A Spatial Autoregressive Dirichlet Process Mixture Model for Crop Yield Distribution[*]

Xiaotian Liu[†] Yong Bao[‡] Xuewen Yu[§]

February 16, 2025

Abstract

The distribution of crop yields plays an important role in agricultural economics. This paper proposes a spatial autoregressive Dirichlet process mixture model for crop yield distribution, which accommodates both cross-sectional interactions and latent group structures. Using a Bayesian approach, we jointly estimate the number of latent groups and model parameters, including heterogeneous spatial dependence parameters. Monte Carlo simulations, based on characteristics of county-level crop yield data in the U.S., demonstrate that our model has good finite-sample performance. We apply the proposed model in two empirical studies. The first one assesses the potential loss of global corn production across five major corn-producing countries. The results suggest that the potential loss varies across the five countries and risk pooling can enhance global food security. The second application compares the proposed model with the USDA's current rating method for area-yield crop insurance contracts and finds that the proposed model may lead to more accurate premium rates.

Keywords: spatial autoregressive; Dirichlet process; Bayesian approach; food security; crop insurance

JEL classification: C21, C51, Q18

[*] The authors thank the seminar and conference participants at Huazhong Agricultural University, Hunan University, and the 2024 Annual Conference of CAASS for their valuable comments.

[†] Corresponding author. Department of Economics, Huazhong Agricultural University, Wuhan, Hubei 430070, China. E-mail: xiaotianliu@mail.hzau.edu.cn.

[‡] Department of Economics, Purdue University, 403 Mitch Daniels Boulevard, West Lafayette, IN 47907, USA. E-mail: ybao@purdue.edu.

[§] Department of Applied Economics, Fudan University, Shanghai, Shanghai 200433, China. E-mail: xuewenyu@fudan.edu.cn.

1 Introduction

In agricultural economics, the distribution of crop yields has significant implications. For example, to determine the premium of a crop insurance policy, we need to use the crop yield distribution to estimate both the probability of future yield falling below a specified threshold and the conditional mean of future yield given this shortfall (Skees et al., 1997; Goodwin and Ker, 1998; Ker and Coble, 2003; Ozaki et al., 2008; Harri et al., 2011; Woodard and Sherrick, 2011; Ker et al., 2016; Yvette Zhang, 2017; Park et al., 2019; Yi et al., 2020; Liu and Ramsey, 2023). To assess the risk of food security, we can employ the distribution of crop yields to analyze the risks of crop failure among different crop-producing regions (Chavas et al., 2022; Caparas et al., 2021; Wang et al., 2024). Other applications include evaluating technological changes in crop yields (Tolhurst and Ker, 2015), analyzing farmers' insurance decisions under uncertainty (Ceballos and Robles, 2020), and assessing the impact of climate change on agricultural output (Barnwal and Kotani, 2013).

Likewise, the distributions of many economic variables play vital roles in different areas of economics. For example, asset return distribution can be used to calculate the so-called Value-at-Risk and expected shortfall of a financial position (Billio and Pelizzon, 2000; Kuester et al., 2005; Lazar et al., 2024); different productivity distributions may be associated with distinct growth paths for different economies (Benhabib et al., 2021); the distribution of pay rates has different impact on output quantity and quality (Cardella and Roomets, 2022); firm size distribution can affect economic growth and productivity (Garicano et al., 2016; Silveira, 2022).

A recurring theme in these different strands of literature is that the baseline Gaussian distribution is not appropriate in modeling a wide class of economic variables. A popular approach to modeling nonnormal distributions is to use finite mixture models, which can be flexible enough to accommodate the (non-zero) skewness and (excess) kurtosis as well as other features (e.g., multi-mode) that are not shared by the Gaussian distribution and other commonly used parametric distributions. For example, in the context of crop yield, Woodard and Sherrick (2011) introduce a method for estimating mixture models using cross-validation optimization to accurately model crop yield distribution; Tolhurst and Ker (2015) propose embedding trend functions within mixture models, allowing for different rates of technological change across components of the yield distribution; Ker et al. (2016) find that using Bayesian model averaging can further improve the efficiency of mixture models when estimating possibly similar yield densities; Chemeris et al. (2022) use mixture models to study the nexus of insurance subsidies, changing climate, and genetically modified seeds on measure of both technological change and yield resiliency; Schuurman and Ker (2024) propose integrating a neural network within the expectation maximization (EM) algorithm to estimate the mixing probabilities of mixture model and find that rising temperatures affect multiple higher moments of the crop yield distribution. In the context of housing market, Belasco et al. (2012) use a finite mixture model to identify latent submarkets from household demographics and estimate a separate hedonic regression equation for each submarket. In urban studies, Su (2020) uses an urban growth theory to explain how heterogeneous growth factors form a mixture that shapes the aggregate city size distribution. Kondo et al. (2023) find that a mixture distribution provides a better fit for describing the U.S. firm and establishment size than the frequently used Pareto and lognormal distributions. Durham (2007) proposes using a single-factor model of stochastic volatility in conjunction with a mixture distribution to study the conditional distribution of returns. Applications of mixture models can naturally arise to account for unobserved heterogeneity, random utility, multiple equilibria, measurement error, regime switching, and so on, see Compiani and Kitamura (2016) for an excellent survey.

In addition to the distributional features that deviate substantially from a baseline normal distribution, cross-sectional correlation has also been increasingly catching the attention of researchers. Crop yields are spatially correlated due to various factors such as similar geographic and meteorological conditions in neighboring areas and technology spillover effects (Anselin, 2001; Long, 1998; Ker et al., 2016; Yvette Zhang, 2017; Park et al., 2019; Bao et al., 2024). The phenomena of financial contagion, where a shock in one market can be transmitted to others, have been well documented in the literature (Aït-Sahalia et al., 2015; Miled et al., 2022), see also the somewhat related sovereign risk contagion (Arellano et al., 2017; Telila, 2023). Countries interact strategically each other in their spending on research and development (R&D) and Hammadou et al. (2014) find that closeness in terms of sectoral specialisation affects countries' strategic interactions on public R&D spending. House price shocks can propagate from one local market to another and thus cause spatial spillovers across metropolitan housing markets (DeFusco et al., 2018).

This paper proposes a spatial autoregressive Dirichlet process mixture (SAR-DPM) model that is more general and flexible than the finite mixture model and at the same time accommodates cross-sectional correlation via a spatial autoregressive structure. The Dirichlet process mixture (DPM) model can be interpreted as a mixture model with a countably infinite number of components or groups. Similar to finite mixture models, it is flexible enough to account for nonnormal features. However, it is different from finite mixture models that use a fixed number of groups to model the data. The actual number of groups used to model the data by DPM is not fixed and can be automatically inferred from the data based on the usual Bayesian posterior inference framework. As such, DPM is capable of capturing a very rich spectrum of latent group structure. Finite mixture models usually rely on model selection methods to decide the number of groups and thus provide only a single point estimate of the group count, see Woodard and Sherrick (2011), Tolhurst and Ker (2015), Ker et al. (2016), Chemeris et al. (2022), and Schuurman and Ker (2024) on the study of crop yield distribution. The DPM model allows us to directly estimate the group count for reach region and also provide its posterior distribution. The spatial autoregressive (SAR) structure is used to capture the cross-sectional correlation in the dependent variable. It is well known that ignoring spatial correlation can lead to biased parameter estimates and misleading inference (Anselin, 2001). The SAR structure allows us to directly quantify the degree of spatial dependence in the dependent variable, which on many occasions is crucial for public policy design. For example, to what extent crop yields from different regions are correlated with each other is critical for public policies involving food security and subsidies for agricultural insurance. Our proposed SAR-DPM model aims to take into account both group structure heterogeneity and cross-sectional correlation.

The plan of this paper is as follows. In the next section, we introduce the model specification and develop the Bayesian posterior sampler for estimating and inference. Section 3 reports our Monte Carlo simulation results, demonstrating the satisfactory finite-sample performance of the Bayesian estimation method through simulations that reflect the characteristics of county-level crop yield data in the U.S. Section 4 provides two empirical applications of the SAR-DPM model. The first application assesses the potential loss of corn failure at different risk levels in five global major corn-producing countries. The findings suggest that the potential loss varies across the five countries and risk pooling can enhance global food security. In the second application, we compare the rating approach based on the SAR-DPM model with the current rating methodology by the United States Department of Agriculture (USDA) for area-yield crop insurance contracts. We find that SAR-DPM model may lead to more accurate rates. Finally, Section 5 concludes the paper.

Throughout, we adopt the following set of notation: $\mathrm{Dg}(\cdot)$ denotes an operator that creates a diagonal matrix by diagonally stacking its scalar arguments in order; $\mathbb{1}(\cdot)$ is an indicator function that takes the value of 1 when its argument (of a statement) is true and zero otherwise; $\mathcal{B}(\alpha, \beta)$ is a beta distribution with the first and second shape parameters α and β; $\mathcal{G}(a, b)$ is a gamma distribution with shape parameter a and rate parameter b; $\mathcal{IG}(a, b)$ is an inverse gamma distribution with shape parameter a and scale parameter b; $\mathcal{U}(a, b)$ is a uniform distribution on the interval (a, b).

2 Model Specification and Bayesian Estimation

The SAR-DPM model is specified as follows:

$$y_{it} = \lambda_i \sum_{j=1}^{n} w_{ij} y_{jt} + \boldsymbol{x}_{it}' \boldsymbol{\beta}_{i,g_{it}} + u_{it}, \tag{1}$$

where y_{it} is the observation on the outcome of unit i at time t and u_{it} is the disturbance term, $i = 1, \cdots, N$, $t = 1, 2, \cdots, T$. The term $\sum_{j=1}^{n} w_{ij} y_{jt} \equiv y_{it}^*$ is the weighted average of the outcomes of units i's neighbors, where w_{ij}'s are spatial weights between units i and j, typically based on geographical proximity. In a broad sense, we use the term "spatial" loosely, recognizing that connectivity between units is not always restricted to the spatial dimension. The parameter λ_i is the spatial autoregressive coefficient for unit i. It measures the influence of the weighted average of the outcomes from neighboring units on the outcome of unit i. It captures the degree of spatial correlation and is allowed to vary across individual units. In other words, it allows for the situation where a unit's outcome may be barely related to those of other neighboring units, whereas another unit's outcome may be easily influenced by its neighbors.

The set of regressors denoted by the $k \times 1$ vector $\boldsymbol{x}_{it}$ contains a constant term and other covariates (e.g., weather conditions and soil quality when y_{it} is crop yield). It may also include a time trend t, indicating that, for instance, crop yield increases over years due to technological advancements (e.g., improvement of seed quality). [1] The $g_{it} \in \{1, 2, \cdots, G_i\}$ in equation (1) is an index that signals the latent group. Note that it has double subscripts i and t and the group count G_i can be infinite. Consider the case of crop yield, where $\boldsymbol{x}_{it}$ includes a constant term and a time trend. If nature endows region i with the same growing conditions in years t_1 and t_2, then the parameters $\beta_{i,g_{it_1}}$ and $\beta_{i,g_{it_2}}$ are the same; otherwise,

[1] For crop yield, when the time trend is included, y_{it} reflects the actual yield. Otherwise, y_{it} represents the detrended yield, which may also be adjusted for potential heteroskedasticity. In the empirical applications discussed in Section 4, we employ a two-knot linear spline method to adjust for the trend and then correct for heteroskedasticity, following the methodology of Harri et al. (2011). This is also the approach used by the Risk Management Agency (RMA) of the USDA.

they are different. This is in line with the scenario under which growing conditions can fluctuate from year to year.

The idiosyncratic error term's variance, denoted by $\sigma^2_{i,g_{it}}$, is also allowed to follow a latent group structure. For instance, companies have different degrees of exposure to market shocks over time. Since the responsiveness of the dependent variable to different covariates and the idiosyncratic error may follow different group structures, we may interpret G_i as the sum of all possible group counts for unit i.

In matrix form, equation (1) can be rewritten as

$$\boldsymbol{y}_t = \boldsymbol{\Lambda}\boldsymbol{W}\boldsymbol{y}_t + \boldsymbol{X}_t\boldsymbol{\beta}_t + \boldsymbol{u}_t, \tag{2}$$

where $\boldsymbol{y}_t = (y_{1t}, y_{2t}, \cdots, y_{Nt})'$, $\boldsymbol{u}_t = (u_{1t}, u_{2t}, \cdots, u_{Nt})'$ is normally distributed with a diagonal variance matrix, $\boldsymbol{W}$ is an $N \times N$ weight matrix, consisting of w_{ij} in its (i, j)-th position, $\boldsymbol{\Lambda} = \mathrm{Dg}(\lambda_1, \lambda_2, \cdots, \lambda_N)$ diagonally collects all the spatial coefficients, $\boldsymbol{X}_t$ is an $N \times kN$ block diagonal matrix, with $\boldsymbol{x}'_{it}$ as its i-th diagonal block, and $\boldsymbol{\beta}_t = (\boldsymbol{\beta}'_{1,g_{1t}}, \cdots, \boldsymbol{\beta}'_{N,g_{Nt}})'$. Note that here the normal assumption on $\boldsymbol{u}_t$ does not imply that the actual data is normally distributed. In fact, in the presence of the latent group structure, the unconditional distribution is nonnormal. The parameters of our main interest can be collected as $\boldsymbol{\theta} = (\lambda_1, \lambda_2, \cdots, \lambda_N, \boldsymbol{\beta}'_1, \cdots, \boldsymbol{\beta}'_T, \boldsymbol{\sigma}^{2\prime}_1, \cdots, \boldsymbol{\sigma}^{2\prime}_T)'$, where $\boldsymbol{\sigma}^2_t = (\sigma^2_{1,g_{1t}}, \sigma^2_{2,g_{2t}}, \cdots, \sigma^2_{N,g_{Nt}})'$.

Let $\boldsymbol{S} = \boldsymbol{S}(\lambda_1, \lambda_2, \cdots, \lambda_N) = \boldsymbol{I}_N - \boldsymbol{\Lambda}\boldsymbol{W}$, consisting of rows $\boldsymbol{S}_{i\circ}$, $i = 1, \cdots, N$. Then at each time t, $\boldsymbol{y}_t = \boldsymbol{S}^{-1}\boldsymbol{X}_t\boldsymbol{\beta}_t + \boldsymbol{S}^{-1}\boldsymbol{u}_t$, where $\boldsymbol{S}^{-1}\boldsymbol{u}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{S}^{-1}\mathrm{Var}(\boldsymbol{u}_t)\boldsymbol{S}^{-1\prime})$ and $\mathrm{Var}(\boldsymbol{u}_t) = \mathrm{Dg}(\sigma^2_{1,g_{1t}}, \cdots, \sigma^2_{N,g_{Nt}})$. It follows that the density function of $\boldsymbol{y}_t$, given the complete set of model parameters $\boldsymbol{\theta}$ and $\boldsymbol{g}_t = (g_{1t}, \cdots, g_{Nt})'$, is

$$P(\boldsymbol{y}_t|\boldsymbol{\theta}, \boldsymbol{g}_t) = \exp\left[-\frac{N}{2}\ln(2\pi) + \ln|\boldsymbol{S}| - \frac{1}{2}\sum_{i=1}^{N}\left(\ln\sigma^2_{i,g_{it}} + \frac{\xi^2_{it}}{\sigma^2_{i,g_{it}}}\right)\right], \tag{3}$$

where $\xi_{it} = y_{it} - \lambda_i y^*_{it} - \boldsymbol{x}'_{it}\boldsymbol{\beta}_{i,g_{it}}$.

Given that we do not limit or fix the number of possible groups, we use the Dirichlet process (DP) prior to model the latent group structure. On the other hand, the DP prior have countably infinite discrete masses, which makes estimation of equation (1) challenging (Neal, 2000). To overcome this technical difficulty, we follow Walker (2007) to use the slice-sampling approach. Specifically, the probability of a region i at time t belonging to group g_{it} is determined by a latent variable η_{it} relative to a threshold $\omega_{i,g_{it}} = z_{i,g_{it}} \prod_{l=1}^{g_{it}-1}(1 - z_{i,l})$, where $z_{i,l}$, $l = 1, \cdots, g_{it}$, determine the threshold value, such that it falls into group g_{it} only when $\eta_{it} < \omega_{i,g_{it}}$. We treat all the relevant variables, including g_{it}, η_{it}, $z_{i,g_{it}}$, as well as h_i (used for the prior of $z_{i,g_{it}}$) and τ_i (a latent variable for h_i), which are used in determining the latent group structure as hyper-parameters to be estimated. They are all collected in $\boldsymbol{\delta}$. In the next three subsections, we discuss the Markov chain Monte Carlo (MCMC) cycle that we use for deriving the Bayesian posterior sampler.

2.1 Conditional Posterior Distributions of $\boldsymbol{\beta}_{i,g}$ and $\sigma_{i,g}^2$

This subsection outlines the conditional posterior distributions of the coefficients $\boldsymbol{\beta}_{i,g} = (\beta_{i1,g}, \cdots, \beta_{ik,g})'$ and variance parameter $\sigma_{i,g}^2$. For ease of presentation, we have dropped the subscript in g_{it}. In other words, we present the conditional distribution of each element of the i-th unit's exogenous covariate parameter vector and error variance, given all the other parameters and the observable data $\boldsymbol{y} = (\boldsymbol{y}_1', \cdots, \boldsymbol{y}_T')'$, that come from from a given group g at time t.

We assume a conjugate normal prior for $\boldsymbol{\beta}_{i,g}$, namely, $\boldsymbol{\beta}_{i,g} \sim \mathcal{N}(\bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_{i,g}}, \bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i})$. Suppose we use $\boldsymbol{\theta}_{-\boldsymbol{\beta}_{i,g}}$ to denote the parameter vector with $\boldsymbol{\beta}_{i,g}$ excluded, and similar notation is used for the other conditional distributions to be presented. Given $\boldsymbol{\theta}_{-\boldsymbol{\beta}_{i,g}}$, the conditional posterior distribution of $\boldsymbol{\beta}_{i,g}$ is also normal, with updated mean vector $\underline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_{i,g}}$ and covariance matrix $\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}$, namely,

$$(\boldsymbol{\beta}_{i,g} | \boldsymbol{y}, \boldsymbol{\theta}_{-\boldsymbol{\beta}_{i,g}}, \boldsymbol{\delta}) \sim \mathcal{N}(\underline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_{i,g}}, \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_{i,g}}), \tag{4}$$

where

$$\underline{\boldsymbol{\mu}}_{\boldsymbol{\beta}_{i,g}} = \underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_{i,g}} \left(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{-1} \bar{\boldsymbol{\mu}}_{\boldsymbol{\beta}_{i,g}} + \sum_{t=1}^{T} \frac{\boldsymbol{S}_{i\circ} \boldsymbol{y}_t \boldsymbol{x}_{it} \mathbb{1}(g_{it} = g)}{\sigma_{i,g_{it}}^2} \right),$$

$$\underline{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_{i,g}} = \left(\bar{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}_i}^{-1} + \sum_{t=1}^{T} \frac{\boldsymbol{x}_{it}\boldsymbol{x}_{it}' \mathbb{1}(g_{it}=g)}{\sigma^2_{i,g_{it}}}\right)^{-1}.$$

We use a conjugate inverse gamma prior for $\sigma^2_{i,g}$, with shape and rate hyper-parameters $\bar{a}$ and $\bar{b}$, respectively. Given $\boldsymbol{\theta}_{-\sigma^2_{i,g}}$, the conditional posterior distribution of $\sigma^2_{i,g}$ is also an inverse gamma distribution,

$$(\sigma^2_{i,g}|\boldsymbol{y}, \boldsymbol{\theta}_{-\sigma^2_{i,g}}, \boldsymbol{\delta}) \sim \mathcal{IG}\left(\bar{a} + \frac{1}{2}\sum_{t=1}^{T}\mathbb{1}(g_{it}=g),\ \bar{b} + \frac{1}{2}\sum_{t=1}^{T}\xi^2_{it}\mathbb{1}(g_{it}=g)\right). \tag{5}$$

2.2 Conditional Posterior Distribution of λ_i

For the prior of λ_i, we use $\mathcal{U}(-1,1)$. The conditional posterior distribution of λ_i, given $\boldsymbol{\theta}_{-\lambda_i}$, is

$$P(\lambda_i|\boldsymbol{y}, \boldsymbol{\theta}_{-\lambda_i}, \boldsymbol{\delta}) \propto \exp\left[T\ln|\boldsymbol{S}| - \frac{1}{2}\sum_{t=1}^{T}\frac{\xi^2_{it}}{\sigma^2_{i,g_{it}}}\right]. \tag{6}$$

This distribution is nonstandard and we adopt the Metropolis-Hastings sampling procedure following the approach suggested in LeSage and Pace (2009). Specifically, for the parameter λ_i, we draw a proposal candidate λ_i^c as $\lambda_i^c = \lambda_i + c_{i,\lambda}\epsilon$, where $c_{i,\lambda}$ is a tuning parameter and ϵ is a standard normal variable. We restrict the proposal candidates to lie within the admissible parameter range of $(-1,1)$. Subsequently, we evaluate the conditional distribution in equation (6) at $\{\lambda_i, \lambda_i^c\}$ and calculate an acceptance probability to decide whether to accept the new candidate λ_i^c or retain the current λ_i. We adapt the tuning parameter $c_{i,\lambda}$ to ensure an acceptance probability between 40% and 60%.

2.3 Conditional Posterior Distributions of Hyper-Parameters

To complete our MCMC cycle, we need to know all the g_{it}'s, though they are not the parameters of our primary interest. Recall that we use a latent variable η_{it} relative to a threshold $\omega_{i,g_{it}} = z_{i,g_{it}}\prod_{l=1}^{g_{it}-1}(1-z_{i,l})$ to determine the probability of a unit i at time t belonging to group g_{it}. We treat all the g_{it}'s as unknowns and use data to estimate, for each

i and t, the position of each unit in the relevant group structure. Define $g_i^\star$ as the smallest integer g such that $\sum_{j=1}^{g} \omega_{i,j}$ is greater than $1 - \eta_i^\star$, where $\eta_i^\star = \min\{\eta_{i,1}, \eta_{i,2}, \cdots, \eta_{i,T}\}$. Walker (2007) proves that $g_i^\star$ is the maximum number of groups we should consider when updating the model parameters for the i-th unit in the MCMC cycle.

We assume a beta distribution prior $\mathcal{B}(1, h_i)$ for $z_{i,g_{it}}$. The corresponding conditional posterior distribution of $z_{i,g_{it}}$ is

$$(z_{i,g_{it}} | \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\delta}_{-z_{i,g_{it}}}) \sim \mathcal{B}\left(1 + \sum_{s=1}^{T} \mathbb{1}(g_{is} = g_{it}),\ h_i + \sum_{s=1}^{T} \mathbb{1}(g_{is} > g_{it})\right), \tag{7}$$

where $g_{it} = 1, 2, \cdots, g_i^\star$. This implies that if a group is empty, $z_{i,g_{it}}$ is sampled from the prior distribution. Moreover, if $g_{it} > g_i^\star$, updating $z_{i,g_{it}}$ is unnecessary since the extended likelihood function will always be 0.

With a $\mathcal{U}(0,1)$ prior for η_{it}, given $\omega_{i,g_{it}} = z_{i,g_{it}} \prod_{l=1}^{g_{it}-1}(1 - z_{i,l})$, the conditional posterior distribution of η_{it} is uniform on the interval $(0, \omega_{i,g_{it}})$:

$$(\eta_{it} | \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\delta}_{-\eta_{it}}) \sim \mathcal{U}(0, \omega_{i,g_{it}}). \tag{8}$$

Once we have sampled g_{it} and η_{it}, $t = 1, \cdots, T$, we can determine $g_i^\star$. Given $g_i^\star$, the conditional posterior distribution of g_{it} is $P(g_{it} = g | \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\delta}_{-g_{it}}) \propto P(y_{it} | \boldsymbol{\theta}, \boldsymbol{g}_{g_{it}=g}) \mathbb{1}(\eta_{it} < \omega_{i,g})$ for $g \leq g_i^\star$, and $P(g_{it} = g | \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\delta}_{-g_{it}}) = 0$, for $g > g_i^\star$, where $\boldsymbol{g}_{g_{it}=g}$ is $\boldsymbol{g}_t$ with g_{it} replaced by g. Note that g_{it} does not appear in $\ln |\boldsymbol{S}|$ that is contained in $P(\boldsymbol{y}_t | \boldsymbol{\theta}, \boldsymbol{g}_t)$ (see equation (3)), so $P(y_{it} | \boldsymbol{\theta}, \boldsymbol{g}_{g_{it}=g}) \propto \exp[-(\ln \sigma_{i,g}^2 + \xi_{it}^2 / \sigma_{i,g}^2)/2]$, where now $\xi_{it} = y_{it} - \lambda_i y_{it}^* - \boldsymbol{x}_{it}' \boldsymbol{\beta}_{i,g}$. Therefore, we can compute the conditional probability of $g_{it} = g$ by self-normalization using the following equation:

$$P(g_{it} = g | \boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\delta}_{-g_{it}}) = \frac{P(y_{it} | \boldsymbol{\theta}, \boldsymbol{g}_{g_{it}=g}) \mathbb{1}(\eta_{it} < \omega_{i,g})}{\sum_{l=1}^{g_i^\star} P(y_{it} | \boldsymbol{\theta}, \boldsymbol{g}_{g_{it}=l}) \mathbb{1}(\eta_{it} < \omega_{i,l})}, \quad g \leq g_i^\star. \tag{9}$$

The number of groups G_i for each unit i is available once we have sampled g_{it}, $t = 1, \cdots, T$. Under the DP prior, G_i can be sensitive to h_i, and therefore we follow Escobar and West (1995) to treat h_i as an unknown parameter. With a gamma prior with shape $\bar{a}_{h_i}$

and rate $\bar{b}_{h_i}$ for h_i, Escobar and West (1995) show that, given G_i, the conditional posterior distribution of h_i is the marginal distribution from a joint for h_i and a continuous quantity τ_i on the unit interval. Accordingly, Escobar and West (1995) demonstrate that, given h_i, we can sample τ_i from a beta distribution:

$$(\tau_i|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\delta}_{-\tau_i}) \sim \mathcal{B}(h_i + 1, T), \tag{10}$$

With G_i and τ_i available, we can then sample h_i from a mixed gamma distribution:

$$\begin{aligned}(h_i|\boldsymbol{y}, \boldsymbol{\theta}, \boldsymbol{\delta}_{-h_i}) \sim & \frac{\bar{a}_{h_i} + G_i - 1}{\bar{a}_{h_i} + G_i - 1 + T(\bar{b}_{h_i} - \ln(\tau_i))} \mathcal{G}\left(\bar{a}_{h_i} + G_i,\ \bar{b}_{h_i} - \ln(\tau_i)\right) \\ & + \frac{T(\bar{b}_{h_i} - \ln(\tau_i))}{\bar{a}_{h_i} + G_i - 1 + T(\bar{b}_{h_i} - \ln(\tau_i))} \mathcal{G}\left(\bar{a}_{h_i} + G_i - 1,\ \bar{b}_{h_i} - \ln(\tau_i)\right).\end{aligned} \tag{11}$$

3 Monte Carlo Simulations

This section provides simulation results to demonstrate the finite-sample performance of the Bayesian estimator, drawing upon the characteristics of U.S. county-level crop yield data. The complete MCMC algorithm cycles through equations (4)–(11) by sequentially updating $\boldsymbol{\beta}_{i,g_{it}}$, $\sigma^2_{i,g_{it}}$, λ_i, $z_{i,g_{it}}$, η_{it}, g_{it}, τ_i, and h_i from their corresponding conditional posterior distributions. Parameter estimates and standard errors can be constructed using the means and standard deviations of the sampled parameters from the post-burn-in draws.

In our simulation setup, we consider diverse model configurations, distinguished by unit counts ($N = 25, 50$), time spans ($T = 60, 120$), and the maximum number of groups ($G = 1, 2, \text{or } 3$, where $G = \max\{G_i|i = 1, 2, ...N\}$). Under each configuration, we simulate sample data 1,000 times. For each simulated sample, our MCMC sampler undertakes 30,000 draws, where the first 20,000 are discarded as burn-in draws.

The regressors include a constant term and a time trend. Our choice of weight matrix is a third-order contiguity matrix, defined such that $w_{ij} = 1$ if $|i - j| \leq 3$ and $i \neq j$, and zero otherwise. Group count G_i is uniformly distributed over $\{1, 2, ..., G\}$. For the i-th unit, if $G_i = 1$, it consistently belongs to group 1 in all periods. If $G_i = 2$, then in each period, there is a 0.5 probability the i-th unit is assigned to group 1 and a 0.5 probability it is assigned to group 2. If $G_i = 3$, the respective probabilities of each period being assigned to groups 1, 2, and 3 are 0.5, 0.25, and 0.25. We may think of the grouping structure indicating growing conditions for crop yield – group 1 corresponds to median, group 2 to good, and group 3 to bad conditions.

When $g_{it} = 1$, parameters are generated from independent and identically distributed uniform (IIDU) and normal (IIDN) distributions: $\lambda_{i,1} \sim \text{IIDU}(0.05, 0.95)$, $\beta_{i1,1} \sim \text{IIDN}(20, 1)$, $\beta_{i2,1} \sim \text{IIDU}(1, 2)$ and $\sigma^2_{i,1} \sim \text{IIDU}(0, 1)$. For $g_{it} = 2$, parameters are adjusted to $\beta_{i1,2} = \beta_{i1,1} + 10$, $\beta_{i2,2} = \beta_{i2,1} + 1$, and $\sigma^2_{i,2} = 2\sigma^2_{i,1}$. When $g_{it} = 3$, they are set to $\beta_{i1,3} = \beta_{i1,1} - 10$, $\beta_{i2,3} = \beta_{i2,1} - 1$, and $\sigma^2_{i,3} = 2\sigma^2_{i,1}$. We also consider a homogeneous-coefficient case, which by definition has a group count of 1 for all i. In this case, $\lambda = 0.5$, $\beta_1 = 10$, $\beta_2 = 1.5$, $\sigma^2 = 0.5$. These parameter configures are largely in line with the characteristics of U.S. county-level crop yield data.

Table 1 presents the average bias and root mean squared error (RMSE) in estimating $\boldsymbol{\theta}$ from the MCMC algorithm, as well as the average probability of the estimated group count matching the true count.[2] Notably, the Bayesian estimator exhibits commendable precision, providing nearly unbiased results across diverse experimental configurations. An interesting observation is the influence of temporal duration, T, on bias and RMSE. Specifically, when extending T from 60 to 120, both bias and RMSE demonstrate a decreasing trend, with the exception of the estimated intercept when there is only one group.[3] However, alterations in unit counts, N, do not significantly influence these metrics.

[2] Average over 1,000 Monte Carlo runs, N observations, and T periods. For ease of exposition, we simply drop the subscripts i and g_{it} in parameters in the table.

[3] When there is only one group, whether the model contains homogeneous or heterogeneous coefficients, our MCMC algorithm still tries to estimate g_{it}, though it is unnecessary. The intercept parameter is of relatively large magnitude in our experimental design and the extra uncertainty introduced by estimating redundant group structure when there is only one group appears to carry over to the intercept parameter in our experiments.

Table 1: Finite-Sample Performance

		$T = 60$				$T = 120$			
		HomoCoef	$G = 1$	$G = 2$	$G = 3$	HomoCoef	$G = 1$	$G = 2$	$G = 3$
$N = 25$	λ	0.0087	0.0080	0.0010	0.0003	0.0015	0.0008	0.0003	0.0000
		(0.0766)	(0.2396)	(0.0147)	(0.0089)	(0.0313)	(0.2514)	(0.0114)	(0.0037)
	β_1	−0.3512	−0.0468	−0.0505	−0.0245	−0.0599	0.2866	−0.0166	−0.0076
		(3.0666)	(9.7960)	(0.6917)	(0.5113)	(1.2593)	(10.3170)	(0.5255)	(0.2835)
	β_2	−0.0261	−0.0034	−0.0029	−0.0004	−0.0045	0.0220	−0.0010	−0.0000
		(0.2298)	(0.7348)	(0.0507)	(0.0320)	(0.0940)	(0.7747)	(0.0367)	(0.0133)
	σ^2	0.0341	0.0462	0.0533	0.0749	0.0173	0.0305	0.0328	0.0440
		(0.1020)	(0.1246)	(0.2006)	(0.2658)	(0.0678)	(0.0931)	(0.1863)	(0.2296)
	$\Pr(G_i = 1)$	0.9957	0.9963	0.4943	0.3353	0.9977	0.9980	0.4974	0.3282
	$\Pr(G_i = 2)$	0.0043	0.0037	0.4967	0.3250	0.0023	0.0020	0.4953	0.3337
	$\Pr(G_i = 3)$	0.0000	0.0000	0.0088	0.3254	0.0000	0.0000	0.0051	0.3303
	$\Pr(G_i \geq 4)$	0.0000	0.0000	0.0001	0.0142	0.0000	0.0000	0.0022	0.0078
$N = 50$	λ	0.0097	0.0071	0.0009	0.0004	0.0019	0.0013	0.0003	0.0000
		(0.0765)	(0.2393)	(0.0163)	(0.0090)	(0.0315)	(0.2524)	(0.0126)	(0.0041)
	β_1	−0.3900	0.0008	−0.0449	−0.0280	−0.0754	0.2925	−0.0179	−0.0070
		(3.0651)	(9.8411)	(0.7455)	(0.5103)	(1.2644)	(10.3988)	(0.5654)	(0.2920)
	β_2	−0.0292	0.0001	−0.0025	−0.0007	−0.0057	0.0220	−0.0010	−0.0000
		(0.2296)	(0.7393)	(0.0547)	(0.0333)	(0.0945)	(0.7806)	(0.0400)	(0.0144)
	σ^2	0.0338	0.0461	0.0532	0.0947	0.0168	0.0320	0.0302	0.0523
		(0.1027)	(0.1263)	(0.2021)	(0.4129)	(0.0684)	(0.0955)	(0.1639)	(0.3057)
	$\Pr(G_i = 1)$	0.9960	0.9963	0.4971	0.3304	0.9978	0.9981	0.4969	0.3309
	$\Pr(G_i = 2)$	0.0040	0.0036	0.4941	0.3277	0.0022	0.0019	0.4959	0.3310
	$\Pr(G_i = 3)$	0.0000	0.0000	0.0087	0.3278	0.0000	0.0000	0.0054	0.3303
	$\Pr(G_i \geq 4)$	0.0000	0.0000	0.0001	0.0141	0.0000	0.0000	0.0019	0.0078

Note: This table presents the average bias, root mean squared error (RMSE), and the average probability associated with group counts from 1,000 Monte Carlo simulations.

The probability that the estimated group count matches the true group count, namely, $\Pr(G_i = 1) + \cdots + \Pr(G_i = G)$, are all above 95% across all configurations. (Recall that G_i is uniformly distributed over $\{1, 2, ..., G\}$.) This probability amplifies as T goes up and contracts with an increase in G. Again, changes in N appear to affect little this probability.

4 Empirical Applications

In this section, we provide two empirical applications of our proposed SAR-DPM model, where the posterior predictive distribution is used to calculate the magnitude of potential corn production loss, relative of the average, in the first subsection. It is used to calculate premium rates for area-yield insurance policies in the second subsection. The estimation results are based on the MCMC cycle outlined in Section 2 with a sampler of 30,000 draws, where the initial 20,000 simulations are discarded as burn-in draws. To follow the existing literature, we include a constant and possibly a time trend as covariates (Tolhurst and Ker, 2015; Ker et al., 2016; Schuurman and Ker, 2024). When the time trend is included, y_{it} reflects the actual yield. Otherwise, y_{it} represents the detrended and heteroskedasticity corrected yield (Harri et al., 2011).

4.1 Failure of Global Corn Production

Simultaneous crop failure across different global regions can cause significant threats to global food security, particularly impacting developing countries (Mehrabi and Ramankutty, 2019). For example, a dramatic increase in world food prices during 2007 and early 2008 led to widespread food insecurity and sparked civil unrest in several nations. Adverse weather conditions have been considered to be the principal factor (Lazear, 2008). Prolonged droughts in major grain-producing regions such as Ukraine, Russia, parts of Africa, and Australia have resulted in lower-than-average yields over the past several years, straining world grain supplies (Mueller et al., 2011). Following the 2007–2008 world food price crisis, food prices temporarily dropped in 2009 but surged again in 2010. This surge was once again due to droughts that caused yields reduction in various regions like Russia, Southwest Australia, and Southwestern China.

Therefore, it is important to quantitatively assess the potential loss due to global crop failure. For this purpose, we first compare three models, both with and without a time trend, in fitting corn yield data for five major corn-producing countries. The five breadbaskets are the United States (US), China (CN), Argentina (AR), Brazil (BR), and India (IN). In 2020, they accounted for more than 30% of the global corn production. We use annual historical corn yield data spanning from 1970 to 2020. The data for the United States, China, and Brazil used are at the sub-national level, whereas the data for Argentina and India are at the national level.[4] The three models under consideration are: (i) a SAR model with heterogeneous coefficients (LeSage and Chih, 2018), (ii) a DPM model (Walker, 2007), and (iii) the proposed SAR-DPM model in this paper. The spatial weight matrix $\boldsymbol{W}$ in both SAR and SAR-DPM is based on inverse squared distance[5].

Table 2 presents the estimated spatial coefficients from SAR and SAR-DPM as well as the DIC values from the three models. For countries with sub-national data, each MCMC iteration produces multiple estimates of λ_i, one for each sub-national unit. As a first step, we average these sub-national estimates within each iteration to obtain a single country-level estimate. These averages are then treated as the new MCMC draws for that country, from which we compute the country-level spatial coefficient, posterior standard deviations, and the 95% highest posterior density Bayesian credible intervals. The overall average ("All") is computed in the same manner, but aggregates across every national and sub-national unit in the sample. It can be seen that under both the SAR-DPM and

[4] Crop yield data for the period 1970 to 2017 is obtained from (Anderson et al., 2022), whereas data for 2018 to 2020 is retrieved from respective national statistical agencies, and the ministries or departments of agriculture in each country. Observations with missing values are excluded, resulting a total of 55 regions. The term sub-national here refers to a geographical level below the national level. Figure 1 in the appendix presents the annual average crop yields for the five countries over the period from 1970 to 2020.

[5] To construct the weight matrix, we first compute pairwise distances $d(i,j)$ between the centroid of regions i and j. For $i \neq j$, the unnormalized weight is defined as $w_{ij} = 1/d(i,j)^2$, and for $i = j$, $w_{ij} = 0$. Each row is then normalized by dividing its elements by their row sum so that the weights in each row sum to one.

Table 2: Estimation Results: λ and DIC Values

	W/ Trend			W/O Trend		
λ	SAR-DPM	SAR	DPM	SAR-DPM	SAR	DPM
All	0.6280^*	0.6830^*	-	0.7298^*	0.8140^*	-
	(0.0164)	(0.0143)	-	(0.0092)	(0.0081)	-
US	0.7750^*	0.8058^*	-	0.8266^*	0.8904^*	-
	(0.0143)	(0.0125)	-	(0.0102)	(0.0067)	-
CN	0.4777^*	0.5663^*	-	0.6115^*	0.7282^*	-
	(0.0412)	(0.0337)	-	(0.0172)	(0.0172)	-
BR	0.5913^*	0.6420^*	-	0.6576^*	0.7347^*	-
	(0.0662)	(0.0477)	-	(0.0287)	(0.0287)	-
AR	0.1248	0.1664	-	0.6257^*	0.7477^*	-
	(0.1469)	(0.1338)	-	(0.0702)	(0.0704)	-
IN	-0.4573	-0.2647	-	0.2944^*	0.3774^*	-
	(0.1242)	(0.1466)	-	(0.0519)	(0.0774)	-
DIC	14617	19735	20958	19197	20444	20525

Note: Posterior standard deviations are in parentheses and * indicates significance based on 95% highest posterior density Bayesian credible intervals.

SAR models, with and without a trend, the average spatial coefficients over all countries are positively significant. When averaged over individual countries, the majority of the spatial coefficients also remain significantly positive, with the exception of Brazil and India under the SAR-DPM and SAR model with a trend. The estimation results suggest that spatial dependence in corn yields is present. Table 2 also shows that the SAR-DPM model consistently achieves the lowest DIC values, with or without a time trend, indicating its better fit relative to the other two models.

Table 3 reports the average posterior probabilities of the number of groups from the SAR-DPM model. For countries with sub-national data, we record how many times each group count appears among their sub-national units across all post–burn-in MCMC draws. We then divide this frequency by the total number of post–burn-in MCMC draws and the total number of sub-national units in that country, thereby obtaining the average posterior

Table 3: Posterior Probabilities of Group Counts

Model	Country	$\Pr(G_i = 1)$	$\Pr(G_i = 2)$	$\Pr(G_i = 3)$	$\Pr(G_i = 4)$	$\Pr(G_i \geq 5)$
W/ Trend	US	0.7248	0.2466	0.0260	0.0024	0.0001
	CN	0.7595	0.1940	0.0447	0.0018	0.0000
	BR	0.2643	0.6362	0.0911	0.0082	0.0003
	AR	0.9028	0.0892	0.0075	0.0005	0.0000
	IN	0.8965	0.1035	0.0000	0.0000	0.0000
W/O Trend	US	0.4608	0.3006	0.1391	0.0596	0.0398
	CN	0.4498	0.2469	0.1618	0.0764	0.0651
	BR	0.2169	0.3866	0.2184	0.1083	0.0698
	AR	0.5813	0.2691	0.1131	0.0365	0.0100
	IN	0.6605	0.2591	0.0620	0.0139	0.0045

Note: This table presents the average posterior probabilities of the number of groups, where the average is over the different countries.

probability of each group count. For countries without sub-national data, the country itself is treated as a single unit. The results indicate that the number of groups is higher when the trend is not included. Furthermore, in all cases, the probability of group count exceeding one is always positive, suggesting the presence of latent group structures.

With the SAR-DPM model in mind, now we assess the potential loss of synchronized and country-specific corn production failure at different risk levels. We use the posterior predictive distribution from the SAR-DPM model to calculate the following measure of relative potential loss:

$$\mathcal{L}_i(\alpha) = \left[1 - \frac{\inf\{Y \in \mathbb{R} : \mathrm{P}(Y_{iT} \leq Y) \geq \alpha\}}{\mathrm{E}(Y_{iT})}\right] \times 100\%.$$

Here, Y_{iT} denotes the production of country i in year T, which equals the sum of corn yield times harvest area in region i in year 2022.[6] We use α to signal the risk level, serving as a scaled indicator of vulnerability by quantifying how severe the worst $\alpha\%$ of losses are

[6] Specifically, for each iteration of the MCMC simulation, we generate a posterior predictive sample of the regional yield y_{jT}. The posterior predictive sample of the total production for each country i is computed by summing the products of these sampled yields and their corresponding harvest areas, using the formula $Y_{iT} = \sum_{j \in \text{Country}_i} y_{jT} \times \text{Area}_j$, where Area_j is the harvest area in the j-th region in year 2020. For the estimate of pooled loss, denote $Y_T^{pool} = \sum_{i=1}^{5} Y_{iT}$ and $\mathcal{L}_{pool}(\alpha)$ is defined analogously.

relative to the average production.

Figure 1: Estimated Relative Potential Loss

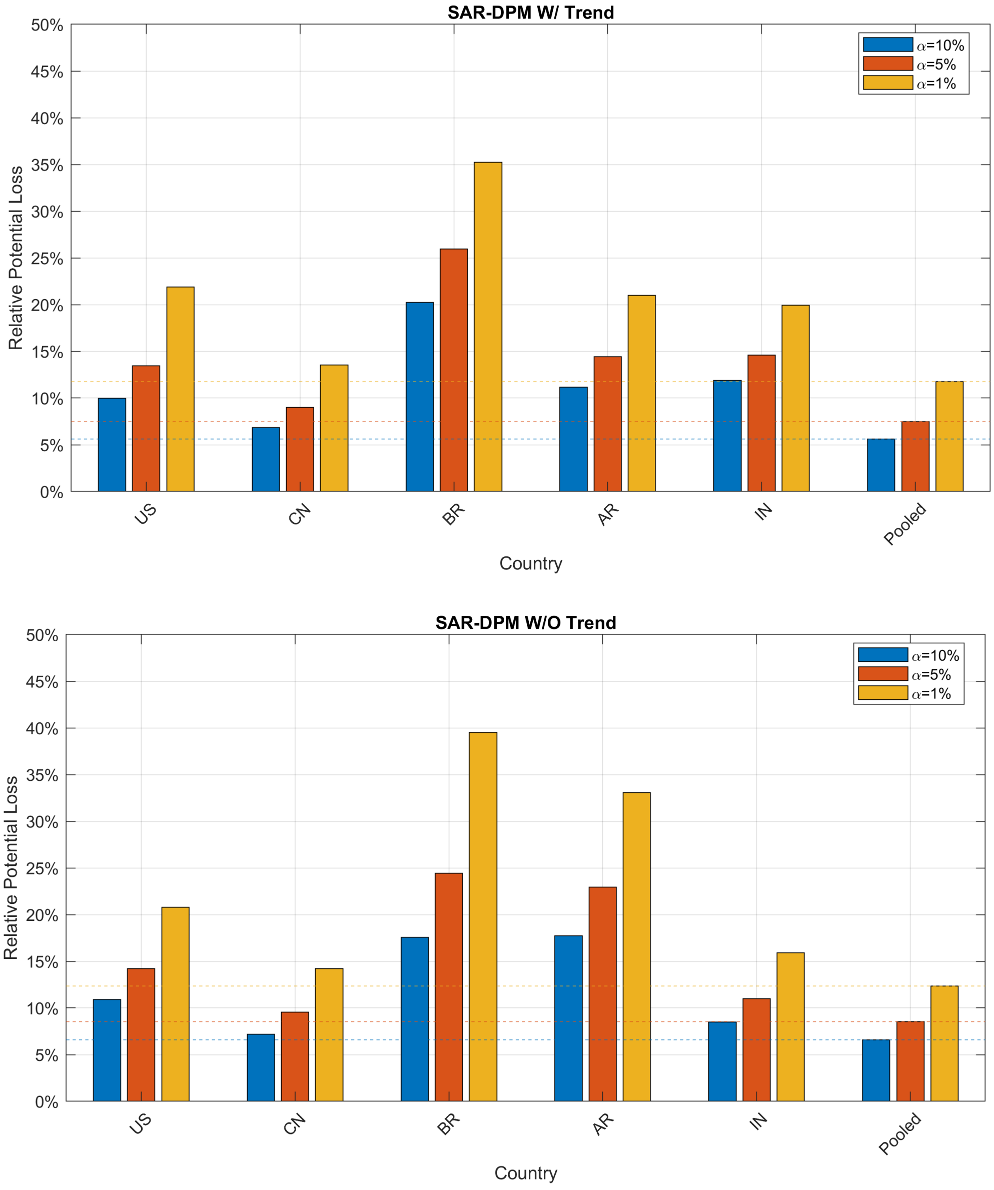

Figure 1 illustrates the estimated $\mathcal{L}_i(\alpha)$ under the SAR-DPM model for $\alpha = 1\%$, 5%, and 10%, corresponding to relative potential losses that might occur once every 100, 20, and 10 years, respectively. This result provides evidence-based recommendations on optimal corn storage capacities for governments to consider in order to buffer against potential production losses. The relative potential loss values across the five countries range from about 7% to 20%, 9% to 26% , and 14% to 35%, respectively, at the three risk levels when the SAR-DPM model includes a time trend. Without a time trend, they range from about 7% to 18%, 10% to 24%, and 14% to 40%. Among all countries assessed, Brazil exhibits the highest degree of relative potential production loss, whereas China consistently demonstrates the lowest potential losses.

We also notice that the pooled potential losses, obtained by aggregating production across the five major corn-producing regions, are consistently lower than those of any individual country across all three risk levels. This finding aligns with the rationale for risk pooling, wherein risks are jointly managed across multiple regions. This kind of strategy, when applied on a regional or global scale, can reduce the impact of yield reduction on food security (IPCC, 2012). Examples of risk pooling strategies in action include the Caribbean Catastrophe Insurance Facility (CCRIF), the African Risk Capacity (ARC), and the European Union Solidarity Fund (EUSF).

4.2 Crop Insurance Ratemaking

Crop insurance, a costly federal program overseen by the U.S. government through the USDA's Risk Management Agency (RMA), relies heavily on accurate estimations of crop yield distribution to set appropriate premium rates. Federal crop insurance policies are available at both the farm and aggregate levels. Area-yield insurance is based on county-level yields, and while farm-level insurance is more popular, area-based policies protect against most of the moral hazard and adverse selection problems (Miranda, 1991). In this subsection, we compare the SAR-DPM model with the current RMA methodology to evaluate their effectiveness in pricing area-based crop insurance policies.

For an area-yield insurance policy, the premium rate is given by

$$R_i = \frac{\mathrm{E}(I_i)}{L_i} \tag{12}$$

where both indemnity (I_i) and liability (L_i) are dependent on the coverage level. The subscript i denotes the county over which the policy is written. For a given coverage level $\alpha \in (0, 1)$, the liability under the policy can be calculated as $L_i = \alpha y_i^\star$, where $y_i^\star$ is the the predicted yield for a future time. The expected indemnity can be computed as $\mathrm{E}(I_i) = \Pr(y_i < \alpha y_i^\star)(\alpha y_i^\star - \mathrm{E}(y_i | y_i < \alpha y_i^\star))$, where y_i represents the realized yield.[7] The loss ratio for a portfolio of policies is given by the summation of $\max(\alpha y_i^\star - y_i, 0)$ divided by the sum of premiums, i.e.,

$$LR = \frac{\sum_{i=1}^{N} \max(\alpha y_i^\star - y_i, 0)}{\sum_{i=1}^{N} \mathrm{E}(I_i)}. \tag{13}$$

To evaluate the effectiveness of different models for pricing crop insurance policies, we utilize an out-of-sample rating game that assesses the predictive performance of these models. This rating game, akin to Ker and McGowan (2000) and Ker et al. (2016), mimics the Standard Reinsurance Agreement that governs the relationship between private insurers and the federal government in the federal crop insurance program. During the game, a private insurer cedes a certain number of policies to the government and relies on an internal rating system for pricing these policies that may be more accurate than the RMA's method. Based on this information, the insurer uses a decision rule to decide whether to retain or cede a policy. If the insurer's rate is higher than the government's rate, the policy is ceded as the government is underestimating the inherent risk. Conversely, if the insurer's rate is lower than the government's rate, the policy is retained. We limit both the insurance company and the government to retaining half of the contracts and employ the statistical test proposed by Ker et al. (2016). This test accounts for the benefits that private insurance

[7] Practically, the expected indemnity under an area-yield policy is adjusted by a specified price. However, this rating exercise does not consider it since prices are not stochastic.

companies gain under the Standard Reinsurance Agreement, where they can react to the premiums that the government agency suggested.

Our data includes county-level corn and soybeans yields from Illinois (IL), Indiana (IN), Iowa (IA), Minnesota (MN), and Nebraska (NE), with corn data from 1960 to 2022 and soybeans data from 1960 to 2020[8]. We adopt coverage levels of 90% and 80% and repeat the out-of-sample game over 20 years. We consider three rating systems. The first is the rating approach currently employed by RMA, which models the temporal process of yields with a two-knot robust linear spline and then makes adjustments to the estimated residuals for heteroskedasticity.[9] The resulting RMA rate is an empirical rate derived from these adjusted yields. The next two rating systems are based on the SAR-DPM model with and without a time trend. The calculation of premium rates employs equation (12), where both the expected values and probability measures are estimated from the predictive posterior distribution.[10] The spatial weight matrix we use is an inverse-squared distance weight matrix, constructed in the same way as described in the previous subsection.

Table 4 presents the average posterior probabilities of the number of groups, where the average is over the different states and the repeated 20 years[11]. It shows that in all cases, there is a probability that the number of groups is greater than one, suggesting the presence of latent group structures. Furthermore, Figures 2–9 in the appendix display the box plots of the posterior distributions of the average spatial coefficients λ_i for different states over repeated years 1 to 20 for corn and soybeans, under the SAR-DPM model with and without a trend. It can be seen that the posterior distributions of the average spatial coefficients in all cases are positive. The estimation results confirm that applying the proposed SAR-DPM model is appropriate.

[8] There are a total of 222 county-level observations for corn and 246 county-lelvel observations for soybeans.

[9] Our use of robust two-knot splines is based on Harri et al. (2011), but unlike the current RMA methodology, we do not conduct visual inspections or make ex-post adjustments during our analysis.

[10] We use two-year ahead predictions to be consistent with RMA.

[11] We compute the average posterior probabilities of the number of groups in the same way as in the previous subsection, except that we further average these probabilities over the repeated 20 years.

Table 4: Posterior Probabilities of Group Counts

Crop	Model	State	$\Pr(G_i = 1)$	$\Pr(G_i = 2)$	$\Pr(G_i = 3)$	$\Pr(G_i = 4)$	$\Pr(G_i \geq 5)$
Corn	W/ Trend	IL	0.6956	0.2438	0.0533	0.0066	0.0006
		IN	0.7767	0.2000	0.0213	0.0017	0.0002
		IA	0.6378	0.3038	0.0530	0.0049	0.0005
		MN	0.6523	0.2887	0.0538	0.0047	0.0004
		NE	0.7409	0.2207	0.0346	0.0035	0.0003
	W/O Trend	IL	0.3850	0.2852	0.1559	0.0848	0.0891
		IN	0.3780	0.2832	0.1712	0.0880	0.0796
		IA	0.3606	0.2712	0.1660	0.0949	0.1073
		MN	0.3123	0.2865	0.1750	0.1038	0.1224
		NE	0.4291	0.2832	0.1470	0.0714	0.0693
Soybeans	W/ Trend	IL	0.8878	0.1085	0.0037	0.0001	0.0000
		IN	0.9129	0.0826	0.0043	0.0001	0.0000
		IA	0.8319	0.1612	0.0068	0.0002	0.0000
		MN	0.8571	0.1301	0.0125	0.0004	0.0000
		NE	0.8982	0.0971	0.0045	0.0001	0.0000
	W/O Trend	IL	0.4304	0.2870	0.1544	0.0738	0.0544
		IN	0.4488	0.2848	0.1432	0.0679	0.0553
		IA	0.3699	0.3134	0.1689	0.0824	0.0655
		MN	0.4009	0.3138	0.1679	0.0732	0.0443
		NE	0.5677	0.2453	0.1098	0.0471	0.0301

Note: This table presents the average posterior probabilities of the number of groups, where the average is over the different states and the repeated 20 years.

Table 5 reports the out-of-sample rating game results. For each crop and method, there are 10 coverage-state combinations. In all combinations, the ceded to retained loss ratios, which represents the ratio of ceded loss ratio to retained loss ratio, are above 1 for the SAR-DPM model, both with and without a time trend. For the significance test at the 5% level, the SAR-DPM model with a trend produces 7 significant results for corn and 8 significant results for soybeans. The SAR-DPM model without a trend produces 8 significant results for corn and 9 significant results for soybeans. Overall, the game results demonstrate that the SAR-DPM model outperforms the RMA method regardless of whether a time trend is incorporated, which suggests that there are opportunities for insurers to extract rents by employing the SAR-DPM model to set up their premium rates.

Table 5: Out-of-sample Rating Game Results

Crop	Model	Coverage rate	State	Payouts (%)	Loss Ratio (Retained)	Loss Ratio (Ceded)	Ceded to Retained Loss Ratio	p-value
Corn	W/ Trend	0.9	IL	0.1522	0.6620	1.2173	1.8388	0.1796
			IN	0.1556	0.6435	1.0557	1.6406	0.0013
			IA	0.1636	0.4921	1.2702	2.5811	0.0176
			MN	0.1711	0.4678	0.7321	1.5651	0.0207
			NE	0.1139	0.5282	1.6009	3.0310	0.7483
		0.8	IL	0.0630	0.6744	1.6572	2.4571	0.0096
			IN	0.0722	0.7049	1.4870	2.1096	0.0000
			IA	0.0606	0.2660	1.1793	4.4327	0.0000
			MN	0.0553	0.2358	0.3912	1.6588	0.0318
			NE	0.0472	0.4821	2.0899	4.3354	0.8204
	W/O Trend	0.9	IL	0.1652	0.6300	1.3613	2.1606	0.0059
			IN	0.1472	0.6583	1.0213	1.5515	0.0002
			IA	0.1879	0.5595	1.2609	2.2535	0.0002
			MN	0.2026	0.4695	0.7718	1.6437	0.0002
			NE	0.1861	0.8280	1.4947	1.8053	0.2517
		0.8	IL	0.0630	0.6495	1.9466	2.9972	0.0013
			IN	0.0694	0.6961	1.5643	2.2472	0.0000
			IA	0.0788	0.3527	1.1184	3.1713	0.0000
			MN	0.0737	0.2412	0.4069	1.6870	0.0059
			NE	0.0833	0.6378	5.2418	8.2189	0.0577
Soybeans	W/ Trend	0.9	IL	0.1694	0.5182	1.2140	2.3427	0.0318
			IN	0.1667	0.5363	1.4395	2.6843	0.0059
			IA	0.2059	0.5632	1.3116	2.3290	0.0096
			MN	0.2395	0.8059	1.2042	1.4943	0.5000
			NE	0.1156	0.5634	1.4446	2.5640	0.0064
		0.8	IL	0.0629	0.4190	1.3618	3.2503	0.0059
			IN	0.0688	0.4284	1.7228	4.0213	0.0002
			IA	0.0853	0.4145	1.4235	3.4342	0.0096
			MN	0.0868	0.5023	0.9721	1.9352	0.2517
			NE	0.0594	0.5720	2.1244	3.7139	0.0207
	W/O Trend	0.9	IL	0.1952	0.5901	1.1389	1.9299	0.0059
			IN	0.1792	0.5933	1.4292	2.4091	0.0002
			IA	0.2265	0.6335	1.1842	1.8693	0.0013
			MN	0.2447	0.8105	1.1861	1.4633	0.0577
			NE	0.1812	0.6811	1.7687	2.5967	0.0013
		0.8	IL	0.0726	0.4855	1.2016	2.4751	0.0013
			IN	0.0729	0.4393	1.8516	4.2144	0.0000
			IA	0.0926	0.4607	1.2823	2.7835	0.0013
			MN	0.0868	0.4109	1.1626	2.8297	0.0207
			NE	0.0750	0.7341	3.3252	4.5298	0.0000

Note: This table is based on out-of-sample simulation results. "Ceded to Retained Loss Ratio" refers to the ratio of "Loss Ratio (Ceded)" to "Loss Ratio (Retained)" and p-value is calculated by following Ker et al. (2016). A small p-value suggests evidence of the model being statistically superior to the RMA method.

5 Conclusion

The distributions of many economic variables play vital roles in various areas of economics, such as agricultural, financial and urban economics. In this paper, we propose a SAR-DPM model that accommodates both cross-sectional interactions and latent group structures. Relatively to finite mixture models commonly employed in the literature for modeling the distributions of economic variables, our proposed SAR-DPM model offers at least two advantages. First, by employing a DP prior to model the latent group structure, our method jointly estimates the model parameters and the number of latent groups without relying on model selection methods. This approach also provides the posterior distribution of group counts rather than only a point estimate. Second, we incorporate an SAR structure into the dependent variable to directly capture cross-sectional correlation, a feature that is often crucial for public policy design. Monte Carlo simulations, designed to reflect the characteristics of U.S. county-level crop yield data, demonstrate the good finite-sample performance of the Bayesian estimation approach. We apply the SAR-DPM model in two empirical applications. The first application examines the failure of global corn production across five major corn-producing countries. The findings show different degrees of relative potential losses across the five countries and provide support to risk pooling strategies in enhancing global food security. The second application compares the proposed SAR-DPM model with the USDA's current rating methodology for area-yield crop insurance contracts. The results indicate that the proposed model may lead to more accurate premium rates. Our findings regarding the USDA's current rating methodology are also relevant for many developing countries, where area-yield insurance is viewed as a means to help small scale farmers boost living standards and adopt improved technologies (Carter et al., 2016).

For future empirical research, it woule be of interest to incorporate weather forecast information into the SAR-DPM model to further enhance its performance. This could involve adding variables such as the El Niño–Southern Oscillation index, degree days, and precipitation to the regressors, following the approaches of Yi et al. (2020) and Liu and Ramsey (2023). In terms of methodology, it would be interesting to develop model selection and model averaging methods as in Zhang and Yu (2018) when there are possibly multiple channels of cross-sectional correlation.

References

Aït-Sahalia, Y., J. Cacho-Diaz, and R. J. Laeven (2015). Modeling financial contagion using mutually exciting jump processes. *Journal of Financial Economics 117*(3), 585–606.

Anderson, W., W. Baethgen, F. Capitanio, P. Ciais, B. Cook, C. da Cunha, L. Goddard, B. Schauberger, K. Sonder, G. Podestá, et al. (2022). Twentieth century crop statistics, 1900–2017.

Anselin, L. (2001). Spatial effects in econometric practice in environmental and resource economics. *American Journal of Agricultural Economics 83*(3), 705–710.

Arellano, C., Y. Bai, and S. Lizarazo (2017). Sovereign risk contagion. NBER Working Papers 24031, National Bureau of Economic Research, Inc.

Bao, Y., G. Li, and X. Liu (2024). A spatial sample selection model. *Oxford Bulletin of Economics and Statistics*.

Barnwal, P. and K. Kotani (2013). Climatic impacts across agricultural crop yield distributions: An application of quantile regression on rice crops in Andhra Pradesh, India. *Ecological Economics 87*, 95–109.

Belasco, E., M. Farmer, and C. Lipscomb (2012). Using a finite mixture model of heterogeneous households to delineate housing submarkets. *Journal of Real Estate Research 34*(4), 577–594.

Benhabib, J., J. Perla, and C. Tonetti (2021). Reconciling models of diffusion and innovation: A theory of the productivity distribution and technology frontier. *Econometrica 89*(5), 2261–2301.

Billio, M. and L. Pelizzon (2000). Value-at-Risk: A multivariate switching regime approach. *Journal of Empirical Finance 7*(5), 531–554.

Caparas, M., Z. Zobel, A. D. Castanho, and C. R. Schwalm (2021). Increasing risks of crop failure and water scarcity in global breadbaskets by 2030. *Environmental Research Letters 16*(10), 104013.

Cardella, E. and A. Roomets (2022). Pay distribution preferences and productivity effects: An experiment. *Journal of Behavioral and Experimental Economics 96*, 101814.

Carter, M. R., L. Cheng, and A. Sarris (2016). Where and how index insurance can boost the adoption of improved agricultural technologies. *Journal of Development Economics 118*, 59–71.

Ceballos, F. and M. Robles (2020). Demand heterogeneity for index-based insurance: The case for flexible products. *Journal of Development Economics 146*, 102515.

Chavas, J.-P., G. Rivieccio, S. Di Falco, G. De Luca, and F. Capitanio (2022). Agricultural diversification, productivity, and food security across time and space. *Agricultural Economics 53*(S1), 41–58.

Chemeris, A., Y. Liu, and A. P. Ker (2022). Insurance subsidies, climate change, and innovation: Implications for crop yield resiliency. *Food Policy 108*, 102232.

Compiani, G. and Y. Kitamura (2016). Using mixtures in econometric models: A brief review and some new results. *The Econometrics Journal 19*(3), C95–C127.

DeFusco, A., W. Ding, F. Ferreira, and J. Gyourko (2018). The role of price spillovers in the American housing boom. *Journal of Urban Economics 108*, 72–84.

Durham, G. B. (2007). SV mixture models with application to S&P 500 index returns. *Journal of Financial Economics 85*(3), 822–856.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Garicano, L., C. Lelarge, and J. Van Reenen (2016). Firm size distortions and the productivity distribution: Evidence from France. *American Economic Review 106*(11), 3439–79.

Goodwin, B. K. and A. P. Ker (1998). Nonparametric estimation of crop yield distributions: implications for rating group-risk crop insurance contracts. *American Journal of Agricultural Economics 80*(1), 139–153.

Hammadou, H., S. Paty, and M. Savona (2014). Strategic interactions in public R&D across european countries: A spatial econometric analysis. *Research Policy 43*(7), 1217–1226.

Harri, A., K. H. Coble, A. P. Ker, and B. J. Goodwin (2011). Relaxing heteroscedasticity assumptions in area-yield crop insurance rating. *American Journal of Agricultural Economics 93*(3), 707–717.

IPCC (2012). *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation: Special Report of the Intergovernmental Panel on Climate Change.* Cambridge University Press.

Ker, A. P. and K. Coble (2003). Modeling conditional yield densities. *American Journal of Agricultural Economics 85*(2), 291–304.

Ker, A. P. and P. McGowan (2000). Weather-based adverse selection and the US crop insurance program: The private insurance company perspective. *Journal of Agricultural and Resource Economics 25*(2), 386–410.

Ker, A. P., T. N. Tolhurst, and Y. Liu (2016). Bayesian estimation of possibly similar yield densities: Implications for rating crop insurance contracts. *American Journal of Agricultural Economics 98*(2), 360–382.

Kondo, I. O., L. T. Lewis, and A. Stella (2023). Heavy tailed but not Zipf: Firm and establishment size in the United States. *Journal of Applied Econometrics 38*(5), 767–785.

Kuester, K., S. Mittnik, and M. S. Paolella (2005). Value-at-Risk prediction: A comparison of alternative strategies. *Journal of Financial Econometrics 4*(1), 53–89.

Lazar, E., J. Pan, and S. Wang (2024). On the estimation of Value-at-Risk and expected shortfall at extreme levels. *Journal of Commodity Markets 34*, 100391.

Lazear, E. P. (2008). *Responding to the Global Food Crisis: Hearings before the Senate Foreign Relations Committee, 110th Congress (testimony of Edward P. Lazear).*

LeSage, J. and R. K. Pace (2009). *Introduction to Spatial Econometrics.* New York: Chapman and Hall/CRC.

LeSage, J. P. and Y.-Y. Chih (2018). A Bayesian spatial panel model with heterogeneous coefficients. *Regional Science and Urban Economics 72*, 58–73.

Liu, Y. and A. F. Ramsey (2023). Incorporating historical weather information in crop insurance rating. *American Journal of Agricultural Economics 105*(2), 546–575.

Long, D. S. (1998). Spatial autoregression modeling of site-specific wheat yield. *Geoderma 85*(2-3), 181–197.

Mehrabi, Z. and N. Ramankutty (2019). Synchronized failure of global crop production. *Nature Ecology & Evolution 3*, 780–786.

Miled, W., Z. Ftiti, and J.-M. Sahut (2022). Spatial contagion between financial markets: New evidence of asymmetric measures. *Annals of Operations Research 313*(2), 1183–1220.

Miranda, M. J. (1991). Area-yield crop insurance reconsidered. *American Journal of Agricultural Economics 73*(2), 233–242.

Mueller, S. A., J. E. Anderson, and T. J. Wallington (2011). Impact of biofuel production and other supply and demand factors on food price increases in 2008. *Biomass and Bioenergy 35*(5), 1623–1632.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

Ozaki, V. A., B. K. Goodwin, and R. Shirota (2008). Parametric and nonparametric statistical modelling of crop yield: implications for pricing crop insurance contracts. *Applied Economics 40*(9), 1151–1164.

Park, E., B. W. Brorsen, and A. Harri (2019). Using bayesian kriging for spatial smoothing in crop insurance rating. *American Journal of Agricultural Economics 101*(1), 330–351.

Schuurman, D. and A. Ker (2024). Heterogeneity, climate change, and crop yield distributions: Solvency implications for publicly subsidized crop insurance programs. *American Journal of Agricultural Economics 107*, 248–268.

Silveira, F. (2022). Firm size distribution and growth: An empirical investigation. *Structural Change and Economic Dynamics 63*, 422–434.

Skees, J. R., J. R. Black, and B. J. Barnett (1997). Designing and rating an area yield crop insurance contract. *American Journal of Agricultural Economics 79*(2), 430–438.

Su, H.-L. (2020). On the city size distribution: A finite mixture interpretation. *Journal of Urban Economics 116*, 103216.

Telila, H. F. (2023). Frontier markets sovereign risk: New evidence from spatial econometric models. *Finance Research Letters 58*, 104665.

Tolhurst, T. N. and A. P. Ker (2015). On technological change in crop yields. *American Journal of Agricultural Economics 97*(1), 137–158.

Walker, S. G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation 36*(1), 45–54.

Wang, T., F. Yi, X. Wu, H. Liu, and Y. Y. Zhang (2024). Calamitous weather, yield risk and mitigation effect of harvest mechanisation: Evidence from china's winter wheat. *Australian Journal of Agricultural and Resource Economics 68*(2), 386–412.

Woodard, J. D. and B. J. Sherrick (2011). Estimation of mixture models using cross-validation optimization: Implications for crop yield distribution modeling. *American Journal of Agricultural Economics 93*(4), 968–982.

Yi, F., M. Zhou, and Y. Y. Zhang (2020). Value of incorporating ENSO forecast in crop insurance programs. *American Journal of Agricultural Economics 102*(2), 439–457.

Yvette Zhang, Y. (2017). A density-ratio model of crop yield distributions. *American Journal of Agricultural Economics 99*(5), 1327–1343.

Zhang, X. and J. Yu (2018). Spatial weights matrix selection and model averaging for spatial autoregressive models. *Journal of Econometrics 203*(1), 1–18.