tool for producing model-based estimates of cash rental rates with measures of uncertainty. We fit the cash rental rate model for each land type (non-irrigated land, irrigated land and pasture land) for 49 states in the US.

Convergence diagnostics are conducted. The convergence for parameters involved is monitored using trace plots, the Geweke test of stationarity(Geweke (1992)) and the effective sample sizes. There is a single run of 25,000 iterates to fit the model and a "burn in" of 5,000 iterates. In order to eliminate the correlations among neighboring iterations, those iterations are thinned by taking a systematic sample of 1 in every 20 samples. Finally, 1,000 MCMC samples are obtained to construct the posterior distributions of parameters and make inferences for the current year cash rental rate estimates $\theta_{2s},\ s = 1, \ldots, 49$. Table 1 shows that the Gibbs sampler is mixing very well. It shows that the Geweke tests for all the parameters in the model

are not significant and the effective sample sizes are all near the actual sample size of 1,000 (mostly all of them are 1,000). Computation time is an additional factor when candidate models are evaluated for use in NASS production. In the case study, the computation time for models fitted for all states by three land types was less than 10 mins, which is acceptable for the production process.

Table 1: P-value (pval) of Geweke test and the effective sample size (ESS)

Parameter	Pval	ESS
β_0	0.86	1000.00
β_1	0.72	1000.00
β_2	0.68	1000.00
β_3	0.85	1000.00
β_4	0.90	1000.00
a	0.51	1000.00
ρ	0.93	1000.00
p_1	0.98	807.76
p_2	0.90	1000.00
δ^2	0.94	1000.00
γ	0.31	1000.00

To check the model fit, we have computed the Bayesian predictive p-value (BPP) for entire model and for the part of the model for the current year. Letting $a_1 = a$ and $a_2 = 1$, we note that

$$\hat{\theta}_{ts} \mid \Omega \overset{ind}{\sim} (1-p_s)N(\underset{\sim}{x}'_{t,s}\underset{\sim}{\beta} + \phi_{ts}, \rho\hat{\sigma}^2_{ts}/a_s) + p_s N(\underset{\sim}{x}'_{ts}\underset{\sim}{\beta} + \phi_{ts}, \hat{\sigma}^2_{ts}/a_s), t = 1, 2, s = 1, \ldots, S.$$

It is easy to show that

$$\mathrm{E}(\hat{\theta}_{ts} \mid \Omega) = \underset{\sim}{x}'_{t,s}\underset{\sim}{\beta} + \phi_{ts}, \;\; \mathrm{Var}(\hat{\theta}_{ts} \mid \Omega) = (1-p_s)\rho\hat{\sigma}^2_{ts}/a_s + p_s\hat{\sigma}^2_{ts}/a_s.$$

Using the discrepancy function,

$$T(\underset{\sim}{\hat{\theta}}; \underset{\sim}{\theta}) = \sum_{t=1}^{2}\sum_{s=1}^{S} \frac{\left\{\hat{\theta}_{ts} - \mathrm{E}(\hat{\theta}_{ts} \mid \Omega)\right\}^2}{\mathrm{Var}(\hat{\theta}_{ts} \mid \Omega)},$$

we can calculate the BPP,

$$\mathrm{BPP} = Pr\{T(\underset{\sim}{\hat{\theta}}^{(rep)}; \underset{\sim}{\theta}) \geq T(\underset{\sim}{\hat{\theta}}^{(obs)}; \underset{\sim}{\theta}) \mid \underset{\sim}{\hat{\theta}}\}.$$

For both years, the BPP is 0.837 and for the current year it is 0.872, showing that the model is reasonable.

In Table 2, we present posterior summaries of the hyperparameters with 95% highest posterior density intervals (HPDIs). Two of the regression parameters are significant; a 95% HPDIs for β_1 and β_4, respectively, are $(0.522, 2.635)$ and $(1.348, 2.031)$. In addition, the 95% HPDI for the spatial 'correlation' parameter γ is $(0.002, 0.854)$ and the posterior mean (PM) is 0.488 with a posterior standard deviation (PSD) of 0.242, indicating that the spatial correlation is significant. As for the outliers, the 95% HPDIs for p_1 and p_2 are, respectively, $(0.080, 0.500)$ and $(0.150, 0.500)$ and the PMs (PSDs) are respectively 0.293 (0.123) and 0.354 (0.104). It is useful that the model is able to detect states that are outliers.

Table 2: Posterior summaries for hyper-parameters

Parameter	PM	PSD	95% HPDI
β_0	-5.031	14.669	(-34.574, 23.396)
β_1	1.469	0.563	(0.522, 2.635)
β_2	-0.013	0.013	(-0.038, 0.010)
β_3	5.511	3.656	(-0.368, 13.479)
β_4	1.726	0.179	(1.348, 2.031)
a	0.546	0.271	(0.099, 0.99)
ρ	0.564	0.276	(0.095, 0.998)
p_1	0.293	0.123	(0.080, 0.500)
p_2	0.354	0.104	(0.150, 0.499)
δ^2	6556.999	2260.816	(2956.877, 11510.399)
γ	0.488	0.242	(0.002, 0.854)

4.2 Numerical Summaries and Comparisons

The model of Section 3 was fit to the cash rental rates reported on the 2021 and 2022 Cash Rent Surveys (CRSs) at the state level for 49 states in the US. Survey estimates, model posterior means, and previous published estimates for pasture land, respectively, are displayed in Figure 2 from left to right. The colors indicate the magnitude of the point estimates ($/Acres). The darker the color, the higher the value is. It is straightforward to see that the patterns of the map based on survey estimates are different from those based on the model and previously published estimates. This is due to the spike in states with few number of reports. However, the model estimates are adjusted when integrating with other data through the models. In addition, the spatial patterns can be seen on the maps as well. The states with higher values are concentrated in the middle of the US and spread to the east. Based on visual inspection of Figure 2, the model adjusted

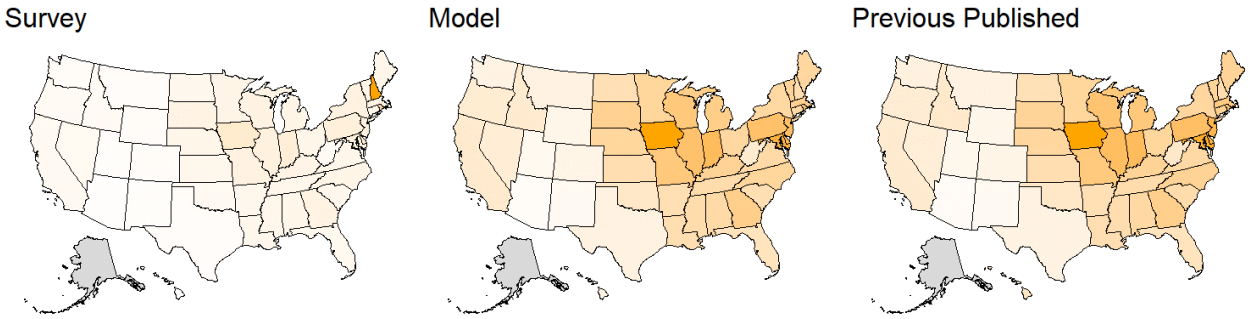

Figure 2: The plots of state-level maps based on survey direct estimates, model point estimates, and previous year published estimates for pasture land.

outliers are also maps also indicated that the model also adjusted outliers.

One way of illustrating the differences between the model-based estimates and the survey's direct estimates is to use the ratios of the model-based estimate to the corresponding survey's direct estimate as a function of the sample size. The ratios can provide further insights into the differences in the two methods of estimation (see Figure 3 for plots of the ratios of model estimates to survey estimates for the estimated cash rental rates). The widest range of ratios between the model estimates and the survey's direct estimates was for states with small sample sizes, and the ratios tended to become closer to one as the sample size increased for all three estimates. This illustrates the shrinkage of the direct estimates toward the modeled (regression) estimates obtained by using all available sources of information.

To demonstrate the gain in reliability of estimates based on the model relative to the survey, we compare the posterior coefficients of variation (CVs) from the model to the CVs from the survey. The CV comparisons between model and survey in all three different land types are displayed in Figure 4. The posterior CVs have a greater reduction when compared to the CVs of the survey estimates. The results demonstrate the tendency of the small area model to improve the reliability of estimates when compared to the reliability of survey estimates. The overall average reductions in the CVs are approximately 12%, 15%, and 18%, respectively, for non-irrigated, irrigated, and pasture land.

5. Conclusion

We have proposed a spatial two-component mixture model that accounts directly for outliers, and it uses the power prior to input how much of the past data one wants to use. Instead of using all the past data, as is

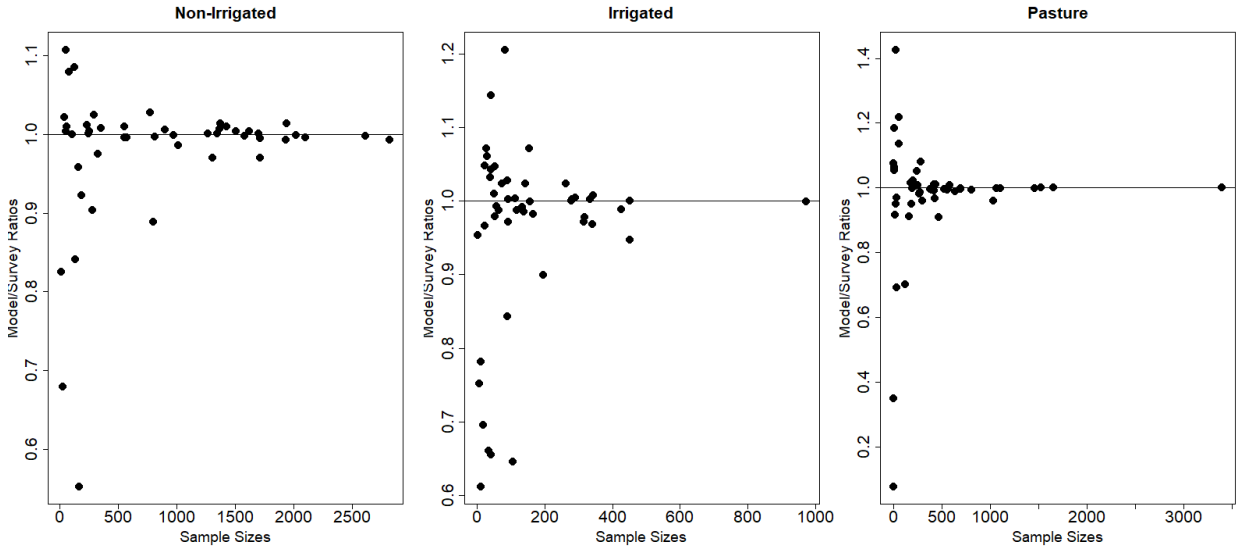

Figure 3: Relative measures of model estimates versus survey's direct estimates of cash rental rates in all three different land types.

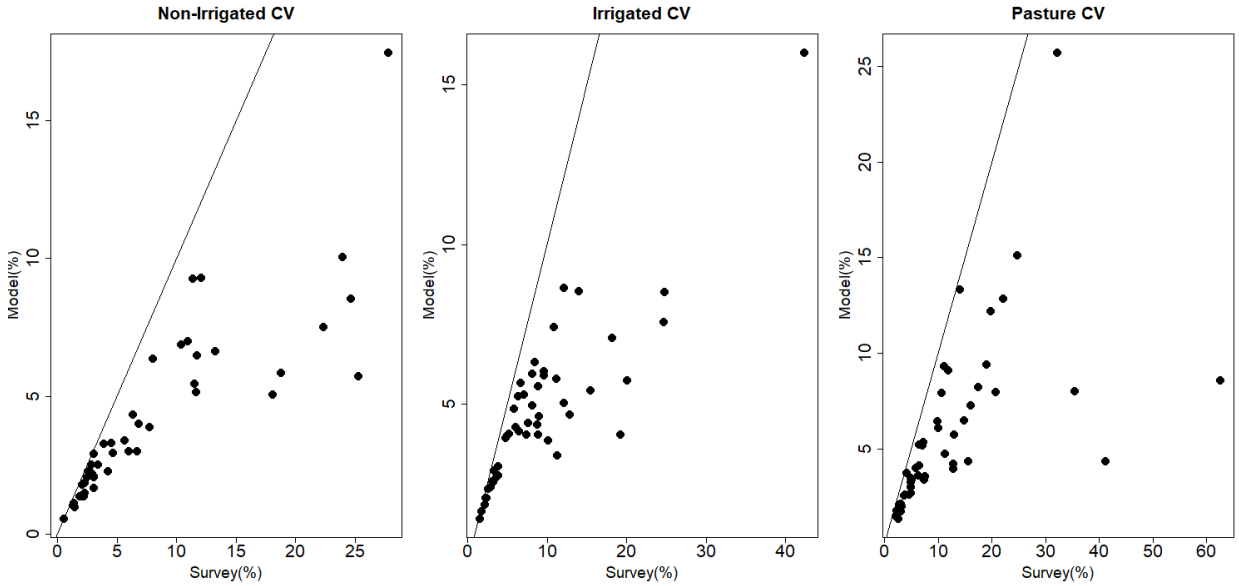

Figure 4: Coefficients of variation (%) for state-level model and survey estimates of cash rental rates in all three different land types.

suggested by the two published papers (Berg et al., 2014; Erciulescu et al., 2019) on this topic, we model the percentage of the usage of the past data similar to the county-level model proposed by Chen and Nandram (2022). One contribution discussed in this paper is that the state-level model takes the neighborhood structure into consideration based on the CAR model prior. The spatial model is preferable since it accommodates one additional physical feature (spatial) of the data. The area-level model provides anther way to operationalize the unit-level SAE model (Erciulescu et al., 2019). Therefore, the new proposed area-level model is practical in terms of computation time when compared with the unit-level model, which is a key factor for the evaluation of a model to be used in production.

In the case study using data from 2021 and 2022 CRSs, we compare between model-based estimates and survey estimates. First, the model diagnostics are good in different aspects, showing good mixing and a reasonable model. The maps illustrate the spatial pattern and the difference between survey estimates and the model estimates. The model could adjust the outliers. Moreover, the associated measures of uncertainty (CVs) from models are generally smaller than the CVs of the survey estimates. The model can reduce the CVs while borrowing strength from auxiliary information and all counties within one region. Therefore, the performance of the model with a respect to accuracy, precision, and reliability, illustrates significant improvement of the state-level estimates of cash rental rates for all three land types.

Ongoing and future research related to the model involve the investigation of different auxiliary information. The auxiliary information considered here is the nationwide data sources available at the state level. Future efforts will be on investigating and applying other useful data sources to strengthen the model. Tax data, farm intensity data and other census data are available at the county level in specific states.

In addition, for the spatial model, it is possible to move into the more desirable nonparametric approach via the stick-breaking prior with Gaussian process. Nearest neighbor type spatial models may over shrink, creating 'a false sense of security' with artificially small CVs.

On the other hand, the future research is to combine both state-level models and county-level models. Since the state estimates of cash rental rates are determined prior to setting county estimates, NASS employs a "top-down" strategy, first setting state-level estimates and then setting corresponding county-level estimates. For the county-level estimates, the CAR models discussed in the paper are not quite appropriate. States with the highest production (speculative states) should be modeled separately from those with lower production (non-speculative states). However, for coherence, the two models should be connected into a single one. Geographically closed states may not be so with respect to the study variable. For example, a

speculative state may be a geographical neighbor of a non-speculative one, but they may be very different in terms of the study variable.

References

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Berg, E., Cecere, W., and Ghosh, M. (2014). Small Area Estimation for County-Level Farmland Cash Rental Rates. *Journal of Survey Statistics and Methodology*, 2(1):1–37.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Chakraborty, A., Datta, G. S., and Mandal, A. (2019). Robust hierarchical bayes small area estimation for the nested error linear regression model. *International Statistical Review*, 87(S1):S158–S176.

Chen, L., Cruze, N. B., and Young, L. J. (2022a). Model-based estimates for farm labor quantities. *Stats*, 5(3):738–754.

Chen, L. and Nandram, B. (2022). Combining survey and administrative data to produce official statistics. *In JSM Proceedings, Survey Research Methods Section*, pages 1823–1839.

Chen, L., Nandram, B., and Cruze, N. B. (2022b). Hierarchical bayesian model with inequality constraints for us county estimates. *Journal of Official Statistics*, 38(3):709–732.

Chen, M.-H. and Ibrahim, J. G. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46 – 60.

Erciulescu, A., Berg, E., Cecere, W., and Ghosh, M. (2019). Bivariate hierarchical bayesian model for estimating cropland cash rental rates at the county level. *Survey methodology*, 45(2):199–216.

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(366a):269–277.

Gershunskaya, J. and Lahiri, P. (2017). Robust empirical best small area finite population mean estimation using a mixture model. *Calcutta Statistical Association Bulletin*, 69(2):183–204.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *IN BAJ. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, Eds., Bayesian Statistics*, 4:169–193.

Goyal, S., Datta, G. S., and Mandal, A. (2021). A hierarchical bayes unit-level small area estimation model for normal mixture populations. *Sankhya B*, 83(1):215–241.

Hodges, J. S., Carlin, B. P., and Fan, Q. (2003). On the precision of the conditionally autoregressive prior in spatial models. *Biometrics*, 59(2):317–322.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). The power prior: theory and applications. *Statistics in medicine*, 34(28):3724–3749.

Moran, P. A. (1950). Notes on continuous stochastic phenomena. *Biometrika*, 37(1/2):17–23.

Nandram, B., Erciulescu, A., Cruze, N. B., and Chen, L. (2022). Hierarchical bayesian model with inequality constraints for us county estimates. *Research Report RDD-22-02, National Agricultural Statistics Service, USDA*.

National Academies of Sciences, Engineering, and Medicine (2018). *Improving Crop Estimates by Integrating Multiple Data Sources*. The National Academies Press.

Sun, D., Tsutakawa, R. K., and Speckman, P. L. (1999). Posterior distribution of hierarchical models using car (1) distributions. *Biometrika*, 86(2):341–350.

Young, L. J. and Chen, L. (2022). Using small area estimation to produce official statistics. *Stats*, 5(3):881–897.