Spatio-Temporal Stick-Breaking Process

Clara Grazian[*],[†]

Abstract. Dirichlet processes and their extensions have reached a great popularity in Bayesian nonparametric statistics. They have also been introduced for spatial and spatio-temporal data, as a tool to analyze and predict surfaces. A popular approach to Dirichlet processes in a spatial setting relies on a stick-breaking representation of the process, where the dependence over space is described in the definition of the stick-breaking probabilities. Extensions to include temporal dependence are still limited, however it is important, in particular for those phenomena which may change rapidly over time and space, with many local changes. In this work, we propose a Dirichlet process where the stick-breaking probabilities are defined to incorporate both spatial and temporal dependence. We will show that this approach is not a simple extension of available methodologies and can outperform available approaches in terms of prediction accuracy. The approach results in a predictive model that does not rely on the assumptions of Gaussianity or the separability of time and space.

MSC2020 subject classifications: Primary 62F15, 62G05; secondary 62M30.

Keywords: Dirichlet process, spatio-temporal modeling, stick-breaking processes, mixture models, rainfalls.

1 Introduction

Spatial data is data that describes a surface, usually through variables on a two-dimensional plane, i.e. data that references a specific geographical area or location. The analysis and prediction of surfaces are essential to a diverse range of fields, including ecology, meteorology, sociology, image recognition, real estate, etc.

Spatial data is defined as a realisation of a stochastic process over space

$$y(s) = \{y(s), s \in \mathcal{D}\},$$

where $\mathcal{D} \subset \mathbb{R}^d$. The collected data comes as $y = \{y(s_1), \ldots, y(s_n)\}$, where $(s_1, \ldots, s_n)$ is a set of spatial locations, such that $s_i = (s_{i,1}, s_{i,2})$ for $i = 1, \ldots, n$; for example, s_i may represent latitude and longitude. Depending on $\mathcal{D}$ being a continuous surface or a countable collection of d-dimensional spatial units, the problem can be specified as a spatially continuous or discrete random process, respectively (Gelfand et al., 2010). For example, data can be obtained by monitoring stations located in the set $(s_1, \ldots, s_n)$ of n points. Alternatively, data may be observed in a set $(s_1, \ldots, s_n)$ of n areas, defined, for example, by counties; we refer to this case as areal data.

---

[*]Carslaw Building, University of Sydney, Camperdown Campus, Sydney, Australia, clara.grazian@sydney.edu.au

[†]ARC Training Centre in Data Analytics for Resources and Environments (DARE), Sydney, Australia

The first step in defining a spatial model within a Bayesian framework is to identify a probability distribution for the observed data. We usually select a distribution from the exponential family, indexed by a set of parameters θ accounting for spatial correlation; for instance

$$y(s) = \theta(s) + x(s)^T\beta + \varepsilon(s), \tag{1.1}$$

where $\theta(s) \in \Theta$ is the spatial effect, $x(s)^T$ is a set of p covariates which may be dependent on the location s and β is a column-vector of p regression coefficients; finally, $\varepsilon(s)$ is a normal error such that $\varepsilon(s) \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ and errors are considered independent and identically distributed (i.i.d.) over space.

A common approach to the problem of modelling surfaces is kriging (Stein, 2012), which interpolates and smooths spatial data; in kriging, prediction $y(s^*)$ at a spatial location s^* is obtained as a weighted average, where observations at closer locations are given higher weights than observations at distant locations. The parameters $\theta(s)$ are defined as a latent stationary Gaussian field. This is equivalent to assuming that $\theta(s)$ has a multivariate normal distribution with mean μ and spatially structured covariance matrix Σ, whose generic element is $\Sigma_{ij} = \mathbb{C}\text{ov}(\theta(s_i), \theta(s_j))$ (Guttorp and Gneiting, 2006; Banerjee et al., 2003; Cressie, 2015; Rue and Held, 2005).

The concept of spatial process can be extended to the spatio-temporal case including a time dimension. The data are then defined by a process

$$y(s, t) = \{y(s, t), (s, t) \in \mathcal{D} \times \mathcal{T} \subset \mathbb{R}^d \times \mathbb{R}\}$$

and are observed at n spatial locations or areas and at T time points. When spatio-temporal geostatistical data is considered (Gelfand et al., 2010; Cressie and Wikle, 2015), the kriging model can be extended by defining a spatio-temporal covariance function given as $\mathbb{C}\text{ov}(\theta(s_i, t_\ell), \theta(s_j, t_h))$. Assuming stationarity in space and time, the space-time covariance function can be written as a function of the spatial Euclidean distance $d(s_i, s_j)$ and of the temporal lag $|t_\ell - t_h|$.

In practice, to overcome the computational complexity of these models, some simplifications are introduced. For example, under the separability hypothesis the space-time covariance function is decomposed into the sum (or the product) of a purely spatial and a purely temporal term, e.g. $\mathbb{C}\text{ov}(\theta(s_i, t_\ell), \theta(s_j, t_h)) = \sigma^2 C_1(d(s_i, s_j))C_2(|t_\ell - t_h|)$ (Finkenstadt et al., 2006). Alternatively, it is possible to assume that the spatial correlation is constant in time, giving rise to a space-time covariance function that is purely spatial when $t_\ell = t_h$ and zero otherwise. In this case, the temporal evolution could be introduced by assuming that the spatial process evolves in time following autoregressive dynamics (Harvill, 2010). It is also possible to avoid the assumption of separability between time and space; several examples of non-separable space-time covariance functions are reported in Cressie and Huang (1999) and Gneiting (2002).

The assumption of Gaussianity in kriging may be inappropriate or difficult to verify, and overly restrictive, e.g. by oversmoothing the prediction. A first relaxation may be represented by finite mixture models. The use of mixture models in the context of spatio-temporal data is useful because it allows to better describe the heterogeneity of the data, which is essential to optimise the allocation of resources and to perform better

prediction without the need to include a complete list of covariates (which may not be available). Fernández and Green (2002) and Neelon et al. (2014) propose models for areal (possibly multivariate) data, Frühwirth-Schnatter and Kaufmann (2008) introduce a clustering approach for dynamic regression models, while Hossain et al. (2014) analyze space-time mixtures with a particular focus on model selection. However, finite mixture models assume it is possible to consistently select the number of components, which may be difficult, in particular, in presence of covariates. Therefore, extensions to consider infinite mixture models have also been considered.

Mixture models can be generalised by considering an infinite number of components, through Dirichlet process mixture models. A random distribution F is distributed according to a Dirichlet process (DP) if its marginal distributions are Dirichlet distributed (Ferguson, 1973, 1974). When analyzing spatial data, we want the Dirichlet process to depend on a covariate represented by the location s. Dirichlet processes are then generalised to dependent Dirichlet processes (DDPs) (MacEachern, 2000; Barrientos et al., 2012), which model a collection of unknown distributions indexed by a covariate, through the definition of weights $\pi = (\pi_1, \pi_2, \ldots)$ and atoms $\theta = (\theta_1, \theta_2, \ldots)$ to vary according to the covariate. Spatial dependence can be introduced in the base distribution F_0, by keeping the weights fixed: this is usually called single-π DDP (Gelfand et al., 2005; Dunson and Herring, 2006). It can also be introduced in the DP weights by keeping the atoms fixed: this is usually called single-atom DDP (Reich and Fuentes, 2007; Dunson and Park, 2008). Finally it can be introduced in both the atoms and the weights (Duan et al., 2007).

Introducing the possibility of analysing data evolving over time and space incurs in additional modelling and computational challenges. For this reason, many proposed approaches assume that the DDP only depends on time or space. This work presents a novel Bayesian nonparametric method applied towards problems of prediction in spatio-temporal settings. The method relies on the stick-breaking representation of DPs, with a kernel-based method along the line of Dunson and Park (2008). The model takes advantage of the natural clustering property of the DP to construct clusters evolving with time and space for spatio-temporal data and is able to describe and use for prediction the heterogeneity of the observed data, without the need to introduce a large number of covariates. Observations are allocated so that data points which are observed at nearby locations and at consecutive times are assumed to be more similar. In practice, this means that the model is able to adapt to rapid changes over time and space, to learn more from points close in space and time to improve the prediction accuracy. The case where the DP atoms are not varying with the location and the time allows to define computationally simple and efficient algorithms; extensions to consider both varying weights and varying atoms are conceptually straightforward, but come at an increased computational cost. Both these versions will be considered in this work.

The remaining of the paper is organized as follows: Section 2 reviews the DP and the stick-breaking process, focusing on the introduction of spatial and temporal dependence; the spatio-temporal proposal, which is the main contribution of this work, is introduced in Section 3. The performance of the model is tested on simulation studies in Section 4 and on a real-data example in Section 5. Section 6 concludes the paper.

2   Spatial and temporal Dirichlet processes

Dirichlet processes have been widely applied in Bayesian nonparametric statistics due to their property of conjugacy, their flexibility, and ease of implementation. Taking F_0 to be a distribution on Θ, and some real positive number α, consider a finite number of sets $A_1, \ldots, A_r \in \Theta$; then F is considered to have a DP prior with base distribution F_0 and concentration parameter α, i.e. $F \sim DP(\alpha, F_0)$, if

$$(F(A_1), \ldots, F(A_r)) \sim Dir(\alpha F_0(A_1), \ldots, \alpha F_0(A_r)).$$

As $\alpha \to \infty$, $F(A) \to F_0(A)$, for $A \subset \Theta$. The properties of the DP are clear from its stick-breaking construction (Sethuraman, 1994), for which, given a concentration parameter $\alpha > 0$ and a non-atomic base distribution F_0 on $(\Theta, \mathcal{B}(\Theta))$, where Θ is the parameter space and $\mathcal{B}(\Theta)$ the corresponding σ-algebra, we have that $F \sim DP(\alpha, F_0)$ if

$$F = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k},$$

$$\pi_1 = V_1 \qquad \text{and} \qquad \pi_k = V_k \prod_{j=1}^{k-1}(1 - V_j) \quad \text{for } k = 2, \ldots$$

$$V_k \sim Beta(1, \alpha) \qquad \perp \qquad \theta_k \sim F_0 \qquad k = 1, 2, \ldots,$$

where δ_z is a probability measure concentrated at z (a Dirac mass). This construction is not limited to the Dirichlet process, but generalisation of the beta distribution, such as $V_k \sim Beta(a_k, b_k)$, are also available (Ishwaran and James, 2001).

In problems of density estimation, in many situations data may be better described by a continuous distribution, therefore the mixing proposed by Antoniak (1974) is usually applied, to produce an infinite mixture of (often Gaussian) kernels $P(\cdot)$, so that $y_i \sim \sum_{k=1}^{\infty} \pi_k p(y_i | \theta_k)$ for $i = 1, \ldots, n$, i.e. the model for y_i is assumed to be an infinite mixture model, where the weights of the mixture are defined through the stick-breaking construction.

Consider a random field $\{y(s) : s \in \mathcal{D}\}$ for $\mathcal{D} \subseteq \mathbb{R}^2$, and let $s = (s_1, \ldots, s_n)$ be a set of distinct locations in $\mathcal{D}$. Let $y(s)$, the observation at location s, be characterized by the model in (1.1). The spatial effect $\theta(s)$ is assumed to follow an unknown distribution F_s, such that $F_{\mathcal{D}} = \{F_s : s \in \mathcal{D}\}$ and $F_{\mathcal{D}} \sim \mathcal{P}$ where $\mathcal{P}$ is a probability measure on $(\Gamma, \mathcal{C})$, Γ is the space of uncountable collections of probability measures on the Polish space $(\Theta, \mathcal{B})$ indexed by $s \in \mathcal{D}$ and $\mathcal{C}$ is the corresponding σ-algebra. Therefore F_s is a probability measure over a measurable Polish space $(\Theta, \mathcal{B})$. Then, the density of $y(s)$ is

$$g(y(s)|x(s)) = \int_{\Theta} p\left(y(s)|\theta(s) + x(s)^T \beta\right) dF_s,$$

where $p(\cdot)$ is some kernel, possibly Gaussian.

The spatial generalisation of the DP can be introduced in the definition of the base distribution, such that $F_{0\mathcal{D}}$ is indexed by s, and the atoms θ_k of the DP are replaced with

$\theta_{k\mathcal{D}} = \{\theta_k(s) : s \in \mathcal{D}\}$, for $k = 1, 2, \ldots$. This approach can be considered as a particular case of the approach based on regression in the base measure first proposed by Cifarelli and Regazzini (1978). For instance, the spatial Dirichlet process (SDP) introduced by Gelfand et al. (2005) uses a base measure $F_{0\mathcal{D}}$ which is a Gaussian process with a (possibly isotropic) covariance function. Then $F_{\mathcal{D}}$ is defined such that $F_s = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k(s)}$ for each $s \in \mathcal{D}$. The distribution of the observations $y(s) = (y(s_1), \ldots, y(s_n))$ is almost surely a mixture of Gaussian distributions. This is a model in the class of single-π DDPs (MacEachern, 2000), and maintains the same weights $(\pi_1, \pi_2, \ldots,)$ for all locations $s \in \mathcal{D}$. This means that it is difficult to identify the spatial contribution of each component of the mixture. Examples of this approach can be found in Kottas et al. (2012) and Park et al. (2022).

One of the drawbacks of this approach is that, since the observations on the surface defined by the spatial locations constitute one single observation from the random field, replications are needed in order to conduct inference. Such replications are usually taken to be temporal replications. A dynamic model can be introduced in the DP atoms $\theta_{k\mathcal{D}}$ (Kottas et al., 2008), however this represents an extension of the modelling of the DP parameters, and the clustering probabilities are still not let evolving with time.

When observations are indexed by time instead of space, Dunson (2006) propose a dynamic mixture of Dirichlet processes (DMDP), where a latent variable dynamically varies across groups; Zhu et al. (2005), Caron et al. (2007) and Caron et al. (2017) propose a time-varying DPM with dynamic evolution of the atoms; Rodriguez and Ter Horst (2008) propose a model where atoms are allowed to evolve over time. Lau and So (2008) consider a model for temporal observations, where each atom can be expressed as an infinite mixture of autoregressions of order p. Di Lucca et al. (2013) propose a single-π DDP where the atom processes are expressed as linear autoregression $\theta_k(t) = \beta_k + \alpha_k \theta_k(t-1)$. They also consider the case where the atom process is defined as a Ornstein-Uhlenbeck process. An interesting variation of a dynamic DDP construction is proposed by Ascolani et al. (2021) who define a family $\mathcal{F} = \{F_t, t \geq 0\}$ of dependent random probability measures F_t which share some, but not all atoms and the set of atoms which are shared in F_t is defined as a pure death process over time, in particular a Fleming-Viot process.

Models that incorporate dependent weights have a number of theoretical and practical advantages over models that only use dependent atoms (MacEachern, 2000). Therefore, a second characterization of the DDP for spatial problems relies on the definition of a spatial stick-breaking prior, where the stick-breaking weights are allowed to vary with the spatial locations. Taddy (2010) proposes an approach for multivariate correlated time series, allowing for marginal Dirichlet process priors at each time. Nieto-Barajas et al. (2012) propose an approach where random proabilities measure are marginally Dirichlet processes, and temporal dependence is introduced via multivariate beta distributions for the unnormalized weights of the stick-breaking representation. The spatial stick-breaking model was first proposed by Reich and Fuentes (2007) and, more generally, by Dunson and Park (2008). The spatial stick-breaking prior assumes $y(s)$ to be a set of observations at site s such that model (1.1) is still valid. A prior is assigned to the spatial random effects $\theta(s) \sim F_s$. Differently from the SDP of Gelfand et al. (2005),

this framework utilizes marginal models F_s, $F_{s'}$ rather than a joint model $F_{s,s'}$. In fact, the distribution for $\theta(s)$ is given by

$$F_s = \sum_{k=1}^{\infty} \pi_k(s)\delta_{\theta_k} \qquad s \in \mathcal{D} \qquad \text{where}$$

$$\pi_1(s) = V_1(s) \qquad \text{and} \qquad \pi_k(s) = V_k(s)\prod_{j=1}^{k-1}(1 - V_j(s)) \qquad \text{for } k = 2,\ldots$$

$$V_k(s) = w_k(s,\psi_k)V_k, \qquad V_k \sim Beta(a,b) \qquad \perp \qquad \theta_k \sim F_0 \qquad \text{for } k = 1,2,\ldots$$

The distributions F_s and $F_{s'}$ are related through the dependence among $\pi_k(s)$ and $\pi_k(s')$ which arises from dependence among $V_k(s)$ and $V_k(s')$, according to the bounded functions $w_k(s,\psi_k)$, which are bounded kernels $w : \mathcal{D} \times \mathbb{R}^2 \to [0,1]$, allowing the masses of the DP to vary with space. An example of kernel function is the squared exponential kernel

$$w_k(s,\psi_k) = \prod_{j=1}^{2}\exp\left[-\frac{(s_j - \psi_{jk})^2}{h_{jk}^2}\right]. \tag{2.1}$$

The kernel function is centred around knots $\psi_k = (\psi_{1k},\psi_{2k})$ for $k = 1,2,\ldots$, such that $\psi \sim Q$, where Q is a probability measure on the Polish space $(\Psi,\mathcal{A})$, where Ψ is a Lebesgue-measurable subset of $\mathbb{R}^2$ (not necessarily corresponding to $\mathcal{D}$) and $\mathcal{A}$ is the corresponding Borel σ-algebra. The kernel spread is controlled by a bandwidth parameter $h_k = (h_{1k},h_{2k})$ which can be fixed to a value or allowed to vary according to a prior distribution, e.g. an inverse gamma distribution. The kernel function is defined so that knots closer to s and having a smaller index k have, on average, higher probability weight. F_s and $F_{s'}$ allocate similar probabilities to atoms $\theta = (\theta_1,\theta_2,\ldots)$ if s and s' are close in space. An application of the spatial stick-breaking prior was discussed by Reich et al. (2011) and Reich and Bondell (2011), within the field of landscape genetics. Reich and Fuentes (2007) and Dunson and Park (2008) show that the covariance of $(y(s),y(s'))$ is stationary when integrating with respect to (V_k,ψ_k,h_k), however the conditional covariance can be non-stationary, so this model can be more robust than standard kriging. Moreover, the correlation function tends to one as $s \to s'$.

When the focus is on a distribution indexed on time, F_t, Mena and Ruggiero (2016) construct a single-atom dynamic DDP by setting up a Wrights-Fisher diffusion on the fractions $V_k(t)$ in the stick-breaking construction of the marginal DP prior for F_t. Gutiérrez et al. (2016) propose a time-dependent Bayesian nonparametric model, where the stick-breaking weights are defined on beta variables evolving over time. In discrete time, Dunson (2006) and Griffin and Steel (2011) propose an AR-process type model allowing for efficient inference, such that the marginal distribution of $V_k(t)$ is $Be(a(t),b(t))$ for all k and t. An alternative construction relies on the arrivals contribution of the m-DPP (Griffin and Steel, 2006).

There are few DDP models in the literature where both weights and atoms are allowed to vary according to a covariate. DeYoreo and Kottas (2018) introduce a DDP prior with temporal dependence in both the atoms and weights for ordinal regression.

Duan et al. (2007) describe a model where atoms $\theta_k(s)$ are i.i.d. from F_0 and the weights $\pi_{k(s_1),\dots,k(s_n)}$ determine the site-specific joint probabilities, however the contribution of time and space to the specific component through the mixing weight is integrated out. Rodriguez and Dunson (2011) propose an approach based on a probit representation of the weights

$$F_s = \sum_{k=1}^{K} \pi_k(s)\delta_{\theta_k(s)}$$

$$\pi_k(s) = \Phi(\alpha_k(s)) \prod_{j<k} (1 - \Phi(\alpha_j(s))),$$

where $\alpha_k(s)$ have Gaussian marginals and $\theta_k(s)$ are independent and identically distributed sample paths from a given stochastic process. Arbel et al. (2016) and Barcella et al. (2017) empirically show the good predictive performance of models using a probit representation of the weights. Nevertheless, spatial or spatio-temporal dependence among the parameters α_k, and indirectly the mixture weights, can be introduced by assuming a Gaussian process over $\mathcal{D}$, which means that the computational cost increases and these models may be difficult to fit in cases with large sample sizes, as the examples in Section 5.

To the best of our knowledge, the only work introducing a spatio-temporal varying kernel is proposed by Hossain et al. (2013), who, however, use the logistic normal kernel as in Fernández and Green (2002), where spatio-temporal dependence is introduced through spatio-temporal varying covariates in a logistic function. This approach has limitations, including the need for covariates (which may not be available), and the need for a truncation of the possible number of weights (and atoms). Moreover, this approach has been implemented only for areal data, while in this work we focus on point referenced data. In general, the approach of Hossain et al. (2013) has a more difficult interpretation of the interaction of the spatial and temporal components. The contribution of this work is to introduce spatio-temporal dependence in the clustering probabilities, in order to increase the prediction accuracy in highly heterogeneous phenomena.

3 Spatio-temporal stick-breaking process

In this work, we suppose the observation $y(s,t)$ is a response variable observed at time t, with $t \in \{1,\dots,T\} = \mathcal{T}$, where $T < \infty$ and $\mathcal{T}$ is s distinct set of points in time, and location $s \in \mathcal{D} \subset \mathbb{R}^2$, where $\mathcal{D}$ is the spatial region of interest. While it is easy to define the stick-breaking weights in the case of continuous time, the complete model would imply important modifications and we leave it for further research; see, for example, Mena et al. (2011) and Arndt (2019).

Here we propose a class of stick-breaking processes that can be seen as a prior for $F_{\mathcal{D},\mathcal{T}}$ and induces a spatio-temporal dependence on the DP weights. Similarly to model (1.1), $y(s,t)$ can be modelled according to

$$y(s,t) = \theta(s,t) + x(s,t)^T\beta + \varepsilon(s,t),$$

where $\theta(s,t)$ is a spatial effect evolving over time, $x(s,t)$ is a column-vector of p covariates which may be dependent on the location s and the time t and β is a column-vector of p regression coefficients; finally, $\varepsilon(s,t)$ are i.i.d. normal errors. For ease of notation, in what follows, we assume that no covariate is available, simplifying the model to:

$$y(s,t) = \theta(s,t) + \varepsilon(s,t). \tag{3.1}$$

However, it is straightforward to extend the model to include covariates. Model (3.1) aligns with the data application described in Section 5.

The spatio-temporal effect $\theta(s,t)$ is assumed to follow an unknown distribution $F_{s,t}$, such that $F_{\mathcal{D},\mathcal{T}} = \{F_{s,t} : s \in \mathcal{D}, t \in \mathcal{T}\}$. $F_{\mathcal{D},\mathcal{T}} \sim \mathcal{P}$ where $\mathcal{P}$ is a probability measure on $(\Gamma, \mathcal{C})$, Γ is the space of uncountable collections of probability measures on the Polish space $(\Theta, \mathcal{B})$ indexed by $s \in \mathcal{D}$ and $t \in \mathcal{T}$, and $\mathcal{C}$ is the corresponding σ-algebra. Therefore $F_{s,t}$ is a probability measure over a measurable Polish space $(\Theta, \mathcal{B})$. In particular

$$g(y(s,t)|F_{s,t}) = \int p\left(y(s,t)|\theta(s,t)\right) dF_{s,t},$$

where $p(\cdot)$ is a (possibly Gaussian) kernel.

The distribution of the observation $y(s,t)$ can be represented as an infinite mixture. The mixing weights are allowed to vary over time and space. They are built similarly to the spatial stick-breaking:

$$F_{s,t} = \sum_{k=1}^{\infty} \pi_k(s,t)\delta_{\theta_k} \qquad s \in \mathcal{D}, \, t \in \mathcal{T} \qquad \text{where}$$

$$\pi_1(s,t) = V_1(s,t) \qquad \text{and} \qquad \pi_k(s,t) = V_k(s,t)\prod_{j=1}^{k-1}(1 - V_j(s,t)) \qquad \text{for } k = 2, \ldots$$

$$V_k(s,t) = w_k(s,\psi_k,t,\zeta_k)V_k, \qquad V_k \sim Beta(a,b) \quad \perp \quad \theta_k \sim F_0 \qquad \text{for } k = 1, 2, \ldots$$

Then, we consider a countable sequence of mutually independent random variables $(\psi_1, \psi_2, \ldots)$ such that $\psi_k \sim Q$ for each $k = (1, 2, \ldots)$, where Q is a probability measure on a Polish space $(\Psi, \mathcal{A})$, where $\mathcal{A}$ is a Borel σ-algebra and Ψ is a Lebesgue measurable subset of $\mathbb{R}^2$ which may or may not correspond to $\mathcal{D}$; moreover, we consider a countable sequence of mutually independent random variables $(\zeta_1, \zeta_2, \ldots)$ such that $\zeta_k \sim L$ for each $k = (1, 2, \ldots)$ and where L is a probability measure on a Polish space $(\Xi, \mathcal{E})$, where Ξ is a Lebesgue measurable subset of $\mathbb{R}$ and $\mathcal{E}$ is the corresponding Borel σ-algebra.

Similarly to the spatial stick-breaking, the spatio-temporal stick-breaking is proper if $\mathbb{E}[V_k]$ and $\mathbb{E}[w_k(s,\psi_k,t,\zeta_k)]$ are both positive. These are not restrictive requirements: V_k is usually assumed beta distributed and the prior distribution for a, b, ψ_k and ζ_k can be chosen to assure this feature. We propose $V_k \sim Be(a,b)$, where a and b are fixed or can be assumed to follow prior distributions (for example, uniform distributions in a reasonable compact space), the spatial knots $\psi_k = (\psi_{k,1}, \psi_{k,2})$ can be fixed, or given uniform prior distributions over the spatial domain, and, similarly, the temporal knots ζ_k can be fixed or given uniform prior distributions over the temporal domain.

$F_{s,t}$ and $F_{s',t'}$ allocate similar probabilities to elements of the set θ_k if $s \to s'$ and $t \to t'$, and in this way the process induces dependence on the response variable. $F_{s,t}$ is well defined with $\sum_{k=1}^{\infty} \pi_k(s,t) = 1$ almost surely for each $(s,t) \in \mathcal{S} \times \mathcal{T}$ (Ishwaran and James, 2003). In order to deal with the sparseness that may be present in the data, it is possible to choose the hyper-parameters (a,b) so that they favour values of V_k close to one or a value (or prior distribution) for the DP concentration parameter α so that the DP assigns high mass to few atoms.

When dealing with spatial and spatio-temporal data, it is important to study the spatio-temporal correlation or covariance functions assumed by the model. First, we derive the covariance function conditionally on the beta variables V_k, the knots ψ_k and ζ_k.

It is easy to see that $\mathbb{C}\text{ov}(y(s,t), y(s',t')|V, \psi, \zeta) = \mathbb{C}\text{ov}(\theta(s,t), \theta(s',t')|V, \psi, \zeta)$, given that $\varepsilon(s,t)$ and $\theta(s',t')$ are independent for any (t,s) and (t',s'), and $\varepsilon(s,t)$ and $\varepsilon(s',t')$ are independent for any (t,s) and (t',s'). Moreover, the covariance between $\theta(s,t)$ and $\theta(s',t')$ can be seen as dependent on the mixture weights, i.e. $\mathbb{C}\text{ov}(y(s,t), y(s',t')|V, \psi, \zeta) = \sigma_\theta^2 \Pr(\theta(s,t) = \theta(s',t')|V, \psi, \zeta)$. See the Supplementary Material (Grazian, 2024) for details of this derivation.

The covariance between $y(s,t)$ and $y(s',t')$, conditional on the mixture weights, but marginalised with respect to the atoms can be written as

$$
\begin{aligned}
\mathbb{C}\text{ov}(y(s,t), y(s',t')|V, \psi, \zeta) &= \sigma_\theta^2 \Pr(\theta(s,t) = \theta(s',t')|V, \psi, \zeta) = \sigma_\theta^2 \sum_{k=1}^{\infty} \pi_k(s,t)\pi_k(s',t') \\
&= \sigma_\theta^2 \sum_{k=1}^{\infty} \Bigg[w_k(s, \psi_k, t, \zeta_k) w_k(s', \psi_k, t', \zeta_k) V_k^2 \\
&\qquad \times \prod_{j<k} [1 - w_j(s, \psi_j, t, \zeta_j)V_j][1 - w_j(s', \psi_j, t', \zeta_j)V_j] \Bigg] \\
&= \sigma_\theta^2 \sum_{k=1}^{\infty} \Bigg[w_k(s, \psi_k, t, \zeta_k) w_k(s', \psi_k, t', \zeta_k) V_k^2 \\
&\qquad \times \prod_{j<k} [1 - (w_j(s, \psi_j, t, \zeta_j) + w_j(s', \psi_j, t', \zeta_j))V_j \\
&\qquad + w_j(s, \psi_j, t, \zeta_j) w_j(s', \psi_j, t', \zeta_j)V_j^2] \Bigg].
\end{aligned}
$$

The covariance function, as expected, depends on the choice of the kernel, and this means that it is maximised when choosing a kernel for which $w(s, \psi, t, \zeta) \to w(s', \psi, t', \zeta)$ when $s \to s'$ and $t \to t'$. It is also interesting to notice that, the conditional covariance function does not depend on the specific value of the response variable. This covariance can be a function of only the distances between spatial locations and between time points or can depend on the specific spatial locations and time points, therefore the fact that the process is stationary depends on the choice of kernel.

To better understand the properties of the covariance function between $y(s,t)$ and $y(s',t')$, it is useful to marginalise over the random beta variables V_k and the knots to obtain the unconditional covariance. It is reasonable to use independent priors: for example, uniform independent priors for ψ_k over the spatial domain and uniform independent priors for ζ_k over the temporal window of the observations. Then, the unconditional covariance function is

$$\mathbb{C}\mathrm{ov}(y(s,t), y(s',t')) = \mathbb{E}_{[V,\psi,\zeta]}\{\mathbb{C}\mathrm{ov}(y(s,t), y(s',t')|V,\psi,\zeta) \\ + \mathbb{C}\mathrm{ov}(\mathbb{E}(y(s,t)|V,\psi,\zeta), \mathbb{E}(y(s',t')|V,\psi,\zeta)\}$$

which is $\mathbb{E}_{[V,\psi,\zeta]}\{\mathbb{C}\mathrm{ov}(y(s,t), y(s',t')|V,\psi,\zeta)\}$ given that $\mathbb{C}\mathrm{ov}(\mathbb{E}(y(s,t)|V,\psi,\zeta), \mathbb{E}(y(s',t')|V,\psi,\zeta) = \mathbb{C}\mathrm{ov}(\mathbb{E}(\theta(s,t)|V,\psi,\zeta), \mathbb{E}(\theta(s',t')|V,\psi,\zeta)\} = 0$, under the assumption of independence of the stick-breaking atoms. Then,

$$\mathbb{C}\mathrm{ov}(y(s,t), y(s',t')) = \sigma_\theta^2 \mathbb{E}_{[V,\psi,\zeta]} \left\{ \sum_{k=1}^\infty \left[w_k(s,\psi_k,t,\zeta_k) w_k(s',\psi_k,t',\zeta_k) V_k^2 \cdot \right. \right.$$

$$\left. \left. \prod_{j<k} [1 - (w_j(s,\psi_j,t,\zeta_j) + w_j(s',\psi_j,t',\zeta_j))V_j + w_j(s,\psi_j,t,\zeta_j) w_j(s',\psi_j,t',\zeta_j)V_j^2] \right] \right\}$$

$$= \sigma_\theta^2 \int_\Psi \int_\Xi w(s,\psi,t,\zeta) w(s',\psi,t',\zeta) \pi(\psi,\zeta) \mathbb{E}[V_k^2]$$

$$\cdot \sum_{k=1}^\infty \left[1 - \int_\Psi \int_\Xi w(s,\psi,t,\zeta) \pi(\psi,\zeta) \mathbb{E}[V_k] + \right.$$

$$\left. - \int_\Psi \int_\Xi w_t(s',\psi,t',\zeta) \pi(\psi,\zeta) \mathbb{E}[V_k] + \int_\Psi \int_\Xi w(s,\psi,t,\zeta) w(s',\psi,t',\zeta) \pi(\psi,\zeta) \mathbb{E}[V_k^2] \right]^{k-1}.$$

If we use $V_k \sim Be(a,b)$ for all k, then $\mathbb{E}[V_k]$ and $\mathbb{E}[V_k^2]$ have a fixed form, depending on a and b, and

$$\mathbb{C}\mathrm{ov}(y(s,t), y(s',t')) = \sigma_\theta^2 \frac{g(s,s',t,t')}{2\left(1 + \frac{b}{a+1} - g(s,s',t,t')\right)}. \tag{3.2}$$

Derivation of these formulas are provided in the Supplementary Material (Grazian, 2024). The unconditional covariance does not depend on the value of $y(s,t)$ or $y(s',t')$ and it is equal to one only when $w(s,\psi,t,\zeta) \to w(s',\psi,t',\zeta)$ when $s \to s'$ for every $\psi \in \Psi$ and $t \to t'$ for every $\zeta \in \Xi$. Moreover, the covariance does not depend on F_0. The covariance structure depends on the expectations of the kernels, i.e. the expectation of $w_k(s,\psi_k,t,\zeta_k)$, the expectation of $w_k(s',\psi_k,t',\zeta_k)$, and the expectation of $w_k(s,\psi_k,t,\zeta_k) w_k(s',\psi_k,t',\zeta_k)$, therefore it depends on the form of the chosen kernel. It is important to notice that, similarly to all single-atom DDPs, the covariance function presents a lower bound. However, this does not influence the fact that the process has full support with respect to the weak topology, and the corresponding DP mixture model has large Hellinger support (Barrientos et al., 2012).

The kernel $w(s,\psi,t,\zeta)$ is a bounded kernel, $w : \mathcal{D} \times \mathbb{R}^2 \times \mathcal{T} \times \mathbb{R} \to [0,1]$, allowing masses to vary over time and space. The easiest way to define a spatio-temporal kernel
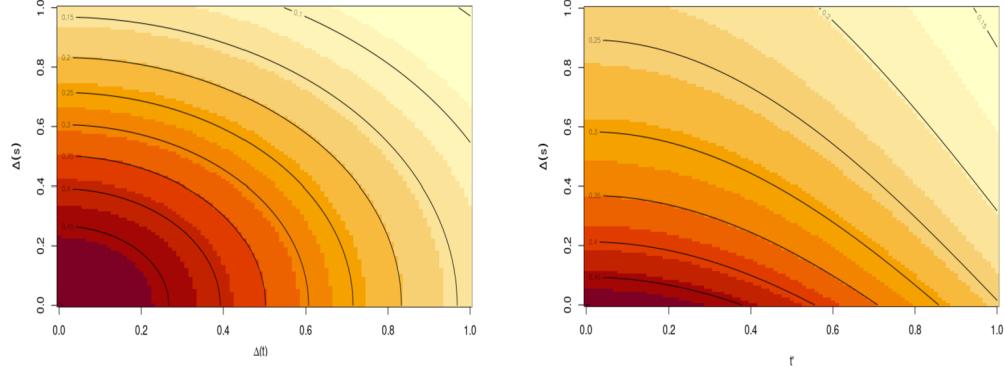
Figure 1: Covariance function $\mathbb{C}\mathrm{ov}(y(s,t), y(s't'))$ of the spatio-temporal process (marginalised over V, ψ, ζ) when separable exponential kernels are used for space and time (left), and when the non-separable kernel (3.3) is used (right). In the second case, the covariance structure does not only depend on the distance between time points, so in this plot we fix $t = 0$, and compute the covariance by varying t'.

is by considering a separable kernel

$$w_k(s, \psi_k, t, \zeta_k) = w_k(s, \psi_k) \cdot w_k(t, \zeta_k) \qquad k = 1, 2, \ldots$$

The hypothesis of separability in this setting implies that the temporal autocorrelation of the mixing weights for two locations at distance d is proportional to the temporal autocorrelation of the mixing weights for two locations at distance rd for some constant r. While the kernel is separable, the covariance structure of the process is not separable. Indeed, in this setting

$$g(s, s', t, t') = \frac{\int_\Psi w(s, \psi) w(s', \psi) p(\psi) d\psi}{\int_\Psi \int w(s, \psi) p(\psi) d\psi} \cdot \frac{\int_Z w(t, \zeta) w(t', \zeta) p(\zeta) d\zeta}{\int_Z w(t, \zeta) p(\zeta) d\zeta}.$$

Even if the kernel is separable, the covariance function is not separable, in the sense that it is not a product of functions in time and space, however it is a combination of separable factors. While the model is already flexible, leading to a non-separable covariance function, it implies a symmetry between time and space, as shown in Figure 1. For example, consider using an exponential kernel for both time and space as in (2.1); Figure 1 shows that there is a symmetry between time and space, such that the covariance between the observations decreases symmetrically as the distance between time and spaces increases, and this may represent a strong assumption depending on the phenomenon, in particular when this interaction is not known. Importantly, the conditional covariance function is non-stationary, but the unconditional covariance function only depends on the distance between locations and time points, i.e. it is stationary, when choosing separable exponential kernels.

The impact of the hypothesis of separability in the definition of the mixing weights in complicated, given that it relates to non-observable variables. There are several non-separable kernels proposed in the literature: see, for example, Cressie and Huang (1999),

Gneiting (2002) and Genton (2007). We propose to use the class of kernels proposed by
Gneiting (2002):

$$w(s, \psi, t, \zeta) = \frac{1}{\xi(|t - \zeta|^2)^{d/2}} \phi \left(\frac{(s_1 - \psi_1)^2 + (s_2 - \psi_2)^2}{\xi(|t - \zeta|^2)} \right),$$

where $\phi(\cdot)$ is a completely monotone function, and $\xi(\cdot)$ is a positive function with a
completely monotone derivative. For example,

$$w(s, \psi, t, \zeta) = \frac{1}{\gamma|t - \zeta| + 1} \exp \left(-\frac{(s_1 - \psi_1)^2 + (s_2 - \psi_2)^2}{(\gamma|t - \zeta| + 1)^{\lambda/2}} \right), \qquad (3.3)$$

where γ is a non-negative scaling parameter and $\lambda \in [0, 1]$ is a parameter controlling
the space-time interaction: if $\lambda = 0$, the kernel is separable; as $\lambda \to 1$, the space-time
interaction in the definition of the stick-breaking probabilities increases and the spatial
dependence at positive lags decreases more slowly. The parameter γ can be fixed or
given an inverse-gamma or a uniform prior distribution.

The definition of λ allows an automatic testing about it being equal to zero (i.e.
separability of space and time in the definition of the mixing weights), by assuming a
spike-and-slab prior (Mitchell and Beauchamp, 1988; Ishwaran and Rao, 2005; Andersen
et al., 2017), so that

$$\lambda \sim \omega_\lambda f(\lambda) + (1 - \omega_\lambda)\delta_0, \qquad (3.4)$$

i.e. the prior distribution results in a mixture between a continuous distribution $f(\lambda)$
(for example, uniform or beta) and a Dirac mass at zero δ_0, with some mixing weights
ω_λ that can be fixed or allowed to follow a prior distribution. In this way, the posterior
distribution for λ will be again a mixture, and it is possible to test the hypothesis of
separability by looking at the weight of the mass component at zero. In more detail,
it is possible to examine the posterior probability $\Pr(\lambda = 0|\mathbf{y})$ and compare it to an
experimenter's chosen threshold, such as 0.5, in order to make a decision regarding
whether time and space are separable in the definition of the mixture weights.

The structure of the covariance function among observations is similar to the one
obtained with separable kernels in Equation (3.2), however the function $g(\cdot, \cdot)$ is not
defined on separable factors anymore. More specifically,

$$g(s, s', t, t') = \frac{\int_\Psi \int_\Xi w_k(s, \psi_k, t, \zeta_k) w_k(s', \psi_k, t', \zeta_k) p(\psi, \zeta) d\psi_k d\zeta_k}{\int_\Psi \int_\Xi w_k(s, \psi_k, t, \zeta_k) p(\psi, \zeta) d\psi_k d\zeta_k},$$

when considering γ and λ fixed. In the case of kernel (3.3), while the integral with
respect to ψ_1 and ψ_2 results in a function that depends on the distance among space
only, the integral with respect to ζ is a function that does not depend on the distance
of time points only, but depends on the specific points t and t', see also Figure 1. When
λ is estimated to be zero, then

$$\begin{aligned} g(s, s', t, t') = \exp \left(-((s_1 - s_1')^2 + (s_2 - s_2')^2) \right) \\ \cdot \left(-\frac{\log(t') - \log(t' - 1) - \log(t) - \log(t - 1)}{\gamma^2(t - t')} \right), \end{aligned}$$
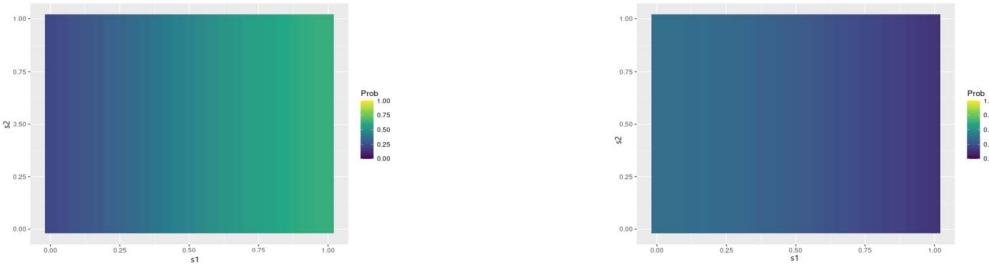
Figure 2: Map of π_1 (left) and π_2 (right) for each location and $t = 1$.

then the covariance is again non-separable among time and space, but made by separable factors, and it depends on the distance among spatial locations, and distances in standard scale and in log scale for the time. Therefore, depending on the choice of the kernel, the covariance can be stationary or non-stationary, and it is always non-separable with respect to space and time. In the particular choice of kernel (3.3), the covariance function obtained by marginalising over the parameters of the kernel is stationary with respect to space, but not with respect to time. Therefore, the covariance obtained for the spatio-temporal stick-breaking can be non-stationary, differently from the spatial stick-breaking.

An example of covariance structure of the process is illustrated in Figure 1, which shows that the covariance can decrease in a non-symmetric way with respect to the direction of time and space, for instance, by decreasing more slowly along the time direction. Figure 1 is defined for a fixed value of γ and λ, however, in real situations, the parameters are estimated, therefore the type of interaction between time and space is left to be defined by the data. Supplementary Material (Grazian, 2024) shows additional examples of the covariance function of the process by varying the parameters (γ, λ, ψ, ζ).

Observations which are close in space and time are more likely to be allocated to similar components, because the mixing weights for each component are more similar. Figure 2 shows a map of the weights $\pi_1(s, t = 1)$ and $\pi_2(s, t = 1)$ over the space domain, which represents how locations tend to be more likely allocated to component one or two.

Finally, Figure 3 shows the behavior of the number of allocated clusters of four processes: classic DP, spatial stick-breaking, spatio-temporal stick-breaking and probit stick-breaking. For all the processes, as expected, the number of populated clusters increases with the number of observations. The spatio-temporal stick-breaking on average is associated with a larger number of populated clusters than the other processes: in terms of prediction and density estimation, this property can be associated with a higher ability to take into account local heterogeneity, and can lead to higher prediction scores in more complex situations, as it will be shown in Section 5.

Under the single-atom setting, the dependence on space and time is expressed through the weights of the stick-breaking representation. One advantage of doing so
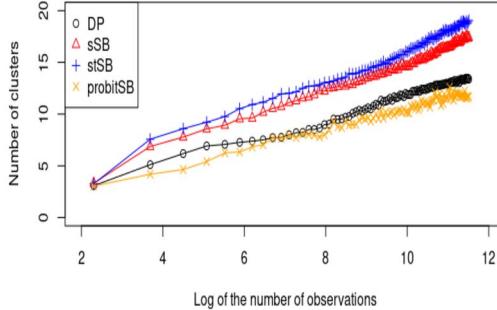
Figure 3: Clustering structure generated by the spatio-temporal stick-breaking. Comparison with the classic DP, the spatial stick-breaking and the probit stick-breaking process. The figure shows the expected number of clusters with respect to the log number of observations.

is that the implied partition model changes with the values of $s \in \mathcal{S}$ and $t \in \mathcal{T}$. This is an essential property, in particular when the allocation of the observations to the mixing component is also of interest. Another important feature is that problems related to inference of $\theta_k(s,t)$ are avoided: this has computational consequences, i.e. the approaches are less computationally intensive, and predictive consequences, i.e. the atoms $\theta_k(s,t)$ should include values which are not linked with any observed data in the case of inference for new values of (s,t). However, extensions are also possible.

3.1 Allowing the atoms to depend on space and time

The single-atom stick-breaking process (stSB) is capable of capturing features of distributions that vary over time and space, even whithout the need for replicates for each location in the index space. However, any nonparametric model that does not allow both weights and atoms to vary over the covariate space assumes a lower bound to the covariance between the distributions (Rodriguez and Dunson, 2011). This is not necessarily a drawback of the model, if the assumption of dependence is a reasonable one, but it may be an unreasonable assumption in the case we want to take into account the possibility of no dependence over time or space.

Atoms may also be allowed to vary by introducing a modification of the stick-breaking representation of each element

$$F_{s,t} = \sum_{k=1}^{\infty} \pi_k(s,t) \delta_{\theta_k(s,t)}, \tag{3.5}$$

where $\pi_k(s,t)$ are defined as in the previous section, and $\theta_k(s,t)$ are independent stochastic processes with index set $\mathcal{S} \times \mathcal{T}$ and $F_{s,t,0}$ marginal distributions. The processes associated to the weights and atoms are independent. We refer to the process with varying weights and atoms as stSB (VA).

The atoms $\theta_k(s,t)$ can be generated from a Gaussian process characterised by a covariance matrix considering space and time distances. In this work, we use a spatio-temporal covariance function defined as the product of one covariance function depending on space and one covariance function depending on time. More specifically, if $d(s,s')$ is the distance between two spatial locations, for instance the Euclidean distance, and $d(t,t_0)$ is the temporal lag with respect to the initial time point, then the covariance function is

$$C(d(s,s'),d(t,t_0)) = C_s(d(s,s')) \times C_t(d(t,t_0)),$$

where C_s is a spatial covariance function and C_t is a temporal covariance function. In this work, we use an exponential covariance function for $C_s(\cdot)$ and the temporal covariance corresponding to an $AR(1)$ process $C_t(\cdot) = \rho(t - t_0)$. Such covariance matrix is implemented in the `SpatialTools` package available in `R` (French, 2022).

Figure 4 shows the temporal evolution of the process over time and space; on the left, the single-atom process is represented, and on the right the process with both varying weights and atoms is represented. Figure 4 shows a case where locations are uniformly generated in a unit-square set; the Supplementary Material (Grazian, 2024) includes additional simulation, considering more structured locations and a different prior for the parameters V_k. The process seems to vary more smoothly over space and with a higher variation over time by introducing spatio-temporal dependence also among the atoms with respect to the process only introducing dependence among weights. From the simulations, it seems that the model with both weights and atoms dependent on (s,t) can better describe heterogeneity and situations where the process varies more quickly over time. However, the main drawback of this process is that inference for it is more computationally expensive. For this reasons, in some simulation settings and the real-data applications we fit the model on a subset of the datasets.

3.2 Computational details

Posterior inference can be obtained through Markov Chain Monte Carlo (MCMC). Usually, two different finite approximation strategies are adopted for stick-breaking processes. The first strategy is a marginal approach which uses a Pólya urn representation to marginalise over the Dirichlet process prior. The second strategy is a conditional approach (Ishwaran and Zarepour, 2000; Papaspiliopoulos and Roberts, 2008), which avoids marginalization. We use this second strategy with the retrospective sampling and the generalized Pòlya urn sampling scheme of Dunson and Park (2008). We present details of the MCMC scheme for the case of both atoms and weights varying over (s,t), while the single-atom case results as a simplification of this scheme.

It is possible to use a truncation scheme or to consider the possibility to have an infinite number of components. Earlier works have investigated the truncation level, see, for example, Muliere and Tardella (1998) and Ishwaran and James (2001). In our simulation studies and real-data applications, we have set the maximum number of clusters to a very large number, i.e. $M = 500$. Details of the algorithms are given in the Supplementary Material (Grazian, 2024).

$t = 1$ $t = 2$ $t = 1$ $t = 2$

$t = 3$ $t = 4$ $t = 3$ $t = 4$

$t = 5$ $t = 6$ $t = 5$ $t = 6$

$t = 7$ $t = 8$ $t = 7$ $t = 8$

$t = 9$ $t = 10$ $t = 9$ $t = 10$
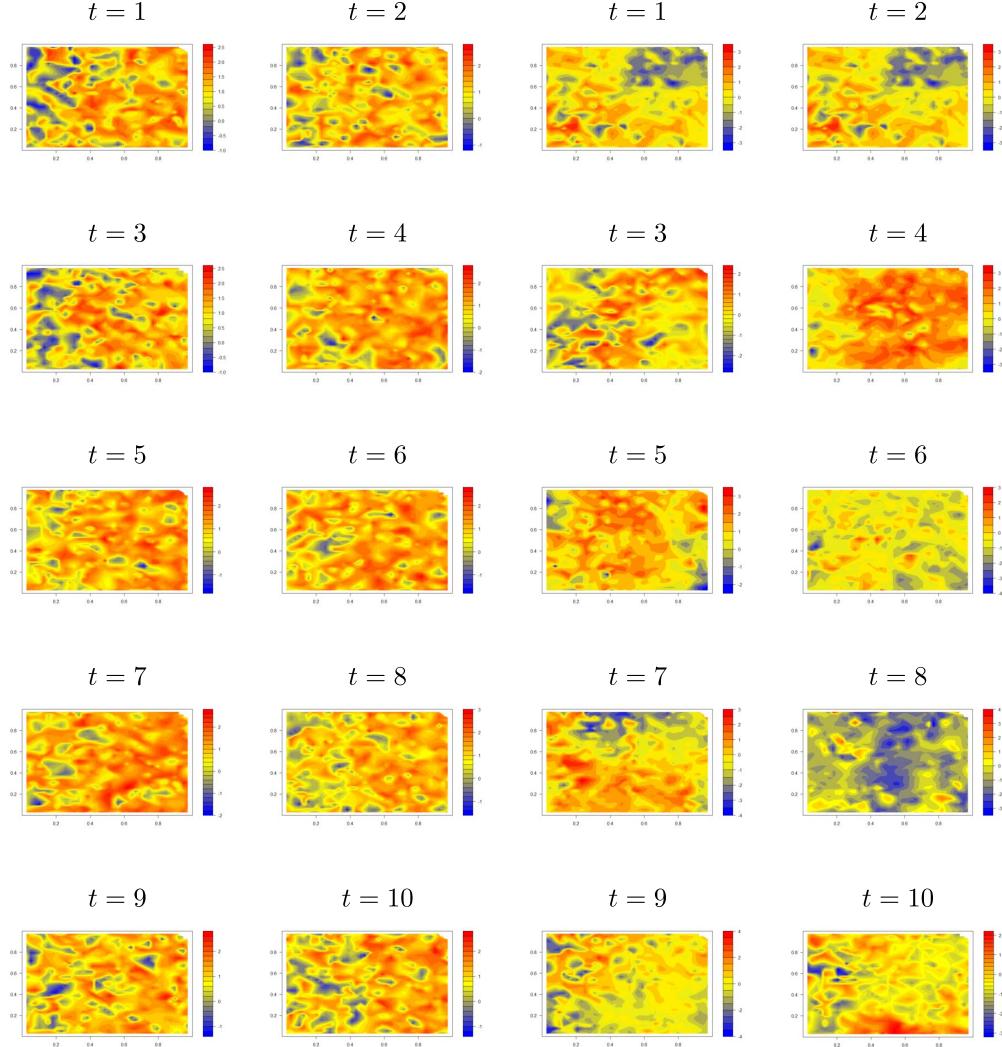
Figure 4: Temporal evolution of the observations for uniformly selected locations. The spatial locations are simulated uniformly over the unit-square set; after having fixed the locations, the process is simulated using the spatio-temporal stick breaking and the colour represents the value of $y(s, t)$. Process with only varying weight (stSB) on the left (columns 1 and 2) and process with both varying weights and atoms (stSB (VA)) on the right (columns 3 and 4). We assume $V_k \sim Be(a, b)$ and $a = b = 1$.

For full specification of the prior distributions for the model described in Section 3, please refer to the Supplementary Material (Grazian, 2024).

4   Simulation study

4.1   Simulation study with a complex temporal structure

We first illustrate the performance of the model following the simulation study set-up used by Gutiérrez et al. (2016), with the difference that we include spatial dependence.

We simulate $n = 7,200$ observations for each time point $t = 1, 2, \ldots, 24$ from

$$
y(s,t) = \begin{cases} \mathcal{N}(f(t), \sigma_1^2 C(s; \rho)) & 1 \leq t < 8 \\ 0.3\mathcal{N}(f(t), \sigma_1^2 C(s; \rho))) + 0.7\mathcal{N}(f(t), \sigma_2^2 C(s; \rho))) & 8 \leq t < 16 \\ 0.5\mathcal{N}(f(t), \sigma_3^2 C(s; \rho))) + 0.5\mathcal{N}(0.1t + f(t), \sigma_3^2 C(s; \rho))) & t \geq 16, \end{cases}
$$

where $\sigma_1^2 = 0.04$, $\sigma_2^2 = 1$, $\sigma_3^2 = 0.09$,

$$
f(t) = \cos(t) + 2 \times \sin(t) + \frac{t}{2} - \min(t, 16),
$$

and $C(s; \rho)$ is a covariance matrix defined by a squared exponential kernel with length-scale $\rho > 0$. $\mathcal{N}(\mu, \Sigma)$ stands for a multivariate normal distribution, where Σ is given an inverse-Wishart prior distribution with expected value $\mathbb{E}(\Sigma) = \frac{A}{\nu - 3}$ and $\nu = 8$. We assume no covariates are available.

For this setting and the relative sample size, it is not possible to apply the version of the spatio-temporal stick-breaking with varying atoms of Section 3.1, for reasons associated to both memory and computational time. Therefore it has been fitted by using only 50% of the observations as training set.

The results of the single-atom spatio-temporal stick-breaking (stSB), and its extension with varying atoms (stSB (VA)), are compared with two main alternatives available in the literature. The benchmark model is the Bayesian spatio-temporal model (sp):

$$
y(s,t) = \theta(s,t) + \varepsilon(s,t),
$$

where $\theta(s,t)$ is a spatio-temporal process with a separable covariance function $\Sigma = C_{st} = C(y(s,t), y(s',t')) = C_s(s,s') \cdot C_t(t,t')$. $C_s(s,s')$ and $C_t(t,t')$ represent the spatial and temporal covariance functions, respectively. Then $\theta(s,t) \sim \mathcal{GP}(0, \sigma_\theta^2 C_{st})$, where $\mathcal{GP}$ stands for a Gaussian process with a column vector of zero-elements as mean. Moreover, $\varepsilon(s,t) \sim \mathcal{N}(0, \sigma_\varepsilon^2 \mathbf{I})$, where $\mathbf{I}$ is the identity matrix. The model can be fitted by using the `spTimer` package (Bakar and Sahu, 2015) available for the R software (R Core Team, 2020). In particular, we choose to use exponential correlation functions $C_s(s,s') = \exp(-\rho d(s,s')^2)$ and $C_t(t,t') = \exp(-\xi(t-t')^2)$, where $d(s,s')$ is the Euclidean distance, and $\mathbb{V}\mathrm{ar}(\theta(s,t)) = \sigma_\theta^2$. Prior distributions are defined as $\rho, \xi \sim \mathcal{U}(3, 30)$, $\sigma_\varepsilon^2 \sim \mathcal{IG}(0.01, 0.01)$, and $\sigma_\theta^2 \sim \mathcal{IG}(0.01, 0.01)$, where $\mathcal{IG}$ stands for an inverse gamma distribution.

The second model is the one implemented by Reich and Fuentes (2007) (sSB), described in Section 2, with prior distributions $a \sim Unif(0,10)$, $b \sim Unif(0,10)$, $\sigma_\varepsilon^2 \sim \mathcal{IG}(0.01, 0.01)$ and $\sigma_\theta^2 \sim IG(0.01, 0.01)$, the knots are uniform over the spatial domain; an exponential kernel (2.1) is used for $w_k(s, \psi_k)$, with $h_{jk} \sim IG\left(1.5, \frac{\nu^2}{2}\right)$ and $\nu \sim Unif(0, \nu_{max})$ where ν_{max} is the maximum distance between any pair of points in the spatial grid. Temporal dependence is introduced in the atoms of the stick-breaking process, by adding a temporal structure through a transition equation for $\theta(s,t)$, say $\theta(s,t) = \zeta\theta(s, t-1) + \eta(s,t)$, where $|\zeta| < 1$ and $\eta_{\mathcal{D},\mathcal{T}} = \{\eta(s,t) : s \in \mathcal{D}, t \in \mathcal{T}\}$ are independent realizations from a spatial stochastic process; see Kottas et al. (2008).

We run the MCMC algorithm for 20,000 iterations, with the first 10,000 iterations discarded for burn-in. We set the maximum number of components to $M = 500$. Data simulation is repeated 50 times.

We use the expected mean squared prediction error (ESPE) to compare the prediction accuracy for each considered model; we keep 70% of the simulated data as training dataset, and use the estimated model to predict the remaining 30% of the observations. The ESPE is computed over repetitions of the simulations and over all the time points and spatial locations:

$$\text{ESPE} = \mathbb{E}\left[\sum_{t \in \mathcal{T}}\sum_{s \in \mathcal{S}}[y(s,t) - \tilde{y}(s,t)]^2\right],$$

which is approximated by the corresponding sample average. Here, $y(s,t)$ is the observed value, while $\tilde{y}(s,t)$ is the predicted mean value in the testing dataset.

The mean squared error is 175.69 for sp, 196.04 for sSB, and 52.33 for stSB; for stSB, 97.3% of accepted values of λ are equal to zero and the posterior mean of values different from zero is 0.037. The model stSB (VA) has ESPE equal to 142.15, with 100% of accepted values of λ equal to zero; however, it has to be noticed that this last model is fitted with a reduced training set with respect to the other models. By considering the prediction (not shown here), it is evident that the spatio-temporal model smooths too much over time, while the sSB is associated with the highest uncertainty in the prediction. The stSB is able to follow the data more closely.

4.2   Simulation study with a complex spatial structure

We now consider the case where data are simulated from different spatio-temporal random fields with consecutive 15 time points and a variable number of locations which are simulated in the following way. Locations are generated according to a Thomas point process, i.e. a cluster point process, so that locations tend to get clustered: first a homogeneous Poisson point process with intensity $\omega > 0$ is simulated on a spatial rectangle (parent process), then for each simulated point another Poisson point process with intensity $\delta > 0$ is simulated (daughter processes). On these simulated daughter points, a random field is generated. The daughter point processes are simulated within a cluster disk (of radius 0.1), which has the centre in each parent point. This way of

simulating spatial points has the aim of recreating a realistic situation where observations are locally closed; for instance, meteorological variables are often recorded by monitoring stations which are present in urban areas in a large number, while are less present or totally absent in rural areas.

Some of the models include temporal dependence and some do not; this decision is meant to show the ability of the proposed approach to adapt to less complicated situations. We considered the following models:

1. a stationary isotropic covariance model with Gaussian covariance $C(s, s') = \sigma_\theta^2 e^{-\frac{||s-s'||^2}{h}}$ with variance $\sigma_\theta^2 = 1$, scale parameter $h = 0.4$, and time-independent observations;

2. a stationary isotropic exponential model $C(s, s') = \sigma_\theta^2 e^{-\frac{|s-s'|}{h}}$ with variance $\sigma_\theta^2 = 4$ and scale $h = 10$, with a nugget component and a mean described by a trend of mean 0.5, and time-independent observations;

3. a stationary isotropic covariance model belonging to the stable family, with covariance function given by $C(s, s') = \sigma_\theta^2 e^{-\frac{|s-s'|^\alpha}{h}}$ with $\alpha = 1.9$ so that the Gaussian sample paths are not differentiable, scale parameter $h = 0.4$, and time-independent observations;

4. a nugget effect model with variance $\sigma_\theta^2 = 1$, zonal anisotropy, and time-independent observations;

5. a Stein non-separable space-time model (Stein, 2005), where the covariance function is given by

$$C(h, t) = W_\nu(z) - \frac{(<h, c> t)}{((\nu - 1)(2\nu + d)} \cdot W_{\nu-1}(z)),$$

where $h = ||s - s'||$, $W_\nu(\cdot)$ is a Whittle-Matern covariance model with smoothness parameter ν, $z = ||(h, t)||$ is the norm of the vector (h, t), and d is the dimension of the space on which the random field is defined; here $\nu = 1.5$; t is intended as the difference between the observed time and the starting point and $< \cdot >$ is the inner product with respect to a reference vector of coordinates $c = (0.9, 0.1)$;

6. a stationary space-isotropic covariance model with covariance function given by

$$C(s, s', t, t') = (\psi(t, t') + 1)^{d/2} \phi \left(\frac{|s - s'|}{\sqrt{\psi(t, t') + 1}} \right),$$

where d depends on the spatial dimension of the field and here is set equal to 2; $\phi(\cdot)$ is a Gaussian covariance model; $\psi(t, t') = |t - t'|^\alpha$ with $\alpha = 1$.

Simulations are obtained via the R package `RandomFields` (Schlather et al., 2015). For every model, we repeat simulations 50 times. We compare two exponential models

(Model 2) with nugget and trend, using $\omega = 10$ (Model 2a) and $\omega = 30$ (Model 2b) respectively in the Thomas point process.

Results are shown in Table 1. The stSB outperforms, in terms of ESPE, sSB for simulations from a Gaussian model (Model 1), an anysotropy model (Model 4), a Stein model (Model 5), and a non-separable space-time model (Model 6). Results are very similar for the sp, sSB and stSB for simulations from exponential models (Models 2a and 2b). Overall, stSB seems to be a good prior process in more complex situations, where space and time are not separable or where the covariance function is complicated, as for the Stein model. The approach still performs well in comparison with a simpler kriging model in more standard situations. In this sense, stSB seems to represent a robust prior process; this may be due to the ability to estimate the parameter a of the beta prior for V_k from the data, so that if a increases the process is smoother (see Supplementary Material (Grazian, 2024)), while if a is close to one the process is more heterogeneous. stSB (VA) tends to perform slightly worse than the other processes with simulations from the exponential models (Model 2a and 2b), the stable model, the Stein model and the non-separable model. This may be due to the additional complexity of the model, which requires an additional number of MCMC iterations to provide good estimates and an additional number of observations to reach the same level of accuracy.

Table 2 presents results on the estimation of the separability parameter λ in Equation (3.3). When data comes from a separable model and observations are time-indepen-

	sp	sSB	stSB	stSB (VA)
1. Gaussian	0.93611	0.96797	0.92212	**0.89518**
2a. Trend, $\omega = 10$	**1.77601**	1.78522	1.86372	2.86420
2b. Trend, $\omega = 30$	**1.66367**	1.68756	1.71742	1.79734
3. Stable	**0.11978**	0.31019	0.90373	0.98556
4. Zonal Anisotropy	1.00187	0.99860	**0.90475**	0.98629
5. Stein	0.63084	0.68333	**0.57722**	0.77151
6. Nonseparable	0.60420	0.64485	**0.53416**	0.93089

Table 1: Comparison of the estimated ESPEs for the four considered models and for simulations from models described in the list with the relative number. The bold number represents the preferred models.

	stSB	stSB (VA)	stSB	stSB (VA)
	$\Pr(\lambda = 0\vert\mathbf{y})$	$\Pr(\lambda = 0\vert\mathbf{y})$	$\mathbb{E}[\lambda\vert\mathbf{y}]$	$\mathbb{E}[\lambda\vert\mathbf{y}]$
1. Gaussian	0.97476	0.00000	0.04098	0.86898
2a. Trend, $\omega = 10$	0.70008	0.15749	0.34658	0.94788
2a. Trend, $\omega = 30$	0.69074	0.50000	0.33236	0.99458
3. Stable	0.87393	0.10000	0.13651	0.91893
4. Zonal Anisotropy	0.02957	0.08295	0.94224	0.89265
5. Stein	0.98938	0.10000	0.01257	0.86846
6. Nonseparable	0.98252	0.00000	0.03513	0.97822

Table 2: Comparison of posterior measures on the separability parameter λ in the spatio-temporal stick-breaking models - posterior probability that $\lambda = 0$ and posterior expectation of λ when different from zero.
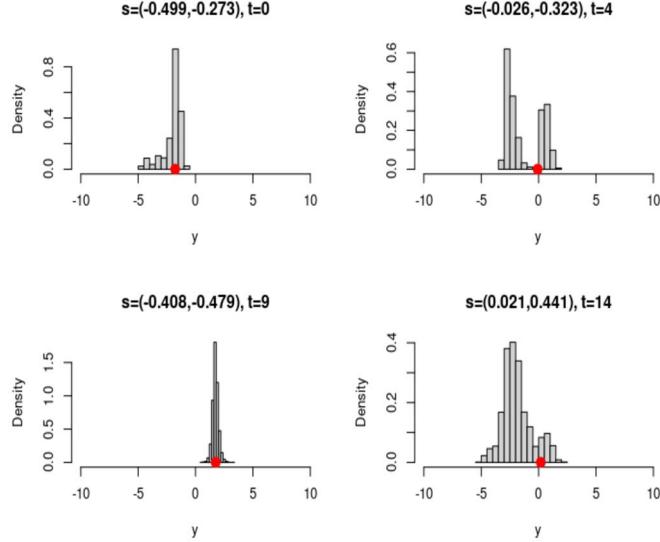
Figure 5: Mixture densities obtained by choosing the modal means μ, weights π and standard deviation σ parameters for four observations, characterized by different spatial locations and time points and for data generated from a Stein model. The red dots on the x-axis represent the observed value.

dent, there is a strong support that weights obtained using a separable kernel are preferred. It is, however, interesting to notice, that the support for a non-separable kernel is never strong, except for the anisotropic case: data are evidently not strongly informative about the parameter defining separability within the kernel. This is not surprising, since the covariance among observations obtained by using either a separable kernel or a non-separable kernel is non-separable in both cases. However, it represents an additional level of flexibility for the model, allowing for the covariance function to be non-stationary. The Supplementary Material (Grazian, 2024) provides results with an alternative prior distribution for λ.

From the simulation study, it seems that stSB is able to adapt to spatial and temporal changes, by modifying the mixing weights according to the information carried by the data. Figure 5 shows the mixture densities obtained via posterior estimate of the parameters relative to four locations and time points from data generated from a Stein model: the process seems to be adaptable to different situations and produce very variable posterior densities.

5   Real-data application

We now implement the methodology proposed in this work on a real-data application, the Supplementary Material (Grazian, 2024) includes an alternative example on temperature data.

Rainfalls in Australia are known to be highly variable (Nicholls et al., 1997), with low rainfall over most of the country and intense seasonal falls in the tropical regions in the North. The rainfall pattern is concentric from the coast, in particular the North and East coasts can be characterized by large rainfall volumes, while the centre is more arid. Around the centre there is another humid area, particularly in the East (Dey et al., 2019). The temporal and spatial variability of Australia's rainfall can lead to droughts and floods that have large impact on various sectors, e.g. agriculture, hydrology, resource management and allocation, human health and housing.

The dataset contains 10 years of daily weather observations in 49 locations across all Australia, from 31 October 2007 to 24 June 2017. The spatial area $\mathcal{D}$ corresponds to the entire Australia, with spatial locations identified by the 49 meteorological stations included in the dataset. Observations were drawn from weather stations, and include the amount of rainfall recorded for the day in mm. The dataset is publicly available on the website of the Bureau of Meteorology of the Australian Government http://www.bom.gov.au/climate/data/. The code and the dataset used for this example are available at https://github.com/bayesgra/spatiotempoSB. For illustration, Figure 6 shows the monthly rainfall volumes for each location from 2009 to 2017, however daily
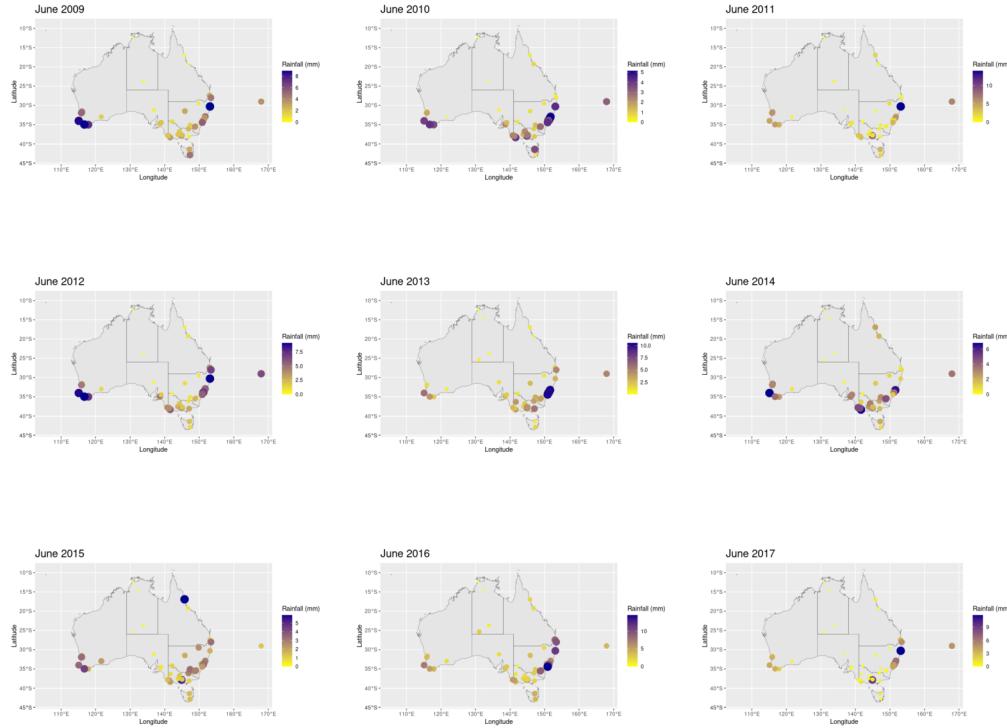
Figure 6: Monthly rainfalls for the month of June from 2009 to 2017 in Australia. A larger and darker dot represents a larger volume of rainfall.

data were used as training dataset. Years 2007 and 2008 were left out of the illustration because data were highly missing, however they were included in the training dataset. The complete dataset comprises 140,793 data points. The training dataset, consisting of data from 2007 to 2016, includes 132,420 observations, while the testing dataset, which pertains to the year 2017, contains 8,373 data points. Predictions are made at the same 49 spatial locations as those available in the training dataset, but at different time points. The model is capable of handling situations where there is no rainfall, so zeros were not modified.

Four models have been compared: sp, sSB, stSB and stSB (VA). The first three models were trained on the daily observations from 2007 to 2016 to predict data for 2017; stSB (VA) could not be fit on the full dataset, for computational reasons. Therefore, it was fitted on weakly data, instead of daily data, obtained by avaraging the daily rainfall data for each week at each location.

Table 3 presents a comparison of prediction accuracy and model preference among sp, sSB, and stSB. For this application, we also assessed the models using the Deviance Information Criteria as defined by DIC_1 in Celeux et al. (2006) and the Watanabe–Akaike information criterion as defined by $WAIC_2$ in Gelman et al. (2014). Once again, stSB appears to produce more reliable predictions compared to the other models in terms of ESPE, DIC, and WAIC, providing strong support for it in both prediction accuracy and model preference.

	ESPE	**DIC**	**WAIC**
sp	109.74	39,960,564	3,109,830
sSB	115.47	495,345	524,083
stSB	71.05	233,070	257,581

Table 3: Model comparison for the three models.

The ESPE for stSB(VA) on the same testing dataset is 104.67 for stSB (VA). However, conducting a fair model comparison is not possible because the model was trained on a relatively different dataset.

Figure 7 displays the posterior means of the weights of the fitted mixture model, illustrating how the component with the highest posterior probability of allocation varies over time and space. We have chosen to focus on Sydney and Perth due to their geographical differences and the significant distance between the two cities. Additionally, we have selected two different seasons to showcase: summer for Sydney and winter for Perth.

The majority of the components have weights close to zero. Specifically, components associated with large indices are never selected. For any location and time point, no component beyond the 100th exhibits a posterior mean significantly different from zero.

Figure 8 shows the average squared residuals for predictions relative to June 2017 (values based on the posterior predictive mean) for each spatial location. The first two models, sp and sSB, tend to oversmooth over the area, with high volatility, in particular in the area relative to Perth (West) and the area on the coast of New South Wales
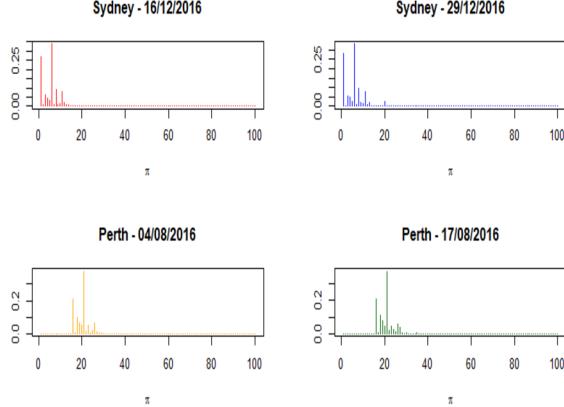
Figure 7: Posterior means for the weights of the mixture, truncated at 50 components for stSB.
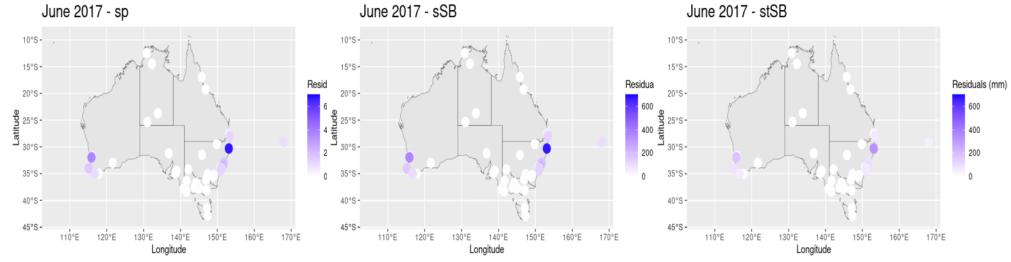
Figure 8: Residuals plots for the month of June 2017 obtained with the three methods, sp (left), sSB (centre), stSB (right). The size of dots is fixed.

(East), especially around Byron Bay. On the other hand, stSB is able to capture the peaks of rainfall and the squared residuals show less spatial structure. It is interesting to note that the spatio-temporal stick-breaking process appears to adapt quickly to changes over time. With access to daily rainfall data, the model adjusts to seasonal changes without the necessity of including a seasonal regression effect in the model.

In this application, 100% of the posterior samples for the separability parameter λ equal zero. In this scenario, the hypothesis of separability between time and space for the mixing weights is not rejected. Therefore, the covariance between observations recorded at different spatial locations and time points can be interpreted as dependent on the distance between spatial locations and the distance between time points, as described in Section 3.

6   Conclusion

Modelling spatio-temporal phenomena is an important and challenging problem. Sometimes there is not enough information to define a parametric model and nonparametric methods may be preferred. This work presents a semi-parametric method to model spatio-temporal data, which can be used to obtain higher predictive accuracy without the need of introducing a large number of covariates.

The spatio-temporal stick-breaking prior proposed in this paper is a generalisation of Reich and Fuentes (2007) and Dunson and Park (2008). This model avoids over-smoothing, but, differently from the spatial stick-breaking of Reich and Fuentes (2007), can model non-stationary covariance structures (with respect to either time or space, or both of them) and with a covariance among observations that can be symmetric or asymmetric with respect to time and space. Moreover, this model avoids the need for space-time varying covariates as in Hossain et al. (2013) to model the weight components.

Two models have been considered: a single-atom model where only the mixing weights are dependent on time and space, and a model where both atoms and weights are allowed to vary over time and space. The two models have shown to often lead to similar results, however the latter is definitely more computationally intensive, and this can lead to the need to train the model on a reduced number of observations.

The model has been validated by using simulated data and real data, and compared to other spatio-temporal models. Two dataset has been analyzed: a dataset including temperature recordings between 1985 and 2004 in California, USA (available in the Supplementary Material (Grazian, 2024)), and a dataset of daily rainfall measurements in Australia. With respect to the spatio-temporal kriging model, the spatial stick-breaking and the spatio-temporal stick-breaking avoid oversmoothing, and with respect to the spatial stick-breaking, the spatio-temporal stick-breaking allows to better differentiate among areas over time: for example, predicting larger volumes of rain in areas that have become more humid in recent years, and which used to be drier in the past.

An interesting current line of research focuses on situations where samples share some instead of all mixture components (with or without differing weights) and the components themselves may show some misalignment. For example, Lopes et al. (2003) and Müller et al. (2004) allows some clusters to be shared among samples, Teh et al. (2004) and Cron et al. (2013) allows all clusters to vary across samples, Rodriguez et al. (2008) allows a structures where clusters in each sample must be either identical or completely different, and Soriano and Ma (2019) propose a hierarchical mixture modelling framework allowing for both varying weights across samples and distribution misalignment. Such proposal often relies on a stick-breaking construction, however they also tend to assume independence within samples. The possibility to introduce spatial or temporal dependence on the mixing weights in a hierarchical way is left for further research.

The model is open to several generalisations. First, we have intentionally considered cases where no covariates were available, but it is easy to incorporate covariate infor-

mation. Similarly, it is possible to introduce seasonality components in the regression part of the model. Moreover, the model is easily generalisable to a multivariate setting.

Supplementary Material

Supplementary material for "Spatio-Temporal Stick-Breaking Process" (DOI: 10.1214/24-BA1419SUPP; .pdf). The Supplementary Material includes the derivation of Equation (3.2), figures describing more features of the correlation structure between observations and the stick-breaking weights, additional simulations from the spatio-temporal stick-breaking process (with single atoms and with varying atoms), the results of the simulation study introduced in Section 4.2 using a discrete prior distribution for λ instead of the spike-and-slab prior distribution presented in the paper, an additional application to a temperature dataset, details of the MCMC algorithm used for the implementation, and the prior specification of the parameters of the model.

References

Andersen, M. R., Vehtari, A., Winther, O., and Hansen, L. K. (2017). "Bayesian inference for spatio-temporal spike-and-slab priors." *The Journal of Machine Learning Research*, 18(1): 5076–5133. MR3763773. 12

Antoniak, C. E. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 1152–1174. MR0365969. 4

Arbel, J., Mengersen, K., and Rousseau, J. (2016). "Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity." *The Annals of Applied Statistics*, 1496–1516. MR3553233. doi: https://doi.org/10.1214/16-AOAS944. 7

Arndt, P. F. (2019). "Sequential and continuous time stick-breaking." *Journal of Statistical Mechanics: Theory and Experiment*, 2019(6): 064003. MR4007067. doi: https://doi.org/10.1088/1742-5468/ab1dd8. 7

Ascolani, F., Lijoi, A., and Ruggiero, M. (2021). "Predictive inference with Fleming–Viot-driven dependent Dirichlet processes." *Bayesian Analysis*, 16(2): 371–395. MR4255334. doi: https://doi.org/10.1214/20-ba1206. 5

Bakar, K. S. and Sahu, S. K. (2015). "spTimer: Spatio-temporal bayesian modelling using R." *Journal of Statistical Software*, 63(15): 1–32. 17

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2003). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC. MR3362184. 2

Barcella, W., De Iorio, M., and Baio, G. (2017). "A comparative review of variable selection techniques for covariate dependent Dirichlet process mixture models." *Canadian Journal of Statistics*, 45(3): 254–273. MR3691836. doi: https://doi.org/10.1002/cjs.11323. 7

Barrientos, A. F., Jara, A., and Quintana, F. A. (2012). "On the support of MacEachern's dependent Dirichlet processes and extensions." *Bayesian Analysis*, 7(2): 277–310. MR2934952. doi: https://doi.org/10.1214/12-BA709. 3, 10

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2007). "Bayesian inference for linear dynamic models with Dirichlet process mixtures." *IEEE Transactions on Signal Processing*, 56(1): 71–84. MR2439814. doi: https://doi.org/10.1109/TSP.2007.900167. 5

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(27). MR3634894. 5

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). "Deviance information criteria for missing data models." *Bayesian Analysis*, 1: 651–673. MR2282197. doi: https://doi.org/10.1214/06-BA122. 23

Cifarelli, D. and Regazzini, E. (1978). "Nonparametric statistical problems under partial exchangeability: The role of associative means." Technical report, Tech. rep., Quaderni Istituto Matematica Finanziaria dell'Università di Torino. 5

Cressie, N. (2015). *Statistics for spatial data*. John Wiley & Sons. MR3559472. 2

Cressie, N. and Huang, H. C. (1999). "Classes of nonseparable, spatio-temporal stationary covariance functions." *Journal of the American Statistical Association*, 94(448): 1330–1339. MR1731494. doi: https://doi.org/10.2307/2669946. 2, 11

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons. MR2848400. 2

Cron, A., Gouttefangeas, C., Frelinger, J., Lin, L., Singh, S. K., Britten, C. M., Welters, M. J., van der Burg, S. H., West, M., and Chan, C. (2013). "Hierarchical modeling for rare event detection and cell subset alignment across flow cytometry samples." *PLoS computational biology*, 9(7): e1003130. 25

Dey, R., Lewis, S. C., Arblaster, J. M., and Abram, N. J. (2019). "A review of past and projected changes in Australia's rainfall." *Wiley Interdisciplinary Reviews: Climate Change*, 10(3): e577. 22

DeYoreo, M. and Kottas, A. (2018). "Modeling for dynamic ordinal regression relationships: An application to estimating maturity of rockfish in California." *Journal of the American Statistical Association*, 113(521): 68–80. MR3803440. doi: https://doi.org/10.1080/01621459.2017.1328357. 6

Di Lucca, M. A., Guglielmi, A., Müller, P., and Quintana, F. A. (2013). "A simple class of Bayesian nonparametric autoregression models." *Bayesian Analysis*, 8(1): 63. MR3036254. doi: https://doi.org/10.1214/13-BA803. 5

Duan, J. A., Guindani, M., and Gelfand, A. E. (2007). "Generalized spatial Dirichlet process models." *Biometrika*, 94(4): 809–825. MR2416794. doi: https://doi.org/10.1093/biomet/asm071. 3, 7

Dunson, D. B. (2006). "Bayesian dynamic modeling of latent trait distributions." *Biostatistics*, 7(4): 551–568. MR2490545. doi: https://doi.org/10.1016/S0169-7161(05)25025-3. 5, 6

Dunson, D. B. and Herring, A. H. (2006). "Semiparametric Bayesian latent trajectory models." *Proceedings ISDS Discussion Paper*, 16. 3

Dunson, D. B. and Park, J.-H. (2008). "Kernel stick-breaking processes." *Biometrika*, 95(2): 307–323. MR2521586. doi: https://doi.org/10.1093/biomet/asn012. 3, 5, 6, 15, 25

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 209–230. MR0350949. 3

Ferguson, T. S. (1974). "Prior distributions on spaces of probability measures." *The Annals of Statistics*, 2(4): 615–629. MR0438568. 3

Fernández, C. and Green, P. J. (2002). "Modelling spatially correlated data via mixtures: a Bayesian approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4): 805–826. MR1979388. doi: https://doi.org/10.1111/1467-9868.00362. 3, 7

Finkenstadt, B., Held, L., and Isham, V. (2006). *Statistical methods for spatio-temporal systems*. Chapman and Hall/CRC. MR2307967. 2

French, J. (2022). "Package 'SpatialTools'." 15

Frühwirth-Schnatter, S. and Kaufmann, S. (2008). "Model-based clustering of multiple time series." *Journal of Business & Economic Statistics*, 26(1): 78–89. MR2422063. doi: https://doi.org/10.1198/073500107000000106. 3

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010). *Handbook of Spatial Statistics*. CRC press. MR2761512. doi: https://doi.org/10.1201/9781420072884. 1, 2

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(471): 1021–1035. MR2201028. doi: https://doi.org/10.1198/016214504000002078. 3, 5

Gelman, A., Hwang, J., and Vehtari, A. (2014). "Understanding predictive information criteria for Bayesian models." *Statistics and computing*, 24(6): 997–1016. MR3253850. doi: https://doi.org/10.1007/s11222-013-9416-2. 23

Genton, M. G. (2007). "Separable approximations of space-time covariance matrices." *Environmetrics*, 18(7): 681–695. MR2408938. doi: https://doi.org/10.1002/env.854. 12

Gneiting, T. (2002). "Nonseparable, stationary covariance functions for space–time

data." *Journal of the American Statistical Association*, 97(458): 590–600. MR1941475. doi: https://doi.org/10.1198/016214502760047113. 2, 12

Grazian, C. (2024). "Supplementary material for "Spatio-Temporal Stick-Breaking Process"." *Bayesian Analysis*. doi: https://doi.org/10.1214/24-BA1419SUPP. 9, 10, 13, 15, 17, 20, 21, 25

Griffin, J. E. and Steel, M. F. J. (2006). "Order-based dependent Dirichlet processes." *Journal of the American statistical Association*, 101(473): 179–194. MR2268037. doi: https://doi.org/10.1198/016214505000000727. 6

Griffin, J. E. and Steel, M. F. J. (2011). "Stick-breaking autoregressive processes." *Journal of Econometrics*, 162: 383–396. MR2795625. doi: https://doi.org/10.1016/j.jeconom.2011.03.001. 6

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. MR3425946. doi: https://doi.org/10.1016/j.csda.2015.10.002. 6, 17

Guttorp, P. and Gneiting, T. (2006). "Studies in the history of probability and statistics XLIX on the Matérn correlation family." *Biometrika*, 93(4): 989–995. MR2285084. doi: https://doi.org/10.1093/biomet/93.4.989. 2

Harvill, J. L. (2010). "Spatio-temporal processes." *Wiley interdisciplinary reviews: Computational Statistics*, 2(3): 375–382. 2

Hossain, M. M., Lawson, A., Cai, B., Choi, J., Liu, J., and Kirby, R. (2014). "Space-time areal mixture model: relabeling algorithm and model selection issues." *Environmetrics*, 25(2): 84–96. MR3233278. doi: https://doi.org/10.1002/env.2265. 3

Hossain, M. M., Lawson, A. B., Cai, B., Choi, J., Liu, J., and Kirby, R. S. (2013). "Space-time stick-breaking processes for small area disease cluster estimation." *Environmental and ecological statistics*, 20(1): 91–107. MR3039588. doi: https://doi.org/10.1007/s10651-012-0209-0. 7, 25

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173. MR1952729. doi: https://doi.org/10.1198/016214501750332758. 4, 15

Ishwaran, H. and James, L. F. (2003). "Generalized weighted Chinese restaurant processes for species sampling mixture models." *Statistica Sinica*, 1211–1235. MR2026070. 9

Ishwaran, H. and Rao, J. S. (2005). "Spike and slab variable selection: frequentist and Bayesian strategies." *The Annals of statistics*, 33(2): 730–773. MR2163158. doi: https://doi.org/10.1214/009053604000001147. 12

Ishwaran, H. and Zarepour, M. (2000). "Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models." *Biometrika*, 87(2): 371–390. MR1782485. doi: https://doi.org/10.1093/biomet/87.2.371. 15

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(1): 29–42. MR2414636. doi: https://doi.org/10.1002/bimj.200610375.   5, 18

Kottas, A., Wang, Z., and Rodríguez, A. (2012). "Spatial modeling for risk assessment of extreme values from environmental time series: A Bayesian nonparametric approach." *Environmetrics*, 23(8): 649–662. MR3019057. doi: https://doi.org/10.1002/env.2177.   5

Lau, J. W. and So, M. K. (2008). "Bayesian mixture of autoregressive models." *Computational Statistics & Data Analysis*, 53(1): 38–60. MR2528591. doi: https://doi.org/10.1016/j.csda.2008.06.001.   5

Lopes, H. F., Müller, P., and Rosner, G. L. (2003). "Bayesian meta-analysis for longitudinal data models using multivariate mixture priors." *Biometrics*, 59(1): 66–75. MR1978473. doi: https://doi.org/10.1111/1541-0420.00008.   25

MacEachern, S. N. (2000). "Dependent dirichlet processes." *Unpublished manuscript, Department of Statistics, The Ohio State University*, 1–40.   3, 5

Mena, R. H. and Ruggiero, M. (2016). "Dynamic density estimation with diffusive Dirichlet mixtures." *Bernoulli*, 22(2): 901–926. MR3449803. doi: https://doi.org/10.3150/14-BEJ681.   6

Mena, R. H., Ruggiero, M., and Walker, S. G. (2011). "Geometric stick-breaking processes for continuous-time Bayesian nonparametric modeling." *Journal of Statistical Planning and Inference*, 141(9): 3217–3230. MR2796026. doi: https://doi.org/10.1016/j.jspi.2011.04.008.   7

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian variable selection in linear regression." *Journal of the American Statistical Association*, 83(404): 1023–1032. MR0997578.   12

Muliere, P. and Tardella, L. (1998). "Approximating distributions of random functionals of Ferguson-Dirichlet priors." *Canadian Journal of Statistics*, 26(2): 283–297. MR1648431. doi: https://doi.org/10.2307/3315511.   15

Müller, P., Quintana, F., and Rosner, G. (2004). "A method for combining inference across related nonparametric Bayesian models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3): 735–749. MR2088779. doi: https://doi.org/10.1111/j.1467-9868.2004.05564.x.   25

Neelon, B., Gelfand, A. E., and Miranda, M. L. (2014). "A multivariate spatial mixture model for areal data: examining regional differences in standardized test scores." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(5): 737. MR3269410. doi: https://doi.org/10.1111/rssc.12061.   3

Nicholls, N., Drosdowsky, W., and Lavery, B. (1997). "Australian rainfall variability and change." *Weather*, 52(3): 66–72.   22

Nieto-Barajas, L. E., Müller, P., Ji, Y., Lu, Y., and Mills, G. B. (2012). "A Time-Series

DDP for Functional Proteomics Profiles." *Biometrics*, 68(3): 859–868. MR3055190. doi: https://doi.org/10.1111/j.1541-0420.2011.01724.x. 5

Papaspiliopoulos, O. and Roberts, G. O. (2008). "Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models." *Biometrika*, 95(1): 169–186. MR2409721. doi: https://doi.org/10.1093/biomet/asm086. 15

Park, J., Yi, S., Chang, W., and Mateu, J. (2022). "A Spatio-Temporal Dirichlet Process Mixture Model for Coronavirus Disease-19." *arXiv preprint arXiv:2207.06587\/.* MR4680802. doi: https://doi.org/10.1002/sim.9925. 5

R Core Team (2020). "R: A Language and Environment for Statistical Computing." URL https://www.R-project.org/ 17

Reich, B. J. and Bondell, H. D. (2011). "A spatial Dirichlet process mixture model for clustering population genetics data." *Biometrics*, 67(2): 381–390. MR2829007. doi: https://doi.org/10.1111/j.1541-0420.2010.01484.x. 6

Reich, B. J., Bondell, H. D., and Li, L. (2011). "Sufficient dimension reduction via Bayesian mixture modeling." *Biometrics*, 67(3): 886–895. MR2829263. doi: https://doi.org/10.1111/j.1541-0420.2010.01501.x. 6

Reich, B. J. and Fuentes, M. (2007). "A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields." *The Annals of Applied Statistics*, 1(1): 249–264. MR2393850. doi: https://doi.org/10.1214/07-AOAS108. 3, 5, 6, 18, 25

Rodriguez, A. and Dunson, D. B. (2011). "Nonparametric Bayesian models through probit stick-breaking processes." *Bayesian Analysis*, 6(1). MR2781811. doi: https://doi.org/10.1214/11-BA605. 7, 14

Rodriguez, A., Dunson, D. B., and Gelfand, A. E. (2008). "The nested Dirichlet process." *Journal of the American Statistical Association*, 103(483): 1131–1154. MR2528831. doi: https://doi.org/10.1198/016214508000000553. 25

Rodriguez, A. and Ter Horst, E. (2008). "Bayesian dynamic density estimation." *Bayesian Analysis*, 3(2): 339–365. MR2407430. doi: https://doi.org/10.1214/08-BA313. 5

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC. MR2130347. doi: https://doi.org/10.1201/9780203492024. 2

Schlather, M., Malinowski, A., Menck, P. J., Oesting, M., and Strokorb, K. (2015). "Analysis, simulation and prediction of multivariate random fields with package RandomFields." *Journal of Statistical Software*, 63: 1–25. 19

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 639–650. MR1309433. 4

Soriano, J. and Ma, L. (2019). "Mixture modeling on related samples by ψ-stick breaking and kernel perturbation." *Bayesian Analysis*, 14(1): 161–180. MR3910042. doi: https://doi.org/10.1214/18-BA1106. 25

Stein, M. L. (2005). "Space–time covariance functions." *Journal of the American Statistical Association*, 100(469): 310–321. MR2156840. doi: https://doi.org/10.1198/016214504000000854. 19

Stein, M. L. (2012). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media. MR1697409. doi: https://doi.org/10.1007/978-1-4612-1494-6. 2

Taddy, M. A. (2010). "Autoregressive mixture models for dynamic spatial Poisson processes: Application to tracking intensity of violent crime." *Journal of the American Statistical Association*, 105(492): 1403–1417. MR2796559. doi: https://doi.org/10.1198/jasa.2010.ap09655. 5

Teh, Y., Jordan, M., Beal, M., and Blei, D. (2004). "Sharing clusters among related groups: Hierarchical Dirichlet processes." *Advances in neural information processing systems*, 17. MR2279480. doi: https://doi.org/10.1198/016214506000000302. 25

Zhu, X., Ghahramani, Z., and Lafferty, J. (2005). "Time-sensitive Dirichlet process mixture models." Technical report. 5