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
Notice: Accepted version subject to English editing.	

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 many applications such as economics and environmental science. In this paper, we focus on regression models for spatial panel data analysis, where repeated measurements are collected over time at various spatial locations. We propose a novel class of nonparametric priors that combines Markov random field (MRF) with the product partition model (PPM) and show that the resulting prior, called by MRF-PPM, is capable of identifying the latent group structures among the spatial locations while efficiently utilizing the spatial dependence information. We derive a closed-form conditional distribution for the proposed prior and introduce a new way to compute the marginal likelihood that renders efficient Bayesian inference. We further 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 via extensive simulation studies and applications to a US precipitation data and a California median household income data study.

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

1. Introduction

Panel data has been widely studied in many applications such as economy (Pesaran, 2015) and climate science (Hao et al., 2016) since it represents a common data format where observations are collected for each subject at different time points. Our focus in this paper is to model a special type of panel data, called by *spatial panel data*, where each subject represents a spatial location and there is a need to take account for the spatial dependence among those locations. Spatial panel data analysis has received a growing interest in recent years (Elhorst, 2014; Belotti et al., 2017) and a central question is to model the relationship between variables measured repeatedly over a study time period at various spatial locations. For example, in economic studies, it is of interest to quantify the association between the 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 naturally formulated 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., 2007; Su and Chen, 2013). To account for such spatial heterogeneity pattern, it is common to assume a *latent group structure*, i.e., spatial locations are grouped into clusters and those assigned to the same cluster share the same set of regression parameters. This strategy has several advantages in practice. First, for the obvious reason that neglecting the unobserved heterogeneity may lead to inconsistent parameter estimation and severally misleading results as demonstrated by the famous Simpson's paradox and other examples for spatial panel data (Wagner, 1982; Su et al., 2016; Hsiao, 2014). Secondly, the obtained latent group structure is usually informative for empirical analysis, such as finding possible unobserved confounders and performing secondary analysis. Another benefit is that the latent group structure allows a convenient way to incorporate the spatial dependence information in the model and in turn helps improve the accuracy/efficiency of model fit and interpretation (Miao et al., 2020).

Several approaches have been introduced in the frequentist literatures to study panel data regression model with latent group structure. For example, Lin and Ng (2012) considered a panel data linear regression model with group-varying slopes. This method was further extended to allow both group-varying intercept and slopes by Su et al. (2016). Several more complicated models have also been proposed with different features, e.g., 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 those frequentist approaches, limited effort has been made under the Bayesian framework until very recently (Zhang, 2020; Ma et al., 2020; Hu et al., 2021; Geng and Hu, 2021). Teixeira et al. (2019) has introduced a Bayesian spatio-temporal clustering method, but not suitable for clustering locations in panel data analysis. Conceptually, an ideal Bayesian approach would naturally be able to incorporate the spatial dependence and latent group structure information in the prior distribution. The inference may also be conveniently conducted without pursing complicated procedures such as bootstrap or post model selection. The main goal of our paper is to pursue under this direction by introducing a new class of nonparametric priors and exploring their computational and theoretical properties.

Our first step towards constructing a prior distribution for spatial panel data with 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, and this is usually achieved by specifying a class of partition probability functions. Among this class, the product partition model (PPM), which was first introduced by Hartigan (1990) and studied from a Bayesian point of view by Quintana and Iglesias (2003), has received a considerable interest over the years. PPM is defined by taking the product of some non-negative 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 Dirichlet Process prior (DP, Green and Richardson (2001)) and the Mixture of Finite Mixture Model prior (MFM, Miller and Harrison (2018)). Lately, PPM was extended to include covariates (Park and Dunson, 2010; Page et al., 2015) and spatial information (Page et al., 2016). But it still remains unclear how to *systematically* incorporate the spatial dependence information into the PPM.

To solve this issue, in this paper, we introduce a new class of priors called by a Markov random field constrained product partition model (MRF-PPM) prior. This prior is generated by taking the product of two priors, a Markov Random Field (MRF) prior and a PPM prior. There is a long history of using MRF priors defined on undirected graphs in the literature 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, MRF-PPM prior, as a general class of priors that combines MRF and PPM, has not been systemically studied in the literature in terms of its theoretical and computational properties; and it is our goal to fill this gap in this paper. 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 MRF-PPM. The clustering consistency result states that with posterior probability tending to one, the posterior distribution of MRF-PPM is capable of identifying the correct unknown partition structure in the spatial panel data. This result, to the best of our knowledge, is new in the Bayesian spatial panel model literature, and is generally applicable to regression models with well-defined posterior contraction rate under mild identifiability conditions.

2. Methodology

2.1 Markov Random Field Constrained 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, i.e., $i \in c$, 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 latent groups structure can be written in the following way,

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

where f_{θ_c} is the regression likelihood function for cluster c. For the rest of the paper, we also use $Y_i(t)$ to denote the observation collected at time t for location i for the simplicity of notation. Note that model (2.1) allows the temporal correlation between $Y_i(t_j^{(i)})$ and $Y_i(t_k^{(i)})$ for every $j \neq k$. To model the clustering structure, we consider a prior on the partition of the index set $[N] = \{1, \ldots, N\}$ and the associated parameters sets θ_c as follows,

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

where G_0 is a non-atomic base measure for θ_c with a 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 spatial correlation among different locations, we propose to incorporate a Markov random field (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. Recall that 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

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

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 θ_i's distribution only depends on its neighbors, i.e., vertices that are connected to θ_i. The graphical information $\mathcal{G}_N$ is usually determined by the network structure in real-world applications where vertices $V_{\mathcal{G}_N}$ represent subjects and edges $E_{\mathcal{G}_N}$ represent their relationships. Some examples include the 51 States and their adjacency matrix in the United States, different users and their friendship connection in social media network, and international airports and the airlines among them.

Inspired by the Markov property, we propose to 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 non-negative function, $k(\cdot \mid \mathcal{G}_N)$ and $k_i(\cdot \mid \mathcal{G}_N)$ are non-negative cohesion functions defined for every $c \subseteq [N]$ given the graphical information; and 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, since the cohesion value of $c \cup \{i\}$ is related to the joint density of $c \cup \{i\}$, while 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. 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 MRF-PPM, we proceed to let $P(\theta_1, \ldots, \theta_N)$ (P for short) be the prior on $\{\theta_1, \ldots, \theta_N\}$ defined in (2.2) and (2.3), which is proportional to $\Pi_{c \in \mathcal{C}} g(\theta_c) h(c)$. A MRF-PPM prior Π can hence be constructed by taking the product of P and the MRF cost function M with some positive normalizing constant K_0 as follows,

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

It can be shown that 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 equals to $k(\cdot \mid \mathcal{G}_N)h(\cdot)$ and a probability density function of base measure being $K_1 l(\cdot)g(\cdot)$ 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.

(P3) It is exchangeable since the cohesion function is invariant under permutation (it only depends 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 MRF-PPM prior in Theorem 1. The proof is given in the Supplementary File.

Theorem 1. *Suppose that $l(\theta)g(\theta)$ in MRF-PPM prior is integrable as a function of θ, then the conditional distribution of θ_i given $\theta_{(-i)}$, 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 $K_1 l(\theta)g(\theta)$.

Apparently, if $l(\theta) = 1$, we have the base measure $L_0 = G_0$. From the second term in (2.7), we can see that MRF-PPM is able to account for the spatial correlation since location i would have 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).

In addition to 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 i.i.d sampled from a base measure G_0, then the resulting MRF-EPPF satisfies Properties (P1)-(P3) and its full conditional distribution can be obtained similarly as in* (2.7).

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

It is also worthy mentioning 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 extra 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 consequence, the Kolmogorov's extension theorem (Durrett, 2019) cannot be directly applied to show the existence of $\Pi(\theta_1, \ldots, \theta_N \mid \mathcal{G}_N)$ as $N \to \infty$. For the same reason, the Pólya urn scheme is not available for MRF-MFM. However, this will not affect our method because our main focus is on the fixed N situation, that is, the number of spatial locations of interest is fixed in the study.

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, $K_{\sigma_c,\ell_c}(t_k^{(i)}, t_l^{(i)}) = \sigma_c^2 \exp\{-\frac{1}{2\ell_c}(t_k^{(i)} - t_l^{(i)})^2\}$. 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. Also, $a_0, a_1, a_2, b_0, b_1, b_2$ are hyperparameters in their associated gamma distributions. These yield an MRF-EPPF prior as defined in Theorem 2. For simplicity, we refer this prior as MRF-MFM for 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 $p_\lambda(\mathcal{C} \mid \mathcal{G}_N)$, equals 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 $x^{(m)} = \Gamma(x+m)/\Gamma(x)$, $x_{(m)} = \Gamma(x+1)/\Gamma(x-m+1)$, $x^{(0)} = x_{(0)} = 1$.

In practice, we let $\lambda \geq 0$, with a larger value of λ representing a higher spatial correlation. It can be seen that 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; and when $\lambda \to \infty$, it will degenerate to the Dirac delta function $\delta_{[N]}$, i.e., there is only one cluster. The term $\exp\{\lambda E_c\}$ will change the prior's preference on different partitions, and the prior mass will concentrate on those partitions with more within-cluster edges. Therefore, λ is referred as *spatial smoothness parameter* in Zhao et al. (2020).

As the partition probability function of MFM is an EPPF, the closed-form full conditional distribution for our model can be conveniently obtained by Theorem 2 in the following lemma. We omit the proof here because it is based on a very similar calculation with 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 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

We investigate the asymptotic property for the proposed method and show a clustering consistency result in this section. Note that the asymptotics in our model refers to the situation that 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 are collecting more data over time for each spatial location, the proposed method will be able to correctly identify the true unknown clustering structure with posterior probability tending to one.

We first introduce some notations. 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, i.e., $\mathcal{P}_1=\{\mathcal{C}_1:\mathcal{C}_1\neq\mathcal{C}_0,\ \text{and}\ \forall c'\in\mathcal{C}_1,\exists\ c\in\mathcal{C}_0,\text{s.t.},\ c'\subseteq c\}$. 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

$$p_\lambda(\mathcal{C} \mid \mathcal{G}_N) \propto p(\mathcal{C})\Pi_{c\in\mathcal{C}} \exp\{\lambda E_c\}. \tag{3.11}$$

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, since the location number N is finite. Let $n_{\min} = \min\{n_1, \ldots, n_N\}$. We make the following assumptions.

(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 $\theta = \theta'$.

(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 the Bayes factor between the true and incorrect models to converge to 0, which is why Assumptions (A2) and (B3) are needed. In particular, (A2) can be interpreted as, one only needs to find an upper bound $q_{\mathcal{C}}(n_{\min})$ for the contraction rate of the Bayes factor between the true and incorrect models, which is a reasonable assumption since the true model usually has a faster posterior contraction rate than that of an incorrect model if the Bayes Factor is consistent (Chib and Kuffner, 2016).

Assumption (A1) is needed for consistent estimation of the cluster-wise parameters. This assumption is satisfied for generalized linear models with a link function being identity, logarithm, and logistic. A proof is given in the Supplementary File, Section S5. Assumption (A0) and (A3) are alternative replacements for (B3) that allows a weaker rate condition on the Bayes factor between the true and over-clustered models. Assumption (A1) is needed for model identifiability purpose, and it 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 spatial graphic information, 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 a prior* $p_\lambda(\mathcal{C} \mid \mathcal{G}_N)$ *being specified in* (3.11). *Assume that* $p(\mathcal{C}_0) > 0$, *and Assumptions (A1), (A2), (B3) hold. Then for any* $\lambda \geq 0$, *we have*

$$\Pi(\mathcal{C} = \mathcal{C}_0 \mid \mathcal{G}_N, \{Y_i, X_i\}_{i=1}^N) \xrightarrow{p} 1, \quad \text{as } n_{\min} \to \infty. \tag{3.12}$$

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* (3.12) *holds.*

Theorem 3 implies that if the Bayes factor is consistent (for definition, see Chib and Kuffner (2016)) and the true model contracts at a faster rate than 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 the spatial information is available, then we can achieve the same clustering consistency result even when the true model contracts at the same rate with the over-fitted model. This finding provides a theoretical explanation of the advantage (in terms of weaker conditions needed for obtaining the same clustering consistency result) by

appropriately modeling the spatial information.

In 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 File.

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

$$\begin{aligned} &Y_i(t_j^{(i)}) \mid X_i(t_j^{(i)}) \overset{ind}{\sim} \mathcal{N}(X_i(t_j^{(i)})\beta_c, \sigma_c^2), \text{ for } j = 1, \ldots, n_i,\ i \in c, \\ &\beta_c \mid \sigma_c^2 \sim \mathcal{N}(\mu_0, \sigma_c^2 \Lambda_0^{-1}), \sigma_c^2 \sim IG(\text{shape} = a_0, \text{rate} = b_0), \text{ for } c \in \mathcal{C}, \\ &\mathcal{C} \sim p_\lambda(\mathcal{C} \mid \mathcal{G}_N). \end{aligned} \tag{3.13}$$

Under additional assumptions on the design matrix as listed in Section 1.4 of the Supplementary File, Assumptions (A1), (A2), and (B3) hold. As 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 Theorem 3 and 4, we can also obtain the usual posterior consistency result for the regression parameters within each cluster. Note that model (3.13) that we consider in Theorem 4 is slightly different from (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 are taking small values, e.g., model (2.8) becomes equivalent to (3.13) if $l_c = 0$. It will be of interest in a future work to extend results in Theorems 3 and 4 by allowing temporal random effects.

4. Bayesian Inference

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

Next, we decide 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 criteria, whose value, however, 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) (Lewis et al., 2014) and marginal likelihood computed by harmonic mean (Newton and Raftery, 1994). However, these methods usually suffer from the Pseudo-Bias issue (Lenk, 2009), which tends to prefer the model with higher complexity. Sequential Importance Sampling Method (Basu and Chib, 2003) is another popular approach for marginal likelihood estimation, but it cannot be applied to MRF-PPM since the Pólya urn scheme is not available.

Our solution is to consider a prior sampling based approach for estimating the marginal likelihood as follows,

$$\widehat{m}(Y \mid X) = \frac{1}{M - M'} \sum_{k=(M'+1)}^{M} \Pi_{c \in \mathcal{C}_{(k)}} f(Y_c \mid X_c, \alpha_k, \ell_k), \tag{4.14}$$

where $\mathcal{C}_{(k)}$,α_k and ℓ_k are the associated parameters sampled from the partition probability function at the k-th iteration, and $f(Y_c \mid X_c, \alpha_k, \ell_k)$ is the likelihood function by integrating out (β, σ^2) on its prior. More specifically, $\mathcal{C}_{(k)}$ at each iteration is sampled via 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) by letting $\mathcal{C}_{(0)}$ (the initial partition) equal to the last sample from our algorithm, with $n_{iter} = 1000$ and the first 500 iterations be burn-in iterations, $M = 10^6$, $M' = 10^4$ (burn-in procedure) and set the same random seed for different λ value. In both simulation and real data analysis, we choose λ from $\{0, 0.1, \ldots, 1\}$ and find out that this range works quite well since the selected optimal λ is always inside $(0, 1)$ in the results.

5. Simulation

In this section, we study the empirical performance of our MRF-MFM model and compare with four approaches, including two Bayesian methods without accounting for spatial correlation, namely, the Dirichlet Process (DP) and MFM, and two frequentist methods in Lin and Ng (2012) and Su et al. (2016). In the numerical analysis, Dahl's method (Dahl, 2006) will be used to summarize the posterior samples and obtain a deterministic result for both cluster assignment and cluster-wise parameter. Rand index

(RI; Rand (1971)) will be adopted as a metric to evaluate the discrepancy between different partitions. All computations were performed on 10 computing servers. Each server has 94.24GB RAM and 24 processing cores. We distributed our simulation tasks to 100 workers (10 cores for each server), and it took approximately 20 hours to finish each simulation scenario with 100 Monte Carlo replications including the tuning procedure.

5.1 Simulation Setting

For simulation data generation, 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 spatial correlation, as 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 since it allows two spatially non-contiguous blocks belonging to the same cluster. For each partition

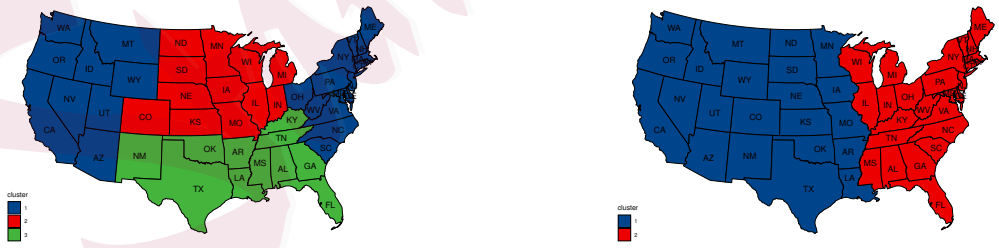

Figure 1: Simulation partition scenarios 1 and 2

scenario, we generate data from the following model,

$$\begin{aligned} &Y_i(t_j^{(i)}) \mid \left\{e_i(t_j^{(i)}), X_i(t_j^{(i)})\right\} \overset{\text{ind}}{\sim} \mathcal{N}(X_i(t_j^{(i)})\beta_{c_i} + e_i(t), \sigma_{c_i}^2 \cdot \alpha_{c_i}), \\ &e_i(t_j^{(i)}) \sim \mathcal{N}(0, K_{\sigma_{c_i}^2, \ell_{c_i}}(\cdot,\cdot)), \text{ for } j = 1, \cdots, n_i,\ i \in c, \end{aligned} \tag{5.15}$$

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 (DGP) with different cluster-wise parameters, detailed in Section S1 of the Supplementary File.

Among them, DGPs 1, 2, 5, 6 are for partition scenario 1, and the other four are for scenario 2. Because of different variance (σ_i^2, for $i = 1, 2, 3$) magnitude of the random error, we call DGPs 1, 3, 5, 7 the strong noise design, and DGPs 2, 4, 6, 8 the weak noise design.

In both simulation and 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 cluster size evenly distributed. We set the hyper-parameters as $\mu_0 = \mathbf{0}_{p\times 1}$, $\Lambda_0 = 10^{-6} \cdot \mathrm{I}_{p\times p}$, $a_0 = 0.1$, $b_0 = 1$. Throughout the numerical studies, we find that the results are not sensitive to the choice of those values. In addition, we set $a_1 = a_2 = 2$ and $b_1 = b_2 = 1$ to encourage α and ℓ 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 rand index 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 DP, and set the parameters of MFM to be the same with those of MRF-MFM in Section 5.1. For the method in Lin and Ng (2012), we follow the default setting of their code, which assumes the number of clusters $|\mathcal{C}|$ is within $\{2, 3, 4\}$, and selects $|\mathcal{C}|$ based on BIC. The partition is determined following Conditional K-means (CK-means) criteria, which is pointed out in their paper to be more robust than the other methods when $n_{\min}$ is small. For the method in Su et al. (2016), we use penalized least squares approach (PLS) to fit the model, and follow the default setting of their code, which assumes $|\mathcal{C}|$ is within $\{1, \ldots, 5\}$. Since both frequentist models can only discriminate latent groups when they have different slopes, they will only be implemented for DGPs 5–8.

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 a higher value in RI under all scenarios, which confirms the benefit of appropriately incorporating the spatial correlation across different locations. All three Bayesian

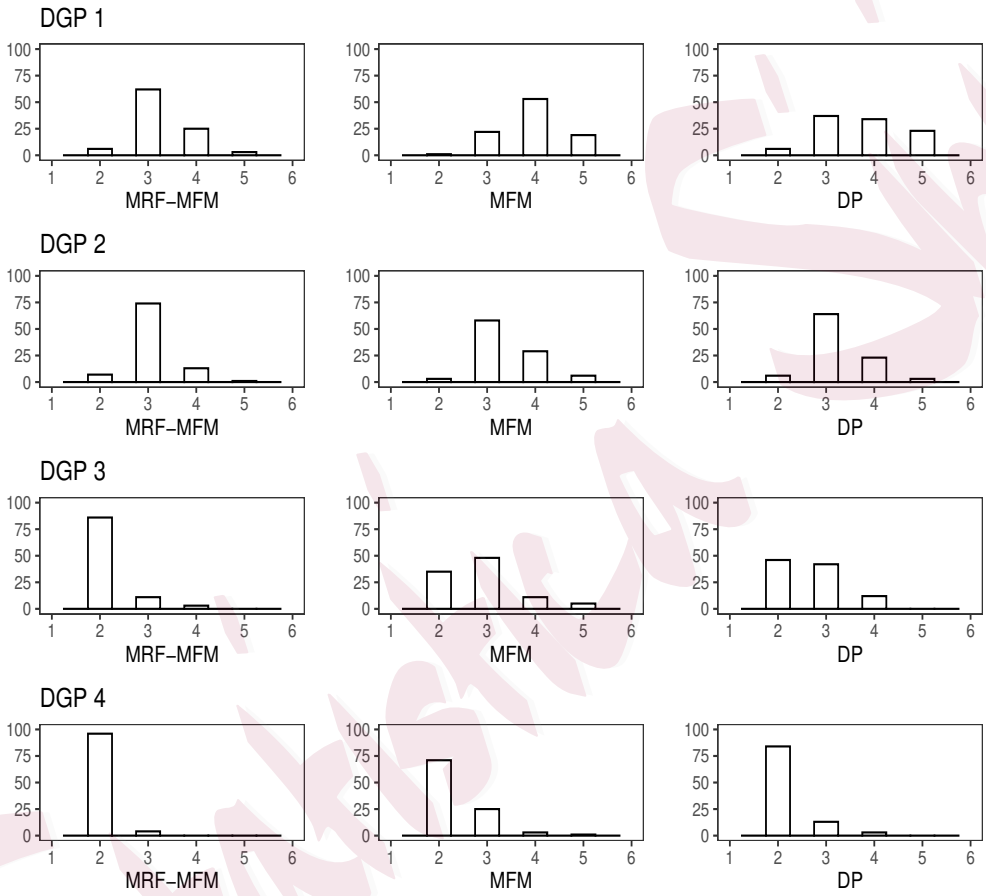

Figure 2: Histograms for selected number of clusters by MRF-MFM and four competing methods (DGPs 1-4).

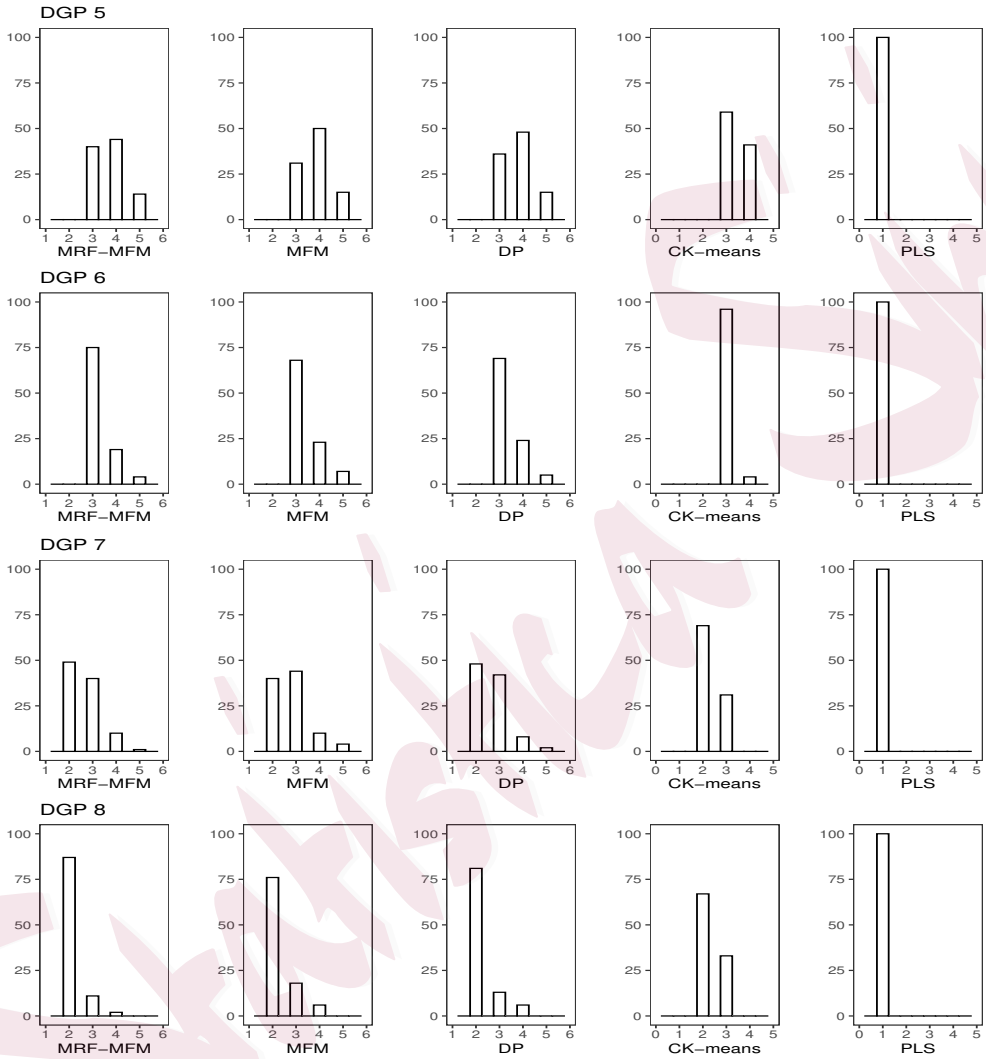

Figure 3: Histograms for selected number of clusters by MRF-MFM and four competing methods (DGPs 5-8).

Table 1: Median (Mean) of 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)

methods provide a more accurate clustering partition result than that of the two frequentist methods for the reason that those 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 in the table 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 MRF-MFM and other two Bayesian methods. We find that our method has a smaller coefficient estimation error than the other two Bayesian methods in most scenarios. For the first two DGPs, MFM has a smaller average ℓ_1 error than our method, although that advantage is minor. Among eight DGPs, DGP 5 is the most challenging case because the cluster-wise regression parameters are less separable compared to those in other DGPs. This is also reflected by the lowest average random index in Table 1 and the highest estimation error in Table 2.

We also compare the CK-means 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 has a strong temporal correlation at each location due to large ℓ and small α values. This creates trouble for PLS method (Su et al., 2016), where they use a z-transformation on both Y and $\mathbf{x}$ and the estimation of slope becomes equivalent to estimating the correlation coefficient. As a consequence, the difference in slopes between clusters cannot be fully captured by their method as the correlation coefficient is close to 1 for all clusters due to high serial correlation. On the other hand, the CK-means method is more robust since it is distance-based. Similar findings are observed in Figure 2 and Figure 3, where we show the

histograms for the selected number of clusters. In general, MRF-MFM has an excellent performance 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 among locations. We have also conducted a sensitivity analysis for the choice of the covariance kernel function and hyperparameter values (α and γ). Based on the results summarized in Section S6 of the Supplementary File, our results (both clustering and parameter estimation) are quite stable under the use of different covariance kernels and hyperparameter values.

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 by 48 states (exclude Washington, D.C., Alaska and Hawaii) from 2000 to 2019. The main goal is to study the relationship between the annual precipitation and the average temperature

and understand its heterogeneity over different states. It is well known that the precipitation is strongly associated with the convection, which is influenced by the 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 which 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 5,000 as burn-in. The final partition is obtained by Dahl's method. The average rand index between the reporting partition and the 100 replications is 0.9362, which indicates that the final partition is representative.

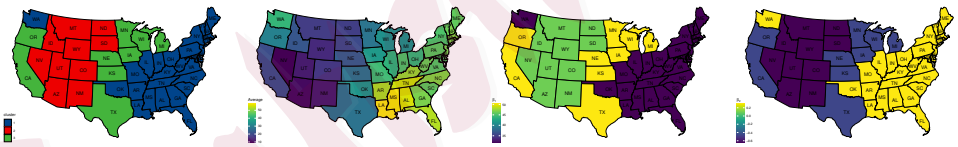

Figure 4: From Left to Right: (a) The estimated partition; (b) The average annual precipitation map; (c) Estimated intercepts; (d) Estimated slopes for the annual temperature.

We summarize the results in Figure 4. We find that the estimated partition in (a) in general matches the pattern observed in the average annual precipitation map in (b) quite well. More specifically, the first cluster

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

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

(blue) contains most states with the climate type of humid continental and humid subtropical, which usually receive plenty of rainfall annually. The climate types of most states in the second cluster (red), on the other hand, are desert and semi-arid, which naturally associate with a low level of precipitation.

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

In our study, we interpret the predictor "the average annual temperature" as a hybrid indicator, e.g., for the second cluster, the annual tempera-

ture seems to indicate the 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 the Supplementary File. By comparing the estimated partition maps obtained from our method and 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 dataset available at `https://www.countyhealthrankings.org/app/`. The data consists of annual measurements of median household income, total gross domestic product (GDP) and the unemployment rate between the year of 2011 and 2018. Our interest here is to conduct a regression analysis of the median income on GDP and unemployment rate and study the heterogeneity pattern in the regression parameters over different counties. Before applying our method, we did 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 described in Section 6.1. The spatial smoothness parameter is selected as $\lambda = 0.1$ based on the maximum marginal likelihood. The average rand index between the final cluster assignment and the ones 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 regression parameter estimates for each of three clusters in Table 4. Here we can see a uniform pattern that the annual household median income is negatively associated with the unemployment rate and positively associated with the GDP in all three clusters, which agrees with the common sense. Among the obtained three clusters, Cluster 1 (see Figure 5) has the strongest negative association between the unemployment and the median income; and it can be observed that 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 the household income, most counties in this cluster are blue counties (Democrats votes $\geq$ 60% during the 2020 presidential election), including Napa, Sonoma, Yolo, Los Angeles, San Diego, and Imperial. Those results suggest that political opinions and industrial structure may be potential confounders that can be included in the future analysis. We also implement the MFM prior (without the spatial consideration) and present the results in Table 2 and Figure 2 of the Supplementary File. By comparing the estimated regression coefficients obtained from our method and MFM, we find that our method is better at differentiating the three clusters in terms of the estimated regression coefficients, e.g., the estimated coefficient for DGP is more distinct across three clusters in our results.

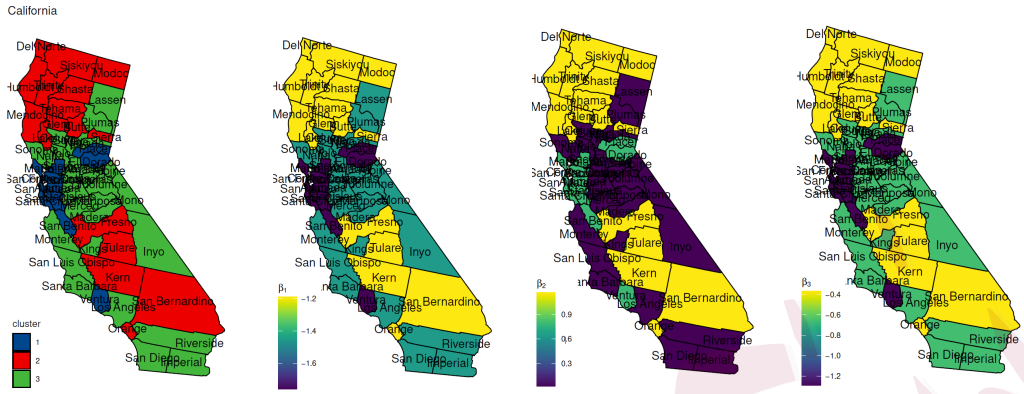

Figure 5: From Left to Right: (a) The estimated partition; (b) Visualized intercepts; (c) Visualized slopes for GDP; (d) Visualized slopes for log odds of the unemployment rate.

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

Cluster	Color	$\hat{\beta}_{\text{intercept}}$	$\hat{\beta}_{\text{GDP}}$	$\hat{\beta}_{\text{unemployment}}$	$\hat{\sigma^2}$	$\hat{\ell}$	$\hat{\alpha}$
1	dark blue	-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	red	-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	green	-1.45 (0.16)	0.02 (0.05)	-0.65 (0.06)	0.15 (0.01)	8.28 (0.29)	0.11 (0.01)

7. Discussion

In this paper, we propose a general Bayesian spatial clustering method based on the product partition model equipped with a Markov random field structure for panel data analysis. We study the fundamental properties of MRF-PPM, and prove a clustering consistency result under mild conditions on the MRF structure. A computationally tractable MCMC algorithm, as well as a model selection method based on the marginal likelihood are introduced. Numerical studies confirm that MRF-PPM effectively avoids the over-clustering issue and is more robust to model mis-specification com-

pared to the classical PPM.

Several future work directions remain open. First, it is challenging to study the asymptotic behavior of MRF-PPM prior when $N \to \infty$, as Kolmogorov's extension theorem does not hold generally after accounting for spatial information. It will 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. Secondly, 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 a more generic form of the regression functions such as nonparametric or semi-parametric models. Developing efficient posterior computation and understanding the theoretical properties for this prior remains a challenging future work direction.

Acknowledgment

The authors thank the editor, the associate editor, and reviewers for their valuable comments, which helped improve the presentation of this paper.

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.

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.

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.

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.

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.

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

University of Missouri - Columbia, Columbia, MO, 65211

E-mail: gh7mr@missouri.edu