Spatio-Temporal Random Partition Models

Garritt L. Page
Brigham Young University, Provo, USA
BCAM - Basque Center of Applied Mathematics, Bilbao, Spain
and
Fernando A. Quintana *
Pontificia Universidad Católica de Chile, Santiago, Chile
and
David B. Dahl
Brigham Young University, Provo, USA.

April 10, 2022

Abstract

The number of scientific fields that regularly collect data that are spatio-temporal continues to grow. An intuitive feature of this type of data is that measurements taken on experimental units near each other in time and space tend to be similar. As such, many methods developed to accommodate spatio-temporal dependent structures attempt to borrow strength among units close in space and time, which constitutes an implicit space-time grouping. We develop a class of dependent random partition models that explicitly models this spatio-temporal clustering by way of a dependent random partition model. We first detail how temporal dependence is incorporated so that partitions evolve gently over time. Then conditional and marginal properties of the joint model are derived. We then demonstrate how space can be integrated. Computation strategies are detailed and we illustrate the methodology through simulations and an application.

Keywords: correlated partitions, spatio-temporal clustering, hierarchical Bayes modeling, Bayesian nonparametrics, time dependent partitions.

*Partially supported by grant FONDECYT 1180034 and by Iniciativa Científica Milenio - Minecon Núcleo Milenio MiDaS

1 Introduction

We introduce a method to directly model spatio-temporal dependence in a sequence of random partitions. Our approach is motivated by the practical problem of modeling a prior distribution for a sequence of random partitions that exhibit substantial overlap over time, and where cluster formation may also be spatially influenced. Traditionally, dependencies in random partitions (i.e., the clustering of units) have been obtained as a by-product of dependent random measures in Bayesian nonparametric (BNP). We will argue, however, that when partitions are the inferential objects of principal interest, then the partition should be modeled *directly* rather than relying on *induced* random partition models such as those originating from temporal, or spatio-temporal dependent BNP models. But first, we review the literature on dependent BNP methods.

BNP methods that incorporate time include Caron et al. (2007), Nieto-Barajas et al. (2012), Antoniano-Villalobos and Walker (2016), Gutiérrez et al. (2016), Jo et al. (2017) and Caron et al. (2017). Those accommodating space include Gelfand et al. (2005), Griffin and Steel (2006), Duan et al. (2007), Petrone et al. (2009), and Gelfand et al. (2010). The BNP literature is more sparse for combined space-time methods, with Kottas et al. (2008) being the first to construct a spatio-temporal BNP model for areal data by adding an AR(1)-like temporal transition structure to the spatial Dirichlet process of Gelfand et al. (2005). Zhang et al. (2016) consider a model for functional magnetic resonance imaging data and model temporal dependence in the error term and spatial dependence through a hierarchical Dirichlet process mixture model on voxel-specific coefficients (whose clustering induces spatial dependence in the partition). Savitsky (2016) apply a spatio-temporal BNP model to the American Community Survey with varying spatial resolution. Cassese et al. (2019) construct a space-time species sampling model that permits the identification of

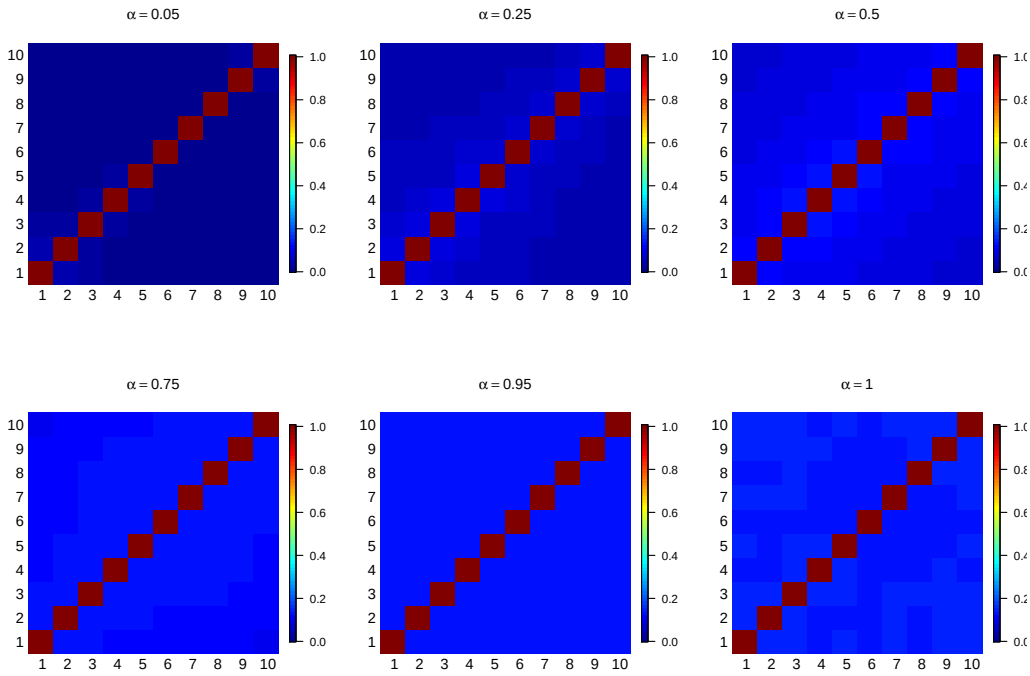

Figure 1: Lagged ARI values using the method of (Caron et al., 2017) based on concentration parameter $M = 0.5$, discount parameter set to zero, and 10,000 Monte Carlo samples. The temporal dependence parameter is $\alpha \in [0, 1]$.

disease outbreaks.

A common aspect of all these methods is that temporal, spatial, or spatio-temporal dependence is accommodated in the sequence of random measures by way of the atoms or weights of the stick-breaking representation (Sethuraman, 1994). The induced random partitions, however, exhibit only weak dependence even when a sequence of random probability measures is highly correlated. To illustrate this point, we conducted a small Monte Carlo simulation where a sequence of partitions were generated with 10 time points and

20 units using the method of Caron et al. (2017). To measure similarity of partitions at different time points, we use the lagged adjusted rand index (ARI). Figure 1 shows these values averaged over 10,000 Monte Carlo samples. Notice that as α increases, the partitions from time period t to $t+1$ only become slightly more similar, such that the dependance between partitions is, at best, only weak. Further, the dependence is not temporally intuitive as it does not decay as a function of lag. This behavior is not unique to Caron et al. (2017)'s approach, as Wade et al. (2014) noticed the same type of behavior when using a linear dependent Dirichlet process mixture model. In fact, all BNP methods that model a sequence of random probability measures will induce a random partition model with similar weak-correlation behavior. This behavior is analogous to trying to induce dependence among random variables from distributions with correlated parameters. There is no guarantee that correlated parameters would produce strong correlations among the random variables themselves.

Paci and Finazzi (2018)'s motivation is more similar to ours as their principal interest is spatially referenced partitions over time. However, their approach is based on a mixture of experts model whose weights depend on space and time. As such, their method retains the same properties as the BNP methods.

Our approach is to consider the sequence of partitions indexed by time as the object of principal interest and propose a method that models it directly. This will provide more control over how "smoothly" partitions evolve over time. Perhaps the work closest to ours (in the sense of explicitly modeling a sequence of partitions) can be found in Zanini et al. (2019). Their modeling approach for temporally-referenced sequence of partitions differs from ours in that they do not focus on smooth evolution of spatial partitions over time.

The rest of the article is organized as follows. Section 2 details our approach to modeling

partitions temporally and spatially. In Section 2 we also provide a few theoretical results and some computational strategies. In Section 3 we detail a number of simulation studies that illustrate the method and highlight its utility. Then we consider a PM_{10} data set that is publicly available. Section 4 contains some concluding remarks.

2 Joint Model for a Sequence of Partitions

We begin with some notation. Let $i = 1, \ldots, m$ denote the m experimental units at time t for $t = 1, \ldots, T$. Let $\rho_t = \{S_{1t}, \ldots, S_{k_t t}\}$ denote a partition of the m experimental units at time $t = 1, \ldots, T$ into k_t clusters. An alternative partition notation is based on m cluster labels at time t denoted by $\boldsymbol{c}_t = \{c_{1t}, \ldots, c_{mt}\}$ where $c_{it} = j$ implies that $i \in S_{jt}$. Notice the one-to-one correspondence between ρ_t and $\boldsymbol{c}_t$. Finally, any quantity with a "$\star$" superscript will be cluster-specific. For example, we will use $\mu^{\star}_{jt}$ to denote the mean of cluster j at time t so that $\mu_{it} = \mu^{\star}_{jt}$ if $c_{it} = j$.

2.1 Temporal Modeling for Sequences of Partitions

We first describe our approach to correlating partitions over time and subsequently, in the next subsection, detail the inclusion of space. Introducing temporal dependence in a collection of partitions requires formulating a joint probability model for $\{\rho_1, \ldots, \rho_T\}$. Generically, we will denote this joint model with $\Pr(\rho_t, \ldots, \rho_T)$. Temporal dependence among the ρ_t's implies that the cluster configurations found in $\rho_{t-1}, \rho_{t-2}, \ldots, \rho_1$ could impact the cluster configuration in ρ_t. However, we assume that the probability model for the sequence of partitions has a Markovian structure. That is, the conditional distribution

of ρ_t given $\rho_{t-1}, \rho_{t-2}, \ldots, \rho_1$ only depends on ρ_{t-1}. Thus, we construct $\Pr(\rho_t, \ldots, \rho_T)$ as

$$\Pr(\rho_1, \ldots, \rho_T) = \Pr(\rho_T|\rho_{T-1})\Pr(\rho_{T-1}|\rho_{T-2}) \cdots \Pr(\rho_2|\rho_1)\Pr(\rho_1). \tag{1}$$

Here $\Pr(\rho_1)$ is an exchangeable partition probability function (EPPF) that describes how the m experimental units at time period 1 are grouped into k_1 distinct groups with frequencies $n_{11}, \ldots, n_{1k_1}$. One characteristic of an exchangeable EPPF that will prove useful in what follows is sample size consistency (or what De Blasi et al. (2015) refer to as the addition rule). This property dictates that marginalizing the last of $m + 1$ elements leads to the same model as if we only had m elements. A commonly encountered EPPF is that induced by a Dirichlet process (DP). This particular EPPF is sometimes referred to as a Chinese restaurant process (CRP) and corresponds to a special case from the family of product partition models (PPM). For more details see De Blasi et al. (2015). Because we employ the CRP-type EPPF in what follows, we provide its form here

$$\Pr(\rho|M) = \frac{M^k}{\prod_{i=1}^{n}(M + i - 1)} \prod_{i=1}^{k}(|S_i| - 1)!, \tag{2}$$

where k is the number of clusters in ρ and M is a concentration parameter controlling the number of clusters. We will denote this random partition distribution as $CRP(M)$.

Although conceptually straightforward, (1) is silent regarding how ρ_{t-1} influences the form of ρ_t. To make this explicit, we introduce an auxiliary variable that guides how similar ρ_t is to ρ_{t-1}. Now, if two partitions are highly correlated, then the cluster configurations between them will change very little and as a result only a few of the m experimental units will change cluster assignment. Conversely, two partitions that exhibit low correlation will likely be comprised of very different cluster configurations. The auxiliary variable we introduce identifies which of the experimental units at time $t - 1$ will be considered for

possible cluster reallocation at time t. Specifically, let γ_{it} denote the following

$$\gamma_{it} = \begin{cases} 1 & \text{if unit } i \text{ is } \textit{not} \text{ reallocated when moving from time } t-1 \text{ to } t \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

By construction we set $\gamma_{i1} = 0$ for all i (i.e., all experimental units are allocated to clusters during the first time period). We then assume that $\gamma_{it} \overset{ind}{\sim} Ber(\alpha_t)$. Note that each of the $\alpha_t \in [0, 1]$ acts as a temporal dependence parameter. Specifically, we will interpret $\alpha_t = 1$ as implying that $\rho_t = \rho_{t-1}$ with probability 1. Conversely, when $\alpha_t = 0$, then ρ_t is independent of ρ_{t-1}. For notational convenience we introduce $\boldsymbol{\gamma}_t = (\gamma_{1t}, \gamma_{2t}, \ldots, \gamma_{mt})$ which is an m-tuple comprised of zeros and ones. The augmented joint model changes (1) to

$$\begin{aligned} \Pr(\boldsymbol{\gamma}_1, \rho_1, \ldots, \boldsymbol{\gamma}_T, \rho_T) = \Pr(\rho_T | \boldsymbol{\gamma}_T, \rho_{T-1}) \Pr(\boldsymbol{\gamma}_T) \times \\ \Pr(\rho_{T-1} | \boldsymbol{\gamma}_{T-1}, \rho_{T-2}) \Pr(\boldsymbol{\gamma}_{T-1}) \cdots \Pr(\rho_2 | \boldsymbol{\gamma}_2, \rho_1) \Pr(\boldsymbol{\gamma}_2) \Pr(\rho_1). \end{aligned} \tag{4}$$

In Section **??** of the online Supplementary Material, we provide a toy example that illustrates how our construction produces intuitive conditional partition distributions. In addition to exhibiting intuitive behavior conditionally, it would be appealing if marginally each of the ρ_t follow the parent EPPF (i.e., the probability model assumed for ρ_1), so that the joint probability model for partitions would become stationary. The following proposition establishes this result which is a consequence of the fact that conditioning on $\boldsymbol{\gamma}_t$ provides a "reduced" EPPF.

Proposition 2.1. *Let $\rho_1 \sim EPPF$ and $\boldsymbol{\gamma}_1 = 0$. If a joint model for $\rho_1 \ldots, \rho_T$ is constructed as described above by introducing $\boldsymbol{\gamma}_t$ for $t = 2, \ldots, T$, then we have that marginally $\rho_1, \ldots, \rho_T$ are identically distributed with law coming from the EPPF used to model ρ_1. Specifically, letting $\rho_{-t} = (\rho_1, \ldots, \rho_{t-1}, \rho_{t+1}, \ldots, \rho_T)$ and $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_T)$, we have that for*

all $\lambda \in P$,

$$\Pr(\rho_t = \lambda) = \sum_{\rho_{-t} \in P^{\otimes}} \sum_{\boldsymbol{\gamma} \in \Gamma^{\otimes}} \Pr(\boldsymbol{\gamma}_1, \rho_1, \ldots, \rho_t = \lambda, \ldots, \boldsymbol{\gamma}_T, \rho_T) = \Pr(\rho_1 = \lambda),$$

where $P^{\otimes} = P \times P \times \ldots \times P$, P *a collection of all partitions of* m *units and* $\Gamma^{\otimes} = \Gamma \times \Gamma \times \ldots \times \Gamma$, Γ *a collection of all possible binary vectors of size* m.

Proof. See the Appendix. □

In what follows we will use $tRPM(\boldsymbol{\alpha}, M)$ to denote our temporal random partition model (4) parameterized by $\alpha_1, \ldots, \alpha_T$ and EPPF (2).

We briefly mention that introducing γ_{it} is similar in spirit to the approach taken by Caron et al. (2007, 2017). However, they use $\boldsymbol{\gamma}_t$ to identify a partial partition at time t that informs how *all* the observational units will be reallocated at time $t + 1$. While this difference may seem benign at first glance, it has drastic ramifications on the type of dependence that exists among the actual sequence of partitions. To see this, similar to what was done in the simulation described in the Introduction, we generate 10,000 sequences of partitions based on our construction an provide the average lagged ARI values in Figure 2. Notice now that the similarity of the partitions behaves in an intuitive way as a function of lag. Mainly, that as lags increase the similarity between partitions decreases. Further, α has a clear impact on the dependence between partitions with large α values resulting in strong dependence. Observe also that the range of ARI values achieved by this construction can be substantially higher than what was described earlier in the discussion leading to Figure 1.

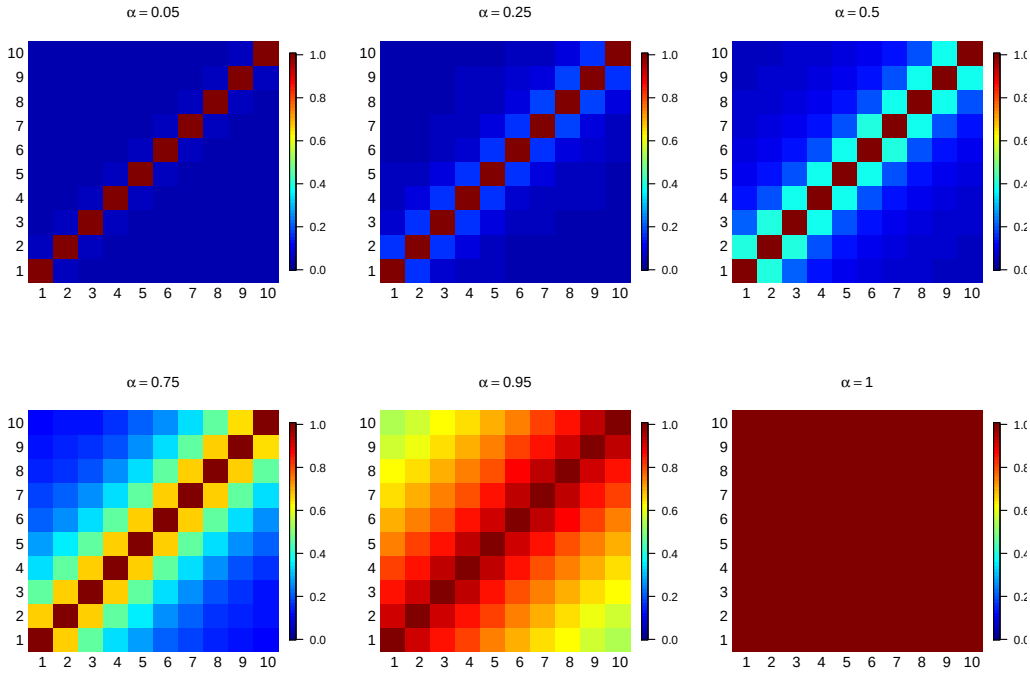

Figure 2: Lagged ARI values using concentration parameter $M = 0.5$ based on 10,000 Monte Carlo samples using our $tRPM$. Our method shows strong temporal dependence, whereas little is seen in Figure 1 for Caron et al. (2017).

2.2 Spatio-Temporal Model for a Sequence of Partitions

Before studying how our joint partition model can be employed in Bayesian modeling, we next describe our approach to incorporating space in the partition model. One possible way of adding a spatial component in the joint model would be to make the auxiliary variables $\boldsymbol{\gamma}_{it}$ spatially referenced. However, sample size consistency would be lost and as a result the marginal property in Proposition 2.1 would not hold. An alternative approach that we adopt is to include spatial information directly in the EPPF. If the spatially referenced

EPPF employed preserves sample size consistency, then Proposition 2.1 still holds. To this end, we consider the spatial product partition model (sPPM) developed in Page and Quintana (2016). As a way of introducing the sPPM, let $\boldsymbol{s}_i$ denote the spatial coordinates of the ith item (note that these coordinates do not change over time) and let $\boldsymbol{s}^{\star}_{jt}$ be the subset of spatial coordinates that belong to the jth cluster at time t. Then we express the EPPF of the tth partition with the following product form

$$\Pr(\rho_t|\nu_0, M) \propto \prod_{j=1}^{k_t} c(S_{jt}|M) g(\boldsymbol{s}^{\star}_{jt}|\nu_0). \tag{5}$$

Here $c(\cdot|M) \geq 0$ is called the cohesion and is a set function that produces cluster weights *a priori*. We consider the cohesion $c(S_{jt}|M) = M \times (|S_{jt}| - 1)!$ as it has connections with the CRP making this version of the sPPM a type of spatially re-weighted CRP. The similarity function $g(\cdot|\nu_0)$ is a set function parametrized by ν_0 that measures the compactness of the spatial coordinates in $\boldsymbol{s}^{\star}_{jt}$ producing higher values if the spatial coordinates in $\boldsymbol{s}^{\star}_{jt}$ are less alike. Not all similarity functions preserve sample size consistency so to ensure this, after standardizing spatial locations, we employ

$$g(\boldsymbol{s}^{\star}_{jt}|\nu_0) = \int \prod_{i \in S_{jt}} N(\boldsymbol{s}_i|\boldsymbol{m}, \boldsymbol{V}) NIW(\boldsymbol{m}, \boldsymbol{V}|\boldsymbol{0}, 1, \nu_0, \boldsymbol{I}) d\boldsymbol{m} d\boldsymbol{V}, \tag{6}$$

where $N(\cdot|\boldsymbol{m}, \boldsymbol{V})$ denotes a bivariate normal density and $NIW(\cdot, \cdot|\boldsymbol{0}, 1, \nu_0, \boldsymbol{I})$ a normal-inverse-Wishart density with mean $\boldsymbol{0}$, scale equal to 1, inverse scale matrix equal to $\boldsymbol{I}$, and ν_0 being the user-supplied degrees of freedom. Note that larger values of ν_0 increase spatial influence on partition probabilities. For more details on why this formulation preserves sample size consistency, see Müller et al. (2011) and Quintana et al. (2018). For more information regarding the impact of ν_0 on product form of the partition model, see Page and Quintana (2016, 2018). We will denote the random partition distribution defined in (5) and (6) using $sPPM(\nu_0, M)$.

We mention briefly that it would be very straightforward to build a partition model based on space and time by extending the sPPM so similarity function g is a function of both space and time. Although this ensures that partitions will be influenced by space and time, the desire for partitions to evolve over time would be lost. In this setting, each spatial location by time point combination would be treated as an observational unit and would create clusters that transect time, which is something we wanted to avoid in our formulation.

We will use $stRPM(\boldsymbol{\alpha}, \nu_0, M)$ to denote our spatio-temporal random partition model (4) parameterized by $\alpha_1, \ldots, \alpha_T$ and EPPF detailed in (5) and (6).

2.3 Hierarchical Data Model

Once a partition model is specified, there is tremendous flexibility regarding how to model space/time (global or cluster-specific) at different levels of a hierarchical model (at the data level or process level or both). Since we are interested to see how including space/time in the partition model impacts clustering and model fits, in the simulations of the next section, we consider a hierarchical model where space/time only appears in the partition model. In particular, using cluster label notation, we will employ the following hierarchical model

$$\begin{aligned} Y_{it}|\boldsymbol{\mu}_t^\star, \boldsymbol{\sigma}_t^{2\star}, \boldsymbol{c}_t &\overset{ind}{\sim} N(\mu^\star_{c_{it}t}, \sigma^{2\star}_{c_{it}t}),\ i = 1, \ldots, m \text{ and } t = 1, \ldots, T, \\ (\mu^\star_{jt}, \sigma^\star_{jt})|\theta_t, \tau^2 &\overset{ind}{\sim} N(\theta_t, \tau^2) \times UN(0, A_\sigma),\ j = 1, \ldots, k_t, \\ (\theta_t, \tau) &\overset{iid}{\sim} N(\phi_0, \lambda^2) \times UN(0, A_\tau),\ t = 1, \ldots, T, \\ \{\boldsymbol{c}_t, \ldots, \boldsymbol{c}_T\} &\sim \textit{joint RPM}, \end{aligned} \tag{7}$$

where Y_{it} denotes the response measured on the ith unit at time t, *joint RPM* denotes some joint random partition model, and UN denotes a uniform distribution. The remaining

assumptions (e.g., independence across clusters and exchangeability within each cluster) are commonly employed. Notice that in this model three entities are in some sense "competing" when determining cluster membership, namely: a) time, b) space, and c) response. This competition, however, is carried out in a probabilistic and coherent fashion.

2.4 Computation

As the posterior distribution implied by the model in (7) is not available in closed form, we build an algorithm that permits sampling from it. The construction of $\Pr(\rho_1, \ldots, \rho_T)$ naturally leads one to consider a Gibbs sampler. In the Gibbs sampler, $\boldsymbol{\gamma}_t$ will need to be updated in addition to ρ_t (by way of $\boldsymbol{c}_t$). But the Markovian assumption reduces some of the cost as we only need to consider ρ_{t-1} and ρ_{t+1} when updating ρ_t. Even though each update of ρ_t and $\boldsymbol{\gamma}_t$ for $t = 1, \ldots, T$ needs to be checked for compatibility (i.e. proposed moves do not violate the prior construction), it is fairly straightforward to adapt standard algorithms, e.g. Algorithm 8 of Neal (2000), with care to make sure that only experimental units with $\gamma_{it} = 0$ are considered when updating cluster labels at time t. In what follows we assume that the *joint RPM* in (7) is the $stRPM(\boldsymbol{\alpha}, \nu_0, M)$ described earlier.

The MCMC algorithm we employ depends on deriving the complete conditionals for ρ_t and γ_t. Before describing them, we introduce some needed notation. Let $N_{0t} = \sum_{j=1}^{m} I[\gamma_{jt} = 0]$ denote the number of units to be reallocated when moving from time $t-1$ to t (note that $N_{0t} \sim \text{Bin}(m, 1-\alpha_t)$) and denote with $\rho_t^{-N_{0t}}$ the "reduced" partition that remains after removing the N_{0t} units that are to be reallocated at time t as indicated by $\boldsymbol{\gamma}_t$. A key result needed to derive the full conditionals of γ_{it} and c_{it} is provided in the following proposition.

Proposition 2.2. *Based on the construction of a joint probability model as described in*

Section 2.1 and $\rho_1 \sim EPPF$*, then we have*

$$\Pr(\rho_t|\boldsymbol{\gamma}_t, \rho_{t-1}) = \begin{cases} \Pr(\rho_t)/\Pr(\rho_t^{-N_{0t}}) & \rho_t \asymp \rho_{t-1} \\ 0 & otherwise, \end{cases} \tag{8}$$

where $\rho_t \asymp \rho_{t-1}$ *indicates that* ρ_t *is compatible with* ρ_{t-1}*.*

Proof. See the Appendix. □

When updating γ_{it} in a Gibbs sampler, one can think of removing γ_{it} from $\boldsymbol{\gamma}_t$, and then reinsert it as either a 0 or 1. To this end, let $N_{0t}^{(-i)} = \sum_{j \neq i} I[\gamma_{it} = 0]$ denote the case when γ_{it} is reinserted as a 1 and $N_{0t}^{(+i)} = N_{0t}^{(-i)} + 1$ denote the case when γ_{it} is reinserted as a 0. Now, the full conditional for $\gamma_{it} = 1$, denoted by $\Pr(\gamma_{it} = 1|-)$, is

$$\begin{aligned} \Pr(\gamma_{it} = 1|-) &\propto \Pr(\rho_t|\boldsymbol{\gamma}_t, \rho_{t-1})\Pr(\boldsymbol{\gamma}_t)\mathrm{I}[\rho_t \asymp \rho_{t-1}], \\ &\propto \frac{\Pr(\rho_t)}{\Pr(\rho_t^{-N_{0t}^{(+i)}})}\alpha_t^{\gamma_{it}}\mathrm{I}[\rho_t \asymp \rho_{t-1}]. \end{aligned}$$

Here $\mathrm{I}[\cdot]$ denotes an indicator function. The resulting normalized full conditional for γ_{it} is

$$\Pr(\gamma_{it} = 1|-) = \frac{\alpha_t \Pr(\rho_t^{-N_{0t}^{(-i)}})}{\alpha_t \Pr(\rho_t^{-N_{0t}^{(-i)}}) + (1-\alpha_t)\Pr(\rho_t^{-N_{0t}^{(+i)}})}\mathrm{I}[\rho_t \asymp \rho_{t-1}]. \tag{9}$$

For a given EPPF that has a closed form (e.g., CRP), it is straightforward to compute $\Pr(\rho_t^{-N_{0t}^{(-i)}})$ and $\Pr(\rho_t^{-N_{0t}^{(+i)}})$. If, however, the EPPF does not have a closed form (e.g., sPPM), then note that (9) can be re-expressed as

$$\Pr(\gamma_{it} = 1|-) = \frac{\alpha_t \mathrm{I}[\rho_t \asymp \rho_{t-1}]}{\alpha_t + (1-\alpha_t)\Pr(\rho_t^{-N_{0t}^{(+i)}})/\Pr(\rho_t^{-N_{0t}^{(-i)}})}. \tag{10}$$

The quantity $\Pr(\rho_t^{-N_{0t}^{(+i)}})/\Pr(\rho_t^{-N_{0t}^{(-i)}})$ is a commonly encountered expression in MCMC methods employed in random partition modeling. See for example Neal's Algorithm 8 (Neal, 2000). Those same methods can be employed to calculate the desired probabilities.

The full conditional for $c_{it} = h$ corresponding to $\gamma_{it} = 0$ is the following

$$\Pr(c_{it} = h|-) \propto N(Y_{it}|\mu^{\star}_{c_{it}=h,t}, \sigma^{2\star}_{c_{it}=h,t})\Pr(c_{1t}, \ldots, c_{it} = h, \ldots, c_{mt})\mathrm{I}[\rho_t \asymp \rho_{t-1}].$$

The case that unit it forms a new cluster must also be considered so that

$$\Pr(c_{it} = h|-) \propto \begin{cases} N(Y_{it}|\mu^{\star}_{c_{it}=h,t}, \sigma^{2\star}_{c_{it}=h,t})\Pr(c_{it} = h)\mathrm{I}[\rho_t \asymp \rho_{t-1}] & h = 1, \ldots, k_t^{-i}, \\ N(Y_{it}|\mu^{\star}_{new_h,t}, \sigma^{2\star}_{new_h,t})\Pr(c_{it} = h)\mathrm{I}[\rho_t \asymp \rho_{t-1}] & h = k_t^{-i} + 1, \end{cases}$$

where $\Pr(c_{it} = h) = \Pr(c_{1t}, \ldots, c_{it} = h, \ldots, c_{mt})$, $\mu^{\star}_{new_h,t}$ and $\sigma^{2\star}_{new_h,t}$ are auxiliary parameters drawn from the prior as in Neal (2000)'s Algorithm 8 (with one auxiliary parameter) and k_t^{-i} are the number of clusters at time t when the ith unit has been removed. Details of computation procedures associated with the sPPM can be found in Page and Quintana (2016). Given ρ_t and $\boldsymbol{\gamma}_t$, the full conditionals of the remaining parameters in model (7) follow standard techniques. A sample can be drawn from the posterior distribution implied by model (7) by iterating through the complete conditionals for $\boldsymbol{\gamma}_t$ and ρ_t and those of other model parameters.

3 Simulation Studies

In this section we describe three simulation studies that explore the performance of our proposal. The first simulation study is focused on the temporal dependence among estimated partitions, the second on the temporal dependence that the joint partition model induces among the $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{iT})$, and the third on the impact that including space in the partition model has on model fit.

3.1 Simulation 1: Temporal Dependence in Estimated Partitions

The purpose of the first simulation is to study the accuracy of partition estimates (i.e., $\hat{\rho}_t$) and how much they change over time. As such, in this study we do not consider spatial clustering. We do however, explore accuracy in estimating μ_{it} and α_t. To this end, we considered model (7) as a data generating mechanism to create one hundred datasets with fifty observations at five time points. For the *joint RPM* in model (7) we used $tRPM(\boldsymbol{\alpha})$ with $\alpha_t = \alpha$ for all t and generate synthetic datasets under $\alpha \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9, 0.999\}$. For all i and t, we set $\sigma^{2\star}_{c_{it}t} = \sigma^2 = 1$, $\tau^2 = 25$, and $\theta_t = 0$.

To each synthetic data set we fit model (7) using the MCMC algorithm detailed in Section 2.4 by collecting 10,000 iterates and discarding the first 5,000 as burn-in and thinning by 5 (resulting in 1,000 MCMC samples) after setting $A_\sigma = 5$ and $A_\tau = 10$. All partition point estimates were estimated using the method developed in the `salso` R package (Dahl 2019) with the Binder loss function (Binder 1978, Lau and Green 2007). To measure similarity between partitions, we employed the adjusted Rand index (Rand 1971; Hubert and Arabie 1985) and we used WAIC (Gelman et al. 2014) to measure model fit.

Table 1 displays the lagged 1 and 4 adjusted Rand index (ARI) as a function of α. As expected, for both lags the ARI increases as α increases. Also as expected lagged 4 ARI increases less as a function of α compared to the lagged 1 ARI. Note that on average the lagged 1 ARI for $\alpha \in \{0.1, 0.25\}$ is smaller than that for $\alpha = 0$. This is because the variability associated with lagged 1 ARI when $\alpha = 0$ is much larger than when $\alpha > 0$ producing a few lagged ARI values that are large. The median of the lagged ARI values increase as a function of α monotonically.

To study the ability to recover μ_{it} and α, 95% credible intervals for each were computed and coverage was estimated. Results are provided in Table 1. Notice that coverage for

Table 1: Adjusted Rand index when comparing $\hat{\rho}_1$ to $\hat{\rho}_2$ and $\hat{\rho}_1$ to $\hat{\rho}_5$. Note that $ARI(\cdot, \cdot)$ denotes the adjusted Rand index as a function of two partitions. Coverage rates for α and μ_{it} and model fit metrics for $tRPM(\alpha, M)$ and $CRP(M)$. These values are averaged over the 100 generated data sets. The values in parenthesis are Monte Carlo standard errors. Note that smaller values of WAIC indicate better fit.

			Coverage		WAIC	
	$ARI(\hat{\rho}_1, \hat{\rho}_2)$	$ARI(\hat{\rho}_1, \hat{\rho}_5)$	α	μ_{it}	tRPM	CRP
$\alpha = 0.0$	0.192 (0.03)	0.182 (0.03)	0.00 (0.00)	0.94 (0.01)	1711	1709
$\alpha = 0.1$	0.122 (0.02)	0.151 (0.02)	0.97 (0.02)	0.94 (0.01)	1689	1694
$\alpha = 0.25$	0.180 (0.02)	0.130 (0.02)	0.95 (0.02)	0.94 (0.01)	1645	1669
$\alpha = 0.5$	0.434 (0.02)	0.132 (0.02)	0.88 (0.03)	0.94 (0.01)	1627	1723
$\alpha = 0.75$	0.714 (0.02)	0.254 (0.02)	0.89 (0.03)	0.96 (0.01)	1576	1636
$\alpha = 0.9$	0.874 (0.01)	0.546 (0.02)	0.91 (0.03)	0.93 (0.01)	1501	1710
$\alpha = 0.9999$	0.980 (0.00)	0.941 (0.01)	0.49 (0.05)	0.93 (0.01)	1502	1611

α is low when the true α is at or near the boundary (e.g., $\alpha \in \{0, 0.9999\}$) which is to be expected. The coverage associated with μ_{it} is close to the nominal rate regardless of the value of α. Therefore, temporal dependence in the partition model does not adversely impact the ability to estimate individual means.

Lastly, to compare model fit when using $tRPM(\alpha, M)$ as the RPM in model (7) relative to $\rho_t \overset{iid}{\sim} CRP(M)$, we calculated the WAIC for each data set when fitting model (7) under both RPMs. Results are provided in Table 1 where each entry is an average WAIC value over all 100 datasets. Notice that, when the independent partitions were used to generate data (i.e., $\alpha = 0$), modeling partitions independently produces slightly better model fit as would be expected. But even if relatively weak temporal dependence exists among the sequence of partitions, there are gains in modeling the sequence of partitions with $tRPM(\alpha, M)$, with gains becoming substantial as α increases.

The upshot from this simulation study is that lagged partition estimates when employ-

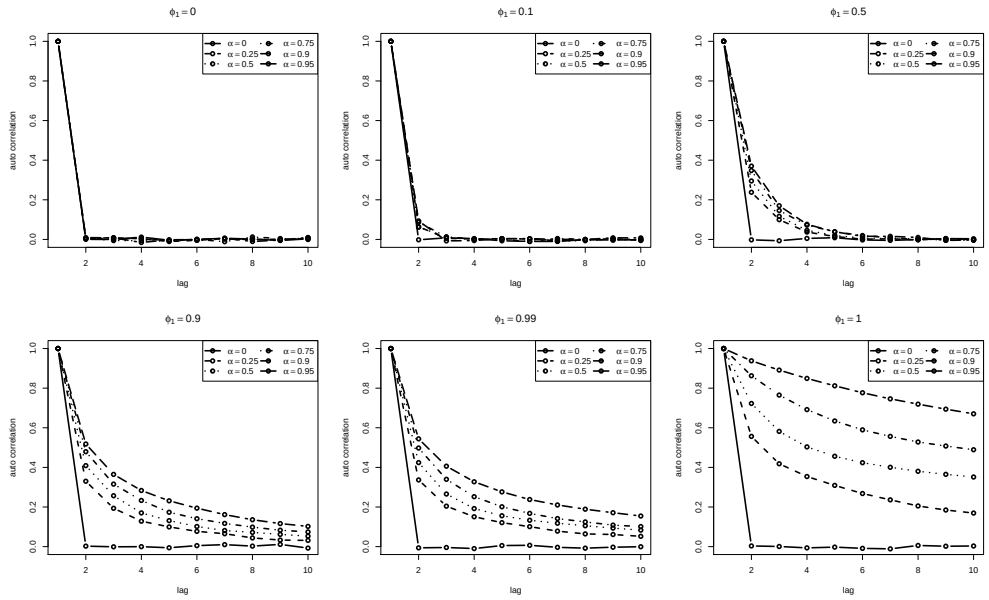

Figure 3: Lagged auto-correlations among the $(Y_{i1}, \ldots, Y_{iT})$ when modeling $\mu^{\star}_{jt}$ with an AR(1) type structure.

ing $tRPM(\alpha, M)$ display intuitive behavior in that similarity between partition estimates decreases as lag increases. In addition, employing the $tRPM(\alpha, M)$ partition model does not negatively impact parameter estimation and produces improved model fits when dependence is present in the sequence of partitions and a minimal cost in model fit when it is not.

3.2 Simulation 2: Induced Correlation at the Response Level

A potential benefit of developing a joint model for partitions is the ability to accommodate temporal dependence that may exist between Y_{it} and Y_{it+1}. To study this, we conducted a small Monte Carlo simulation study that is comprised of sampling repeatedly from the

$tRPM(\alpha, M)$ using the computational approach of Section 2.4. Once the partition is generated, the temporal dependence among the $\boldsymbol{Y}_i$ depends on specific model choices for $\mu^\star_{jt}$. Here we use $\mu^\star_{jt} \sim N(\phi_1 \mu^\star_{jt-1}, \tau^2(1-\phi_1^2))$ for $t > 2$, $j = 1, \ldots, k_t$, and $|\phi_1| \leq 1$. For $t = 1$ we use $\mu^\star_{j1} \sim N(0, \tau^2)$ and if $k_{t+1} > k_t$ new $\mu^\star_{jt+1}$ values are drawn from $N(0, \tau^2)$. Now setting $m = 25$, $T = 10$, $\tau = 10$, and $\sigma = 1$, 100 data sets were generated for $\phi_1 \in \{0, 0.25, 0.5, 075, 0.9, 1\}$. For each data set generated, the lagged auto-correlations among $\boldsymbol{Y}_i$ were computed for $i = 1, \ldots, m$. The results found in Figure 3 are the lagged auto-correlations averaged over the m units for $\alpha \in \{0, 0.25, 0.5, 0.75, 0.9\}$.

As can be seen in Figure 3, when partitions are independent (i.e., $\alpha = 0$), no correlation propagates to the data level. The same can be said if atoms are *iid* (i.e., $\phi_1 = 0$). As the temporal dependance among $\mu^\star_{jt}$ increases (i.e., ϕ_1 increases), there is stronger temporal dependence among $Y_{i1}, \ldots, Y_{iT}$. Notice further that this dependence persists longer in time as α increases as one would expect.

3.3 Simulation 3: Dependence in Estimated Partitions

We now discuss our final simulation study, where we investigated the performance of our procedure when both space and time are considered. To do so, we created synthetic data sets that contain spatio-temporal structure. Each employs a 15×15 regular grid with spatial locations coming from the unit interval. In addition, either 5 or 10 time points were considered resulting in 1,125 or 2,250 total observations. Response values were generated in two ways. The first employs a Gaussian process with a separable spatio-temporal exponential covariance function. We set the spatial scale to 0.3, the temporal scale to 2 and the sill to 1.75 (see Padoan and Bevilacqua 2015 for more details). Note that no "true" partition exists for this data generating mechanism. However, we study it to explore performance of

our method when spatial structure exists among observations but was not induced through partitioning. The second method of generating response values essentially employs model (7) as a data generating mechanism. Spatio-temporal partitions were generated using (6) together with conditional cluster label probabilities of Müller et al. (2011, pg. 265) and setting $\alpha_t = \alpha$ for all t with $\alpha \in \{0, 0.5, 0.9\}$ (note that for $\alpha = 0$ no temporal dependence exists among partitions). In the similarity function (6) we considered $\nu_0 \in \{2, 20\}$ where $\nu_0 = 2$ corresponds to light weight on spatial proximity and $\nu_0 = 20$ moderate weight. Finally, we set $\tau^2 = 1$ and $\sigma^{2\star}_{c_{it}t} = \sigma^2 = 0.04$ for all i and t resulting in smaller with-in cluster variability relative to between-cluster variability.

To determine the impact that each component of our spatio-temporal partition model has on model fit, we fit the hierarchical model (7) to each synthetic data set using a variety of random partition models which are listed below. As a competitor, we consider a linear dependent Dirichlet process (MacEachern, 2000; De Iorio et al., 2009), indexing the random probability measure through the mean function of the atoms by space and time. To ensure sufficient flexibility, B-spline basis functions for both spatial coordinates were employed. The details of each model considered are

Model 1: $(\rho_1, \ldots, \rho_T) \sim stRPM(\boldsymbol{\alpha}, \nu_0, M)$

Model 2: $\rho_t \overset{iid}{\sim} sPPM(\nu_0, M)$ for $t = 1, \ldots, T$.

Model 3: $(\rho_1, \ldots, \rho_T) \sim tRPM(\boldsymbol{\alpha}, M)$

Model 4: $\rho_t \overset{iid}{\sim} CRP(M)$ for $t = 1, \ldots, T$.

Model 5: linear dependent Dirichlet process mixture model (DDPM).

Additionally, for each model that employs the sPPM, we considered both $\nu_0 = 2$ (models 1a, 2a) and $\nu_0 = 20$ (models 1b, 2b). For each data generating scenario, 100 data sets were created and each of the models listed was fit by collecting 1,000 MCMC samples after discarding the first 5,000 as burn-in and thinning by 5 after setting $A_\sigma = 1$ and $A_\tau = 2$. Model fits were compared using WAIC. Results can be found in Figures 4 and 5.

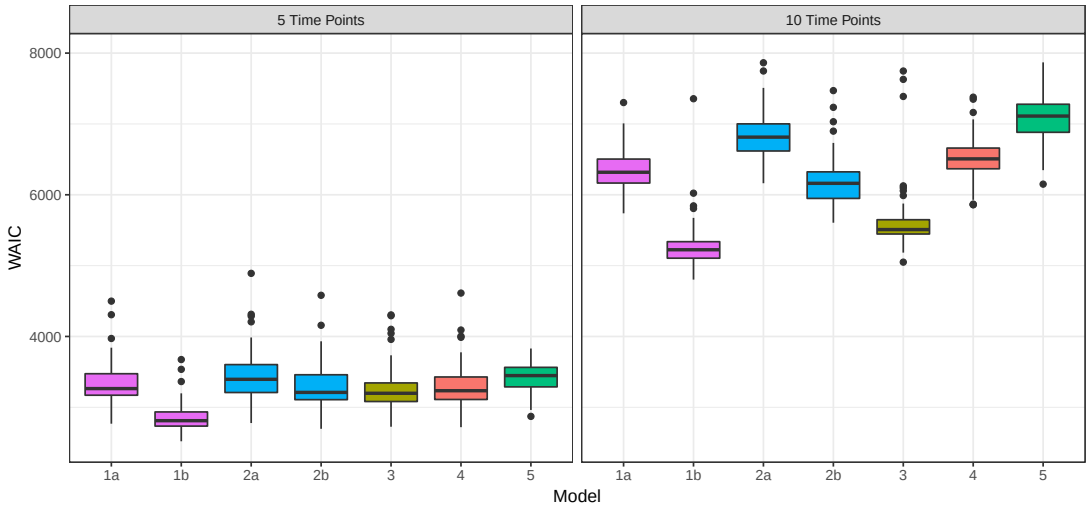

Figure 4: Results from simulation study when observations were generated using a spatio-temporal Gaussian process. Boxplots display the 100 WAIC values that correspond to model fit for each synthetic data set. Note that smaller WAIC values indicate a better fit.

The primary purpose of Figure 4 is to compare model fit from the spatio-temporal partition model we develop to that from the linear DDPM (model 5). It appears that all methods are competitive to the linear DDPM, which is particularly true with 10 time points. Thus, our dependent partition model accommodates temporal dependence more efficiently relative to the linear DDPM under this data generating scenario. Note that regardless of the number of time points, model 1b ($stRPM(\alpha, \nu_0, M)$ with $\nu_0 = 20$) appears to perform best.

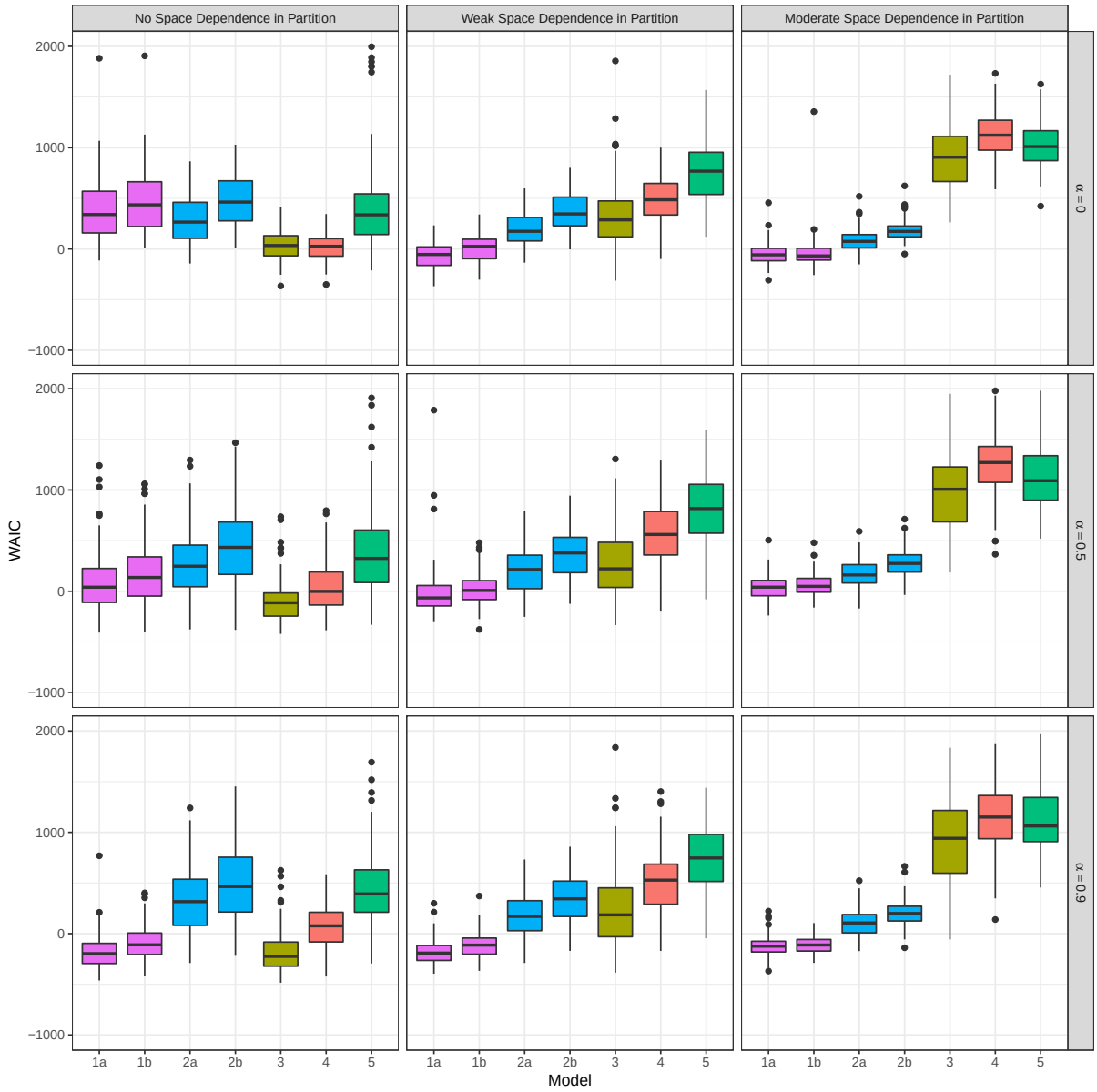

Figure 5: Results from simulation study for the scenario in which partition structure is included in data generation process. Boxplots display the 100 WAIC values that correspond to model fit for each synthetic data generating scenario. Note that smaller indicates better fit.

Surprisingly, $tRPM(\alpha, M)$ (model 4) is quite competitive, particularly with 10 time points. The conclusion here is that employing $stRPM(\alpha, \nu_0, M)$ to model partitions appears to accommodate spatio-temporal dependence even if there is no underlying partition structure.

From Figure 5 we see that when partitions are generated independently, there is very little lost by employing the dependent joint model in terms of model fit (see top left panel for model 3 and 4). However, as spatial and/or temporal structure is introduced in the partition model, there are clear gains in terms of model fit when employing $tRPM(\boldsymbol{\alpha}, M)$ and/or $stRPM(\boldsymbol{\alpha}, \nu_0, M)$. From this simulation it seems that employing the $tRPM(\alpha, M)$ regardless of the strength of temporal dependence among partitions is reasonable as there is minimal cost in terms of model fit even when partitions are generated independently. Finally, it appears that $stRPM(\boldsymbol{\alpha}, \nu_0, M)$ performed best.

3.4 Application

In this section we apply our method to a real-world data set coming from the field of environmental science. A second application in educational measurement is provided in Section **??** of the online Supplementary Material. As mentioned previously, once a partition model is specified there is quite a bit of flexibility regarding how (or if) temporal dependence is incorporated in other parts of a hierarchical model. To illustrate this, we incorporate temporal dependence in three places of the hierarchical model we construct.

As part of preliminary exploratory data analysis (not shown), we examined serial dependence for each experimental (monitoring station), and concluded that they all exhibited a particular type of temporal dependence. Because of this, we introduce a unit-specific temporal dependence parameter $|\eta_{1i}| \leq 1$ and model observations from a single unit over time $(Y_{1i}, \ldots, Y_{iT})$ with an AR(1) structure. In addition, motivated by a desire for parsimony,

we employed a Laplace prior for η_{1i}. Finally, to permit the temporal dependence in the partition model to propagate through the hierarchical model, we model θ_t with an AR(1) structure. The full hierarchical model is detailed in (11).

$$
\begin{aligned}
Y_{it}|Y_{it-1}, \boldsymbol{\mu}_t^\star, \boldsymbol{\sigma}_t^{2\star}, \boldsymbol{\eta}, \boldsymbol{c}_t &\overset{ind}{\sim} N(\mu^\star_{c_{it}t} + \eta_{1i} Y_{it-1}, \sigma^{2\star}_{c_{it}t}(1-\eta_{1i}^2)), \\
Y_{i1} &\overset{ind}{\sim} N(\mu^\star_{c_{i1}1}, \sigma^{2\star}_{c_{i1}1}), \\
\xi_i = \text{Logit}(0.5(\eta_{1i}+1)) &\overset{iid}{\sim} Laplace(a,b), \\
(\mu^\star_{jt}, \sigma^\star_{jt}) &\overset{ind}{\sim} N(\theta_t, \tau^2) \times UN(0, A_\sigma), \\
\theta_t|\theta_{t-1} &\overset{ind}{\sim} N(\phi_0 + \phi_1\theta_{t-1}, \lambda^2(1-\phi_1^2)), \\
(\theta_1, \tau) &\sim N(\phi_0, \lambda^2) \times UN(0, A_\tau), \\
(\phi_0, \phi_1, \lambda) &\sim N(0, s^2) \times UN(-1,1) \times UN(0, A_\lambda), \\
\{\boldsymbol{c}_t, \ldots, \boldsymbol{c}_T\} &\sim stRPM(\boldsymbol{\alpha}, \nu_0, M), \text{ with } \alpha_t \overset{iid}{\sim} Beta(a_\alpha, b_\alpha),
\end{aligned}
\tag{11}
$$

where all Roman letters correspond to parameters that are user supplied. Notice that there are a number of special cases embedded in our hierarchical model. For example, $\eta_{i1} = 0$ for all i results in conditionally independent observations. Further, $\phi_1 = 0$ results in independent atoms and $\alpha_t = 0$ for all t in independent partitions over time. Note that model (7) used in the simulation studies is a special case of (11) ($\phi_1 = 0$ and $\eta_{i1} = 0$ for all i). A_σ may influence partition formation. If this value is selected to be too large, then all observational units could plausibly be allocated to one cluster. If it is too small then many spurious clusters could potentially be formed. Therefore, this parameter must be selected thoughtfully. Our approach is to set A_σ to about half the sample standard deviation computed using all observations.

3.5 Rural Background PM_{10} Data Application

The rural background PM_{10} data is taken from the European air quality database. These data are comprised of the daily measurements of particulate matter with a diameter less than 10 μm from rural background stations in Germany and are publicly available in the `gstat` package (Gräler et al. 2016) found on CRAN in `R` (R Core Team 2018). We focus on average monthly PM_{10} measures from the year 2005. Of the 69 stations, 9 were removed because of missing values.

We fit the hierarchical model (11) to these data and consider all the possible special cases (i.e., $\eta_{1i} = 0$ or not, $\phi_1 = 0$ or not, $\alpha_t = 0$ or not, with and without space). This resulted in 16 total models that were fit by collecting 1,000 MCMC iterates after discarding the first 10,000 as burn-in and thinning by 10. The prior values employed were $A_\sigma = A_\tau = 5$, $s^2 = 100$, $a = 0$, $b = 1$, $a_\alpha = b_\alpha = 1$, and $\nu_0 = 5$. The WAIC and log pseudo marginal likelihood (LPML) for each model are presented in Table 2.

Notice that among all the model fits, employing a variant of $tRPM(\boldsymbol{\alpha}, M)$ (i.e., rows with "Yes" in the "In Partition" column) generally improves model fit. The best performing model in terms of WAIC and LPML includes spatio-temporal dependence in the partition model, temporal dependence among the atoms, and temporal dependence in the likelihood. To see how the different models impact how partitions evolve over time, we provide Figure 6. This figure displays the lagged ARI values for each of the 16 models. Notice that when partitions are modeled independently (first or third rows of Figure 6) then partitions evolve over time quite erratically in the sense that the cluster configuration can change dramatically from one time point to the next. However, when employing $tRPM(\boldsymbol{\alpha}, M)$ (second row of Figure 6) the partitions seemed to evolve much more "smoothly" as there is less drastic changes in cluster configuration. Finally, it appears that employing the $stRPM(\boldsymbol{\alpha}, \nu_0, M)$

Table 2: PM_{10} data: Results of model fitting. The bold font identifies best model fits in terms of LPML and WAIC. Higher values for LPML indicate better fit while lower values for WAIC indicate better fit.

			Space			
Temporal Dependence In			No		Yes	
Partition	Likelihood	Atoms	LPML	WAIC	LPML	WAIC
No	No	No	-1973	3464	-1904	3655
No	No	Yes	-1973	3467	-1899	3653
No	Yes	No	-1762	3071	-1562	3125
No	Yes	Yes	-1770	3070	-1560	3170
Yes	No	No	-1639	3226	-1613	3003
Yes	No	Yes	-1618	3120	-1579	3015
Yes	Yes	No	-1758	3153	-3240	3014
Yes	Yes	Yes	-1590	3016	**-1535**	**2911**

(fourth row of Figure 6) not only produces partitions that evolve "smoothly" over time, but the temporal dependence seems to decay quicker than when employing $tRPM(\boldsymbol{\alpha}, M)$ only. In fact the model that produces the best model fit metrics (right most plot of the bottom row) seems to produce partitions that change quite gently over time as desired.

4 Conclusions

We developed a joint probability model for a sequence of partitions that explicitly considers temporal dependence among the partitions. Further we showed that our methodology is capable of accommodating partitions that evolve slowly over time in that the adjusted Rand index between estimated partitions decays as the lag in time increases. Further, we showed that in the absence of temporal dependence between partitions, the cost in terms

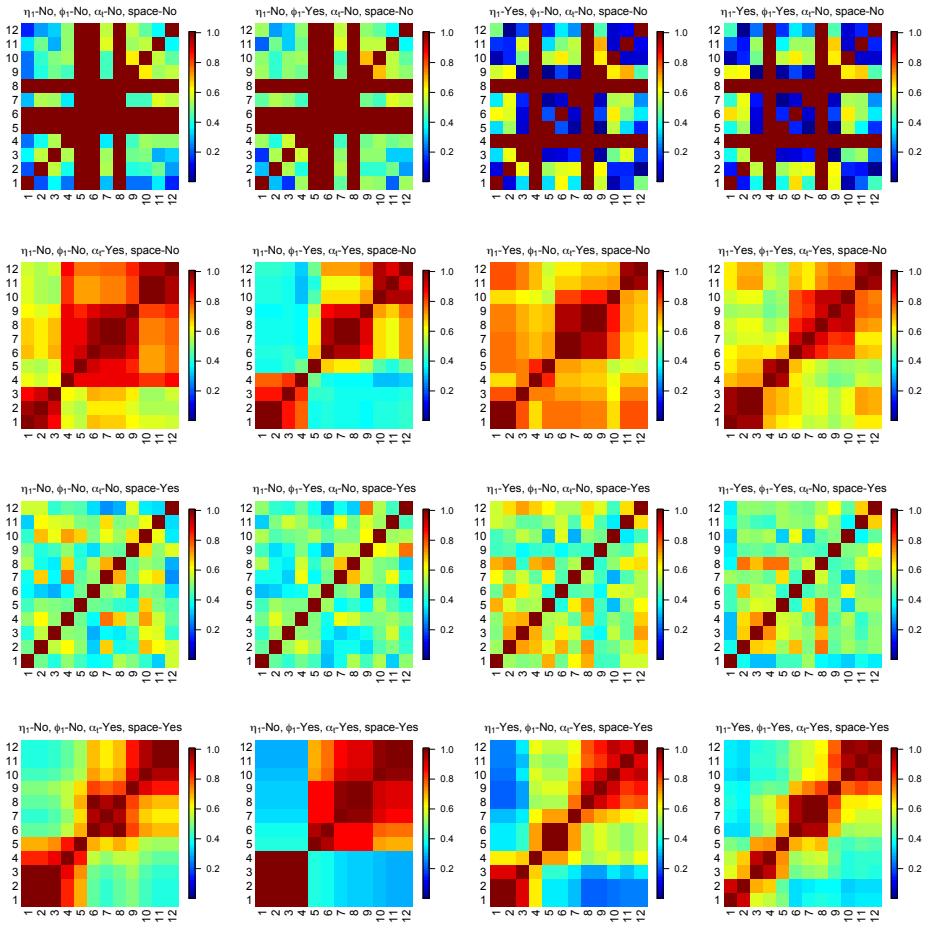

Figure 6: PM_{10} data. Each figure is a summary of the lagged ARI values corresponding to the 16 models in Table 2.

of model fit is minimal.

Even though our main focus is constructing a dependent probability model for a sequence of random partitions, our method, when coupled with a simple hierarchical model, could provide an alternative approach to general space-time modeling that completely avoids inverting matrices. This could result in computation gains compared to employing computationally intense non-separable covariance functions. In addition, assumptions associated with stationarity and/or isotropy can be avoided.

The predictive nature of the spatio-temporal prior on a sequence of random partitions we have presented has a (first-order) Markovian structure. Various extensions can be considered, such as adding higher order dependence across time or dependence in baseline covariates. All of these cases would build on our constructive definition, as extra refinements of the basic idea of carrying smooth transitions on time and space. The Markovian approach can also be used for predictive inference, although that was not our main motivation for the models implemented here, and therefore we have not explored this avenue.

A Proof of Proposition 2.1

Proof. For clarity, here we introduce notation that highlights the dependence of partitions on sample size. For example, $\rho_{t,m} = (S_{1,t}, \ldots, S_{k_t(m),t})$ and $[m] = \{1, \ldots, m\}$. By assumption $\Pr(\rho_{1,m})$ is specified by means of an EPPF which we now construct. Denote $\mathbb{N}^* = \cup_{k=0}^{\infty} \mathbb{N}^k$, and identify any $\boldsymbol{n} = (n_1, \ldots, n_k) \in \mathbb{N}^*$ with the infinite sequence $(n_1, \ldots, n_k, 0, 0, \ldots)$. Given $\boldsymbol{n} \in \mathbb{N}^*$, let $k(\boldsymbol{n})$ denote the number of non-zero entries in $\boldsymbol{n}$ and denote by $\boldsymbol{n}^{j+}$ the result of incrementing $\boldsymbol{n}$'s jth component (i.e., n_j) by 1, with $1 \leq j \leq k(\boldsymbol{n}) + 1$. An EPPF is then any function $r : \mathbb{N}^* \longrightarrow [0, 1]$ that is symmetric in its

arguments and where

$$r(1) = 1 \qquad \text{and} \qquad r(\boldsymbol{n}) = \sum_{j=1}^{k(\boldsymbol{n})+1} r(\boldsymbol{n}^{j+}) \qquad \text{for all } \boldsymbol{n} \in \mathbb{N}^*. \tag{12}$$

Condition (12) implies that a EPPF is sample size consistent, i.e., marginalizing the $(n+1)$st element leads to the model for n elements. The EPPF also implies exchangeability of configurations in the sense that a EPPF is invariant under permutations of the elements that keep the cluster sizes unaltered. We also note that any valid EPPF defines a predictive rule of the form

$$r_j(\boldsymbol{n}) = \frac{r(\boldsymbol{n}^{j+})}{r(\boldsymbol{n})}, \qquad \text{for } 1 \leq j \leq k(\boldsymbol{n}) + 1, \tag{13}$$

where it is assumed that $r(\boldsymbol{n}) > 0$ and $r_j(\boldsymbol{n})$ represents the probability of a new element joining the jth already existing cluster, for $1 \leq j \leq k(\boldsymbol{n})$, or starting a new one (the $k(\boldsymbol{n}) + 1$). The one-step rule (13) can also be extended to predictions of two or more elements by simply iterating the one-step rule as many times as needed. Now, given an EPPF r, we have that

$$\Pr(\rho_{1,m} = (S_{1,1}, \ldots, S_{k_1(m),1})) = r(n_{1,1}, \ldots, n_{k_1(m),1}). \tag{14}$$

To prove the result, it suffices to show that it holds for $\rho_{2,m}$ and then by induction the result holds generally. Denote by $[\Gamma] = \{i \in \{1, \ldots, m\} : \gamma_{i2} = 0\}$ the (random) set of elements removed from $\rho_{1,m}$. Then, $\rho_{1,m}^{-N_{02}}$ is a partition of the elements of $[m] - [\Gamma]$ (where as before $N_{02} = \sum_{j=1}^{m} I[\gamma_{j2} = 0]$). By exchangeability and the fact that an EPPF is sample size consistent, we have that for any partition $S_1^-, \ldots, S_{k([m]-[\Gamma])}^-$ of $[m] - [\Gamma]$:

$$\begin{aligned} \Pr(\rho_{2,m}^{-N_{02}} = (S_1^-, \ldots, S_{k([m]-[\Gamma])}^-) \mid [\Gamma]) &= \Pr(\rho_{1,m}^{-N_{02}} = (S_1^-, \ldots, S_{k([m]-[\Gamma])}^-) \mid [\Gamma]) \\ &= r(|S_1^-|, \ldots, |S_{k([m]-[\Gamma])}^-|), \end{aligned}$$

where $|S_j|$ is the number of elements in S_j. In addition, and again by exchangeability and sample size consistency, the predictive rule starting from $[m] - [\Gamma]$ (or from any subset of $[m]$ for that matter) depends only on the sizes of the subsets in that partition. Thus, conditioning on all reallocation configurations and initial partition after subject removal we have:

$$\begin{aligned}
\Pr(\rho_{2,m} = (S_1, \ldots, S_k)) &= \sum_{[\Gamma]} \sum_{\rho_{2,m}^{-N_{02}}} \Pr(\rho_{2,m} = (S_1, \ldots, S_k) \mid [\Gamma],\, \rho_{2,m}^{-N_{02}}) \times \\
&\qquad \Pr(\rho_{2,m}^{-N_{02}} \mid [\Gamma]) \Pr([\Gamma]), \\
&= \sum_{[\Gamma]} \sum_{\rho_{1,m}^{-N_{02}}} \Pr(\rho_{1,m} = (S_1, \ldots, S_k) \mid [\Gamma],\, \rho_{1,m}^{-N_{02}}) \times \\
&\qquad \Pr(\rho_{1,m}^{-N_{02}} \mid [\Gamma]) \Pr([\Gamma]), \\
&= \Pr(\rho_{1,m} = (S_1, \ldots, S_k)),
\end{aligned}$$

where the second to last equality follows from the constructive description given earlier and the properties of the EPPF. The result then follows. □

B Proof of Proposition 2.2

Proof. Let $P_{C_t} = \{\rho_t \in P : \rho_t \asymp \rho_{t-1}\}$ denote the collection of all partitions of the elements of $[m]$ at time t that are compatible with ρ_{t-1} based on $\boldsymbol{\gamma}_t$. Then by construction, $\Pr(\rho_t | \boldsymbol{\gamma}_t, \rho_{t-1})$ is a random partition distribution whose support is P_{C_t} so that

$$\Pr(\rho_t = \lambda | \boldsymbol{\gamma}_t, \rho_{t-1}) = \frac{\Pr(\rho_t = \lambda) I[\lambda \in P_{C_t}]}{\sum_\lambda \Pr(\rho_t = \lambda) I[\lambda \in P_{C_t}]}.$$

It only remains to show that $\sum_{\lambda \in P_{C_t}} \Pr(\rho_t = \lambda) = \Pr(\rho_t^{-N_{0t}})$ which is more easily seen employing cluster label notation. Let $c_{\gamma_t} = \{c_{it} : \gamma_{it=0}\}$. By iteratively invoking the sample

size consistency property we have that

$$\begin{aligned}\Pr(\rho_t^{-N_{0t}}) &= \sum_{c_{\gamma t}} \Pr(\rho_t = \{c_{1t}, \ldots, c_{mt}\}) \\ &= \sum_{\lambda \in P_{C_t}} \Pr(\rho_t = \lambda),\end{aligned}$$

where the last equality holds since summing over c_{γ_t} is based only on cluster labels that are not fixed from time point $t-1$ to t which results in summing over all possible compatible partitions (i.e., $\lambda \in P_{C_t}$). □

SUPPLEMENTARY MATERIAL

Title: Supplementary Material. This file contains details of our model applied to an additional application in the field of education.

R-package for the *stRPM* routine: An R-package titled `drpm` contains code used to fit model described in (11).

References

Antoniano-Villalobos, I. and Walker, S. G. (2016), "A nonparametric model for stationary time series," *J. Time Series Anal.*, 37, 126–142.

Binder, D. A. (1978), "Bayesian Cluster Analysis," *Biometrika*, 65, 31–38.

Caron, F., Davy, M., and Doucet, A. (2007), "Generalized Polya Urn for Time-varying Dirichlet Process Mixtures," in *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, United States: AUAI Press, UAI'07, pp. 33–40.

Caron, F., Neiswanger, W., Wood, F., Doucet, A., and Davy, M. (2017), "Generalized Pólya Urn for Time-Varying Pitman-Yor Processes," *Journal of Machine Learning Research*, 18, 1–32.

Cassese, A., Zhu, W., Guindani, M., and Vannucci, M. (2019), "A Bayesian Nonparametric Spiked Process Prior for Dynamic Model Selection," *Bayesian Analysis*, advance publication.

Dahl, D. B. (2019), *salso: Sequentially-Allocated Latent Structure Optimization*, r package version 0.1.10.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R. H., Prünster, I., and Ruggiero, M. (2015), "Are Gibbs-Type Priors the Most Natural Generalization of the Dirichlet Process?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 212–229.

De Iorio, M., Johnson, W., Müller, P., and Rosner, G. (2009), "Bayesian Nonparametric Nonproportional Hazards Survival Modeling," *Biometrics*, 65, 762–771.

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007), "Generalized spatial Dirichlet process models," *Biometrika*, 94, 809–825.

Gelfand, A. E., Diggle, P. J., Fuentes, M., and Guttorp, P. (eds.) (2010), *Handbook of Spatial Statistics*, Boca Raton: Chapman and Hall/CRC, 1st ed.

Gelfand, A. E., Kottas, A., and MacEachern, S. N. (2005), "Bayesian Nonparametric Spatial Modeling With Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021–1035.

Gelman, A., Hwang, J., and Vehtari, A. (2014), "Understanding predictive information criteria for Bayesian models," *Statistics and Computing*, 24, 997–1016.

Gräler, B., Pebesma, E., and Heuvelink, G. (2016), "Spatio-Temporal Interpolation using gstat," *The R Journal*, 8, 204–218.

Griffin, J. E. and Steel, M. F. J. (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194.

Gutiérrez, L., Mena, R. H., and Ruggiero, M. (2016), "A time dependent Bayesian nonparametric model for air quality analysis," *Computational Statistics & Data Analysis*, 95, 161 – 175.

Hubert, L. and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193–218.

Jo, S., Lee, J., Müller, P., Quintana, F. A., and Trippa, L. (2017), "Dependent Species Sampling Models for Spatial Density Estimation," *Bayesian Analysis*, 12, 379–406.

Kottas, A., Duan, J. A., and Gelfand, A. E. (2008), "Modeling Disease Incidence Data with Spatial and Spatio-Temporal Dirichlet Process Mixtures," *Biometrical Journal*, 50, 29–42.

Lau, J. W. and Green, P. J. (2007), "Bayesian Model-Based Clustering Procedures," *Journal of Computational and Graphical Statistics*, 16, 526–558.

MacEachern, S. N. (2000), "Dependent Dirichlet processes," Tech. rep., Ohio State University.

Müller, P., Quintana, F., and Rosner, G. L. (2011), "A Product Partition Model With Regression on Covariates," *Journal of Computational and Graphical Statistics*, 20, 260–277.

Neal, R. M. (2000), "Markov Chain Sampling Methods for Dirichlet Process Mixture Models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Nieto-Barajas, L. E., Muller, P., Ji, Y., Lu, Y., and Mills, G. B. (2012), "A Time-Series DDP for Functional Proteomics Profiles," *Biometrics*, 68, 859–868.

Paci, L. and Finazzi, F. (2018), "Dynamic model-based clustering for spatio-temporal data," *Statistics & Computing*, 28, 359 – 374.

Padoan, S. A. and Bevilacqua, M. (2015), "Analysis of Random Fields Using CompRandFld," *Journal of Statistical Software*, 63, 1–27.

Page, G. L. and Quintana, F. A. (2016), "Spatial Product Partition Models," *Bayesian Analysis*, 11, 265–298.

— (2018), "Calibrating Covariate Informed Product Partition Models," *Statistics and Computing*, 28, 1009–1031.

Petrone, S., Guindani, M., and Gelfand, A. E. (2009), "Hybrid Dirichlet Mixture Models for Functional Data," *Journal of the Royal Statistical Society: Series B*, 71, 755–782.

Quintana, F. A., Loschi, R. H., and Page, G. L. (2018), *Bayesian Product Partition Models*, Wiley StatsRef: Statistics Reference Online, pp. 1–15.

R Core Team (2018), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Rand, W. M. (1971), "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, 66, 846–850.

Savitsky, T. (2016), "Bayesian Nonparametric Multiresolution Estimation for the American Community Survey," *Annals of Applied Statistics*, 10, 2157–2181.

Sethuraman, J. (1994), "A constructive definition of Dirichlet priors," *Statistica Sinica*, 4, 639–650.

Wade, S., Walker, S. G., and Petrone, S. (2014), "A Predictive Study of Dirichlet Process Mixture Models for Curve Fitting," *Scandinavian Journal of Statistics*, 41, 580–605.

Zanini, C. T. P., Müller, P., Ji, Y., and Quintana, F. A. (2019), "A Bayesian Random Partition Model for Sequential Refinement and Coagulation," *Biometrics*, 75, 988–999.

Zhang, L., Guindani, M., Versace, F., Engelmann, J. M., and Vannucci, M. (2016), "A Spatiotemporal Nonparametric Bayesian Model of Multi-subject FMRI Data," *The Annals of Applied Statistics*, 10, 638–666.