A time dependent Bayesian nonparametric model for air quality analysis

Luis Gutiérrez

Escuela de Salud Pública, Fac. Medicina, Universidad de Chile

Ramsés H. Mena

IIMAS-UNAM, Mexico

Matteo Ruggiero

University of Torino & Collegio Carlo Alberto

Abstract

Air quality monitoring is based on pollutants concentration levels, typically recorded in metropolitan areas. These exhibit spatial and temporal dependence as well as seasonality trends, and their analysis demands flexible and robust statistical models. Here we propose to model the measurements of particulate matter, composed by atmospheric carcinogenic agents, by means of a Bayesian nonparametric dynamic model which accommodates the dependence structures present in the data and allows for fast and efficient posterior computation. Lead by the need to infer the probability of threshold crossing at arbitrary time points, crucial in contingency decision making, we apply the model to the time-varying density estimation for a $PM_{2.5}$ dataset collected in Santiago, Chile, and analyze various other quantities of interest derived from the estimate.

Keywords: Dirichlet process, density estimation, dependent process, stick-breaking construction, particulate matter.

1. Introduction

Human exposure to high levels of hazardous air pollutants has long been known to have adverse health effects. In October 2013, the International Agency for Research on Cancer, the specialized cancer agency of the World Health Organization, has announced that outdoor air pollution has been formally classified in Group 1, meaning that there is sufficient evidence that the agent is carcinogenic to humans (see IARC, 2013). Urban air pollution is estimated to cause worldwide 9% of lung cancer deaths, 5% of cardiopulmonary deaths, 1% of respiratory infection deaths, and to increase the risk of bladder cancer. Air pollutants associated with health risks include sulfure dioxide (SO_2), carbon monoxide (CO), nitrogen dioxide (NO_2), tropospheric ozone (O_3) and particulate matter (PM). The latter is constituted by solid and liquid particles emitted by

Email address: luisgutierrez@med.uchile.cl (Luis Gutiérrez)

combustion engines and households heating or formed as residual from other products such as vehicles tyres, brakes and road pavement among other sources. Particulate matter was evaluated separately by the International Agency for Research on Cancer and also classified as carcinogenic to humans. Of particular concern, due to their harmful character, are the levels of PM_{10} and $PM_{2.5}$, that is particles with diameter smaller than 10 and 2.5 micrometers respectively, which can reach the lungs through inhalation and even get to the inner organs, where they settle and become cause of serious health problems, including death (Préndez, 1993; Dockery et al., 1992).

The awareness of the hazard caused by air pollution, and PM in particular, generates an emerging concern in overpopulated metropolitan areas around the world. As a response, environmental authorities have set forth a series of actions to reduce and control the levels of pollutants as well as to set policies to enact alerts propitiously. One of the main actions by environmental authorities was the establishment of monitoring networks that collect information about concentrations of different pollutants. These sources of information generate spatially and temporally dependent data, which present asymmetries, heavy tails and sometimes multi-modality. For these reasons, robust and flexible statistical models are needed to accommodate this kind of phenomena with time-varying distributions. Indeed, statistical models constitute a key aspect when elaborating policies to decrease pollution levels (WHO, 2011), to set appropriate thresholds for emission contingencies and similar actions.

The spatial and temporal nature of pollutant measurements rise many statistical questions. One of the most important is a reliable assessment of the probability that a given pollutant concentration is or will be above a given threshold; such information is then used by the environmental authorities in order to call for environmental alerts and initiate policies for decreasing the pollutants levels. Various methods have been used in order to provide answers to statistical questions in air pollution research. These include extreme value theory (Roberts, 1979a,b; Horowitz, 1980; Smith, 1989; Davison and Smith, 1990), multivariate analysis (Guardani et al., 2003), neural networks (Comrie, 1997; Guardani et al., 1999; Pérez et al., 2000; Ordieres et al., 2005), Poisson models (Raftery, 1989; Achcar et al., 2008, 2010) and time series and spatial statistics (Draghicescu and Ignaccolo, 2009) among others. However, most of these either rely on parametric assumptions, such as symmetry or uni-modality on the pollutant level distribution or ignore the temporal dependence inherent to pollution data. The data structure resulting from air pollutant measurements cannot be robustly captured by time-dependent parametric models as the temporal dimension can change structural features of the involved distributions, such as the number of modes, the shape of tails and so on.

These type of datasets instead require models with enough generality to avoid unnecessary prior constraints in the estimation, flexibility to account for particular structures in the data, and relative computational simplicity to avoid excessive algorithmic effort in the presence of multivariate data.

In this paper we develop a flexible, nonparametric model, suitable to analyse multivariate air pollution data which exhibit asymmetries, multi-modality and spatio-temporal dependence. We

provide illustrations of the proposed model with an air quality analysis of the city of Santiago, Chile, where pollution contingencies have recently been enforced as a consequence of long periods during which the daily standard threshold of 50 μg/m^3 of $PM_{2.5}$ has been surpassed. The guiding question will be to estimate the probability that a given pollutant concentration is above a given threshold δ_0 at time t. However, our aim will be rather general, in that the object of inference will be the entire shape of the time dependent data generating distribution. From the time-varying density estimate, other quantities of interest can be derived, such as the mean functional or the probability of exceeding an arbitrary threshold. We emphasize that, unlike other approaches such as extreme value theory and Poisson models, the model and the estimation procedure are by construction invariant to the choice of the threshold, which can be determined *ex post*.

Specifically, we propose to model the pollutant levels through a simple time measure-valued Markov process. This will be a nonparametric mixture of parametric kernels, whose mixing measure is a stochastic process that induces the temporal dependence. The use of a multivariate kernel enables the model to capture the spatial dependence among the observations, and the temporal dependence structure built in the process allows to infer and reproduce that present in the data. The nonparametric approach avoids unrealistic constraints on the shape of the distributions involved, guaranteeing full flexibility in the estimation procedure and the ability to capture features such as asymmetries and multimodality. The relative simplicity of the mechanism which induces the temporal dependence in the mixture makes our proposal particularly appealing for these type of datasets. Unlike similar approaches based on dependent random probability measures, the proposal finds a good trade-off between generality and ease of implementation, in that the simplicity of the induced temporal dependence, together with usual techniques for dealing with the infinite dimensionality of the model, enables to design a fast and efficient algorithm for the posterior computation. Futhermore, the model allows to range along all degrees between fully correlated and incorrelated adjacent probability measures in the collection, thus enabling the researcher to calibrate the use of the procedure for different frameworks.

The paper is organized as follows. Section 2 presents the methodology we introduce for the analysis. After briefly reviewing some general background notions about Bayesian nonparametric density estimation, we develop the model, which falls into the class of dependent Dirichlet process mixtures. We discuss its properties and outline the strategy for posterior computation. In Section 3, we illustrate the performance of the proposed model with a simulated data set, comparing it also with a spline regression alternative. In Section 4, we apply the model to the air quality analysis on a $PM_{2.5}$ dataset collected in Santiago, Chile. The results include estimation of the time-varying density for a four-dimensional spatially correlated time series for the $PM_{2.5}$ levels recorded in different monitoring stations in the metropolitan area, together with other quantities of interest which are derived from the estimate. Furthermore, although these are not explicitly enforced in the model formulation, a study of the seasonality trends for one monitoring location

and the probability of exceeding an arbitrary threshold in single days of the year are also derived as a byproduct.

2. The model

After collecting some considerations on the density estimation problem from a Bayesian standpoint, we present a model for studying the air pollution data with temporal and spatial dependence. The proposed model falls in the realm of Bayesian nonparametric dependent models, and is tailored to multivariate data which exhibit this type of dependence. Since we aim at the entire shape of the time-varying distribution, from which other quantities of interest can be derived, such type of data structure requires a delicate tradeoff between modeling flexibility and ease of implementation, which is the main goal we pursue from a methodological point of view.

2.1. Bayesian nonparametric dependent density estimation

Bayesian nonparametric methods have provided a very flexible and efficient way to carry out density estimation. Starting from the seminal contribution of Lo (1984), a stream of literature has flourished on one of the most used and celebrated models in the discipline: the Dirichlet process mixture model. This assumes the observations come from the random density

$$f_P(y) = \int_\Theta K(y \mid \theta) P(d\theta), \qquad y \in \mathbb{S},$$

where K is a probability kernel density and P is a Dirichlet process (Ferguson, 1973, 1974). The latter admits the following so called *stick-breaking* representation (Sethuraman, 1994)

$$P(B) = \sum_{j=1}^{\infty} \omega_j \delta_{\theta_j}(B), \qquad B \in \mathcal{B}(\Theta), \tag{1}$$

where the ω_j's are weights constructed as

$$\omega_1 = v_1, \qquad \omega_j = v_j \prod_{l<j} (1 - v_l), \tag{2}$$

where $v_j \overset{\text{iid}}{\sim} \text{Beta}(1, M)$, $M > 0$, independent of $\theta_j \overset{\text{iid}}{\sim} F_0$, F_0 is a nonatomic distribution on $(\Theta, \mathcal{B}(\Theta))$ and $\delta_\theta(\cdot)$ a point mass at θ.

As the attention of the researchers, aided by the power of computers, progressively shifts towards datasets of big size or with unconventional structures, new developments are called for. In many applications, for example, it is of interest to study changes in the distribution of the response $y \in \mathbb{S}$ as a function of predictors $x \in \mathcal{X}$, with $\mathcal{X}$ the sample space for the predictors. MacEachern (1999) and MacEachern (2000) proposed the use of predictor-dependent collections of distributions through the use of mixture models. This entails specifying a prior distribution for the collection of random probability measures indexed by the predictors. Following the representation in (1),

MacEachern proposed to construct a collection of dependent random probability measures $\{P_x, x \in \mathcal{X}\}$ as

$$P_x(B) = \sum_{j=1}^{\infty} \omega_j(x)\delta_{\theta_j(x)}(B), \qquad B \in \mathcal{B}(\Theta) \tag{3}$$

where the weights ω_j and the atoms θ_j are allowed to depend on the predictor value $x \in \mathcal{X}$. The dependent Dirichlet process (DDP) corresponds to the case where marginally P_x is a Dirichlet Process. DDPs with fixed weights $\omega_j(x) \equiv \omega_j$ have been successfully applied to the analysis of variance (De Iorio et al., 2004), spatial modeling (Gelfand et al., 2005), classification (De la Cruz et al., 2007b; Gutiérrez and Quintana, 2011) among others. When the space $\mathcal{X} = \mathcal{T}$ indexes time, perhaps the first contribution can be traced back to Feigin and Tweedie (1989), where a Markov chain with DP marginals and associated linear functionals are studied. Subsequently, other proposals, mainly based on the ideas in MacEachern (1999), have been introduced: among others, Dunson (2006) assumes Dirichlet distributed innovations in an autoregression setting; Dunson et al. (2007), Griffin and Steel (2006), Dunson and Park (2008), Rodriguez and Dunson (2011) and Arbel et al. (2014) develop DDPs where predictor dependence is introduced in the weights; Caron et al. (2007) and Caron et al. (2008) propose a time-varying DP mixture with reweighing and movement of atoms; Rodriguez and ter Horst (2008) induce the dependence in time only via the atoms.

In the present framework, the data are assumed to be generated from the random density function $f_{P_t} : \mathbb{S} \times \mathcal{T} \longrightarrow \mathbb{R}_+$ given by

$$f_{P_t}(y) = \int_{\Theta} K(y \mid \theta) P_t(d\theta), \qquad y \in \mathbb{S} \quad \text{and} \quad t \in \mathcal{T} \tag{4}$$

Here $\mathbb{S}$ is the sample space, $K(\cdot \mid \theta)$ is a, possibly multivariate, kernel on $(\mathbb{S}, \mathcal{B}(\mathbb{S}))$, $\theta \in \Theta \subset \mathbb{R}^d$ and $P = \{P_t, t \in \mathcal{T}\}$ is a collection of dependent random probability measures on $(\Theta, \mathcal{B}(\Theta))$. Finally, $\mathcal{T} = (0, T]$ or $\mathcal{T} = \{1, 2, \ldots, T\}$, $T < \infty$, indexing the times at which the observation vectors are collected. With an appropriate specification of the kernel K and the prior distribution for P, virtually any shape for $f_P = \{f_{P_t}, t \in \mathcal{T}\}$ can be recovered from (4). Furthermore, other quantities of interest can be easily estimated given an estimate of f_{P_t}. We will be particularly interested in the mean functional

$$\eta_t = \int_{\mathbb{S}} y f_{P_t}(dy), \tag{5}$$

and the functional which measures the probability of exceeding a fixed threshold δ_0

$$\mathcal{E}_t(\delta_0) := \mathbb{P}_t(y > \delta_0) = \int_{\delta_0}^{\infty} f_{P_t}(y)dy. \tag{6}$$

With the goal of fruitfully combining generality and computational ease, in the next section we will propose a simple dependent Dirichlet process which is analytically tractable and avoids the use of complex simulation algorithms, providing a straightforward way for exploring the posterior quantities of interest. Such model is tailored to use it with multivariate data which show dependence with respect to both time and space, making it particularly appropriate in the research on environmental pollution and related fields.

2.2. A Simple Dependent Dirichlet Process

Dependent processes with fixed weights can lack enough modeling flexibility. On the other hand, letting both the weights and the atoms vary could result in a burdensome computation. Here then we concentrate on models with varying weights and fixed atoms. Such choice proves to be effective on the practical side, as it guarantees fast computation while preserving the ability to capture the data generating distribution.

Consider then

$$P_t = \sum_{j=1}^{\infty} w_j(t)\delta_{\theta_j}, \qquad \theta_j \overset{\text{iid}}{\sim} F_0, \qquad t \in \mathcal{T}, \tag{7}$$

where $\sum_{j=1}^{\infty} w_j(t) = 1$ almost surely for all $t \in \mathcal{T}$ and F_0 is a non atomic probability measure on Θ. Due to the applications considered in Section 3, we will concentrate on the discrete time case, that is when $\mathcal{T} = \{1, 2, \ldots, T\}$, $T < \infty$. A straightforward possible extension to the continuous time case $\mathcal{T} = (0, T]$ is discussed at the end this section.

Aiming at preserving the stick-breaking structure (2) at the marginal level, let $w(\cdot) = \{w(t) = (w_1(t), w_2(t), \ldots), t \geq 0\}$ in (7) be a realization of $W(\cdot) = \{W(t) = (W_1(t), W_2(t), \ldots), t \geq 0\}$, where

$$W_1(t) = V_1(t), \qquad W_j(t) = V_j(t)\prod_{l<j}(1 - V_l(t)). \tag{8}$$

Each component $V(\cdot) = \{V_j(t), t \geq 0\}$ is a Markov chain defined as follows:

$$\begin{aligned} V_j(t_1) &\sim \text{Beta}(1, M), \quad M > 0 \\ V_j(t_k) \mid V_j(t_{k-1}) &= \begin{cases} V_j(t_k) \sim \text{Beta}(1, M) & \text{with probability } \phi, \\ V_j(t_k) = V_j(t_{k-1}) & \text{with probability } 1 - \phi, \end{cases} \qquad k = 2, \ldots, T, \end{aligned} \tag{9}$$

where $\phi \in [0, 1]$.

Thus each stick-breaking component is updated at geometric times with a fresh, uncorrelated value from the stationary distribution $\text{Beta}(1, M)$. This construction clearly guarantees that P_t marginally is a Dirichlet process. The extreme simplicity of the dependent process defined via (9) leads to analytically tractable posterior updates, see Section 2.3. In addition, the parameter ϕ controls the autocorrelation of each weight component, and thus indirectly the autocorrelation of the whole process $P = \{P_t, t \in \mathcal{T}\}$. When ϕ approaches 0, the stick-breaking components tend to spend more and more time on the explored values, whereas when ϕ goes to 1, their trajectory approaches a Beta noise on the interval $(0, 1)$. Figure 1 shows some sample paths of $V_j(t)$ for different values of ϕ, together with the corresponding histograms of the ergodic frequencies and the autocorrelation.

The following result, whose proof can be found in the Appendix, identifies the autocorrelation function for $P = \{P_t, t \in \mathcal{T}\}$.

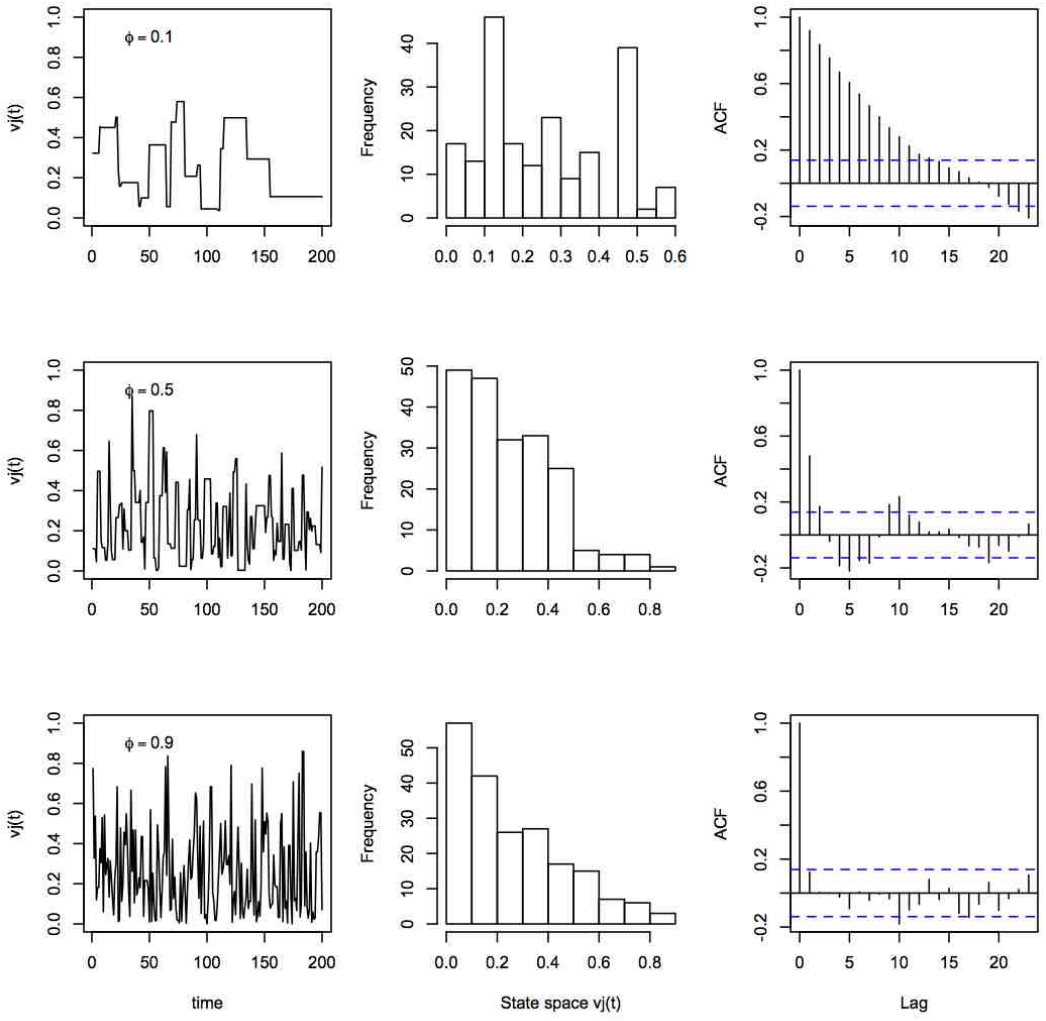

Figure 1: Sample paths, ergodic states frequencies and autocorrelation for $V_j(t)$ with $M = 3$, and $\phi = 0.1$ (top row), $\phi = 0.5$ (middle row), $\phi = 0.9$ (bottom row).

Proposition 2.1. *Let $P = \{P_t, t \in \mathcal{T}\}$ be as above. Then*

$$\mathrm{Corr}(P_t(A), P_{t+s}(A)) = \frac{(1+M)(2+M+(1-\phi)^s M)}{(2+M)(1+2M)-(1-\phi)^s M}.$$

Model (7) can be generalized to the continuous time case, by allowing the time effect to enter via the parameter ϕ. Specifically, by setting $\phi_t := \exp\{-\alpha\Delta_t\}$, where $\alpha > 0$ and Δ_t is the time increment among consecutive and equally spaced observations. The resulting autocorrelation function can be easily derived as a simple extension of Proposition 2.1.

The correlation can be calibrated, to some extent, by means of the parameter ϕ, and can be seen to converge to $(1+M)/(1+2M)$ as $s \to \infty$, which thus gives a lower bound. The existence of such bound is structural and common to all dependent processes with fixed atoms. Although this latter

feature appears to be undesirable, it does not constitute a major inconvenience, in particular due to the fact that it does not prevent the induced prior from having a large support. The fact that the constructed process has full support with respect to the weak topology is implied by Theorems 1 and 3 in Barrientos et al. (2012). On the other hand, fixing the atoms in the model allows to design faster computational schemes, which determines a sure advantage in terms of estimation efficiency.

2.3. Algorithm for posterior inference

In order to perform posterior inference for a set of observed concentration levels of $PM_{2.5}$ modelled via the simple dependent Dirichlet process (4) we resort to a Markov chain Monte Carlo procedure. Specifically, we construct a Gibbs sampler algorithm with slice sampling steps as in Walker (2007) to overcome the infinite-dimensionality inherent to the dependent Dirichlet process. We consider an augmented model given by

$$f_{P_t}(y, u, s) = \mathbb{I}(u < w_s(t))K(y \mid \theta_s), \tag{10}$$

where s denotes the allocation variable of y and u is a uniform random variate on $(0, w_s)$.

Let us assume we observe n trajectories measured at times $(t_1, \ldots, t_T)$. Such multiple-trajectory framework will be used below in Section 3 and 4.2 in order to capture structural changes in the shape of the distribution with respect to the temporal dimension. For $i = 1, \ldots, n$ and $k = 1, \ldots, T$, denote by $y_{i,k}$ the observation of the i-th trajectory at time t_k. Analogously, let $s_{i,k}$ and $u_{i,k}$ the allocation and slice variables corresponding to $y_{i,k}$. Hence, the augmented likelihood can be written as

$$\mathcal{L}_{\mathsf{v},\theta}(\mathbf{y}, \mathbf{u}, \mathbf{s}) = \prod_{i=1}^{n} \prod_{k=1}^{T} \mathbb{I}(u_{i,k} < w_{s_{i,k}}(t_k))K(y_{i,k} \mid \theta_{s_{i,k}}). \tag{11}$$

where $\mathsf{v} := \{(v_{1,k}, v_{2,k}, \ldots); k = 1 \ldots, T\}$ denotes the infinite collection of state-time observations of the stick-breaking components (9), i.e. $\{V_j(t_k) = v_{j,k}\}$, and $\theta := (\theta_1, \theta_2, \ldots)$ is the infinite set of random locations sampled from a non-atomic distribution F_0 with density f_0.

The main variables that need to be sampled at each step of the Gibbs algorithm are $\{v_{j,k}, \theta_j, j = 1, 2, \ldots, N\}$, $s_{i,k}$ and $u_{i,k}$ for $i = 1, \ldots, n$ and $k = 1, \ldots, T$. Here, $N := \max_{i,k}\{N_{i,k}\}$ with $N_{i,k}$ being the largest integer $s_{i,k}$ for which $\{u_{i,k} < w_{s_{i,k}}(t_k)\}$, which is equivalent to find an $N_{i,k}$ such that $\sum_{\kappa=1}^{N_{i,k}} w_\kappa > 1 - u_{i,k}$.

Updating the locations.

For the locations, which are random but independent of time, we obtain

$$\pi(\theta_j \mid \ldots) \propto f_0(\theta_j) \prod_{\{i,k:s_{i,k}=j\}} K(y_{i,k} \mid \theta_j)$$

as in Walker (2007). Note that, for a fixed time t, observations from different trajectories can be allocated to different atoms. For the particular case when $n = 1$ as in Section 4.1 there is a single atom and a single allocation for each time, namely the Gibbs updates are based on s_k instead of $s_{i,k}$.

Updating the weights.

To update the time-dependent weights, we need to update the stick-breaking components; to this end, denote by $p_{j,k}$ the transition density $\mathbb{P}(V_j(t_k) \in A \mid V_j(t_{k-1}) = v_{j,k-1})$ corresponding to the j-th process (9). Hence, it is seen that

$$\begin{aligned}\pi(v_{j,k} \mid \cdots) \propto & \{p_{j,2}\,\mathrm{Beta}(v_{j,1}; 1, M)\,\mathbb{I}(k=1) + p_{j,k+1}p_{j,k}\,\mathbb{I}(k \neq 1, T) + p_{j,T}\,\mathbb{I}(k=T)\} \\ & \times v_{j,k}^{\mathsf{n}_{j,k}}(1-v_{j,k})^{\mathsf{m}_{j,k}} \end{aligned} \tag{12}$$

where $\mathrm{Beta}(\cdot; a, b)$ denotes the density of a Beta distribution with mean $a/(a+b)$ and

$$\mathsf{n}_{j,k} := \sum_{i=1}^{n} \mathbb{I}(s_{i,k} = j), \qquad \mathsf{m}_{j,k} := \sum_{i=1}^{n} \mathbb{I}(s_{i,k} > j). \tag{13}$$

After applying transition (9) and arranging terms one obtains

$$\begin{aligned}\pi(v_{j,k} \mid \ldots) = & \phi M \mathsf{q}_{0,k}(v_{j,k+1})\,\mathrm{Beta}(v_{j,k}; 1+\mathsf{n}_{j,k}, M+\mathsf{m}_{j,k}) \\ & + \phi M\, \mathsf{q}_{1,k}(v_{j,k+1}, v_{j,k-1}, v_{j,k-1})\,\mathbb{I}(v_{j,k} = v_{j,k-1}) \\ & + \phi M\, \mathsf{q}_{1,k}(v_{j,k+1}, v_{j,k+1}, v_{j,k+1})\,\mathbb{I}(v_{j,k} = v_{j,k+1}) \\ & + (1-\phi)\mathsf{q}_{1,k}(0, v_{j,k-1}, v_{j,k-1})\,\mathbb{I}(v_{j,k} = v_{j,k+1} = v_{j,k-1}) \end{aligned} \tag{14}$$

for $k \neq 1, T$ and

$$\pi(v_{j,1} \mid \ldots) = M\mathsf{q}_{0,1}(v_{j,2})\,\mathrm{Beta}(v_{j,1}; 1+\mathsf{n}_{j,1}, M+\mathsf{m}_{j,1}) + M\,\mathsf{q}_{1,1}(v_{j,2}, v_{j,2}, v_{j,2})\mathbb{I}(v_{j,1} = v_{j,2})$$

$$\pi(v_{j,T} \mid \ldots) = \mathsf{q}_{0,T}(0)\,\mathrm{Beta}(v_{j,T}; 1+\mathsf{n}_{j,T}, M+\mathsf{m}_{j,T}) + \mathsf{q}_{1,T}(0, v_{j,T-1}, v_{j,T-1})\,\mathbb{I}(v_{j,T} = v_{j,T-1}),$$

where

$$\mathsf{q}_{0,k}(v) := \frac{\phi M\,(1-v)^{M-1}}{\mathcal{B}(1+\mathsf{n}_{j,k}, M+\mathsf{m}_{j,k})} \tag{15}$$

$$\mathsf{q}_{1,k}(u, v, w) := (1-\phi)\,(1-u)^{M-1}\,v^{\mathsf{n}_{j,k}}\,(1-w)^{\mathsf{m}_{j,k}} \tag{16}$$

with $\mathcal{B}(a, b) = \Gamma(a+b)/(\Gamma(a)\Gamma(b))$.

When $n = 1$, the updates of the $v_{j,k}$'s are performed with equations (14) and (15), using

$$\mathsf{n}_{j,k} := \mathbb{I}(s_k = j), \qquad \mathsf{m}_{j,k} := \mathbb{I}(s_k > j), \tag{17}$$

in place of (13).

Updating the membership and slice latent variables.

The full conditional distributions for the membership and slice latent variables are given by

$$p(s_{i,k} = \kappa \mid \ldots) \propto K(y_{i,k} \mid \theta_\kappa)\mathbb{I}(\{\kappa : w_\kappa > u_{i,k}\}) \tag{18}$$

and

$$\pi(u_{i,k} \mid \ldots) = U(u_{i,k}; 0, w_{s_{i,k}}) \tag{19}$$

respectively .

Updating others hyper-parameters.

The total mass parameter M is updated as in Escobar and West (1995) assuming a gamma prior $\text{Ga}(a, b)$. Finally, for the parameter ϕ, we assume a Beta prior such that $\phi \sim \text{Beta}(\gamma, \nu)$. The posterior distribution for ϕ is not available in closed form, thus a Metropolis-Hasting step is needed. We propose to use a truncated normal distribution as a proposal for ϕ, that is, at iteration τ, $\phi^* \sim N(\phi^* \mid \phi^{\tau-1}, c)\mathbb{1}_{[0,1]}$. Then set:

$$\phi^{\tau} = \begin{cases} \phi^*, & \text{with probability } \min(r, 1) \\ \phi^{\tau-1}, & \text{otherwise} \end{cases}$$

and

$$r = \frac{p(\phi^* \mid \mathbf{y})/N(\phi^* \mid \phi^{\tau-1}, c)\mathbb{1}_{[0,1]}}{p(\phi^{\tau-1} \mid \mathbf{y})/N(\phi^{\tau-1} \mid \phi^*, c)\mathbb{1}_{[0,1]}}. \tag{20}$$

In (20), $p(\phi^\diamond \mid \mathbf{y})$ is given by

$$p(\phi^\diamond \mid \mathbf{y}) \propto \phi^{\diamond(\gamma-1)}(1-\phi^\diamond)^{(\nu-1)} \prod_{l=1}^{n\times T} \sum_{j=1}^{N} w_j^\diamond K(y_l \mid \theta_j)$$

and $w_j^\diamond$ are the weights sampled using $\phi^\diamond$.

An appealing feature of the above Gibbs sampler algorithm is the updating mechanism for the time-dependent weights (14). This is considerably simpler than those corresponding to other time-dependent DP models, where latent processes are needed within the MCMC (see, e.g., Rodriguez and Dunson, 2011; Mena and Ruggiero, 2015). Such simplicity is inherited from the dependence structure induced by the Markovian process (9). Indeed, the availability of a simple transition probability also leads to an appealing way to assess the dependence in the model, e.g. though Proposition 2.1, which does not involve the computation of any integrals, as for other DDP models with dependent weights. The posterior description of the weight processes is an important point for this class of models, since it represents the key component in the time-dependence description of the dynamic density model.

Here, it is worth emphasising that steps (18) and (19) of the above slice sampler algorithm allow the random truncation of the stick-breaking representation to be adaptive, reducing the possibility that the θ's are trapped in a particular value. Another important aspect of our proposal is that the total mass parameter M is randomised and thus, as noted by De Blasi et al. (2015, Remark 2), the resulting DPM model falls into the class of random probability measures where the clustering of observations depend on both, the total number of observations and the number of different parameter values. The model could clearly be extended, without much complication, to Pitman-Yor dependent priors. Such development could be of interest in other scenarios where the clustering structure of the data is considered to be particularly informative. Finally, it is important to highlight that our proposal is specific for time-dependent density estimation, thus it is less general than other approaches where the dependence could be placed in more general covariate spaces, hence providing with an alternative model only for such a case.

3. Illustration with simulated data

In this section we illustrate the use and performance of the proposed DDP model with a simulated data set. We simulate $n = 7200$ observations, corresponding to 30 realisations for each time $t = 0.1, 0.2, \ldots, 24$, from the following model:

$$y(t) \sim \begin{cases} N(f(t), \sigma_1^2), & \text{for} \quad 1 \leq t < 8, \\ 0.3N(f(t), \sigma_1^2) + 0.7N(f(t), \sigma_2^2) & \text{for} \quad 8 \leq t < 16, \\ 0.5N(f(t), \sigma_3^2) + 0.5N(0.1t + f(t), \sigma_3^2), & \text{for} \quad t \geq 16. \end{cases}$$

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

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

To complete the specification of the proposed DDP model, we select a Normal kernel $K(\mathbf{y} \mid \theta) := \mathrm{N}(\mathbf{y} \mid \mu, \sigma^2)$. The centering measure for the Dirichlet process is set to be a Normal/Inverse-Gamma distribution $F_0 := \mathrm{N}(\mu \mid \mu_0, \sigma_\mu^2)\mathrm{IG}(\sigma^2 \mid \alpha, \beta)$. The values of the hyperparameters were set to $\mu_0 = 0$, $\sigma_\mu^2 = 1000$, $\alpha = 1$, $\beta = 1$, $a = 0.5$, $b = 0.2$, $\gamma = 2$ and $\nu = 2$. These hyperparameter values imply proper but vague distributions, in particular for μ and M. The Gibbs sampling algorithm was implemented in R. We generated 50,000 iterations, discarded the first 20,000 as burn-in, and thinned every 30 iterations for a total of 1,000 samples. The convergence diagnostics were performed with the R package CODA (Plummer et al., 2006). We calculated the effective sample size measure and the Gelman and Rubin's convergence diagnostic (Gelman and Rubin, 1992). Concerning the former, the effective sample size was 14,275 for M and 4,131 for ϕ. The smaller effective sample size for ϕ is expected due to the metropolis step involved in its sampling schedule. As for the latter, Figure 10 in the Appendix shows the point estimates of the potential scale reduction factor for the parameters M and ϕ, which control the autocorrelation of the dependent process in (9), to be 1.06 and 1.01 after 50,000 iterations of ten chains started at different points, with their upper 97.5% confidence limits being 1.13 and 1.01 respectively, thus providing evidence of convergence. The other model parameters showed similar results in terms of convergence diagnostics.

With the purpose of comparing our model with an alternative, we also fitted a spline regression model to the data. Specifically, we employed the Bayesian penalized spline regression proposed by Crainiceanu et al. (2005), which can be seen as a mixed model with random and fixed effects. The splines coefficients are the random effects and these are penalized by controlling its variance in the prior distribution.

Figure 2 shows the temporal evolution of the density estimate obtained with the DDP and the spline regression model. Both models follow the non-linear trajectory of the data, but the density is not well captured by the spline model, particularly where this becomes multimodal. Figure 3 compares the performance of the two models at selected time points, indicated in Figure 2 with vertical dashed lines, with the black solid line being the true density, the red solid line being the

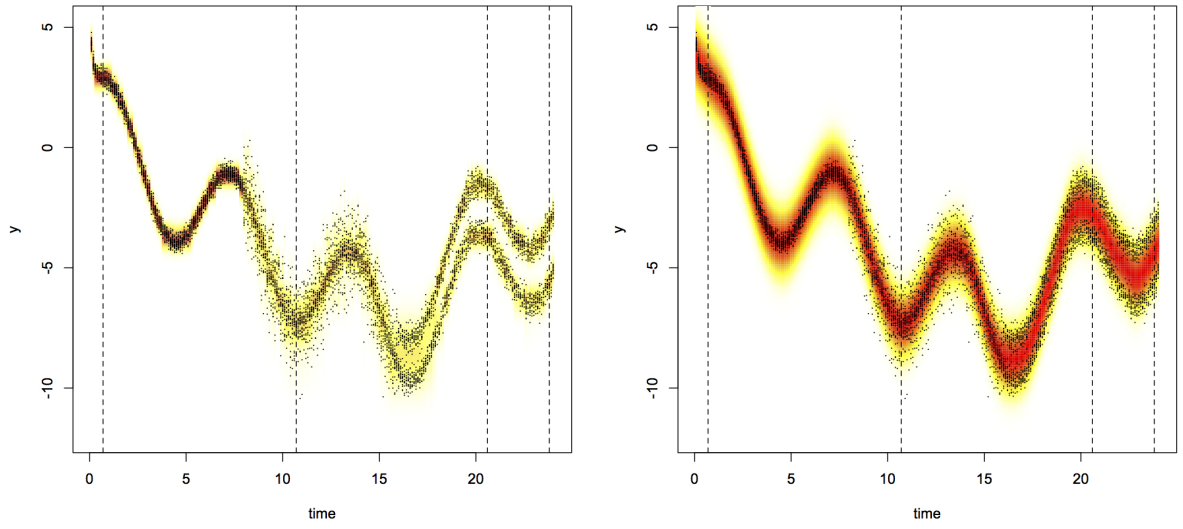

Figure 2: Density estimation for simulated data using the DDP mixture model (left panel), and the penalised spline regression (right panel). The density sections at times indicated by vertical dashed lines are shown in Figure 3.

DDP estimate, with pointwise 95% credible intervals, and the green line being the spline regression estimate. The shape of the density is well captured by the DDP model, especially in temporal regions of multimodality, whereas the spline regression model is not flexible enough to detect and reproduce this structural changes.

4. Application to air pollution studies

We consider two data sets, both relative to the urban area of Santiago, Chile. The first dataset corresponds to ten years of daily observations of $PM_{2.5}$ concentration levels, in $\mu g/m^3$, at Parque O'Higgins station between January 1, 2002 and December 30th, 2011, which is one of the first and most representative stations for pollution levels monitoring in the downtown area of Santiago. The second data set corresponds to daily observations of $PM_{2.5}$ concentration levels registered between December 30th, 2009 and December 30th, 2011 in Santiago for four different stations, located in Parque O'Higgins, Pudahuel, La Florida and Las Condes. These stations are representative of different levels of pollution in Santiago as both the downtown area and the suburbs are included. Specifically, Parque O'Higgins is located in the downtown area (500 meters above sea level (m.a.s.l.)) about 1 km west of a major highway and with a traffic of about 60,000 vehicles per day (Gramsch et al., 2006). Pudahuel is located in the western part of the city (480 m.a.s.l.), with two major

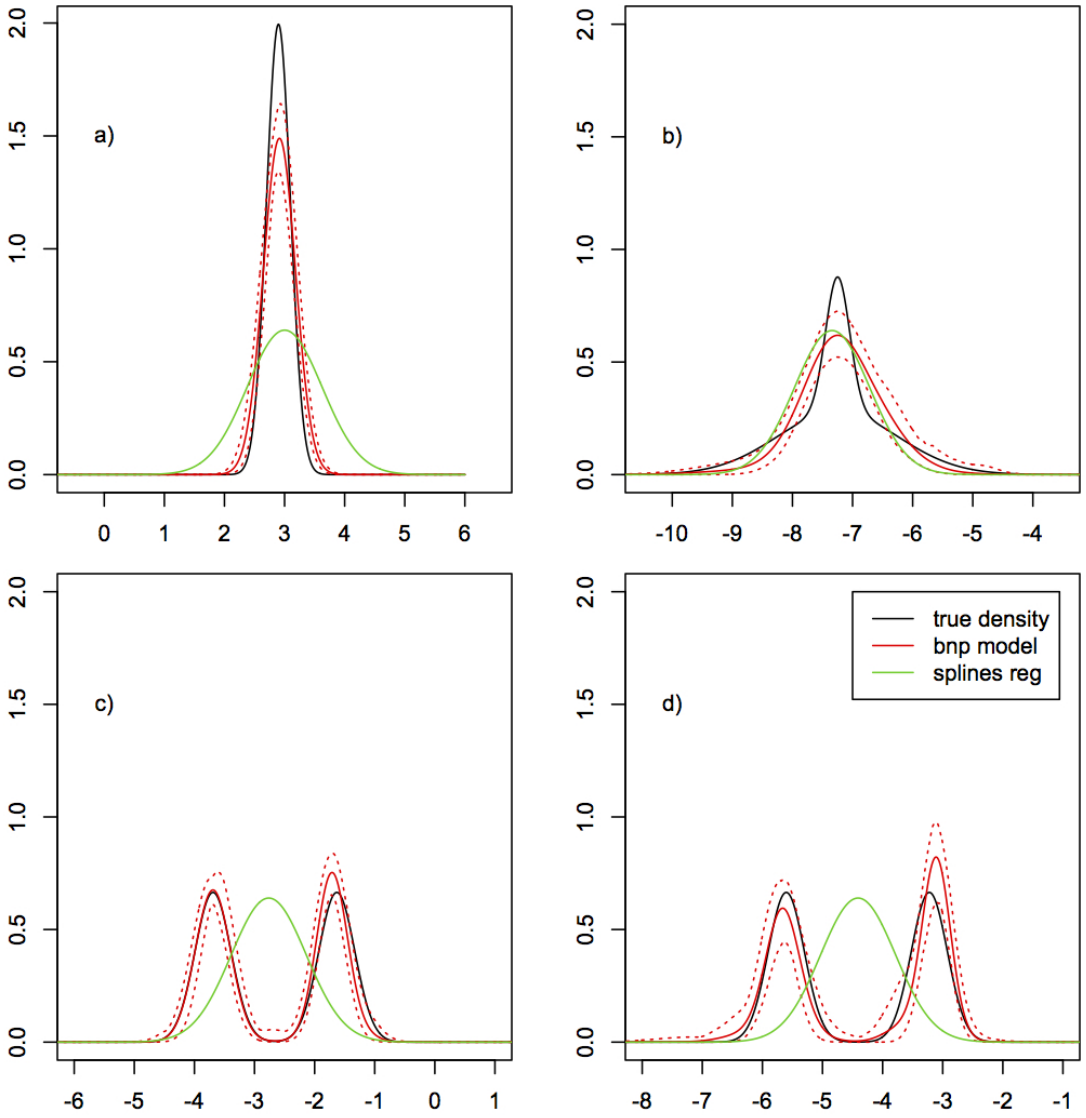

Figure 3: Comparison of density estimates at the time points highlighted in Figure 2. The black solid line is the true density, the red solid line is the DDP estimate, with pointwise 95% credible intervals, and the green line is the spline regression estimate.

roads with about 15,000 to 20,000 circulating vehicles per day. La Florida is placed in the southern part of the city (500 m.a.s.l.), surrounded by three major roads with traffic of 30,000, 35,000 and 55,000 vehicles per day as reported by Gramsch et al. (2006). Finally, Las Condes is placed in the eastern part of the city (700 m.a.s.l.) close to a road with about 15,000 vehicles per day.

4.1. Dynamic density estimation for spatially correlated $PM_{2.5}$ data

We apply the dependent model developed in Section 2 for studying the air pollution data through dynamic density estimation. We consider the $PM_{2.5}$ concentrations reordered at the four

Station	Jan-30-2010	Jan-30-2011	Jun-30-2010	Jun-30-2011
Las Condes	0.0004	0.0004	0.0023	0.0028
La Florida	0.0025	0.0029	0.0443	0.4881
Pudahuel	0.0044	0.0062	0.7536	0.7163
P. O'Higgins	0.0038	0.0037	0.4707	0.7801

Table 1: Probabilities of exceeding 50 $\mu g/m^3$ in 24 hours

stations in a period of two years. Since the stations are located in the same valley, with similar weather conditions, there is a spatial dependence among the pollution levels registered in different stations. To complete the specification of the DDP proposed, since the support of the pollution levels is $\mathbb{S} = [0, +\infty)$, we let $K(\mathbf{y} \mid \theta) = \text{log-N}_4(\mu, \Sigma)$, where log-$\text{N}_4$ denotes a 4-dimensional log-Normal distribution, whose variance-covariance matrix Σ models the spatial dependence among stations. The centering measure is set to be $F_0 := \text{N}_4(\mu \mid \mu_0, \Sigma_\mu)\text{IW}_4(\Sigma \mid \nu, A)$ where, IW_4 denotes a four dimensional inverse-Wishart distribution with expected value $E(\Sigma) = A/(\nu - 3)$. The values of the hyperparameters were fixed at $\mu_0 = (0, 0, 0, 0)^t$, $\Sigma_\mu = 100I_4$, $A = 0.01I_4$, $\nu = 8$, $a = 0.1$, $b = 0.1$, $\gamma = 2$ and $\nu = 2$, implying proper but vague prior distributions and representing lack of genuine prior information about the parameters.

The Gibbs sampler algorithm was implemented in R. We generated 50,000 iterations, discarded the first 20,000 as burn-in, and thinned every 30 iterations for a total of 1,000 samples. Figure 4 shows the posterior density estimates for the air pollution data in the four stations, where the heat contour represents the height of the posterior probability and the solid line the mean functional. From this figure, we can see that the model detects the change points of the data series and accommodates the heteroscedastic behaviour of the $\text{PM}_{2.5}$ concentrations.

Figure 5 shows the marginal density estimates for the four stations at four selected times: January 30th, 2010 and January 30th 2011 (summer in the southern hemisphere); June 30th, 2010 and June 30th, 2011 (winter). The vertical dashed line indicates the threshold level of maximum allowed concentrations in 24 hours (50 $\mu g/m^3$). The results show that in summer the distribution of $\text{PM}_{2.5}$ concentration is very similar between stations and the probability of exceeding the threshold of 50 $\mu g/m^3$ negligible. In winter, the $\text{PM}_{2.5}$ concentration has very different distributions according to the location. Such differences are crucial given the relative position between the curves and the threshold. Furthermore, in winter more variability and heavy tailedness emerge. The results in Figure 5 are in part a consequence of large emissions within the city itself, which in turn combine with a very low ventilation due to low wind speeds and strong inversions. The winter months (May to August) are cold with moderate rain and low wind speeds, whereas the summer is hot and dry and the average wind speed is higher than in other months. The

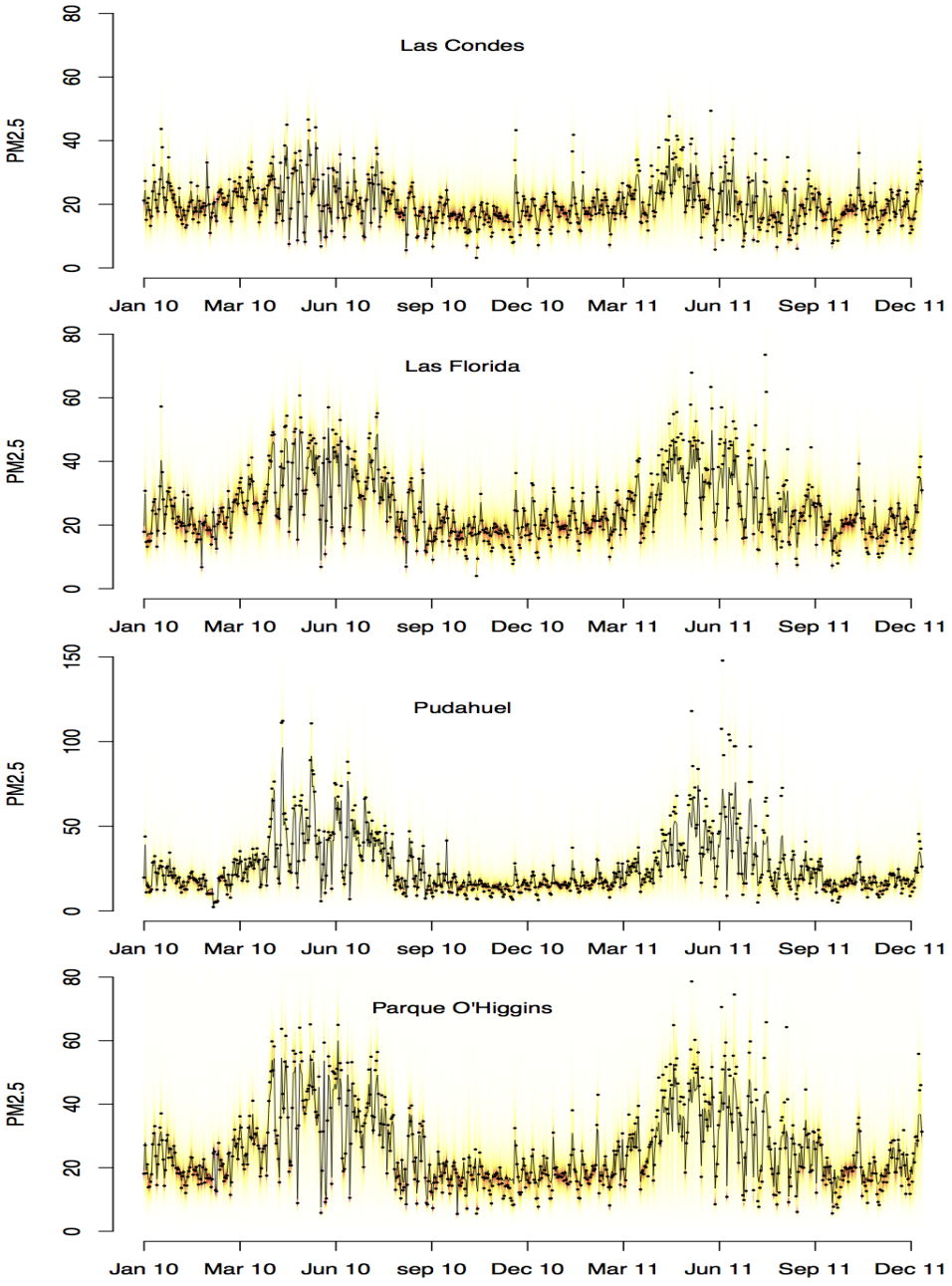

Figure 4: Temporal evolution of the density estimate for air pollution data using the DDP mixture model. The panels show the data points (dots), the posterior density estimate (heat contour) and the mean functional (solid line) for the four stations in the period under study, January 2010 to December 2011.

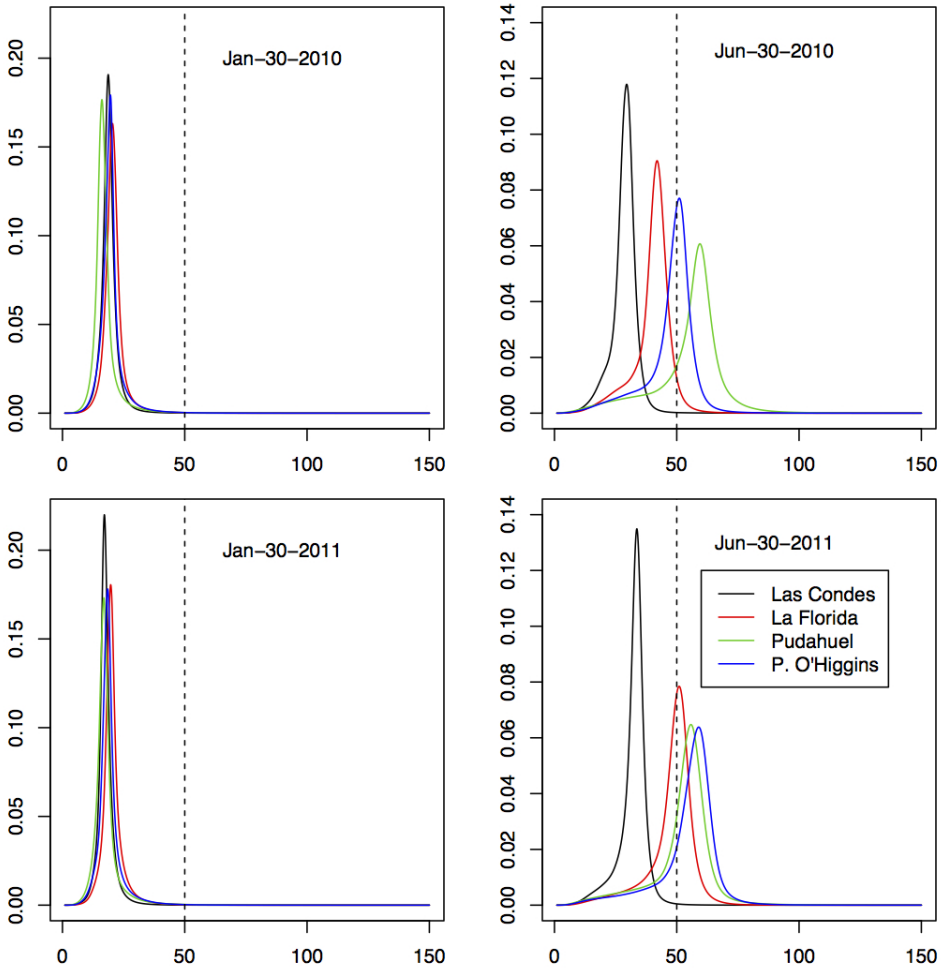

Figure 5: Posterior marginal densities (solid lines) against fixed threshold (dashed).

posterior version of the probabilities (6) of exceeding the given threshold can be easily computed from the model and are reported in Table 1.

Finally, Figure 6 shows the bivariate density estimates for pairs of stations on July 30th, 2011, from which it can be observed that Las Condes and La Florida stations shows less variability compared to Pudahuel and Parque O'Higgins. In both cases the $PM_{2.5}$ concentrations show positive correlation between stations, due to the spatial dependence captured and quantified by the model.

4.2. Seasonality study

The proposed dependent model also allows to investigate seasonality patterns. Here we are interested in the study of the distribution of the $PM_{2.5}$ levels in an average year. To this end, we split the ten years dataset for Parque O'Higgins station and consider every year as a separate trajectory, so that for each day of the year we have 10 data points. To complete the specification of the DDP

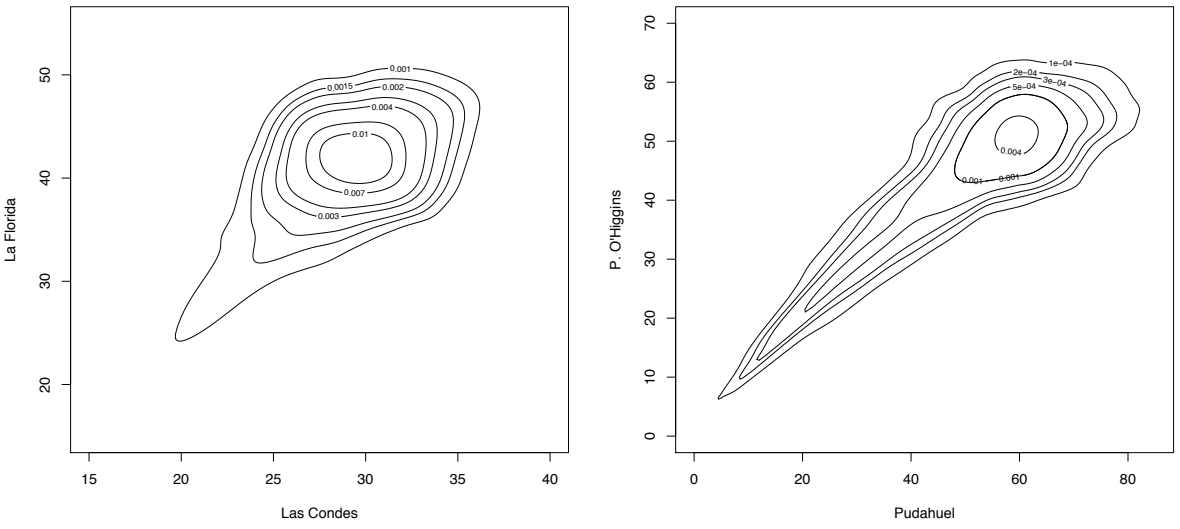

Figure 6: Bivariate density estimates on July 30th, 2011 for Las Condes and La Florida (left) and Pudahuel and Parque O'Higgins (right).

model, for this case we select a kernel $K(\mathbf{y} \mid \theta) := \text{log-N}(\mathbf{y} \mid \mu, \sigma^2)$, where log-N is a univariate log-Normal distribution. The centering measure is set to be $F_0 := \text{N}(\mu \mid \mu_0, \sigma^2_\mu)\text{IG}(\sigma^2 \mid \alpha, \beta)$, where $\text{IG}(\cdot \mid \alpha, \beta)$ denotes the inverse gamma distribution with expected value given by $\frac{\beta}{\alpha - 1}$ for $\alpha > 1$. The values for the hyperparameters were fixed at $\mu_0 = 0$, $\sigma^2_\mu = 100$, $\alpha = \beta = 2$, $a = b = 1$, $\gamma = 2$ and $\nu = 2$. Note that the hyper-parameters a and b, which define the prior distribution on M, must be specified accordingly to the sample size and number of trajectories. Specifically, having more trajectories would require a relatively smaller prior variance on the distribution of M since more information about the local number of groups is available. The general setting of the Gibbs sampler is analogous to that used for the previous example. The results are plotted in Figure 7 which shows the heat contour of the density estimate and the mean functional. The winter seasonality of the data, manifested in the heteroscedastic behaviour of the $\text{PM}_{2.5}$ levels, is captured and rendered by the estimate whose sections are more stretched density functions. This is further highlighted in Figure 8, where the marginal density estimates for May 10th and October 15th are plotted against a fixed threshold. Here is worth noticing the change of the distributional form when viewed at different times. In the present application, the results allow to draw inferences on the probability of exceeding the threshold under a time-varying distribution scenario. This is easily derived from the estimates and, for instance, it equals 0.45 on May, 11th but is negligible on October, 15th.

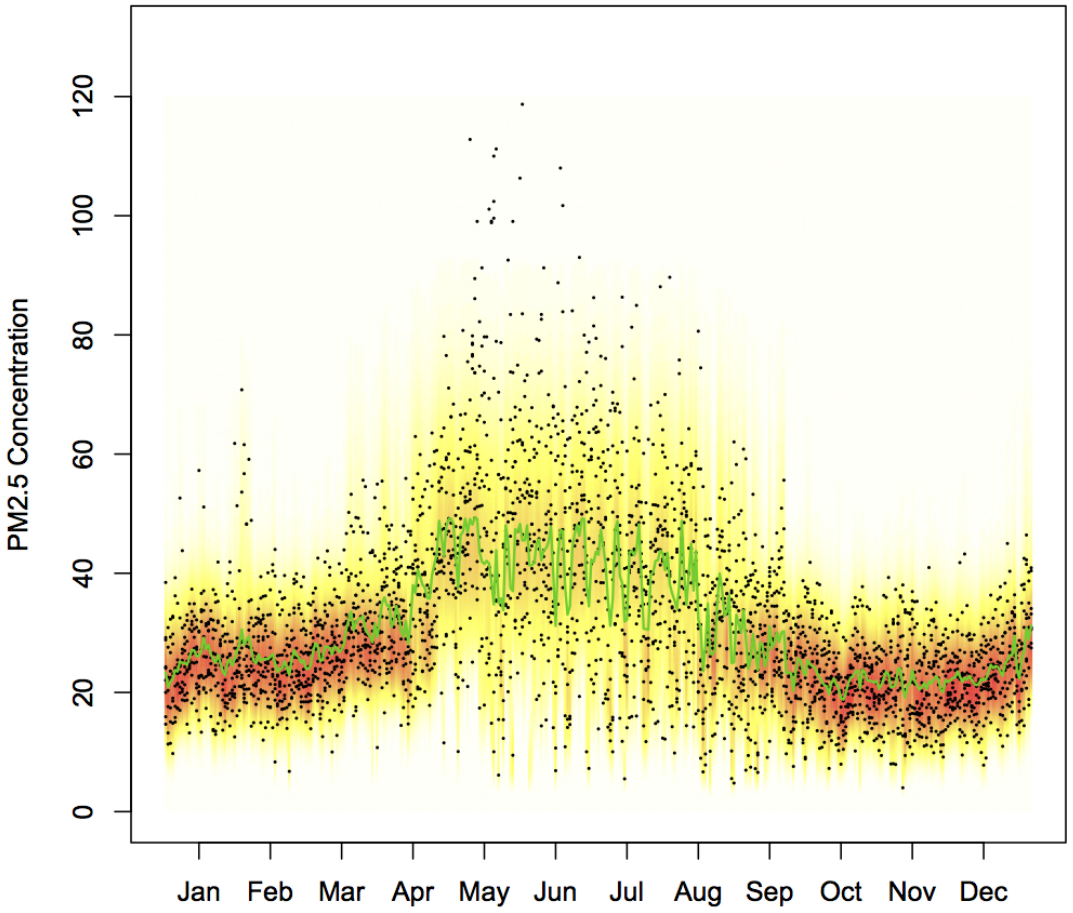

Figure 7: Density estimation for seasonality study on a single station data through several years. The dataset is split and years considered as trajectories for pattern highlighting. The plots shows the posterior estimate of the time-varying density (heat contour) and the mean functional (solid green line).

Elaborating on this point, and as a further illustration of the results that can be derived from the time-varying estimate, Figure 9 plots the probability of exceeding two different thresholds for each day of the year, namely 50 (Chilean) and 35 (US EPA) $\mu g/m^3$, with red bars highlighting those higher than 1/2. This type of information is essential for implementing local policy regulations to contain pollution, decided upon evaluation of the sporadic versus structural occurrence of the threshold crossing and the associated probabilities.

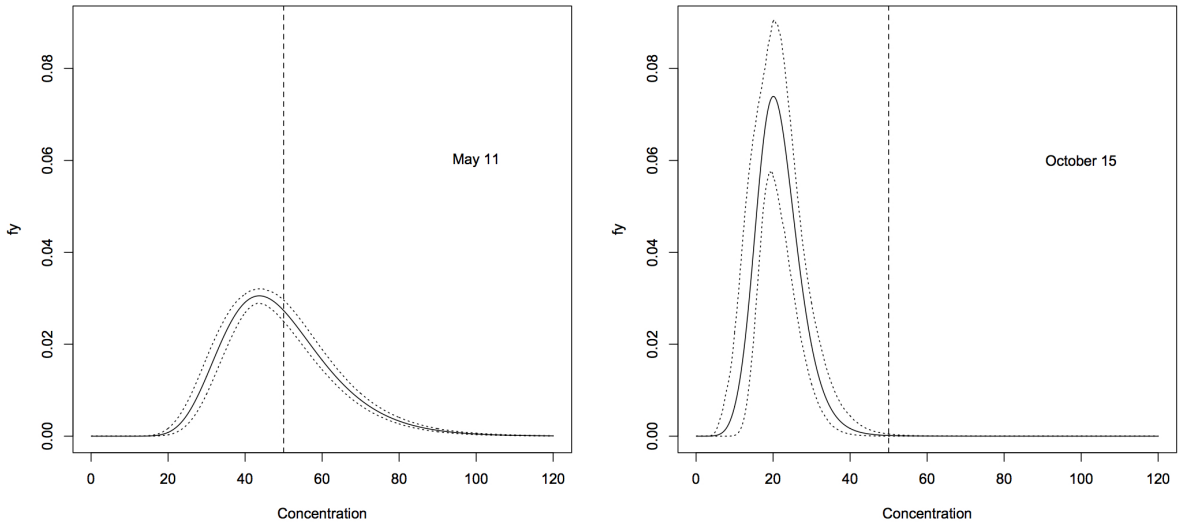

Figure 8: Posterior density estimates with 95% pointwise credible intervals for May, 10th (left) and October, 15th (right). The vertical dashed line highlights the threshold 50 μ_g/m^3.

5. Concluding remarks

We proposed a flexible time-varying density estimation model, which is useful in univariate and multivariate contexts. The proposal is especially designed to capture the distributional dissimilarity, characterizing phenomena such as those encountered in air quality analysis, and at the same time keep the computational complexity low. The guiding interest here was the assessment of the probability that a given pollutant concentration at time t surpasses certain threshold δ_0. This provides information about the number of times that such a threshold is surpassed in a time interval of interest. A robust evaluation of these and related quantities is key to the evaluation of air quality standards as well as the costs-benefits trade-off of the polices for decreasing pollution levels.

The results obtained from ten years of information in Parque O'Higgins station are representative of the behavior of an average year in the downtown area of Santiago city. These are helpful for planning new standards and to quantify the cost and benefit of the current policy. In particular, we have found that the probability of exceeding the current Chilean standard is high between April to August but these probabilities are smaller than 0.5. If Chile adopts the US EPA standard, considering the current levels of PM25, there is a high probability (bigger than 0.5) of exceeding the EPA standard in more than 100 days per year. However, a more demanding standard might not be a realistic policy for Santiago as such will most probably come with high costs in various

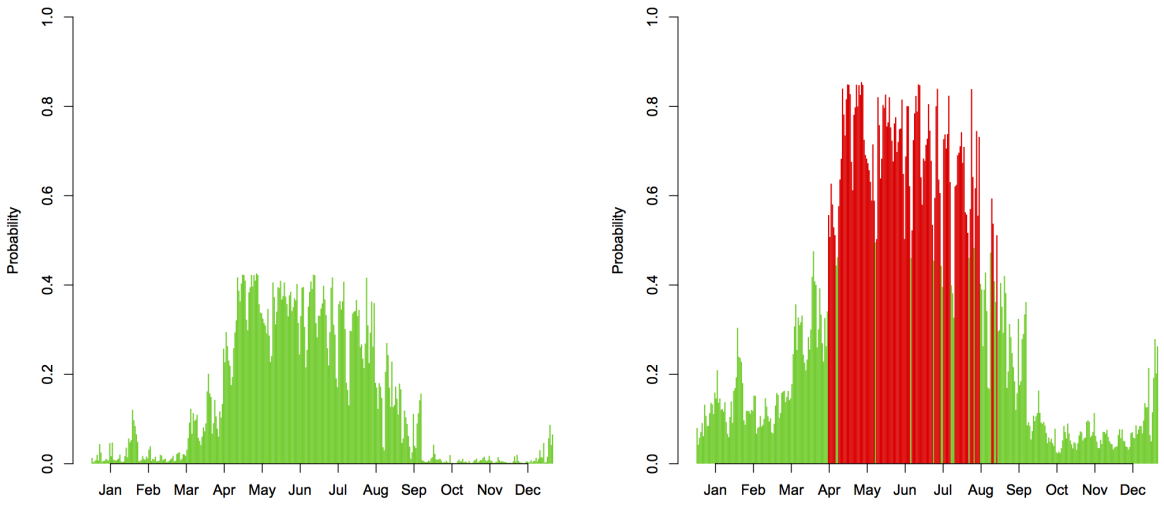

Figure 9: Probability of exceeding the thresholds 50 (left) and 35 (right) $\mu g/m^3$, for each day of the year, with probabilities higher than 1/2 highlighted in red.

sectors of society.

Moreover, the results captured from the multivariate analysis of the four stations allowed us to understand the relationship of $PM_{2.5}$ concentrations between each station. It also allowed us to make informative comparisons between stations. We have found for example, that Pudahuel station almost always showed the highest levels of $PM_{2.5}$ concentrations and the lowest levels were in Las Condes station.

We have illustrated the model in some particular situations, e.g. in the multivariate example we used the information of $p = 4$ stations. However, higher dimensional scenarios can be easily incorporated. In particular, in high dimensions, the variance-covariance matrix Σ of the log-Normal kernel could be simplified and restricted to depend on a few parameters as a function of the distance between stations, that is, a spatial covariance function can be used in order to reduce the computing time and allow for spatial prediction.

6. Acknowledgements

The first author was partially funded by Program U-INICIA VID 2011, grant U-INICIA 02/12A University of Chile and Fondecyt grant 11140013. The second author thanks for the support provided by CONACyT project 241195. The third author is supported by the European Research Council (ERC) through StG "N-BNP" 306406.

References

Achcar, J., Fernández-Bremauntz, A., Rodrigues, E. and Tzintzun, G. (2008). Estimating the number of ozone peaks in Mexico city using a non-homogeneous Poisson model , *Environmetrics* **19**: 469–485.

Achcar, J., Rodrigues, E., Paulino, C. and Soares, P. (2010). Non-homogeneous Poisson models with a change-point: an application to ozone peaks in Mexico city , *Environmental and Ecological Statistics* **17**: 521–541.

Arbel, J., Mengersen, K. and Rousseau, J. (2014). Bayesian nonparametric dependent model for the study of diversity for species data, *arXiv:1402.3093v1* .

Barrientos, A., Jara, A. and Quintana, F. (2012). On the support of MacEachern's Dependent Dirichlet Processes and Extensions, *Bayesian Analysis* **7**: 277–310.

Caron, F., Davy, M. and Doucet, A. (2007). Generalized Polya urn for time-varying Dirichlet process mixtures, *Conf. on Uncertainty in Artificial Intelligence*, Vancouver.

Caron, F., Davy, M., Doucet, A., Duflos, E. and Vanheeghe, P. (2008). Bayesian inference for linear dynamic models with Dirichlet process mixtures, *IEEE Trans. Sig. Proc* **56**: 71–84.

Comrie, A. (1997). Comparing neural networks and regression models for ozone forecasting, *Journal of the Air and Waste Managment Association* **47**: 653–663.

Crainiceanu, C., Ruppert, D. and Wand, M. (2005). Bayesian analysis for penalized spline regression using WinBUGS, *Journal of Statistical Software* **14**: 1–24.

Davison, A. and Smith, R. L. (1990). Models for exceedances over high thresholds, *Journal of the Royal Statistical Society, Series B* **52**(3): 393–442.

De Blasi, P., Favaro, S., Lijoi, A., Mena, R., Pruenster, I. and Ruggiero, M. (2015). Are Gibbs type priors the most natural generalization of the Dirichlet process?, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **37**: 212–229.

De Iorio, M., Müller, P., Rosner, G. and MacEachern, S. (2004). An Anova model for dependent random measures, *Journal of the American Statistical Association* **99**(465): 205–215.

De la Cruz, R., Quintana, F. and Müller, P. (2007b). Semiparametric Bayesian Classification with Longitudinal Markers, *Journal of the Royal Statistical Society, Series C* **56**(2): 119–137.

Dockery, D. W., Schwartz, J. and Spengler, J. D. (1992). Air pollution and daily mortality: Associations with particulates and acid aerosols, *Environmental Research* **59**(2): 362 – 373. **URL:** *http://www.sciencedirect.com/science/article/pii/S0013935105800428*

Draghicescu, D. and Ignaccolo, R. (2009). Modeling threshold exceedance probabilities of spatially correlated time series, *Electronic Journal of Statistics* **3**: 149–164.

Dunson, D. (2006). Bayesian dynamic modelling of latent trait distributions, *Biostatistics* **7**: 551–568.

Dunson, D. and Park, J. (2008). Kernel stick-breaking processes, *Biometrika* **95**(2): 307–323.

Dunson, D., Pillai, N. and Park, J. (2007). Bayesian density regression, *Journal of the Royal Statistical Society,Series B* **69**(2): 163–183.

Escobar, M. and West, M. (1995). Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association* **90**(430): 577–588.

Feigin, P. and Tweedie, R. L. (1989). Linear functionals and markov chains associated with dirichlet processes, *Math. Proc. Camb. Phil. Soc.* **105**: 579–585.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems, *The Annals of Statistics* **1**(2): 209–230.

Ferguson, T. S. (1974). Prior distribution on the spaces of probability measures, *Annals of Statistics* **2**: 615–629.

Gelfand, A., Kottas, A. and MacEachern, S. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing, *Journal of the American Statistical Association* **100**(471): 1021–1035.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences, *Statistical Science* **7**: 457–472.

Gramsch, E., Cereceda-Balic, F., Oyola, P. and von Baer, D. (2006). Examination of pollution trends in Santiago de Chile with cluster analysis of PM10 and Ozone data, *Atmospheric Environment* **40**: 5464–5475.

Griffin, J. and Steel, M. (2006). Order-based dependent Dirichlet processes, *Journal of the American Statistical Association* **101**(473): 179–194.

Guardani, R., Aguiar, J., Nascimento, C., Lacava, C. and Yanagi, Y. (2003). Ground-level ozone mapping in large urban areas using multivariate analysis: application to the Sao Paulo metropolitan area, *Journal of the Air and Waste Managment Association* **53**: 553–559.

Guardani, R., Nascimento, C., Guardani, M., Martins, M. and Romano, J. (1999). Study of atmospheric ozone formation by means of a neural networks based model, *Journal of the Air and Waste Managment Association* **49**: 316–323.

Gutiérrez, L. and Quintana, F. (2011). Multivariate Bayesian semiparametric models for authentication of food and beverages, *Annals of Applied Statistics* **5**(4): 2385–2402.

Horowitz, J. (1980). Extremes values from a nonstationary stochastic process: an application to air quality analysis, *Technometrics* **22**: 469–482.

IARC (2013). *Press release no. 221*, International Agency for Research on Cancer, World Health Organization, Geneva, Switzerland.

Lo, A. (1984). On a class of Bayesian nonparametric estimates I. Density estimates., *Ann. Statist.* **12**: 351–357.

MacEachern, S. (1999). Dependent nonparametric processes, *Proc. Bayesian Statistical Science. Amer. Statistic. Assoc., Alexandria, VA.* pp. 50–55.

MacEachern, S. (2000). Dependent Dirichlet processes, *Tech. rep., Department of Statistics, The Ohio State University* .

Mena, R. H. and Ruggiero, M. (2015). Dynamic density estimation with diffusive Dirichlet mixtures, *Bernoulli* . in press.

Ordieres, J., Vergara, E., Capuz, R. and Salazar, R. (2005). Neural network prediction model for fine particulate matter (PM2.5) on the US - Mexico border in El Paso (Texas) and Ciudad Juárez (Chihuahua), *Environmental Modelling and Software* **20**: 547–559.

Pérez, P., Trier, A. and Reyes, J. (2000). Prediction of PM2.5 concentrations several hours in advance using neural networks in Santiago, Chile, *Atmospheric Environment* **34**: 1189–1196.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc, *R News* **6**(1): 7–11.
URL: *http://CRAN.R-project.org/doc/Rnews/*

Préndez, M. (1993). Características de los contaminantes atmosféricos, *in* H. Sandoval, M. Préndez and P. Ulriksen (eds), *Contaminación Atmosférica de Santiago: Estado Actual y Soluciones.*, Cabo de Hornos, pp. 109–186.

Raftery, A. (1989). Are ozone exceedance rate decreasing ? Comment on the paper :Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone by R. L. Smith., *Statistical Science* **4**: 378–381.

Roberts, E. (1979a). Review of statistics extreme values with application to air quality data. Part I. Review, *Journal of the Air Pollution Control Association* **29**: 632–637.

Roberts, E. (1979b). Review of statistics extreme values with application to air quality data. Part II. Applications, *Journal of the Air Pollution Control Association* **29**: 733–740.

Rodriguez, A. and Dunson, D. (2011). Nonparametric Bayesian models through probit stick-breaking processes, *Bayesian Analysis* **6**(1): 145–178.

Rodriguez, A. and ter Horst, E. (2008). Bayesian dynamic density estimation, *Bayesian Analysis* **3**(2): 339–366.

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

Smith, R. (1989). Extreme value analysis of environmental time series: an application to trend detection in ground-level ozone, *Statistical Science* **4**: 367–393.

Walker, S. (2007). Sampling the Dirichlet mixture model with Slices, *Communications in Statistics-Simulation and Computation* **36**: 45–54.

WHO (2011). *Outdoor air pollution in the world cities*, World Health Organization, Geneva, Switzerland.

Appendix

Proof of Proposition 1

We have

$$\begin{aligned}\mathbb{E}(P_t(A)P_{t+s}(A)) &= \mathbb{E}\Big(\sum_{i\geq 1} W_i(t)\delta_{X_i}(A)\sum_{j\geq 1} W_j(t+s)\delta_{X_j}(A)\Big)\\ &= \mathbb{E}\Big(\sum_{i\geq 1} W_i(t)W_i(t+s)\delta_{X_i}(A)\\ &\qquad + \sum_{i\geq 1}\sum_{j\neq i\geq 1} W_i(t)W_j(t+s)\delta_{X_i}(A)\delta_{X_j}(A)\Big)\\ &= k_s F_0(A) + (1-k_s)F_0^2(A) = k_s F_0(A) + (1-k_s)F_0^2(A)\end{aligned}$$

where

$$k_s = \mathbb{E}\Big(\sum_{i\geq 1} W_i(t)W_i(t+s)\Big) = \sum_{i\geq 1}\mathbb{E}(W_i(t)W_i(t+s)) \tag{21}$$

and $k_0 = \sum_{i\geq 1}\mathbb{E}(W_i^2(t))$. Here k_s is independent of t by stationarity and $1-k_s$ is obtained by subtraction, since

$$1 = \sum_{i\geq 1} W_i(t)\sum_{j\geq 1} W_j(t+s) = \sum_{i\geq 1} W_i(t)W_i(t+s) + \sum_{i\geq 1}\sum_{j\neq i\geq 1} W_i(t)W_j(t+s).$$

Hence

$$\mathrm{Cov}(P_t(A), P_{t+s}(A)) = k_s F_0(A)(1-F_0(A)). \tag{22}$$

Using the independence among the $V_i(\cdot)$'s, we can write

$$\begin{aligned}\mathbb{E}(W_i(t)W_i(t+s)) &= \mathbb{E}\Big[\Big(V_i(t)\prod_{j<i}(1-V_j(t))\Big)\Big(V_i(t+s)\prod_{j<i}(1-V_j(t+s))\Big)\Big]\\ &= \mathbb{E}[V_i(t)V_i(t+s)]\prod_{j<i} E\Big((1-V_j(t))(1-V_j(t+s))\Big)\\ &= \mathbb{E}[V_i(t)V_i(t+s)]\prod_{j<i}\Big(\frac{M-1}{M+1} + \mathbb{E}[V_j(t)V_j(t+s)]\Big).\end{aligned}$$

Since the number of atoms renewals follows a Bernoulli process with parameter ϕ, we have

$$\mathbb{E}[V_j(t)V_j(t+s)] = (1-\phi)^s \frac{2}{(1+M)(2+M)} + (1-(1-\phi)^s)\frac{1}{(1+M)^2}$$

from which we find

$$k_s = \frac{2+M+(1-\phi)^s M}{(2+M)(1+2M)-(1-\phi)^s M}.$$

Setting s to be zero in (22), we have that

$$\mathrm{Var}[P_t(A)] = \frac{F_0(A)(1-F_0(A))}{1+M}$$

as expected, since P_t is marginally a Dirichlet process, and

$$\mathrm{Corr}(P_t(A), P_{t+s}(A)) = \frac{(1+M)(2+M+(1-\phi)^s M)}{(2+M)(1+2M)-(1-\phi)^s M}$$

from which $\mathrm{Corr}(P_t(A), P_{t+s}(A)) \to (1+M)/(1+2M)$ as $s \to \infty$.