Statistica Sinica Preprint No: SS-2021-0247	
Title	Identifying Latent Groups in Spatial Panel Data Using a Markov Random Field Constrained Product Partition Model
Manuscript ID	SS-2021-0247
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0247
Complete List of Authors	Tianyu Pan, Guanyu Hu and Weining Shen
Corresponding Author	Weining Shen
E-mail	weinings@uci.edu, swn1989@gmail.com

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Statistica Sinica

IDENTIFYING LATENT GROUPS IN SPATIAL PANEL DATA USING A MARKOV RANDOM FIELD CONSTRAINED PRODUCT PARTITION MODEL

Tianyu Pan[1], Guanyu Hu[2] and Weining Shen[1]

[1] *Department of Statistics, University of California, Irvine*

[2] *University of Missouri - Columbia, Columbia, MO, 65211*

Abstract: Understanding the heterogeneity over spatial locations is an important problem that has been widely studied in applications such as economics and environmental science. We focus on regression models for spatial panel data analyses, where repeated measurements are collected over time at various spatial locations. We propose a novel class of nonparametric priors that combines a Markov random field (MRF) with the product partition model (PPM), and show that the resulting prior, called MRF-PPM, is capable of identifying latent group structures among the spatial locations, while efficiently using the spatial dependence information. We derive a closed-form conditional distribution for the proposed prior and introduce a new way of computing the marginal likelihood that renders an efficient Bayesian inference. Furthermore, we study the theoretical properties of the proposed MRF-PPM prior and show a clustering consistency result for the posterior distribution. We demonstrate the excellent empirical performance of our method using extensive simulation studies and applications to US precipitation data and a California median household income data study.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Key words and phrases: Marginal Likelihood; Nonparametric Bayesian Method; Posterior Consistency; Spatial Homogeneity.

1. Introduction

Panel data have been widely studied in applications such as economics (Pesaran, 2015) and climate science (Hao et al., 2016), because they represent a common data format in which observations are collected for each subject at different time points. Here, we model a special type of panel data, called *spatial panel data*, where each subject represents a spatial location and we need to account for the spatial dependence between those locations. There is a growing interest in spatial panel data analysis (Elhorst, 2014; Belotti et al., 2017), where a central question is to model the relationship between variables measured repeatedly over a study period at various spatial locations. For example, economic studies, may seek to quantify the association between median household income and other economic indicators, such as gross domestic product (GDP) and unemployment rate over time. In environmental studies, understanding the effect of greenhouse gas emissions on climate change is an important research direction.

The aforementioned question can be formulated naturally as a regression problem in statistics, and it is now well recognized that the regression parameters (e.g., coefficients and variance) can be *highly variable* across different spatial locations (Hsiao and Tahmiscioglu, 1997; Browning et al.,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

2007; Su and Chen, 2013). To account for spatial heterogeneity, it is common to assume a *latent group structure*; that is, spatial locations are grouped into clusters, and those assigned to the same cluster share the same set of regression parameters. This strategy has several practical advantages. First, neglecting unobserved heterogeneity may lead to inconsistent parameter estimations and misleading results, as demonstrated by Simpson's paradox and other examples for spatial panel data (Wagner, 1982; Su et al., 2016; Hsiao, 2014). Second, the obtained latent group structure is usually informative for empirical analysis, such as finding possible unobserved confounders or performing a secondary analysis. Another benefit is that the latent group structure provides a convenient way of incorporating spatial dependence information in the model, which helps to improve the accuracy/efficiency of the model fit and interpretation (Miao et al., 2020).

Several approaches have been introduced in the frequentist literature for studying panel data regression models with a latent group structure. For example, Lin and Ng (2012) considered a panel data linear regression model with group-varying slopes. This method was later extended by Su et al. (2016) to allow for group-varying intercepts and slopes. Several more complicated models have also been proposed, with features such as group-specific time patterns (Bonhomme and Manresa, 2015), time-varying grouped coefficients (Su et al., 2019), and group-varying threshold variables (Miao et al., 2020). Despite the success of these frequentist approaches,

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

studies have only recently begun focusing on Bayesian frameworks (Zhang, 2020; Ma et al., 2020; Hu et al., 2021; Geng and Hu, 2021). Teixeira et al. (2019) introduced a Bayesian spatio-temporal clustering method, but it is not suitable for clustering locations in a panel data analysis. Conceptually, an ideal Bayesian approach would naturally be able to incorporate spatial dependence and latent group structure information in the prior distribution. Inferences can be conducted conveniently without needing to use complicated procedures such as a bootstrap or post-model selection. The main goal of this study is to introduce a new class of nonparametric priors and to explore their computational and theoretical properties.

Our first step toward constructing a prior distribution for spatial panel data with a group structure is to recognize that a latent group structure is essentially equivalent to a *partition* of spatial locations. Therefore, we only need a class of priors assigned to the space of partitions, which is usually achieved by specifying a class of partition probability functions. In this class, the product partition model (PPM), first introduced by Hartigan (1990) and studied from a Bayesian point of view by Quintana and Iglesias (2003), has received considerable interest. The PPM is defined by taking the product of some nonnegative cohesion functions $h(c)$ over different clusters, where $h(c)$ measures the similarity between individual subjects assigned to the same cluster c (Design, 1978). It has been shown that the PPM prior has strong connections to the marginal prior on partitions induced by the

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Dirichlet process (DP) prior (Green and Richardson (2001)) and the mixture of finite mixtures (MFM) prior (Miller and Harrison (2018)). Recently, the PPM was extended to include covariates (Park and Dunson, 2010; Page et al., 2015) and spatial information (Page et al., 2016). However, it remains unclear how to *systematically* incorporate spatial dependence information into the PPM.

To solve this issue, we introduce a new class of priors called Markov random field constrained product partition model (MRF-PPM) priors. These priors are generated by taking the product of two priors, namely, a Markov random field (MRF) prior and a PPM prior. There is a long history of using MRF priors defined on undirected graphs to capture the local homogeneity in image segmentation, spatial statistics, and Bayesian nonparametrics (Geman and Geman, 1984; Orbanz and Buhmann, 2008; Blake et al., 2011). However, to the best of our knowledge, the MRF-PPM prior, as a general class of priors that combines an MRF and a PPM, has not been studied systemically in the literature in terms of its theoretical and computational properties. In particular, we show that several commonly used nonparametric priors (Zhao et al., 2020; Hu et al., 2020; Orbanz and Buhmann, 2008) are special cases of the MRF-PPM. The clustering consistency result states that with posterior probability tending to one, the posterior distribution of the MRF-PPM is capable of identifying the correct unknown partition structure in spatial panel data. To the best of our knowledge, this result

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

is new in the Bayesian spatial panel model literature, and is generally applicable to regression models with well-defined posterior contraction rates under mild identifiability conditions.

2. Methodology

2.1 The MRF-PPM

Consider a total of N spatial locations. For location i, suppose we observe a response $Y_i(t_j^{(i)})$ and a p-dimensional covariate vector $X_i(t_j^{(i)})$ at time point $t_j^{(i)}$, for $j = 1, \ldots, n_i$, where n_i is the total number of time points observed for location i. We use c_i to denote the cluster assignment for location i, and for those locations that belong to the same cluster index set c, that is, $i \in c$, we use θ_c to denote the common set of modeling parameters being shared within the cluster c. Therefore, our spatial panel data regression model with a latent group structure can be expressed as,

$$Y_i(t_j^{(i)}) \mid X_i(t_j^{(i)}) \sim f_{\theta_c}(Y_i(t_j^{(i)}) \mid X_i(t_j^{(i)})), \text{ for } j = 1, \ldots, n_i, \text{ and } i \in c, \quad (2.1)$$

where f_{θ_c} is the regression likelihood function for cluster c. In the remainder of this paper, we also use $Y_i(t)$ to denote the observation collected at time t for location i, for simplicity of notation. Note that model (2.1) allows a temporal correlation between $Y_i(t_j^{(i)})$ and $Y_i(t_k^{(i)})$ for every $j \neq k$. To model

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

the clustering structure, we consider the following prior on the partition of the index set $[N] = \{1, \ldots, N\}$ and the associated parameters sets θ_c:

$$\theta_c \overset{\text{i.i.d}}{\sim} G_0, \text{ for } c \in \mathcal{C}, \quad \mathcal{C} \sim p(\mathcal{C}), \tag{2.2}$$

where G_0 is a non-atomic base measure for θ_c with density function $g(\cdot)$, $\mathcal{C}$ is a partition of $[N]$, and $p(\mathcal{C})$ is a probability mass function over $\mathcal{C}$. It is common to consider a product partition model (PPM) for $p(\mathcal{C})$, that is,

$$p(\mathcal{C}) \propto \Pi_{c \in \mathcal{C}} h(c), \tag{2.3}$$

where $h(c) \geq 0$ is the cohesion function that measures the similarity between individual units assigned to the same cluster c.

To account for the spatial correlation among different locations, we incorporate an MRF structure on $p(\mathcal{C})$. Consider a collection of parameters $\{\theta_1, \theta_2, \ldots, \theta_N\}$ defined on an undirected known graph $\mathcal{G}_N = (V_{\mathcal{G}_N}, E_{\mathcal{G}_N})$, where $V_{\mathcal{G}_N} = \{\theta_1, \theta_2, \ldots, \theta_N\}$ is the vertex set and $E_{\mathcal{G}_N}$ is the set of edges. In our case, θ_i is the regression parameter for location i, which is equivalent to θ_{c_i} defined in (2.2). Given the graphical information, a joint distribution m on $V_{\mathcal{G}_N}$ is called an MRF w.r.t $\mathcal{G}_N$ if

$$m(\theta_i \mid \theta_{(-i)}; \mathcal{G}_N) = m(\theta_i \mid \theta_{\partial(i)}; \mathcal{G}_N), \tag{2.4}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

where $\partial(i) = \{j : (i,j) \in E_{\mathcal{G}_N}\}$ is the collection of node i's neighbors, $\theta_{(-i)} = \{\theta_i\}_{i=1}^N \setminus \{\theta_i\}$, and, $\theta_{\partial(i)} = \{\theta_j : (i,j) \in E_{\mathcal{G}_N}\}$. This Markov property indicates that the distribution of θ_i depends only on its neighbors, that is, the vertices connected to θ_i. The graphical information $\mathcal{G}_N$ is usually determined by the network structure in real-world applications, where the vertices $V_{\mathcal{G}_N}$ represent subjects and the edges $E_{\mathcal{G}_N}$ represent their relationships. Here, examples include the 51 states and their adjacency matrix in the United States, users and their friendship connections in social media networks, and international airports and airlines.

Inspired by the Markov property, we define an MRF joint cost function, which is not necessarily a probability density function, as $M(\theta_1, \ldots, \theta_N \mid \mathcal{G}_N) = \Pi_{c \in \mathcal{C}} l(\theta_c) k(c \mid \mathcal{G}_N)$ (sometimes denoted by M) that satisfies

$$k(c \cup \{i\} \mid \mathcal{G}_N) = k(c \mid \mathcal{G}_N) \cdot k_i(\partial(i) \cap c \mid \mathcal{G}_N), \text{ for every } i \text{ and every } c \subset [N], \tag{2.5}$$

where $l(\cdot)$ is a nonnegative function, and $k(\cdot \mid \mathcal{G}_N)$ and $k_i(\cdot \mid \mathcal{G}_N)$ are nonnegative cohesion functions defined for every $c \subseteq [N]$, given the graphical information. This it satisfies $k(\{i\} \mid \mathcal{G}_N) = 1$ for all i, and $k(\emptyset \mid \mathcal{G}_N) = k_i(\emptyset \mid \mathcal{G}_N) = 1$. Note that (2.5) is conceptually relevant to the Markov property, because the cohesion value of $c \cup \{i\}$ is related to the joint density of $c \cup \{i\}$. The cohesion value of c can then be interpreted as the marginal density by integrating out the parameter of subject i from the joint density.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Consequently, $k_i(\partial(i) \cap c \mid \mathcal{G}_N)$ is associated with the conditional density in the context of the Markov property. A simple example that satisfies (2.5) is $k(c \mid \mathcal{G}_N) = \exp\{E_c\}$, where E_c represents the number of edges among the subjects assigned to cluster c with respect to the adjacency matrix of these N subjects.

To introduce the definition of the MRF-PPM, we let $P(\theta_1, \ldots, \theta_N)$ (P) be the prior on $\{\theta_1, \ldots, \theta_N\}$ defined in (2.2) and (2.3), which is proportional to $\Pi_{c \in \mathcal{C}} g(\theta_c) h(c)$. An MRF-PPM prior Π can then be constructed by taking the product of P and the MRF cost function M, with some positive normalizing constant K_0, as follows:

$$\Pi(\theta_1, \ldots, \theta_N \mid \mathcal{G}_N) = K_0 M(\theta_1, \ldots, \theta_N \mid \mathcal{G}_N) P(\theta_1, \ldots, \theta_N). \tag{2.6}$$

The proposed MRF-PPM prior enjoys the following three attractive properties:

(P1) If $l(\theta)g(\theta)$ is integrable as a function of θ, then $\Pi(\cdot \mid \mathcal{G}_N)$ is still a product partition model, with a cohesion function equal to $k(\cdot \mid \mathcal{G}_N)h(\cdot)$ and a probability density function of the base measure of $K_1 l(\cdot)g(\cdot)$, for some normalizing constant K_1.

(P2) It inherits the ability of clustering, because it provides a full support over the entire space of partitions.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

(P3) It is exchangeable, because the cohesion function is invariant under permutation (it depends only on the clustering configuration), which, by de Finetti's theorem (De Finetti, 1929), justifies the existence of the MRF-PPM prior.

Next, we derive the full conditional distribution of the MRF-PPM prior in Theorem 1. The proof is given in the Supplementary Material.

Theorem 1. *Suppose that $l(\theta)g(\theta)$ in the MRF-PPM prior is integrable as a function of θ. Then the conditional distribution of θ_i given $\theta_{(-i)}$, the induced partition $\mathcal{C}_i$, and distinct values $\{\theta_c\}_{c\in\mathcal{C}_i}$ is proportional to*

$$\frac{k(\{i\} \mid \mathcal{G}_n)h(\{i\})}{K_1}L_0 + \sum_{c\in\mathcal{C}_i} k_i(\partial(i)\cap c \mid \mathcal{G}_n)\frac{h(c\cup\{i\})}{h(c)}\delta_{\theta_c}, \text{ for every } i, \tag{2.7}$$

where L_0 is the base measure associated with the probability density function $K_1 l(\theta)g(\theta)$.

If $l(\theta) = 1$, we have the base measure $L_0 = G_0$. From the second term in (2.7), we can see that the MRF-PPM is able to account for the spatial correlation, because location i has a higher probability of being assigned to a specific cluster that includes more of its neighbors. That probability is determined by the function $k_i(\partial(i)\cap c \mid \mathcal{G}_n)$, which satisfies the Markov property in (2.4).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

In addition to the PPM, we can also impose an MRF structure on an *exchangeable partition probability function* (EPPF) (Pitman et al., 2002), as described in the following theorem.

Theorem 2. *If the partition probability function of P is an EPPF, and the cluster-wise parameters are independent and identically distributed (i.i.d.) and sampled from a base measure G_0, then the resulting MRF-EPPF satisfies Properties (P1)–(P3), and its full conditional distribution can be obtained to* (2.7).

Theorem 2 is widely applicable to many commonly used priors in the Bayesian nonparametric literature. For example, it is well known that the partition probability function of the Dirichlet process is an EPPF. In addition, the partition probability function of the MFM prior is an EPPF (Miller and Harrison, 2018). Therefore, the MRF structure can be conveniently combined with these two priors.

Note that under the MRF structure,

$$\Pi(\theta_1, \ldots, \theta_{N-1} \mid \mathcal{G}_{N-1}) \neq \int \Pi(\theta_1, \ldots, \theta_N \mid \mathcal{G}_N) d\theta_N,$$

because the new observation θ_N will provide additional spatial information to the historical data $\{\theta_i\}_{i=1}^{N-1}$. Hence, the marginal distribution of $\{\theta_i\}_{i=1}^{N-1}$ will change. As a result, we cannnot apply the Kolmogorov's extension theorem (Durrett, 2019) directly to show the existence of $\Pi(\theta_1, \ldots, \theta_N \mid$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

$\mathcal{G}_N)$ as $N \to \infty$. For the same reason, the Pólya urn scheme is not available for the MRF-MFM. However, this does not affect our method because we focus on the case of fixed N; that is, the number of spatial locations of interest is fixed.

2.2 Model Specification

Next, we focus on the linear regression case with Gaussian errors, and demonstrate how the proposed prior works for the model introduced in (2.1). The full model can be formulated in the following hierarchical order:

$$
\begin{aligned}
&Y_i(t_j^{(i)}) \mid \left\{e_i(t_j^{(i)}), X_i(t_j^{(i)})\right\} \overset{\text{ind}}{\sim} \mathcal{N}(X_i(t_j^{(i)})\beta_c + e_i(t_j^{(i)}), \sigma_c^2 \cdot \alpha_c),\ i = 1, \ldots, N,\\
&e_i(t_j^{(i)}) \sim \mathcal{N}(0, K_{\sigma_c,\ell_c}(\cdot,\cdot)), \text{ for every } j = 1, \cdots, n_i,\ i \in c,\\
&\theta_c \equiv \{\beta_c, \sigma_c^2, \alpha_c, \ell_c\} \overset{i.i.d}{\sim} G_0, \text{ for every } c \in \mathcal{C},\\
&\mathcal{C} \sim p_\lambda(\mathcal{C} \mid \mathcal{G}_N),\\
&dG_0 \equiv \pi_0(\beta, \sigma^2)\pi_1(\alpha)\pi_2(l)d\beta d\sigma^2 d\alpha dl,\\
&\beta_c \mid \sigma_c^2 \sim \mathcal{N}(\mu_0, \sigma_c^2\Lambda_0^{-1}),\\
&\sigma_c^{-2} \sim \text{Gamma}(a_0, b_0),\ \alpha_c \sim \text{Gamma}(a_1, b_1),\ \ell_c \sim \text{Gamma}(a_2, b_2),
\end{aligned}
\tag{2.8}
$$

where $e_i(\cdot)$ is the temporal random effect for location i, and K_{σ_c,ℓ_c} is the associated squared exponential covariance kernel, defined by $K_{\sigma_c,\ell_c}(t_k^{(i)}, t_l^{(i)}) =$

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

$\sigma_c^2 \exp\{-\frac{1}{2\ell_c}(t_k^{(i)} - t_l^{(i)})^2\}$. To incorporate an MRF structure, we use the prior in (2.6) for $\mathcal{C}$ by choosing P to be an MFM prior and setting $l(\theta_c) = 1$ and $k(c \mid \mathcal{G}_N) = \exp\{\lambda E_c\}$ for M, where λ is a tuning parameter and E_c denotes the number of edges among the locations assigned to cluster c. Furthermore, a_0, a_1, a_2, b_0, b_1, and b_2 are hyperparameters in their associated gamma distributions. These yield the MRF-EPPF prior defined in Theorem 2. For simplicity, we refer this prior as the MRF-MFM prior in the rest of this paper. Note that the choice of $k(c \mid \mathcal{G}_N) = \exp\{\lambda E_c\}$ satisfies (2.5), and the corresponding $k_i(\cdot \mid \mathcal{G}_N)$ function coincides with the conditional cost function defined in Zhao et al. (2020). The partition probability function induced by the MRF-MFM prior, denoted by $p_\lambda(\mathcal{C} \mid \mathcal{G}_N)$, is equal to

$$p_\lambda(\mathcal{C} \mid \mathcal{G}_N) = \frac{V_N(|\mathcal{C}|)\Pi_{c\in\mathcal{C}}\gamma^{(|c|)} \exp\{\lambda E_c\}}{\sum_{\mathcal{C}'\in\mathcal{P}} V_N(|\mathcal{C}'|)\Pi_{c\in\mathcal{C}'}\gamma^{(|c|)} \exp\{\lambda E_c\}}, \tag{2.9}$$

where $\mathcal{P}$ is the set of all possible partitions of $[N]$. As discussed in Miller and Harrison (2018), γ is the parameter of the symmetric Dirichlet distribution defined in the MFM prior, and $V_N(t) = \sum_{k=1}^{\infty} \frac{k_{(t)}}{(\gamma k)^{(N)}} p_K(k)$, with $x^{(m)} = \Gamma(x+m)/\Gamma(x)$, $x_{(m)} = \Gamma(x+1)/\Gamma(x-m+1)$, and $x^{(0)} = x_{(0)} = 1$.

In practice, we let $\lambda \geq 0$, with a larger value of λ representing a higher spatial correlation. When $\lambda = 0$, $p_\lambda(\mathcal{C} \mid \mathcal{G}_N)$ can recover the partition probability function induced by the MFM prior without any spatial correlation between locations. When $\lambda \to \infty$, it degenerates to the Dirac delta func-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

tion $\delta_{[N]}$, that is, there is only one cluster. The term $\exp\{\lambda E_c\}$ changes the prior's preference on different partitions, and the prior mass concentrates on those partitions with more within-cluster edges. Therefore, λ is referred to as the *spatial smoothness parameter* in Zhao et al. (2020).

Because the partition probability function of an MFM is an EPPF, the closed-form full conditional distribution for our model can be conveniently obtained using Theorem 2, as shown in the following lemma. We omit the proof here because it is based on a very similar calculation to that of Theorem 2.1 in Zhao et al. (2020).

Lemma 1. *For model* (2.8)*, the conditional distribution of* θ_i *given* $\theta_{(-i)}$*, the induced partition* $\mathcal{C}_i$*, and the distinct value* $\{\theta_c\}_{c\in\mathcal{C}_i}$ *is proportional to*

$$\frac{V_N(|\mathcal{C}_i|+1)}{V_N(|\mathcal{C}_i|)}G_0 + \sum_{c\in\mathcal{C}_i}\exp\{\lambda\sum_{j\in c\cap\partial(i)}\mathbf{1}(\theta_j=\theta_i)\}(|c|+\gamma)\delta_{\theta_c}. \tag{2.10}$$

3. Theoretical Properties

In this section, we investigate the asymptotic property of the proposed method and show a clustering consistency result. Note that the asymptotics in our model refers to the situation in which the number of spatial locations N is fixed and the number of observed time points, denoted by n_i for location i, goes to infinity. Then, our clustering consistency result provides a useful justification for our method in the sense that as we collect more

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

data over time for each spatial location, the proposed method correctly identifies the true unknown clustering structure with posterior probability tending to one.

We first introduce some notation. Let $\mathcal{C}_0$ be the true unknown partition (clustering) structure, $\mathcal{P}_0 = \{\mathcal{C}_0\}$, and $\mathcal{P}$ be the collection of all partitions of $[N]$. Let $\mathcal{P}_1$ be the collection of over-clustering partitions, that is, $\mathcal{P}_1 = \{\mathcal{C}_1 : \mathcal{C}_1 \neq \mathcal{C}_0 \text{ and } \forall c' \in \mathcal{C}_1, \exists\, c \in \mathcal{C}_0, \text{s.t.},\ c' \subseteq c\}$, and $\mathcal{P}_2 = \mathcal{P} \setminus (\mathcal{P}_0 \cup \mathcal{P}_1)$ be the collection of mis-clustering partitions. We focus on the model defined in (2.2), and denote the response and covariates for location i by $\{Y_i, X_i\}$. Let $BF_{\mathcal{C},\mathcal{C}_0} = \Pi_{c\in\mathcal{C}} m(Y_c \mid X_c)/\Pi_{c\in\mathcal{C}_0} m(Y_c \mid X_c)$ be the Bayes factor by comparing the regression models given partition $\mathcal{C}$ with the true model $\mathcal{C}_0$, where $m(Y_c \mid X_c)$ is the conditional marginal likelihood of $\{Y_i, X_i\}$ for all $i \in c$. Furthermore, for any partition probability function $p(\mathcal{C})$, we consider its MRF-constrained version modified by the joint cost function introduced in (2.8), namely,

$$p_\lambda(\mathcal{C} \mid \mathcal{G}_N) \propto p(\mathcal{C}) \Pi_{c\in\mathcal{C}} \exp\{\lambda E_c\}. \tag{3.1}$$

Define $p_{\max} = \max_{\mathcal{C}\in\mathcal{P}} p(\mathcal{C})$ and let $E_{\max}$ be the total number of edges among these N locations. Note that both $E_{\max}$ and $p_{\max}$ are finite, because the location number N is finite. Let $n_{\min} = \min\{n_1, \ldots, n_N\}$. We make the following assumptions:

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

(A0) No isolated island: for every $c \in \mathcal{C}_0$, we assume that $|\partial(i) \cap c| \geq 1$ for every $i \in c$.

(A1) Model identifiability: we assume θ and θ' are within the support of G_0. Moreover, it holds that $f_\theta(y \mid x) = f_{\theta'}(y \mid x)$ for any y and x in their domain implies that $\theta = \theta'$.

(A2) Control the mis-clustering partitions: for every $\mathcal{C} \in \mathcal{P}_2$, there exists a sequence of numbers $q_{\mathcal{C}}(n_{\min})$ such that $BF_{\mathcal{C},\mathcal{C}_0} = o_p(q_{\mathcal{C}}(n_{\min}))$ and $q_{\mathcal{C}}(n_{\min}) \to 0$ as $n_{\min} \to \infty$.

(A3) Control the over-clustering partitions: $BF_{\mathcal{C},\mathcal{C}_0} = O_p(1)$ as $n_{\min} \to \infty$, for every $\mathcal{C} \in \mathcal{P}_1$,

(B3) Control the over-clustering partitions: $BF_{\mathcal{C},\mathcal{C}_0} \xrightarrow{p} 0$ as $n_{\min} \to \infty$, for every $\mathcal{C} \in \mathcal{P}_1$.

Selecting a clustering partition structure can be viewed as a model section problem. Under the Bayesian framework, correctly identifying the true model usually requires that the Bayes factor between the true and incorrect models converges to zero, which is why Assumptions (A2) and (B3) are needed. In particular, (A2) can be interpreted as only needing to find an upper bound $q_{\mathcal{C}}(n_{\min})$ for the contraction rate of the Bayes factor between the true and the incorrect models, which is a reasonable assumption, because the true model usually has a faster posterior contraction rate than that

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

of an incorrect model if the Bayes factor is consistent (Chib and Kuffner, 2016).

Assumption (A1) is needed for a consistent estimation of the cluster-wise parameters. This assumption is satisfied for generalized linear models with identity, logarithm, and logistic link functions. A proof is given in the Supplementary Material, Section S5. Assumptions (A0) and (A3) are alternatives replacements for (B3) that allow a weaker rate condition on the Bayes factor between the true and the over-clustered models. Assumption (A1) is needed for model identifiability, and is satisfied for many regression problems. Let $\Pi(\cdot \mid \mathcal{G}_N, \{Y_i, X_i\}_{i=1}^N)$ be the posterior distribution given the collected data and the spatial graphic information. Then, we can state the following clustering consistency theorem.

Theorem 3. *Consider model* (2.1) *with independent samples (i.e., no temporal correlation across different time points) and the prior* $p_\lambda(\mathcal{C} \mid \mathcal{G}_N)$ *specified in* (3.1). *Assume that* $p(\mathcal{C}_0) > 0$, *and that Assumptions (A1), (A2), and (B3) hold. Then, for any* $\lambda \geq 0$, *we have*

$$\Pi(\mathcal{C} = \mathcal{C}_0 \mid \mathcal{G}_N, \{Y_i, X_i\}_{i=1}^N) \xrightarrow{p} 1, \quad \text{as } n_{\min} \to \infty. \tag{3.2}$$

If (A0) and (A3) hold instead of (B3), then there exists a sequence of numbers $\lambda_{n_{\min}} \to \infty$ *as* $n_{\min} \to \infty$ *such that* (3.2) *holds.*

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Theorem 3 implies that if the Bayes factor is consistent (for a definition, see Chib and Kuffner (2016)) and the true model contracts at a faster rate than that of the over-clustered model, then for any partition probability function that assigns a positive probability to the true partition, the weak consistency of clustering holds. Moreover, if spatial information is available, we can achieve the same clustering consistency result, even when the true model contracts at the same rate as that of the over-fitted model. This finding provides a theoretical explanation of the advantage (in terms of weaker conditions needed to obtain the same clustering consistency result) by appropriately modeling the spatial information.

In the next theorem, we choose $f_{\theta_c}(y \mid x)$ as the linear regression model and show that clustering consistency can be obtained under weaker conditions by applying Theorem 3. The proof is given in the Supplementary Material.

Theorem 4. *Consider the following linear regression model:*

$$\begin{aligned} &Y_i(t_j^{(i)}) \mid X_i(t_j^{(i)}) \overset{ind}{\sim} \mathcal{N}(X_i(t_j^{(i)})\beta_c, \sigma_c^2), \text{ for } j = 1, \ldots, n_i,\ i \in c, \\ &\beta_c \mid \sigma_c^2 \sim \mathcal{N}(\mu_0, \sigma_c^2 \Lambda_0^{-1}), \sigma_c^2 \sim IG(\text{shape} = a_0, \text{rate} = b_0), \text{ for } c \in \mathcal{C}, \\ &\mathcal{C} \sim p_\lambda(\mathcal{C} \mid \mathcal{G}_N). \end{aligned} \tag{3.3}$$

Under additional assumptions on the design matrix, listed in Section 1.4 of the Supplementary Material, Assumptions (A1), (A2), and (B3) hold. As

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

a result, as $n_{\min} \to \infty$, $\Pi(\mathcal{C} = \mathcal{C}_0 \mid \mathcal{G}_n, \{Y_i, X_i\}_{i=1}^N) \xrightarrow{p} 1$.

Based on the clustering consistency result in Theorems 3 and 4, we can also obtain the usual posterior consistency result for the regression parameters within each cluster. Note that model (3.3) considered in Theorem 4 is slightly different to (2.8), because of the extra Gaussian process structure in $e_i(t_j^{(i)})$ that accounts for the temporal random effect. However, the clustering consistency results can still shed light on model (2.8), especially when σ_c and l_c take small values, for example, model (2.8) becomes equivalent to (3.3) if $l_c = 0$. In future work, we will extend the results in Theorems 3 and 4 by allowing for temporal random effects.

4. Bayesian Inference

We refer to Algorithm 8 of Neal (2000) by letting $\phi_c = \{\alpha_c, \ell_c\}$, the posterior distribution of which is intractable. The detailed algorithm is provided in Section S1 of the Supplementary Material.

Next, we decide on the value of the spatial smoothness parameter λ, which plays an important role in our model. It is common to use the marginal likelihood function as a selection criterion. However, its value is intractable for many Bayesian complex models, including the one we study here. Several posterior sampling-based approaches have been proposed in the literature, such as logarithm of the pseudo-marginal likelihood (LPML)

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

(Lewis et al., 2014) and the marginal likelihood computed using the harmonic mean (Newton and Raftery, 1994). However, these methods usually suffer from the pseudo-bias issue (Lenk, 2009), which tends to prefer more complex models. The sequential importance sampling method (Basu and Chib, 2003) is another popular approach for marginal likelihood estimation, but it cannot be applied to an MRF-PPM because the Pólya urn scheme is not available.

Our solution is to consider a prior sampling-based approach to estimate the marginal likelihood as follows:

$$\widehat{m}(Y \mid X) = \frac{1}{M - M'} \sum_{k=(M'+1)}^{M} \Pi_{c \in \mathcal{C}_{(k)}} f(Y_c \mid X_c, \alpha_k, \ell_k), \tag{4.1}$$

where $\mathcal{C}_{(k)}$, α_k, and ℓ_k are the associated parameters sampled from the partition probability function at the kth iteration, and $f(Y_c \mid X_c, \alpha_k, \ell_k)$ is the likelihood function after integrating out (β, σ^2) on its prior. More specifically, $\mathcal{C}_{(k)}$ at each iteration is sampled using the Gibbs sampler defined by (2.10). To account for the potential high level of variation in the prior sampling estimate, we follow the suggestion in Basu and Chib (2003), and let $\mathcal{C}_{(0)}$ (the initial partition) be equal to the last sample from our algorithm, with $n_{iter} = 1000$. The first 500 iterations are burn-in iterations, $M = 10^6$, $M' = 10^4$ (burn-in procedure), and we set the same random seed for different λ values. In both the simulation and the real-data analysis, we

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

choose λ from $\{0, 0.1, \ldots, 1\}$. We find that this range works quite well, because the selected optimal λ is always inside $(0, 1)$.

5. Simulation

In this section, we compare the empirical performance of our MRF-MFM model with that of four approaches: including two Bayesian methods that do not account for spatial correlation, namely, the DP and MFM, and the two frequentist methods in Lin and Ng (2012) and Su et al. (2016). In our numerical analysis, we use Dahl's method (Dahl, 2006) to summarize the posterior samples and obtain a deterministic result for both the cluster assignment and the cluster-wise parameter. The Rand index (RI; Rand (1971)) is used as a metric to evaluate the discrepancies between partitions. All computations are performed on 10 servers. Each server has 94.24 GB RAM and 24 processing cores. We distributed our simulation tasks to 100 workers (10 cores for each server). It took approximately 20 hours to finish each simulation scenario with 100 Monte Carlo replications, including the tuning procedure.

5.1 Simulation Setting

To generate the simulation data, we consider two partition scenarios (see Figure 1) with 48 states in the United States, excluding Hawaii, Alaska, and the District of Columbia. Both partition scenarios indicate strong spa-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

tial correlation, because most individual units assigned to the same cluster are spatially contiguous. The main difference between these two partition settings is that the first partition is more complex, because it allows two spatially noncontiguous blocks to belong to the same cluster. For each

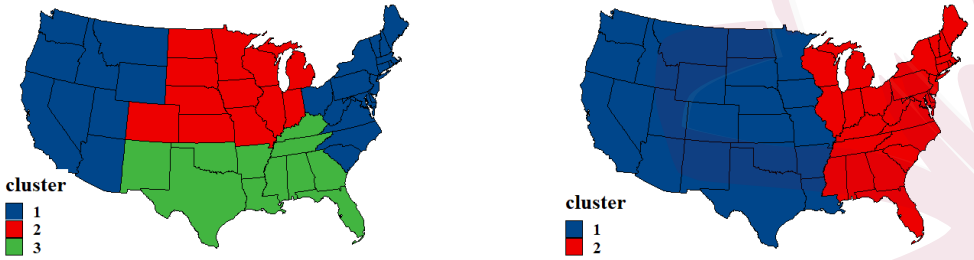

Figure 1: Simulation partition scenarios 1 and 2

partition scenario, we generate data from the following model:

$$
\begin{aligned}
&Y_i(t_j^{(i)}) \mid \left\{ e_i(t_j^{(i)}), X_i(t_j^{(i)}) \right\} \overset{\text{ind}}{\sim} \mathcal{N}(X_i(t_j^{(i)})\beta_{c_i} + e_i(t), \sigma^2_{c_i} \cdot \alpha_{c_i}), \\
&e_i(t_j^{(i)}) \sim \mathcal{N}(0, K_{\sigma^2_{c_i}, \ell_{c_i}}(\cdot, \cdot)), \text{ for } j = 1, \cdots, n_i,\ i \in c,
\end{aligned}
\tag{5.1}
$$

where $K_{\sigma_c^2, \ell_c}$ is the squared exponential kernel, as defined in (2.8), $X_i = [\mathbf{1}, \mathbf{x}_i]$, with each entry in $\mathbf{x}_i$ independently sampled from Unif$(-5, 5)$ and $\{t_i\}_{i=1}^{20}$ equally spaced in $[-1, 1]$. We consider eight data-generating processes (DGPs) with different cluster-wise parameters; see in Section S1 of the Supplementary Material.

DGPs 1, 2, 5, and 6 are for partition scenario 1, and the other four are for scenario 2. Because of the different variance (σ_i^2, for $i = 1, 2, 3$)

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

magnitudes of the random error, we call DGPs 1, 3, 5, and 7 the strong noise design, and DGPs 2, 4, 6, and 8 the weak noise design.

In both the simulation and the real-data analysis, we set $\gamma = 1$ and $p_K(\cdot) = \frac{10^{k-1}e^{-10}}{(k-1)!}$, which corresponds to a Poisson(10) distribution truncated to positive integers. Empirically, the MFM prior with this parameter setting tends to slightly over cluster locations, with the cluster size evenly distributed. We set the hyperparameters as $\mu_0 = \mathbf{0}_{p\times 1}$, $\Lambda_0 = 10^{-6} \cdot \mathrm{I}_{p\times p}$, $a_0 = 0.1$, and $b_0 = 1$. The results of our numerical studies are not sensitive to the choices of these values. In addition, we set $a_1 = a_2 = 2$ and $b_1 = b_2 = 1$ to encourage α and ℓ to concentrate around small values.

5.2 Results

We conduct 100 Monte Carlo replications for eight DGPs, and summarize the mean and the median of the RI obtained by comparing the partition from Dahl's estimate with the ground truth. For the Bayesian methods without spatial smoothness, we let the concentration parameter $\alpha = 1$ for the DP, and set the parameters of the MFM to be the same as those of the MRF-MFM in Section 5.1. For the method in Lin and Ng (2012), we follow their default settings, which assume that the number of clusters $|\mathcal{C}|$ is within $\{2, 3, 4\}$, and select $|\mathcal{C}|$ using the BIC. The partition is determined using the conditional K-means (CK-means) criterion, which the authors note is more robust than the other methods when $n_{\min}$ is small. For the method

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

in Su et al. (2016), we use the penalized least squares approach (PLS) to fit the model, and follow their default settings, which assume that $|\mathcal{C}|$ is within $\{1, \ldots, 5\}$. Because both frequentist models can only discriminate latent groups when they have different slopes, they are only implemented for DGPs 5–8.

Table 1: Median (mean) of the random index over 100 Monte Carlo replications for our MRF-PPM method and four competing methods, MFM, DP, CK-means (Lin and Ng, 2012), and PLS (Su et al., 2016).

DGP	MRF-MFM	MFM	DP	CK-means	PLS
1	**0.973** (**0.924**)	0.879 (0.886)	0.909 (0.894)	-	-
2	**1.000** (**0.929**)	0.968 (0.928)	0.969 (0.926)	-	-
3	**1.000** (**0.981**)	0.915 (0.917)	0.938 (0.934)	-	-
4	**1.000** (**0.990**)	0.979 (0.954)	1.000 (0.966)	-	-
5	**0.914** (**0.916**)	0.889 (0.901)	0.911 (0.909)	0.835 (0.837)	0.373 (0.373)
6	**1.000** (**0.982**)	1.000 (0.975)	1.000 (0.979)	0.969 (0.973)	0.373 (0.373)
7	**0.949** (**0.934**)	0.917 (0.915)	0.936 (0.929)	0.880 (0.881)	0.493 (0.493)
8	**1.000** (**0.984**)	1.000 (0.970)	1.000 (0.975)	0.880 (0.880)	0.493 (0.493)

Table 2: Median (standard error) ℓ_1 error of the regression coefficient estimates over 100 Monte Carlo replications for three Bayesian methods.

DGP	MRF-MFM	MFM	DP
1	**1.78** (3.29)	2.46 (1.88)	2.61 (2.35)
2	**1.64** (3.58)	2.23 (2.35)	2.21 (2.93)
3	**1.44** (0.80)	1.79 (1.15)	1.83 (0.99)
4	**1.31** (0.76)	1.83 (1.00)	1.43 (1.04)
5	**2.21** (0.79)	2.28 (1.01)	2.22 (0.91)
6	1.56 (0.89)	**1.54** (0.90)	1.62 (0.95)
7	**1.60** (1.00)	1.84 (1.11)	1.90 (1.03)
8	**1.23** (0.86)	1.65 (0.94)	1.27 (1.01)

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

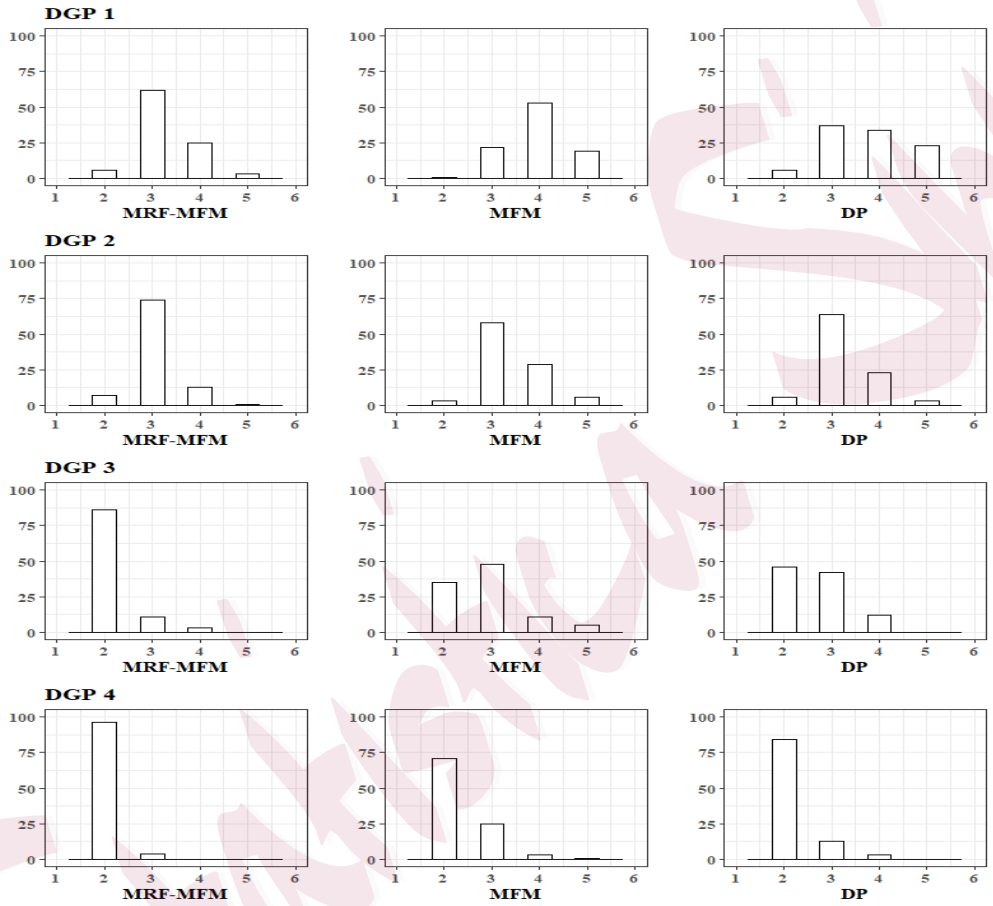

Figure 2: Histograms for the number of clusters selected by the MRF-MFM and four competing methods (DGPs 1–4).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

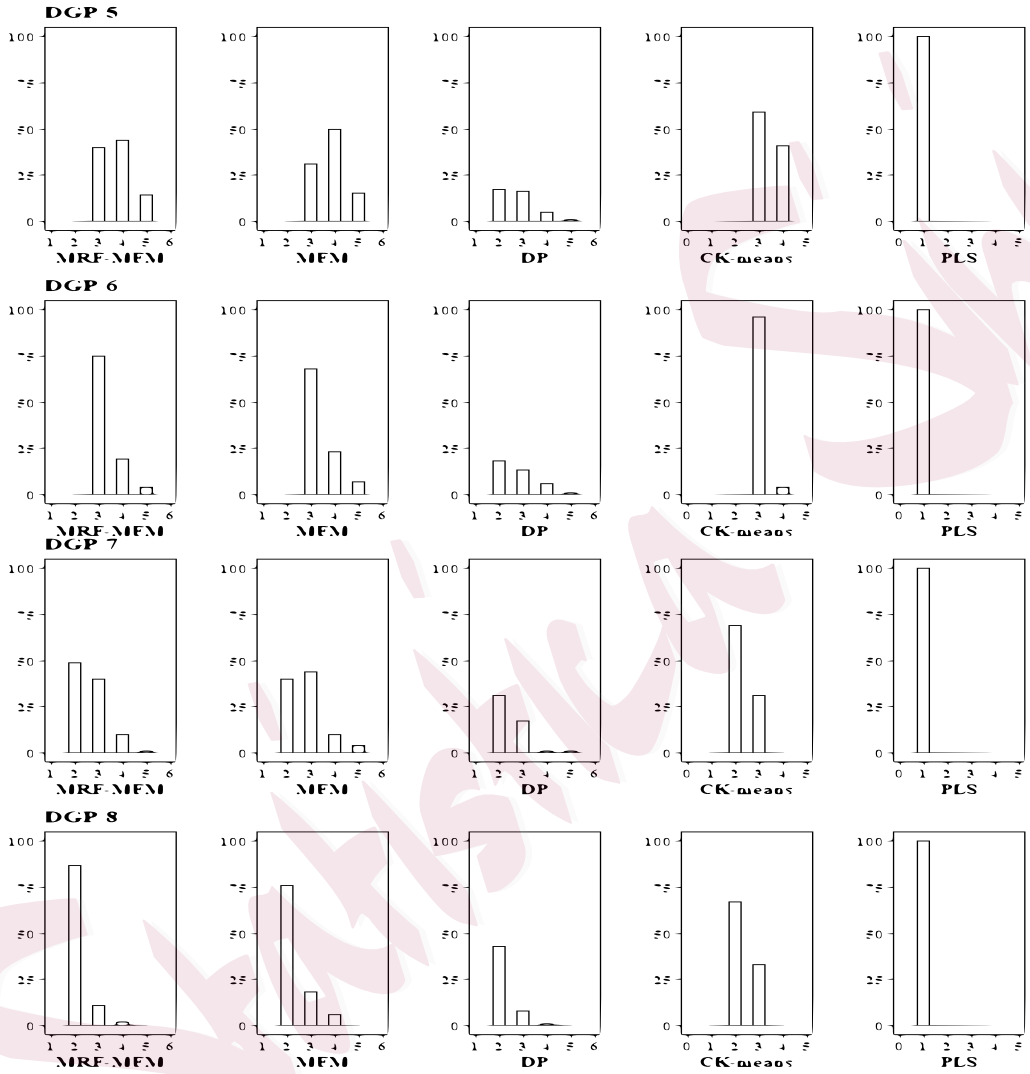

Figure 3: Histograms for the number of clusters selected by the MRF-MFM and four competing methods (DGPs 5–8).

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

The simulation results are summarized in Table 1, Table 2, Figure 2, and Figure 3. In Table 1, we find that the proposed MRF-MRF performs consistently better than the other four methods in terms of having a higher RI value under all scenarios, which confirms the benefit of appropriately incorporating the spatial correlation across locations. All three Bayesian methods provide a more accurate clustering partition result than those of the two frequentist methods, because the Bayesian methods correctly specify the covariance structure. In general, when the noise level is high (DGPs 1,3,5,7) or the true partition structure becomes more complex (DGPs 1,2,5,6), the RI becomes lower, as expected. To evaluate the parameter estimation accuracy, we compute the ℓ_1 error for the estimated regression coefficients, defined by $\frac{1}{N}\sum_{i=1}^{N}\left\|\hat{\beta}_i - \beta_i\right\|_1$, where β_i is the set of true regression coefficients for location i, and $\hat{\beta}_i$ is the corresponding estimate. In Table 2, we summarize the average (median) ℓ_1 error for the MRF-MFM and two other Bayesian methods. We find that our method has a smaller coefficient estimation error than those of the two Bayesian methods in most scenarios. For the first two DGPs, the MFM has a smaller average ℓ_1 error than that of our method, although the advantage is minor. Among the eight DGPs, DGP 5 is the most challenging case because the cluster-wise regression parameters are less separable compared with those in the other DGPs. This is also reflected by the lowest average random index in Table 1 and the highest estimation error in Table 2.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

We also compare the CK-means method with the PLS method in Table 1, and find that PLS, in general, cannot accurately identify the latent group structure in this simulation because the generated data have a strong temporal correlation at each location, owing to large ℓ and small α values. This creates trouble for the PLS method (Su et al., 2016), which uses a z-transformation on both Y and $\mathbf{x}$, and the estimation of the slope becomes equivalent to estimating the correlation coefficient. As a result, the difference between the slopes of the clusters cannot be fully captured using their method, because the correlation coefficient is close to one for all clusters, owing to the high serial correlation. On the other hand, the CK-means method is more robust because it is distance based. Similar findings are observed in Figure 2 and Figure 3, where we show histograms for the selected number of clusters. In general, the MRF-MFM performs well in terms of selecting the correct number of clusters for all scenarios. When the partition structure is complex (e.g., DGPs 1,2,5,6), both the DP and the MFM tend to overestimate the number of clusters, as expected, because they do not account for the spatial correlation between locations. We also conducted a sensitivity analysis for the choices of the covariance kernel function and the hyperparameter values (α and γ). Our results, summarized in Section S6 of the Supplementary Material (both clustering and parameter estimation), are quite stable under different covariance kernels and hyperparameter values.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

6. Real-Data Analysis

We present two real-data applications to demonstrate our proposed methodology. In the data analysis, we choose the spatial smoothness parameter λ from $\{0, 0.1, \ldots, 1\}$ based on the criteria described in Section 4.

6.1 Precipitation Data Analysis

We first consider the annual precipitation and average temperature data available at `https://www.ncdc.noaa.gov/cag/statewide/mapping/110/pcp/201812/12/value`, collected for 48 states (excluding Washington, D.C., Alaska, and Hawaii) for the period 2000 to 2019. The main goal is to study the relationship between annual precipitation and average temperature, and to understand its heterogeneity in different states. It is well known that precipitation is strongly associated with convection, which is influenced by topography (Parsons and Daly, 1983). To account for the spatial heterogeneity, we apply the model in (2.8), treat each state as a spatial location i, rescale the years from 2000 to 2019 onto equally spaced points between $[-1, 1]$, and let Y_i be a 20-by-1 response vector of the annual precipitation and X_i be a 20-by-2 matrix that includes an intercept term and the average temperature as another covariate.

Based on the estimated marginal likelihood, we find the optimal value for λ is $\lambda = 0.1$. We run 10,000 MCMC iterations and discard the first

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

5,000 as a burn-in. The final partition is obtained using Dahl's method. The average RI between the reporting partition and the 100 replications is 0.9362, which indicates that the final partition is representative.

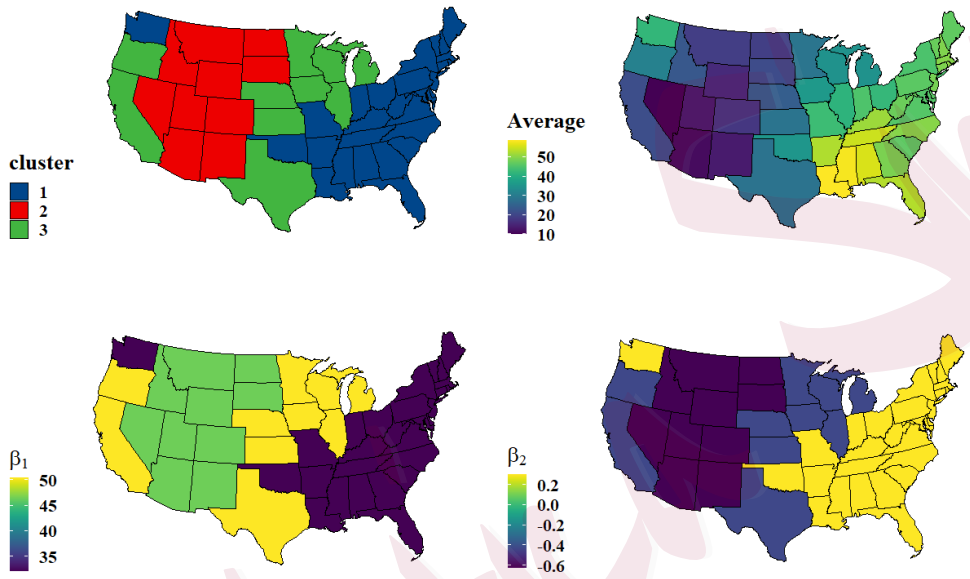

Figure 4: From top to bottom: (a) the estimated partition; (b) the average annual precipitation map; (c) estimated intercepts; (d) estimated slopes for the annual temperature.

We summarize the results in Figure 4. We find that, in general, the estimated partition in (a) matches the pattern observed in the average annual precipitation map in (b) quite well. More specifically, the first cluster contains most states with humid continental and humid subtropical climate types, which usually receive plenty of rainfall annually. On the other hand, the climate types of most states in the second cluster are desert and

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Table 3: Cluster-wise parameter estimates (standard deviation) for the precipitation data.

Cluster	$\hat{\beta}_{\text{intercept}}$	$\hat{\beta}_{\text{tempreature}}$	$\hat{\sigma^2}$	$\hat{\ell}$	$\hat{\alpha}$
1	40.78 (5.99)	0.12 (0.11)	11.91 (6.99)	1.83 (1.65)	4.20 (1.21)
2	47.94 (4.60)	-0.65 (0.10)	6.05 (2.82)	5.27 (1.32)	1.11 (0.39)
3	56.77 (7.54)	-0.53 (0.14)	19.00 (7.27)	5.09 (1.36)	0.88 (0.65)

semi-arid, which naturally associate with a low level of precipitation.

In Table 3, we summarize the estimated regression parameters for each of the three clusters. The results clearly demonstrate a high level of heterogeneity in both the regression coefficients and the variance parameter over the three clusters, which again highlights the benefit and necessity of considering the heterogeneity in spatial panel data. Scientifically, the precipitation mechanism is a complex system, and is known to be more relevant to some other factors, such as the vertical thermal gradient and wind speed. Therefore, for the first and third cluster, $\hat{\sigma}^2$ is considerably large, demonstrating that there may be other latent confounders that are not accounted for in our model.

In our study, we interpret the predictor "average annual temperature" as a hybrid indicator. For example, for the second cluster, the annual temperature seems to indicate aridness, in the sense that a high level of aridness, which is usually implied by a higher annual temperature, usually leads to less annual precipitation. We also implement the MFM prior (without the spatial consideration) and present the results in Table 1 and Figure 1 of

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

the Supplementary Material. By comparing the estimated partition maps obtained using our method and the MFM, we find that our partition map is spatially more "smooth," which naturally allows an easier interpretation.

6.2 Median Household Income Data Analysis

Next, we analyze a California State county-level household income data set, available at `https://www.countyhealthrankings.org/app/`. The data consist of annual measurements of median household income, total gross domestic product (GDP), and the unemployment rate between 2011 and 2018. We conduct a regression analysis of the median income on the GDP and unemployment rate, and study the heterogeneity pattern in the regression parameters across counties. Before applying our method, we perform a logit transformation on the unemployment rate and a z-transform on the median income and the GDP.

We apply our proposed method under the same setting as that described in Section 6.1. The spatial smoothness parameter is selected as $\lambda = 0.1$, based on the maximum marginal likelihood. The average RI between the final cluster assignment and those from 100 replications is 0.9114, which confirms that the final cluster partition is representative. We present the clustering map in Figure 5 and summarize the regression parameter estimates for each of the three clusters in Table 4. Here, we find a uniform pattern in which the annual household median income is negatively associ-

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

ated with the unemployment rate and positively associated with the GDP in all three clusters, which agrees with common sense. Of the three clusters, Cluster 1 (see Figure 5) has the strongest negative association between unemployment and median income; most of the counties in the Bay Area (including Santa Clara and San Mateo) belong to this cluster. For Cluster 3, in which GDP has the lowest impact on household income, most counties are blue counties (Democrat votes $\geq 60\%$ during the 2020 presidential election), including Napa, Sonoma, Yolo, Los Angeles, San Diego, and Imperial. These results suggest that political opinions and industrial structure may be potential confounders that can be included in future analyses. We also implement the MFM prior (without the spatial consideration) and present the results in Table 2 and Figure 2 of the Supplementary Material. By comparing the estimated regression coefficients obtained from our method and the MFM, we find that our method is better at differentiating between the three clusters in terms of the estimated regression coefficients, for example, the estimated coefficient for DGP is more distinct across the three clusters in our results.

Table 4: Cluster-wise parameter estimates (standard deviation) for the income data.

Cluster	$\hat{\beta}_{\text{intercept}}$	$\hat{\beta}_{\text{GDP}}$	$\hat{\beta}_{\text{unemployment}}$	$\hat{\sigma^2}$	$\hat{\ell}$	$\hat{\alpha}$
1	-1.76 (0.39)	0.80 (0.16)	-1.29 (0.15)	0.25 (0.01)	1.79 (0.06)	0.08 (0.01)
2	-1.20 (0.20)	1.17 (0.12)	-0.37 (0.07)	0.12 (0.01)	2.36 (0.11)	0.04 (0.002)
3	-1.45 (0.16)	0.02 (0.05)	-0.65 (0.06)	0.15 (0.01)	8.28 (0.29)	0.11 (0.01)

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

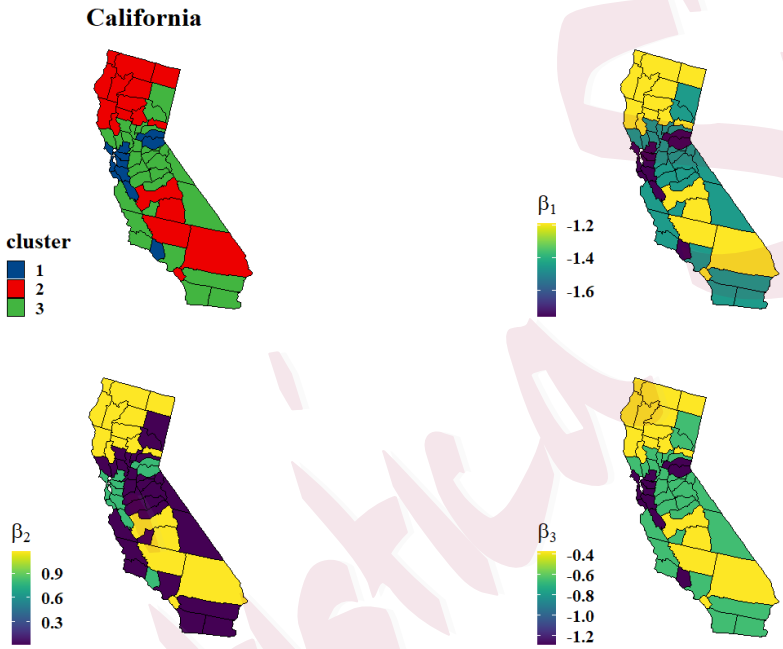

Figure 5: From top to bottom: (a) the estimated partition; (b) visualized intercepts; (c) visualized slopes for GDP; (d) visualized slopes for log odds of the unemployment rate.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

7. Discussion

We have proposed a general Bayesian spatial clustering method based on a PPM equipped with an MRF structure for panel data analyses. We have studied the fundamental properties of the MRF-PPM, and proved a clustering consistency result under mild conditions on the MRF structure. We have also introduced a computationally tractable MCMC algorithm and a model selection method based on the marginal likelihood. Numerical studies confirm that the MRF-PPM effectively avoids the over-clustering issue and is more robust than the classical PPM to model misspecification.

Several work directions remain open. First, it is challenging to study the asymptotic behavior of the MRF-PPM prior when $N \to \infty$, because Kolmogorov's extension theorem does not hold, in general, after accounting for spatial information. It would be of interest to prove a Bayesian clustering consistency result when $N \to \infty$, as obtained in Su et al. (2016) and Bonhomme and Manresa (2015) for their frequentist approaches. Second, we assume no temporal correlation between $Y_i(t_j^{(i)})$ and $Y_i(t_k^{(i)})$, for $j \neq k$, when proving Theorem 3 for general regression models. Relaxing this assumption is of interest. Our prior can also be extended to allow more generic forms of the regression functions, such as nonparametric or semiparametric models. Developing an efficient posterior computation and understanding the theoretical properties of this prior are left to future work.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Supplementary Materials

The online supplementary material contains a detailed description of our proposed algorithm, proof of main theorems, derivation of the full conditional distribution and the marginal likelihood introduced in Section 4, and additional numerical results for the simulation study and real data analysis.

Acknowledgments

The authors thank the editor, associate editor, and reviewers for their valuable comments and suggestions.

References

Basu, S. and Chib, S. (2003). Marginal likelihood and Bayes factors for Dirichlet process mixture models. *Journal of the American Statistical Association*, 98(461):224–235.

Belotti, F., Hughes, G., and Mortari, A. P. (2017). Spatial panel-data models using stata. *The Stata Journal*, 17(1):139–180.

Blake, A., Kohli, P., and Rother, C. (2011). *Markov random fields for vision and image processing*. Mit Press.

Bonhomme, S. and Manresa, E. (2015). Grouped patterns of heterogeneity in panel data. *Econometrica*, 83(3):1147–1184.

Browning, M., Carro, J., et al. (2007). Heterogeneity and microeconometrics modeling. *Econometric Society Monographs*, 43:47.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Chib, S. and Kuffner, T. A. (2016). Bayes factor consistency. *arXiv preprint arXiv:1607.00292*.

Dahl, D. B. (2006). Model-based clustering for expression data via a Dirichlet process mixture model. *Bayesian inference for gene expression and proteomics*, 4:201–218.

De Finetti, B. (1929). Funzione caratteristica di un fenomeno aleatorio. In *Atti del Congresso Internazionale dei Matematici: Bologna del 3 al 10 de settembre di 1928*, pages 179–190.

Design, S. (1978). Fundamentals of a discipline of computer program and systems design.

Durrett, R. (2019). *Probability: theory and examples*, volume 49. Cambridge university press.

Elhorst, J. P. (2014). *Spatial econometrics: from cross-sectional data to spatial panels*, volume 479. Springer.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, (6):721–741.

Geng, L. and Hu, G. (2021). Bayesian spatial homogeneity pursuit for survival data with an application to the seer respiratory cancer data. *Biometrics*.

Green, P. J. and Richardson, S. (2001). Modelling heterogeneity with and without the Dirichlet process. *Scandinavian journal of statistics*, 28(2):355–375.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Hao, Y., Chen, H., Wei, Y.-M., and Li, Y.-M. (2016). The influence of climate change on CO2 (carbon dioxide) emissions: an empirical estimation based on Chinese provincial panel data. *Journal of cleaner production*, 131:667–677.

Hartigan, J. A. (1990). Partition models. *Communications in statistics-Theory and methods*, 19(8):2745–2756.

Hsiao, C. (2014). *Analysis of panel data*. Number 54. Cambridge university press.

Hsiao, C. and Tahmiscioglu, A. K. (1997). A panel analysis of liquidity constraints and firm investment. *Journal of the American Statistical Association*, 92(438):455–465.

Hu, G., Geng, J., Xue, Y., and Sang, H. (2020). Bayesian spatial homogeneity pursuit of functional data: an application to the US income distribution. *arXiv preprint arXiv:2002.06663*.

Hu, G., Xue, Y., and Ma, Z. (2021). Bayesian clustered coefficients regression with auxiliary covariates assistant random effects. *Statistical Modelling*.

Lenk, P. (2009). Simulation pseudo-bias correction to the harmonic mean estimator of integrated likelihoods. *Journal of Computational and Graphical Statistics*, 18(4):941–960.

Lewis, P. O., Xie, W., Chen, M.-H., Fan, Y., and Kuo, L. (2014). Posterior predictive Bayesian phylogenetic model selection. *Systematic biology*, 63(3):309–321.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Lin, C.-C. and Ng, S. (2012). Estimation of panel data models with parameter heterogeneity when group membership is unknown. *Journal of Econometric Methods*, 1(1):42–55.

Ma, Z., Xue, Y., and Hu, G. (2020). Heterogeneous regression models for clusters of spatial dependent data. *Spatial Economic Analysis*, 15(4):459–475.

Miao, K., Su, L., and Wang, W. (2020). Panel threshold regressions with latent group structures. *Journal of Econometrics*, 214(2):451–481.

Miller, J. W. and Harrison, M. T. (2018). Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356.

Neal, R. M. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Newton, M. A. and Raftery, A. E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(1):3–26.

Orbanz, P. and Buhmann, J. M. (2008). Nonparametric Bayesian image segmentation. *International Journal of Computer Vision*, 77(1-3):25–45.

Page, G. L., Quintana, F. A., et al. (2015). Predictions based on the clustering of heterogeneous functions via shape and subject-specific covariates. *Bayesian Analysis*, 10(2):379–410.

Page, G. L., Quintana, F. A., et al. (2016). Spatial product partition models. *Bayesian Analysis*, 11(1):265–298.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Park, J.-H. and Dunson, D. B. (2010). Bayesian generalized product partition model. *Statistica Sinica*, pages 1203–1226.

Parsons, B. and Daly, S. (1983). The relationship between surface topography, gravity anomalies, and temperature structure of convection. *Journal of Geophysical Research: Solid Earth*, 88(B2):1129–1144.

Pesaran, M. H. (2015). *Time series and panel data econometrics*. Oxford University Press.

Pitman, J. et al. (2002). Combinatorial stochastic processes. Technical report, Technical Report 621, Dept. Statistics, UC Berkeley, 2002.

Quintana, F. A. and Iglesias, P. L. (2003). Bayesian clustering and product partition models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):557–574.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850.

Su, L. and Chen, Q. (2013). Testing homogeneity in panel data models with interactive fixed effects. *Econometric Theory*, pages 1079–1135.

Su, L., Shi, Z., and Phillips, P. C. (2016). Identifying latent structures in panel data. *Econometrica*, 84(6):2215–2264.

Su, L., Wang, X., and Jin, S. (2019). Sieve estimation of time-varying panel data models with latent structures. *Journal of Business & Economic Statistics*, 37(2):334–349.

Statistica Sinica: Preprint
doi:10.5705/ss.202021.0247

Teixeira, L. V., Assunção, R. M., and Loschi, R. H. (2019). Bayesian space-time partitioning by sampling and pruning spanning trees. *J. Mach. Learn. Res.*, 20:85–1.

Wagner, C. H. (1982). Simpson's paradox in real life. *The American Statistician*, 36(1):46–48.

Zhang, B. (2020). Forecasting with Bayesian Grouped random effects in panel data. *arXiv preprint arXiv:2007.02435*.

Zhao, P., Yang, H.-C., Dey, D. K., and Hu, G. (2020). Bayesian spatial homogeneity pursuit regression for count value data. *arXiv preprint arXiv:2002.06678*.

Department of Statistics, University of California, Irvine

E-mail: weinings@uci.edu

Department of Statistics, University of Missouri - Columbia, Columbia, MO, 65211

E-mail: gh7mr@missouri.edu