Supplementary Material: Dependent Modeling of Temporal Sequences of Random Partitions

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 21, 2021

This document contains supplementary material to the paper "Dependent Modeling of Temporal Sequences of Random Partitions".

A Proofs of Propositions

In this section of the supplementary material we provide proofs of the two propositions described in the main article

A.1 Proof of Proposition 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

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

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

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

Condition (S.1) 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

$$r_j(\boldsymbol{n}) = \frac{r(\boldsymbol{n}^{j+})}{r(\boldsymbol{n})}, \qquad \text{for } 1 \leq j \leq k(\boldsymbol{n}) + 1, \tag{S.2}$$

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 (S.2) 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

$$\Pr(\rho_{1,m} = (S_{1,1}, \ldots, S_{k_1(m),1})) = r(n_{1,1}, \ldots, n_{k_1(m),1}). \tag{S.3}$$

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. □

A.2 Proof of Proposition 2

Proof. The proof proceeds by direct calculations.

(a) Let $\gamma_i = 1$ if unit $i \in [m]$ is not relocated. Note that $\gamma_1, \gamma_2 \overset{iid}{\sim} Ber(\alpha)$. By definition, $P(c_{11} = c_{21}) = \frac{1}{M+1}$ and $P(c_{11} \neq c_{21}) = \frac{M}{M+1}$. By conditioning on γ_1, γ_2 and c_{11}, c_{21}

we get

$$P\left(c_{12}=c_{22}\mid c_{11}=c_{21},(\gamma_1,\gamma_2)\right)=\begin{cases}1 & \text{if } (\gamma_1,\gamma_2)=(1,1)\\ \frac{1}{M+1} & \text{otherwise.}\end{cases}$$

It then follows that

$$\begin{aligned}P\left(c_{12}=c_{22},c_{11}=c_{21}\right) &= \sum_{\gamma_1,\gamma_2} P\left(c_{12}=c_{22}\mid c_{11}=c_{21},(\gamma_1,\gamma_2)\right)P\left(c_{11}=c_{21}\right)P(\gamma_1)P(\gamma_2)\\ &= \frac{\alpha^2}{M+1}+\frac{(1-\alpha^2)}{(M+1)^2} \qquad \text{(S.4)}\end{aligned}$$

Similarly,

$$P\left(c_{12}\neq c_{22}\mid c_{11}\neq c_{21},(\gamma_1,\gamma_2)\right)=\begin{cases}1 & \text{if } (\gamma_1,\gamma_2)=(1,1)\\ \frac{M}{M+1} & \text{otherwise,}\end{cases}$$

and proceeding as before, we easily get

$$\begin{aligned}P\left(c_{12}\neq c_{22},c_{11}\neq c_{21}\right) &= \sum_{\gamma_1,\gamma_2} P\left(c_{12}\neq c_{22}\mid c_{11}\neq c_{21},(\gamma_1,\gamma_2)\right)P\left(c_{11}\neq c_{21}\right)P(\gamma_1)P(\gamma_2)\\ &= \frac{M\alpha^2}{M+1}+\left(\frac{M}{M+1}\right)^2\left(1-\alpha^2\right) \qquad \text{(S.5)}\end{aligned}$$

The result now easily follows by summing (S.4) and (S.5).

(b) As before, denote by $\gamma_i=1$ if unit $i\in[m]$ is *not* removed from the partition. By conditioning on γ_1,γ_2 and c_{11},c_{21} we get

$$P\left(c_{12}=c_{22}\mid c_{11}=c_{21},(\gamma_1,\gamma_2)\right)=\begin{cases}\frac{6+M}{(M+2)(M+3)} & \text{if } (\gamma_1,\gamma_2)=(1,1)\\ \frac{1}{M+1} & \text{otherwise,}\end{cases}$$

and proceeding as earlier,

$$P\left(c_{12}=c_{22},c_{11}=c_{21}\right)=\frac{(6+M)\alpha^2}{(M+1)(M+2)(M+3)}+\frac{(1-\alpha^2)}{(M+1)^2} \tag{S.6}$$

Also,

$$P\left(c_{12}\neq c_{22}\mid c_{11}\neq c_{21},(\gamma_1,\gamma_2)\right)=\begin{cases}\frac{M(M+4)+2}{(M+2)(M+3)} & \text{if } (\gamma_1,\gamma_2)=(1,1)\\ \frac{M}{M+1} & \text{otherwise,}\end{cases}$$

from which

$$P\left(c_{12}\neq c_{22},c_{11}\neq c_{21}\right)=\left(\frac{M(M+4)+2}{(M+2)(M+3)}\right)\left(\frac{M}{M+1}\right)\alpha^2+\left(\frac{M}{M+1}\right)^2\left(1-\alpha^2\right). \tag{S.7}$$

The result now easily follows by summing (S.6) and (S.7). □

A.3 Proof of Proposition 3

Proof. Let $P_{C_t}=\{\rho_t\in P:\rho_{t-1}^{\mathfrak{R}_t}=\rho_t^{\mathfrak{R}^t}\}$ 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 $\sum_{\lambda\in P_{C_t}}\Pr(\rho_t=\lambda)=\Pr(\rho_t^{\mathfrak{R}_t})$ 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

$$\begin{aligned}\Pr(\rho_t^{\mathfrak{R}_t}) &= \sum_{c_{\gamma_t}} \Pr(\rho_t = \{c_{1t}, \ldots, c_{mt}\}) \\ &= \sum_{\lambda \in P_{C_t}} \Pr(\rho_t = \lambda),\end{aligned}$$

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., $\lambda \in P_{C_t}$). □

B Details Associated With the MCMC Algorithm

Here we provide much more detail associated with the MCMC scheme. We place emphasis on the updating steps for $\boldsymbol{\gamma}_{it}$ and ρ_t as the other steps are straightforward once these parameters have been updated. That said, pseudocode of the entire algorithm is provided in Algorithm 1. A key component of the MCMC algorithm is to check the compatibility between ρ_{t-1} and ρ_t, and between ρ_t and ρ_{t+1} when updating $\boldsymbol{\gamma}_t$ and ρ_t. This is equivalent to ensuring that $\rho_{t-1}^{\mathfrak{R}_t} = \rho_t^{\mathfrak{R}_t}$ and $\rho_t^{\mathfrak{R}_{t+1}} = \rho_{t+1}^{\mathfrak{R}_{t+1}}$. We describe the process of updating each of the c_{it} and γ_{it} sequentially in time so that the entire vector $\boldsymbol{\gamma}_t$ is updated first and then $\boldsymbol{c}_t$.

B.1 Updating γ_{it}

First note that $\boldsymbol{\gamma}_t$ only connects ρ_{t-1} to ρ_t so that when updating γ_{it} only compatibility between ρ_{t-1} and ρ_t needs to be checked (i.e., ρ_t remains compatible with ρ_{t+1} due to γ_{t+1} by construction). We detail updating γ_{it} in an MCMC algorithm based on its full

conditional found in (8) of the main paper and which we provide here for sake of clarity

$$\Pr(\gamma_{it} = 1 \mid -) = \frac{\alpha_t}{\alpha_t + (1-\alpha_t)\Pr(\rho_t^{\mathfrak{R}_t^{(+i)}})/\Pr(\rho_t^{\mathfrak{R}_t^{(-i)}})} \mathrm{I}[\rho_{t-1}^{\mathfrak{R}_t^{(+i)}} = \rho_t^{\mathfrak{R}_t^{(+i)}}]. \tag{S.8}$$

The appeal of this form of the full conditional compared to that found in (7) of the main paper is that the EPPF used to compute $\Pr(\rho_t^{\mathfrak{R}_t^{(+i)}})$ and $\Pr(\rho_t^{\mathfrak{R}_t^{(-i)}})$ need not have a tractable normalizing constant. That said, an exchangeable sequence of cluster labels $(c_{1t}, \ldots c_{mt})$ is necessary. Now let $\gamma_{it}^{(d)}$ and $\rho_t^{(d)}$ be the value of γ_{it} and ρ_t at the dth MCMC iterate. Note that there are four scenarios to consider when moving from $\gamma_{it}^{(d-1)}$ to $\gamma_{it}^{(d)}$. They are

1. $\gamma_{it}^{(d-1)} = 1 \to \gamma_{it}^{(d)} = 0$ (For this move $\rho_{t-1}^{(d)\mathfrak{R}_t^{(-i)}} = \rho_t^{(d-1)\mathfrak{R}_t^{(-i)}}$ continues to hold),
2. $\gamma_{it}^{(d-1)} = 1 \to \gamma_{it}^{(d)} = 1$ (For this move $\rho_{t-1}^{(d)\mathfrak{R}_t^{(+i)}} = \rho_t^{(d-1)\mathfrak{R}_t^{(+i)}}$ continues to hold),
3. $\gamma_{it}^{(d-1)} = 0 \to \gamma_{it}^{(d)} = 0$ (For this move $\rho_{t-1}^{(d)\mathfrak{R}_t^{(-i)}} = \rho_t^{(d-1)\mathfrak{R}_t^{(-i)}}$ continues to hold), and
4. $\gamma_{it}^{(d-1)} = 0 \to \gamma_{it}^{(d)} = 1$ (For this move $\rho_{t-1}^{(d)\mathfrak{R}_t^{(+i)}} = \rho_t^{(d-1)\mathfrak{R}_t^{(+i)}}$ needs to be verified).

Thus compatibility needs to be checked only for (4).

As expected, calculating the ratio $\Pr(\rho_t^{\mathfrak{R}_t^{(+i)}})/\Pr(\rho_t^{\mathfrak{R}_t^{(-i)}})$ in (S.8) is the most challenging part of computing $\Pr(\gamma_{it} = 1 \mid -)$. However, it can be straightforwardly calculated using exchangeability and ideas from Neal (2000). Under the assumption of exchangeability, which permits allocating the ith unit by treating it as if it were the last unit, we have that

$$\frac{\Pr(\rho_t^{\mathfrak{R}_t^{(+i)}})}{\Pr(\rho_t^{\mathfrak{R}_t^{(-i)}})} = \frac{\Pr(c_{it}|\rho_t^{\mathfrak{R}_t^{(-i)}})\Pr(\rho_t^{\mathfrak{R}_t^{(-i)}})}{\Pr(\rho_t^{\mathfrak{R}_t^{(-i)}})} = \Pr(c_{it}|\rho_t^{\mathfrak{R}_t^{(-i)}}). \tag{S.9}$$

Note that the probability in (S.9) is a standard calculation, used in Neal's Algorithm 8, for each c_{it}. When the EPPF does not have a tractable normalizing constant, one may compute the unnormalized probability of allocation to each of the k_t existing clusters and to a new singleton cluster and then normalize to obtain (S.1). Of course, here we know the

value of c_{it} from $\rho_t^{(d-1)}$ and, in constraints to Neal's Algorithm 8, this computation is done for the sake of computing the full conditional distribution of γ_{it}. Once (S.9) is calculated, computing (S.8) and updating γ_{it} is straightforward.

B.2 Updating ρ_t Using Cluster Labels

First note that only those c_{it} that correspond to $\gamma_{it} = 0$ are updated. As a result, the compatibility between ρ_{t-1} and ρ_t is preserved and so only the compatibility between ρ_t and ρ_{t+1} needs to be checked when updating any of the c_{it}. Recall that the full conditional of c_{it} corresponding to $\gamma_{it} = 0$ is

$$\Pr(c_{it} = h \mid -) \propto \begin{cases} N(Y_{it} \mid \mu^{\star}_{c_{it}=h,t}, \sigma^{2\star}_{c_{it}=h,t}) \Pr(c_{it} = h) \mathrm{I}[\rho_t^{h\mathfrak{R}_{t+1}} = \rho_{t+1}^{\mathfrak{R}_{t+1}}] & \text{for } h = 1, \ldots, k_t^{-i}, \\ N(Y_{it} \mid \mu^{\star}_{new_h,t}, \sigma^{2\star}_{new_h,t}) \Pr(c_{it} = h) \mathrm{I}[\rho_t^{h\mathfrak{R}_{t+1}} = \rho_{t+1}^{\mathfrak{R}_{t+1}}] & \text{for } h = k_t^{-i} + 1, \end{cases} \tag{S.10}$$

where $\Pr(c_{it} = h) = \Pr(c_{1t}, \ldots, c_{it} = h, \ldots, c_{mt})$, and k_t^{-i} is the number of clusters at time t when the ith unit has been removed. The partition constructed from $\{c_{1t}, \ldots, c_{it} = h, \ldots, c_{mt}\}$ is denoted as $\rho_t^h = \{S_{1t}^{-i}, \ldots, S_{ht}^{-i} \cup \{i\}, \ldots, S_{k_t^{-i}t}^{-i}\}$ where S_{jt}^{-i} denotes the jth cluster at time t with the ith unit removed. Note that it is possible that $S_{jt}^{-i} = S_{jt}$. Further, abusing notation, for $h = k_t^{-i} + 1$ we have $\rho_t^h = \{S_{1t}^{-i}, \ldots, S_{ht}^{-i}, \ldots, S_{k_t^{-i}t}^{-i}, \{i\}\}$. Additionally, $\mu^{\star}_{new_h,t}$ 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). Then based on a spatial product partition model for ρ_t, for $h = 1, \ldots, k_t^{-i}$ the $\Pr(c_{it} = h)$ becomes

$$\Pr(c_{it} = h) = Pr(\rho_t^h) \propto M\Gamma(|S_{ht}^{-i} \cup \{i\}|) g(\boldsymbol{s}_{ht}^{-i\star} \cup \boldsymbol{s}_i | \nu_0) \prod_{j \neq h}^{k_t^{-i}} M\Gamma(|S_{jt}^{-i}|) g(\boldsymbol{s}_{jt}^{-i\star} | \nu_0), \tag{S.11}$$

while for $h = k_t^{-i} + 1$

$$\Pr(c_{it} = h) = Pr(\rho_t^h) \propto M\Gamma(|\{i\}|)g(\boldsymbol{s}_i|\nu_0)\prod_{j=h}^{k_t^{-i}} M\Gamma(|S_{jt}^{-i}|)g(\boldsymbol{s}_{jt}^{-i\star}|\nu_0), \tag{S.12}$$

where $g(\cdot)$ is the auxiliary similarity function detailed in Page & Quintana (2016) and $\boldsymbol{s}_{jt}^{-i\star} = \{\boldsymbol{s}_{i'} : i' \in S_{jt}^{-i}\}$ are the spatial coordinates from units that belong to the jth cluster at time t. Updating c_{it} can be carried out by evaluating (S.10) based on (S.11) or (S.12) for each $h = 1, \ldots, k_t^{-i} + 1$ and then normalizing.

Once each of the $\boldsymbol{c}_t$ and $\boldsymbol{\gamma}_t$ are updated the MCMC algorithm is completed by cycling through remaining model and latent parameters found in model (9) and updating them on an individual basis using well known approaches. In order to visualize all the moving parts of the MCMC algorithm we provide some pseudocode in Algorithm 1. For sake of simplicity, Algorithm 1 describes an MCMC procedure that can be employed to sample from the joint posterior distribution based on model (5).

C Simulation Studies

In this section we provide more details associated with Simulation 1, the competitors included in Simulation 3, and provide results from additional simulations similar to that described in the Section 3.3 of the main document. We then provide details associated with a simulation study that includes spatial information.

C.1 Simulation 1: Continued

Table S.1 contains the adjusted Rand index values between the estimated partitions and that which was used to generate data. Interestingly, as α increases, the ARI values also tend to increase.

Algorithm 1 : Pseudocode for the MCMC algorithm for model (5) of main article. Let T be the number of time points, m the number of units at each time point, and D the number of MCMC iterations.

for $d = 1, \ldots, D$ **do**
  for $t = 1, \ldots, T$ **do** ▷ For each t, update the entire $\boldsymbol{\gamma}_t$ vector first and then $\boldsymbol{c}_t$
    for $i = 1, \ldots, m$ **do**
      Set $\gamma_{i1}^{(d)} = 0$.
      if $t > 1$ **then**
        - Update γ_{it} based on procedure described in Section B.1. That is,
        if $\gamma_{it}^{(d-1)} = 1$ **then**
          Move to $\gamma_{it}^{(d)}$ using Bernoulli probability in (S.8). Compatibility holds by construction.
        if $\gamma_{it}^{(d-1)} = 0$ **then**
          Move to $\gamma_{it}^{(d)}$ using Bernoulli probability in (S.8). If $\gamma_{it}^{(d)} = 0$, then compatibility holds by construction. If $\gamma_{it}^{(d)} = 1$, the compatibility needs to be checked. If $\rho_{t-1}^{(d)\mathfrak{R}_t^{(+i)}} \neq \rho_t^{(d-1)\mathfrak{R}_t^{(+i)}}$, then set $\gamma_{it}^{(d)} = 0$.
    for $i = 1, \ldots, m$ **do**
      - Update c_{it} based on procedure described in Section B.2.
      for $h = 1, \ldots, k_t^{-i}$ **do**
        Compute the unnormalized multinomial probability in (S.10) based on (S.11).
        if $\rho_t^{h\mathfrak{R}_{t+1}} \neq \rho_{t+1}^{\mathfrak{R}_{t+1}}$ **then**
          Set unnormalized multinomial probability to zero.
      for $h = k_t^{-i} + 1$ **do**
        Compute the unnormalized multinomial probability in (S.10) based on (S.12).
      Sample c_{it} using the normalized $k_t^{-i} + 1$ multinomial probabilities.
    for $j = 1, \ldots, K^{(d)}$ **do** ▷ $K^{(d)}$ = number of clusters at the dth iteration.
      - Update $\mu_{jt}^{\star}$ based on Gaussian full conditional derived using well known arguments.
      - Update $\sigma_{jt}^{2\star}$ using a random walk Metropolis step.
    - Update θ_t based on Gaussian full conditional derived using well known arguments.
    - Update α_t based on beta full conditional derived using well known arguments .
  - Update τ^2 using a random walk Metropolis step.
  - Update ϕ_0 based on Gaussian full conditional derived using well known arguments.
  - Update λ^2 using a random walk Metropolis step.

Table S.1: Adjusted Rand index when comparing $\hat{\rho}_t$ to the true ρ_t for $t = 1, \ldots, 5$. Note that $ARI(\cdot, \cdot)$ denotes the adjusted Rand index as a function of two partitions. These values are averaged over the 100 generated data sets. The values in parenthesis are Monte Carlo standard errors.

	$ARI(\hat{\rho}_1, \rho_1)$	$ARI(\hat{\rho}_2, \rho_2)$	$ARI(\hat{\rho}_3, \rho_3)$	$ARI(\hat{\rho}_4, \rho_4)$	$ARI(\hat{\rho}_5, \rho_5)$
$\alpha = 0.0$	0.58 (0.03)	0.63 (0.03)	0.58 (0.03)	0.54 (0.03)	0.56 (0.03)
$\alpha = 0.1$	0.63 (0.03)	0.56 (0.03)	0.55 (0.03)	0.62 (0.03)	0.57 (0.03)
$\alpha = 0.25$	0.52 (0.03)	0.57 (0.03)	0.55 (0.03)	0.63 (0.03)	0.62 (0.03)
$\alpha = 0.5$	0.60 (0.03)	0.70 (0.03)	0.69 (0.03)	0.66 (0.02)	0.59 (0.03)
$\alpha = 0.75$	0.78 (0.02)	0.77 (0.02)	0.82 (0.02)	0.80 (0.02)	0.75 (0.02)
$\alpha = 0.9$	0.83 (0.02)	0.86 (0.02)	0.87 (0.02)	0.84 (0.02)	0.76 (0.02)
$\alpha = 0.9999$	0.92 (0.01)	0.92 (0.01)	0.92 (0.01)	0.92 (0.01)	0.92 (0.01)

C.2 Simulation 3: Continued

As referenced in the main article, Figure S.1 provides an example of the type of data that is generated in the simulation of Section 3.3 in the main article. To each of the 100 data sets generated, we fit our method and four other procedures. We know provide specific details of the competing methods.

1. weighted DDP (wddp): A complete description of this model can be found in Section 4 of Quintana et al. (2020) (and Müller et al. 2015). We only provide pertinent details here. Let $\boldsymbol{z}_i = (Y_i, t_i)$ be the response and time pair for $i = 1, \ldots, mT$. The wddp models $\boldsymbol{z}_i$ with a Dirichlet process mixture model (DPM) and then derives the conditional model $(Y_i|t_i)$. In hierarchical form (including cluster labels) the model is

$$
\begin{aligned}
\boldsymbol{z}_i|\boldsymbol{\mu}^\star, \boldsymbol{\Sigma}^\star, c_i &\overset{ind}{\sim} N_2(\boldsymbol{\mu}^\star_{c_i}, \boldsymbol{\Sigma}^\star_{c_i}), \quad i = 1, \ldots, mT \\
\boldsymbol{\mu}^\star_j &\overset{iid}{\sim} N_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0), \quad j = 1, \ldots, K \\
\boldsymbol{\Sigma}^\star_j &\overset{iid}{\sim} \text{Inverse-Wishart}_2(\nu, \psi\boldsymbol{I}), \quad j = 1, \ldots, K \\
\boldsymbol{\mu}_0 &\overset{iid}{\sim} N_2(\boldsymbol{m}, s^2\boldsymbol{I}), \\
\boldsymbol{\Sigma}_0 &\overset{iid}{\sim} \text{Inverse-Wishart}_2(\nu_0, \psi_0\boldsymbol{I}), \\
\Pr(c_i = j) = \pi_j \quad &\text{where } \pi_j = V_j \textstyle\prod_{\ell<j}(1 - V_\ell), \\
V_\ell &\overset{iid}{\sim} \text{Beta}(1, M).
\end{aligned}
$$

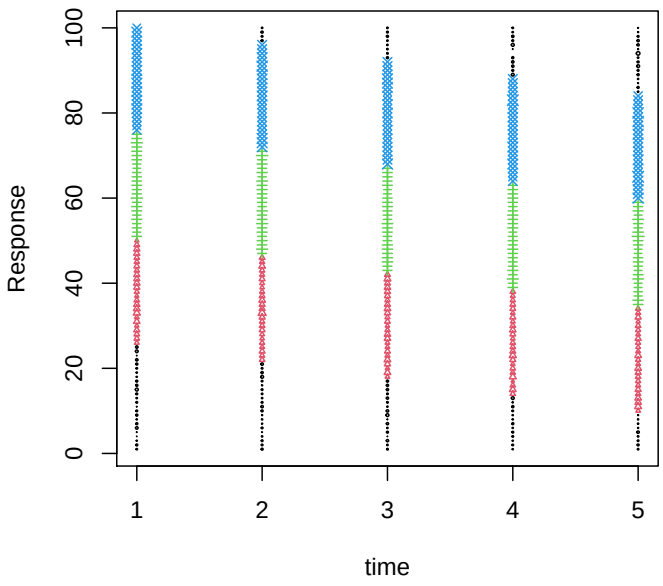

Figure S.1: One realization of a synthetic data set from simulation study in Section 3.3. Each color corresponds to a cluster and the size of plotted symbol is proportional to the value of point being plotted.

We set $\boldsymbol{m} = 0\boldsymbol{j}$, $s^2 = 25$, $\nu = \nu_0 = 4$, $\psi = \psi_0 = 1$, $K = 30$, and $M = 1$. The model induces a weight-dependent mixture model of regressions

$$f(y_i|t_i) = \sum_{j=1}^{K} w_j(t_i) N(y_i|\beta_{0j}^\star + \beta_{1j}^\star t_i, \sigma_j^{2\star}),$$

where

$$w_j(t_i) = \frac{\pi_j N(t_i|\mu_{2j}^\star, \Sigma_{22j}^\star)}{\sum_{\ell=1}^{K} \pi_\ell N(t_i|\mu_{2\ell}^\star, \Sigma_{22\ell}^\star)}, \quad j = 1, \ldots, K,$$

and $\beta_{0j}^\star = \mu_{1j}^\star - \frac{\Sigma_{12j}^\star}{\Sigma_{22j}^\star}\mu_{2j}^\star$, $\beta_{1h} = \frac{\Sigma_{12j}^\star}{\Sigma_{22j}^\star}$, and $\sigma_j^{2\star} = \Sigma_{11j}^\star - \frac{\Sigma_{12j}^\star \Sigma_{21j}^\star}{\Sigma_{22j}^\star}$. Note that time is include in the weights of the the conditional model which is employed to calculate LPML and WAIC. A blocked Gibbs sampler was employed to sample from the posterior where $V_K = 1$ to ensure that $\sum_{j=1}^{K} \pi_j = 1$.

2. linear DDP (lddp): A complete description of this model is provided in chapter 4.4.2 of Müller et al. (2015) (see also Quintana et al. 2020). We only provide pertinent details here. As with the wddp model, for the lddp we consider (Y_i, t_i) for $i = 1, \ldots, mT$. Time is incorporated in the atoms of a Dirichlet process (DP) so that the jth atom is expressed as $\sum_\ell^q B_\ell(t, \boldsymbol{\xi})\beta_{j\ell}$ where $B_\ell(t, \boldsymbol{\xi})$ denotes the ℓ-th B-spline basis function evaluated at t for knots $\boldsymbol{\xi}$. Letting $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$ and $\boldsymbol{B}(t_i, \boldsymbol{\xi})$ the q-dimensional B-spline basis vector for unit i and after introducing cluster labels, the lddp model can be expressed hierarchically as

$$\begin{aligned}
Y_i|\boldsymbol{\beta}^\star, \sigma^{2\star}, c_i &\overset{ind}{\sim} N(\boldsymbol{B}'(t_i, \boldsymbol{\xi})\boldsymbol{\beta}_{c_i}^\star, \sigma_{c_i}^{2\star}), \quad i = 1, \ldots, mT, \\
\boldsymbol{\beta}_j^\star &\overset{iid}{\sim} N_q(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0), \quad j = 1, \ldots, k, \\
\sigma_j^{2\star} &\sim \text{Inverse-Gamma}(a, b), \quad j = 1, \ldots, k, \\
\boldsymbol{\beta}_0 &\overset{iid}{\sim} N_2(\boldsymbol{m}, s^2\boldsymbol{I}), \\
\boldsymbol{\Sigma}_0 &\overset{iid}{\sim} \text{Inverse-Wishart}_2(\nu_0, \psi_0\boldsymbol{I}), \\
\{c_1, \ldots, c_{mT}\} &\sim CRP(M).
\end{aligned}$$

We set $\boldsymbol{m} = 0\boldsymbol{j}$, $s^2 = 25$, $\nu_0 = q + 2$, $\psi_0 = 10$, $a = b = 1$, and $M = 1$. Neal's Algorithm 8 (Neal 2000) was employed to sample from the posterior distribution.

3. Griffiths-Milne dependent Dirichlet process (gmddp) mixture. This is carried out using `DDPdensity` in the `BNPmix` package found in `R`. The function considers partially exchangeable data (Lijoi et al., 2014) such that exchangeability is assumed within each group and the vector of random probability measures at each time point are modeled jointly as a vector of GM-DDP.

4. A temporally independent $CRP(M)$ model (ind_crp): This model is a special case of Caron et al. (2007)'s and our model. Specifically, α is set to 0. For this procedure, the following model was fit separately for each time period.

$$
\begin{aligned}
Y_i \mid \boldsymbol{\mu}^\star, \boldsymbol{\sigma}^{2\star}, c_i &\overset{ind}{\sim} N(\mu^\star_{c_i}, \sigma^{2\star}_{c_i}),\ i = 1, \ldots, m, \\
(\mu^\star_j, \sigma^\star_j) \mid \theta, \tau^2 &\overset{ind}{\sim} N(\theta, \tau^2) \times UN(0, A_\sigma),\ j = 1, \ldots, k, \\
(\theta, \tau) &\overset{iid}{\sim} N(m_0, s_0^2) \times UN(0, A_\tau).
\end{aligned}
$$

   We set $m_0 = 0$, $s_0^2 = 10^2$, $A_\sigma = 0.5sd(Yvec)$, and $A_\tau = 100$. Neal's Algorithm 8 (Neal 2000) was used to sample from the posterior distribution.

5. A temporally static $CRP(M)$ model (static_crp): This procedure is a special case of Caron et al. (2007) ($\alpha = 1$) and is fit to a concatenated version of the data Y_i, for $i = 1, \ldots, mT$. Specifically, the following model was fit

$$
\begin{aligned}
Y_i \mid \boldsymbol{\mu}^\star, \boldsymbol{\sigma}^{2\star}, c_i &\overset{ind}{\sim} N(\mu^\star_{c_i}, \sigma^{2\star}_{c_i}),\ i = 1, \ldots, mT, \\
(\mu^\star_j, \sigma^\star_j) \mid \theta, \tau^2 &\overset{ind}{\sim} N(\theta, \tau^2) \times UN(0, A_\sigma),\ j = 1, \ldots, k, \\
(\theta, \tau) &\overset{iid}{\sim} N(m_0, s_0^2) \times UN(0, A_\tau).
\end{aligned}
$$

   We set $m_0 = 0$, $s_0^2 = 10^2$, $A_\sigma = 0.5sd(Yvec)$, and $A_\tau = 100$. Neal's Algorithm 8 (Neal 2000) was used to sample from the posterior distribution.

C.2.1 Results from Additional Synthetic Data

In addition to the synthetic data generated in Simulation 3 of the main document, we also generated data as described in the following two scenarios.

Scenario 1: For the ith unit, we employ the following as a data generating mechanism

$$y_{it} = \omega y_{it-1} + \epsilon_{it}, \text{ for } i = 1, \ldots, m, \text{ and } t = 1, \ldots, T,$$

where $|\omega| < 1$ and $\epsilon_{it} \sim N(0, v^2)$. For this scenario measurements are are correlated across time, but independent between the m units. Data were generated with $m = 100$ and using the following levels of factors of interest

- $\omega \in \{0, 0.1, 0.25, 0.5, 0.75, 0.9\}$
- $v^2 \in \{0.5^2, 1^2\}$
- $T \in \{5, 10\}$

Notice that for this scenario, there is no "true" partition and so we are interested only in comparing the model fit of our approach to that of the five competitors.

Scenario 2: For the ith unit, we employ the following as a data generating mechanism

$$y_{it} = \omega_{c_i} y_{it-1} + \epsilon_{it}, \text{ for } i = 1, \ldots, m, \text{ and } t = 1, \ldots, T,$$

where as before $c_{it} \in \{1, 2, 3, 4\}$ with $\omega \in \{-0.75, -0.25, 0.25, 0.75\}$ and $\epsilon_{it} \sim N(0, v^2)$. As in the previous scenarios measurements are correlated across time, but independent between the m units. Data were generated with $m = 100$ and using the following levels of factors of interest

- $v^2 \in \{0.5^2, 1^2\}$
- $T \in \{5, 10\}$

In this scenario, there is a "true" partition but our approach, nor the competitors, are parametrized in such a way as to detect it. Indeed, our method models temporal

dependence only through the partition (i.e, there is no temporal correlation parameter in the likelihood). That said, we still compare partition recovery by way of the adjusted Rand index (ARI) in addition to model fit.

In both scenarios the function `arima.sim` in `R` (R Core Team 2020) is used to generate the 100 replicate data sets. We collect 1,000 MCMC iterates after discarding the first 25,000 as burn-in and thinning by 25 (i.e., 50,000 total MCMC draws were collected). The prior parameters that we used are $A_\sigma = 0.5sd(vec(Y))$, $A_\tau = 100$, $A_\lambda = 100$, $m_0 = 0$, $s2_0 = 100$, $a_\alpha = 1$, $b_\alpha = 1$. WAIC is used to compare each of the procedures in terms of model fit and ARI to compare ability of estimating the true partition structure. Results are found in Figures S.2 - S.4. From Figure S.2 we see that our approach produces the smallest WAIC for all factors considered Scenarios 1's data generating schemes. Thus, our approach tends to fit the data best.

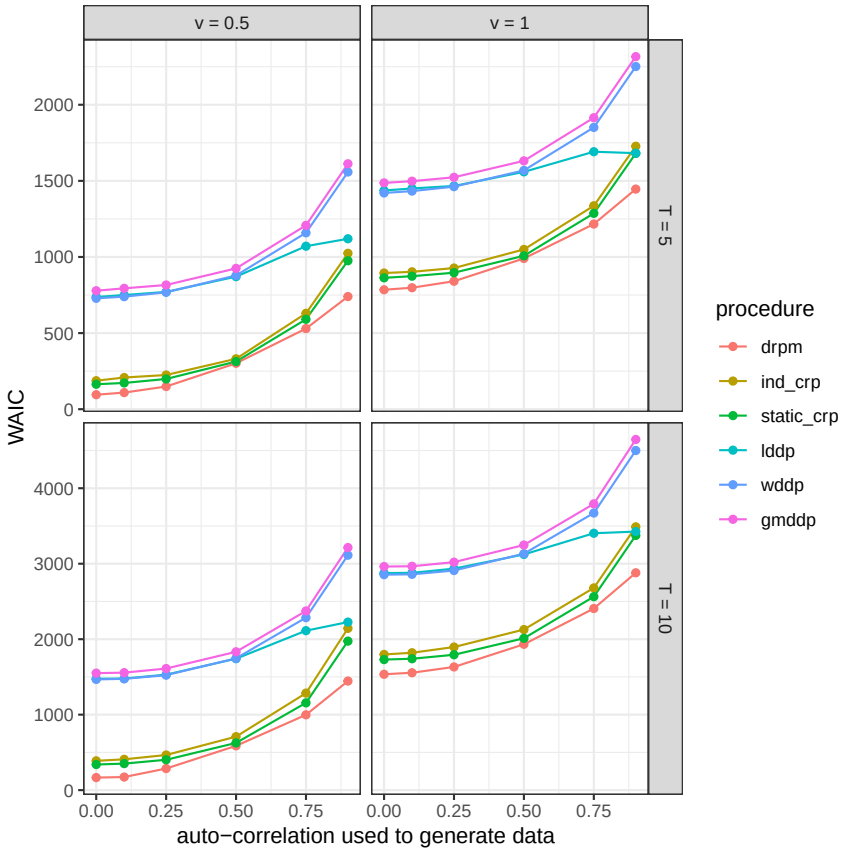

Figure S.2: Results for WAIC from the first data generating scenario.

For Scenario 2, static_crp is quite competitive to our approach and produces similar WAIC values. Apart from that, our approach does better than the other competitors. From Figure S.4 our approach does much better at recovering the partition compared to other procedures. That said, we the ARI values are still quite small (which was expected).

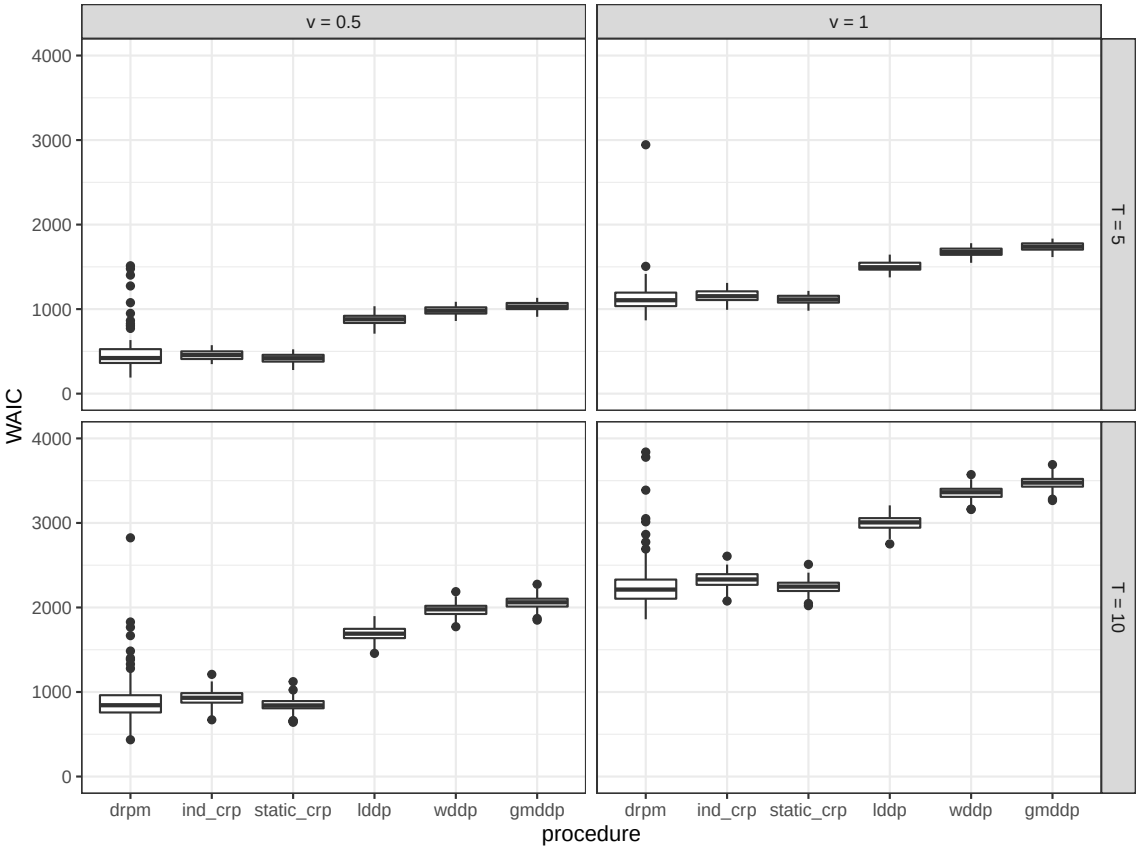

Figure S.3: Results for WAIC from the fourth data generating scenario.

C.3 Simulation 4: Space-Time Data Generation

Here we 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

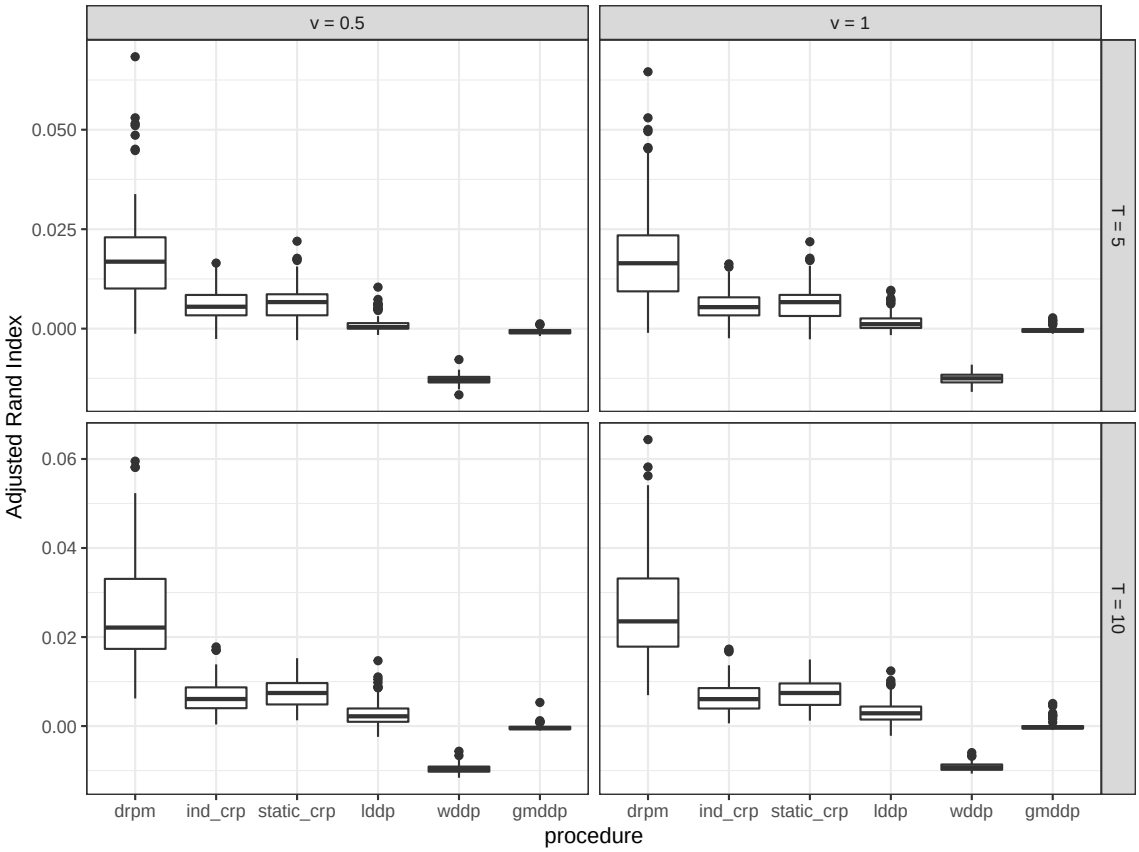

Figure S.4: Results for ARI from the fourth data generating scenario.

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 & 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 (5) as a data generating mechanism. Spatio-temporal partitions were generated using (11) 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 (11) 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 (5) 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 S.5 and S.6.

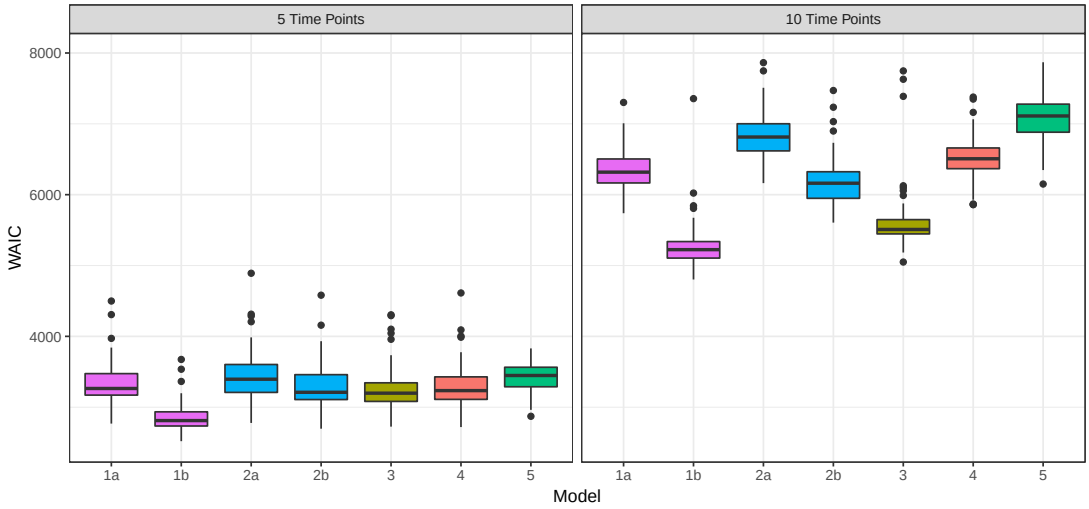

Figure S.5: Results from simulation study when observations were generated using a spatio-temporal Gaussian process. Boxplots display the 100 WAIC values that correspond to model fit for each synthetic data set. Note that smaller WAIC values indicate a better fit.

The primary purpose of Figure S.5 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. 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 S.6 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 $tRPM(\alpha, M)$

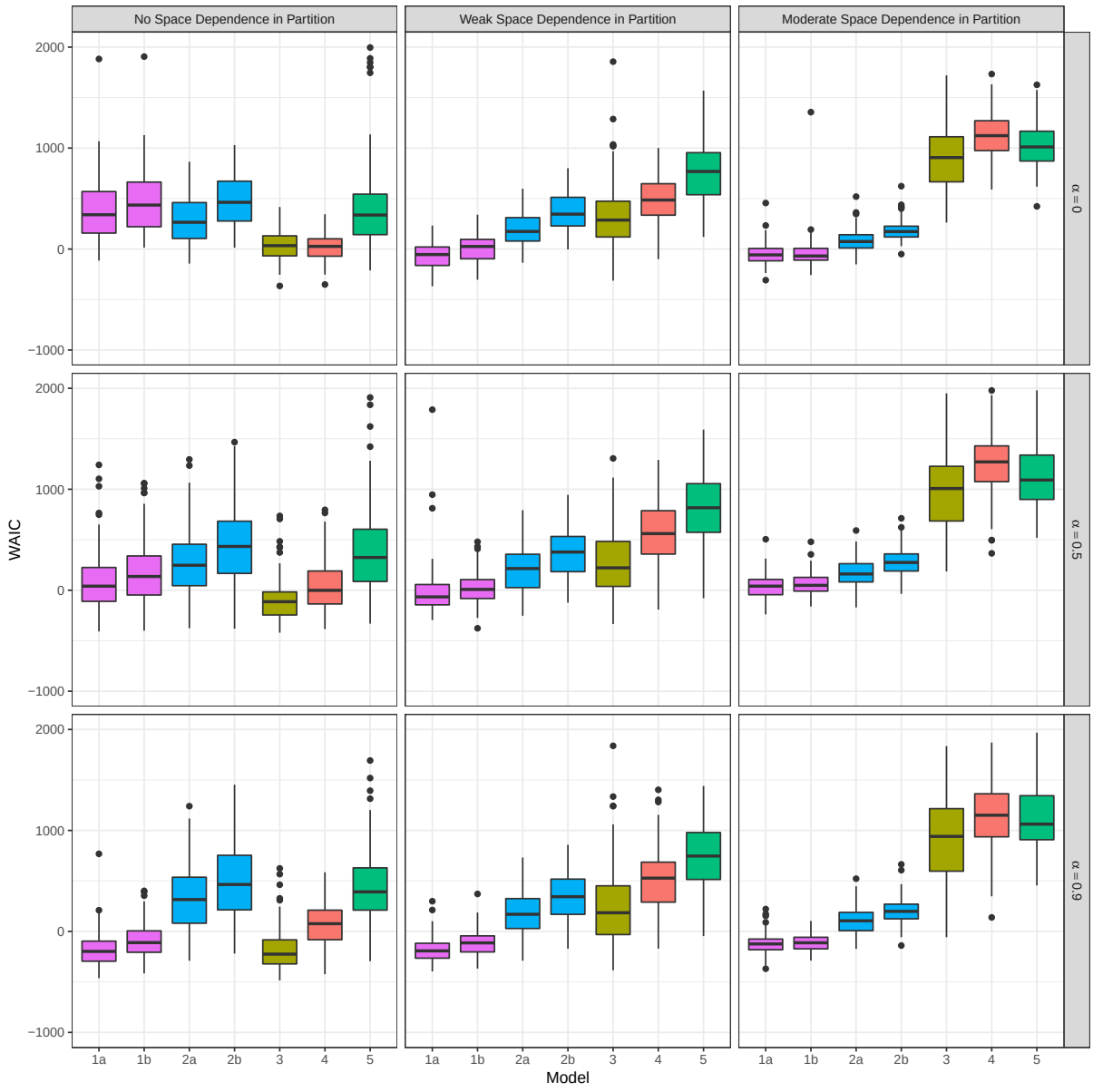

Figure S.6: Results from simulation study for the scenario in which partition structure is included in data generation process. Boxplots display the 100 WAIC values that correspond to model fit for each synthetic data generating scenario. Note that smaller indicates better fit.

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.

D Data Applications

In this section, we detail an additional application in the field of education.

D.1 SIMCE Data Application

Incorporating spatio-temporal structure in education studies has been explored (e.g., Cepeda-Cuervo & Núñez-Antón 2013, Fotheringham et al. 2001). In school assessment and effectiveness studies, temporal persistence in school performance is of principal interest. It seems likely that school performance from year-to-year is relatively stable except in circumstances where a school undergoes many changes in personnel (faculty and students) or curriculum from one year to the next. In addition it seems reasonable that geographic location plays a role in school performance, particularly if communities are segregated socio-economically which happens to be the case in metropolitan area of Santiago, Chile. For these reasons we fit model (9) to these data as well.

In order to formally assess both national and school level education effectiveness in Chile, the Chilean national learning outcome assessment system (Sistema de Medición de Calidad de la Educación, SIMCE) was created to, among other things, administer standardized tests to education institutions in Chile. Each year a standardized test in mathematics and language is administered to 4th grade students. We were granted access to 7 years of data (2005-2011) where the longitude and latitude of most schools were recorded.

For the SIMCE data in addition to the 16 models fit to the PM_{10} data, we also considered an alternative to employing the sPPM at each time period which is more computationally efficient. The alternative approach models only ρ_1 with an $sPPM(\nu_0, M)$ (equation (10) of the main document) and the remaining $T-1$ partitions with an $tRPM(\boldsymbol{\alpha}, M)$ (equation

(4) of the main document) with a $CRP(M)$ EPPF. In this formulation, the strength of α_t would be the only mechanism by which the spatial structure found in ρ_1.

We employed the same prior parameter values here as in Section 4.1 of the main document, except we set $A_\sigma = 15$ and $\nu_0 = 2$ to account for the higher variability present in the SIMCE data. Each of the 24 models were fit to the SIMCE data by collecting 1000 MCMC draws after discarding the first 5000 as burn-in and thinning by 5. The LPML and WAIC results can be found in Table S.2.

Similar to the PM_{10} analysis, the best performing model in terms of WAIC includes spatio-temporal dependence in the partition model, temporal dependence among the atoms, and temporal dependence in the likelihood. The best performing model in terms of LPML assumed atoms are *iid*. Notice further, that generally speaking, incorporating temporal dependence in the model for $(\rho_1, \ldots, \rho_7)$ improves model fit. It appears that there is a cost in model fit associated with employing the $sPPM(\nu_0, M)$ at the first time period and the $CRP(M)$ for subsequent time periods in terms of model fit. However, the cost is not exorbitant relative to extraordinary computation gains (12 hours for model that includes space at time 1 versus 6 days for the model that includes space at each time point). To see how estimated partitions from the two models (that which includes space in the first time point versus that which includes space at each time point) change over time, we provide Figure S.7. Notice that there is a change in dependence from time period 2 and 3 and that the similarity between partitions decays when including space at each time point.

References

Caron, F., Davy, M. & Doucet, A. (2007), Generalized polya urn for time-varying dirichlet process mixtures, *in* 'Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence', UAI'07, AUAI Press, Arlington, Virginia, United States, pp. 33–40.

URL: *http://dl.acm.org/citation.cfm?id=3020488.3020493*

Cepeda-Cuervo, E. & Núñez-Antón, V. (2013), 'Spatial double generalized beta regression

Table S.2: Results of fitting 24 models to the SIMCE data. The bold font identifies the models that produced the best LPML and WAIC values. Higher values for LPML indicate better fit while lower values for WAIC indicate better fit.

			Space					
			No		Yes			
Temporal Dependence In					Each Time		First Time	
Partition	Likelihood	Atoms	LPML	WAIC	LPML	WAIC	LPML	WAIC
No	No	No	-34094	62963	-33543	62416	-34054	62960
No	No	Yes	-34040	62693	-33558	62577	-34044	63043
No	Yes	No	-31214	60087	-30701	60400	-31129	59349
No	Yes	Yes	-31241	59572	-30712	60944	-31115	59686
Yes	No	No	-32457	64835	-30760	61045	-31198	61516
Yes	No	Yes	-31007	61948	-31180	61348	-32690	64578
Yes	Yes	No	-30390	60340	**-29936**	58122	-30573	60132
Yes	Yes	Yes	-30378	60314	-30959	**57834**	-30331	59544

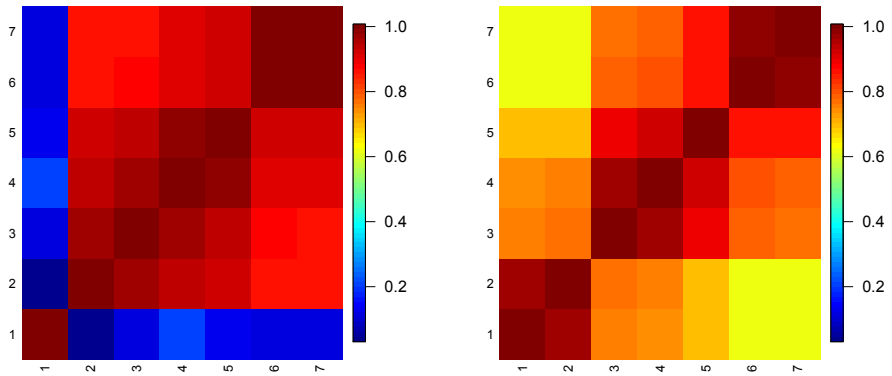

Figure S.7: Each figure is a summary of the lagged *ARI* value corresponding to models that include space in different ways. The left plot corresponds to model that includes space in partition model only at time period 1. The right plot corresponds to model that includes space in partition model at each of the seven time periods.

models: Extensions and application to study quality of education in colombia', *Journal of Educational and Behavioral Statistics* **38**, 604–628.

De Iorio, M., Johnson, W., Müller, P. & Rosner, G. (2009), 'Bayesian nonparametric nonproportional hazards survival modeling', *Biometrics* **65**(3), 762–771.

Fotheringham, A. S., Charlton, M. E. & Brunsdon, C. (2001), 'Spatial variations in school performance: a local analysis using geographically weighted regression', *Geographical & Environmental Modelling* **5**, 43–66.

MacEachern, S. N. (2000), Dependent dirichlet processes, Technical report, Ohio State University.

Müller, P., Quintana, F. A., Jara, A. & Hanson, T., eds (2015), *Bayesian Nonparametric Data Analysis*, 1 edn, Springer International Publishing, Switzerland.

Müller, P., Quintana, F. & Rosner, G. L. (2011), 'A product partition model with regression on covariates', *Journal of Computational and Graphical Statistics* **20**(1), 260–277.

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

Padoan, S. A. & Bevilacqua, M. (2015), 'Analysis of random fields using CompRandFld', *Journal of Statistical Software* **63**(9), 1–27.
URL: *http://www.jstatsoft.org/v63/i09/*

Page, G. L. & Quintana, F. A. (2016), 'Spatial product partition models', *Bayesian Analysis* **11**, 265–298.

Quintana, F. A., Müller, P., Jara, A. & MacEachern, S. N. (2020), 'The dependent dirichlet process and related models'. arXiv:2007.06129v1.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*