UNDERSTANDING HEALTH EFFECTS OF PM SOURCES BY USING STOCHASTIC MACHINE LEARNING MODELS CONFERENCE SUBMISSIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Particulate matter (PM) is a complex mix of organic and inorganic compounds of distinct sources, with a range of physical and chemical properties, which might have a different harmful effects to health. Disentangling total ambient PM concentration into its sources is key for developing strategies to reduce PM through targeted actions. Current methods to identify sources of particulate pollution typically require *a priori* specification of the number of sources and do not include information on covariates in the source allocations. In this work, we develop a comprehensive approach for source apportionment of airborne particles by using machine learning probabilistic models. We proposed a Bayesian nonparametric approach through a Dirichlet process mixture models that enables the better understanding of hidden structures in multi-pollutants and allow to accommodate complex patterns of temporal dependencies as well as for concomitant processes (e.g. meteorology) in the prediction of the source contributions. Then we evaluate the health effects of the sources. To illustrate our model framework, we applied it to the PM_{10} chemical composition data measured at an urban background site (North Kensington) in London, UK, from 2011 to 2012. The health data will be related to cardio-respiratory hospital admission.

1 INTRODUCTION

Atmospheric particulate matter (PM) is a complex mixture of chemically and physically diverse substances, from anthropogenic and natural sources. These source contributions, in combination with other factors, such as meteorological conditions and chemical transformations, determine the pollution concentration in air and the variation in the physio-chemical compounds across space and time. It is hypothesized that mixture of particles from different sources can have different toxicity and health effects (e.g., Dai et al., 2014; Park et al., 2014; Pirani et al., 2015; Samoli et al., 2016). This makes source apportionment (SA) of pollutants all the more important in order for air quality managers to fully understand the potential health outcomes of pollutant mixtures. SA aims at deriving information about ambient air pollution sources based on data registered at monitoring sites, therefore allowing to quantify how much each individual source contributes to the pollution concentrations in the air (Krall & Chang, 2019).

Traditionally methods for SA problem are dominates by two approaches (Viana et al., 2008): source-oriented deterministic models and receptor models. The former relies on the knowledge of emissions and physical and chemical processes of dispersion to predict air quality; the latter is based on statistical procedures for identifying and quantifying the sources of pollutants on the basis of mixture of chemicals measured at receptor sites. Within receptor model approaches, most techniques share a strong assumption that the source contributions or pollution mixtures are independent over time. However, this may be not appropriate and temporal dependence can exist, driven by meteorology or other time-varying factors. Therefore, approaches have been proposed for identifying distinct air pollutant mixtures, mainly framed in clustering-based solutions, leading to the characterization of groups of time points with similar pollutant concentration profiles, while suggesting major source contributions based on the presence of specific markers (e.g., Adame et al., 2012; Austin et al., 2012; Zanobetti et al., 2014; Pirani et al., 2015; Alahamade et al., 2020; Chen et al., 2015; Bellinger et al., 2017; Bousiotis et al., 2021). Krall & Chang (2019) suggest as possible solution to the issue, con-

sisting in the incorporation into the model of temporally-varying variables, such as meteorological factors.
In this conference, my talk will focus on a model-based clustering approach, framed in a nonparametric Bayesian perspective, a Dirichlet process (DP; (Ferguson, 1973)) over the latent generating function, encoding measurement error, and temporal dependencies via the inclusion of wind field (i.e., wind speed and direction). Here, we refer to the class of dependent DP (DDP; MacEachern (1999; 2000)), recently reviewed by Quintana et al. (2020), and we focus on the kernel stick breaking process (KSBP), which was firstly formulated by Dunson & Park (2008) and used in spatial setting by Reich & Fuentes (2007).
We apply our proposed approach for apportioning compositional metrics of particulate with aerodynamic diameter smaller than $10\mu m$ (PM_{10}) in Greater London's urban atmosphere for the period 2011-2012. This dataset was previously apportioned by Beddows et al. (2015) using PMF tool (i.e., PMF-5) provided by the Environmental Protection Agency (US). Beddows and colleagues' study, however, failed to resolve a number of diffuse urban sources including some aspects of the traffic mix and domestic wood burning. To fill the methodological gaps in source characterisation and health effect evaluation, We will explore joint models, which would allow to extend the SA model to evaluate the link between sources and health outcomes.

2 METHODS

2.1 DATA

We use daily (24 hours) filter samples of PM_{10} composition taken over a two-year period, 2011-2012, collected at a monitoring station (North Kensington) in central London, UK. A total of 23 metrics were recorded, covering water-soluble inorganic ions such as sulfate (SO_4^{2-}), nitrate (NO_3^-), chloride (CI^- and ammonium (NH_4^+), elemental and black carbon (EC and BC, respectively), wood burning (WOD), metals or trace elements (Al, Ba, Ca, Cu, Fe, K, Mn, Mo, Na, Ni, Pb, Sb, Ti, V, Zn, Mg). Wind data were imported from the R package `openair` (Carslaw & Ropkins, 2012). The health data related to cardio-respiratory hospital admission are available from the Hospital Episode Statistics registry within the Small Area Health Statistics Unit (SAHSU) at Imperial College London.

2.2 MODEL

In this section we briefly introduce the popular DP mixture model (Neal, 2000), then we describe the extension to the DDP (Quintana et al., 2020) as specified by a kernel stick-breaking process (Dunson & Park, 2008; Reich & Fuentes, 2007). Successively, we show how to apply it for clustering time-series of outdoor chemical species of particles into source profiles, while accounting for temporally-resolved meteorological conditions such as wind field.

2.2.1 GENERAL CHARACTERIZATION OF THE KERNEL DIRICHLET PROCESS

We start with some notation. Consider a set of n observations made at T time points, $\boldsymbol{X}_t = \{X_{ti} : t = 1, 2, \dots, T\}$, $i = 1, \dots, n$, being independently drown from some unknown distribution, modelled as a mixture of distributions of the form $f(\boldsymbol{\theta}_t)$, so that $\boldsymbol{X}_t | \boldsymbol{\theta}_t \sim f(\boldsymbol{\theta}_t)$.

Also, let F be an unknown mixing distribution over the unknown parameters, so that $\boldsymbol{\theta}_t | F \overset{iid}{\sim} F$ (where iid stands for independent and identically distributed). The prior for F is taken to be a DP with non-negative precision (or concentration) parameter α and a continuous distribution function F_0 called as the base (or centering) measure, denoted as $F|\alpha, F \sim \text{DP}(\alpha F_0)$. The random measure F is discrete with probability one, and this is made explicit by the stick-breaking representation of the DP provided by Sethuraman (1994):

$$
\begin{gathered}
F = \sum_{j=1}^{\infty} w_j \delta(\boldsymbol{\theta}_j^*), \\
w_1 = V_1 \;\text{ for }\; j = 1 \;\text{ and }\; w_j = V_j \prod_{k=1}^{j-1} (1 - V_k), \;\text{ for }\; j \geq 2, \\
V_j \overset{iid}{\sim} \text{Beta}(1, \alpha), \;\; \boldsymbol{\theta}_j^* \overset{iid}{\sim} F_0.
\end{gathered} \tag{1}
$$

where $\delta(\theta^*)$ denotes a Dirac measure (point mass) at θ^*. For each component j there is a probability weight w_j. These mixing weights $w_1, w_2, \dots$, and they satisfies $w_j \in (0,1)$ and $\sum_{j=1}^{\infty} = 1$. The name of this construction derives by an analogy given by breaking pieces off from a stick of unit length, where the breakpoints $(V_1, V_2, \dots)$ are randomly sampled from a Beta distribution. The mixture probabilities break the stick into a potentially infinite number of pieces, such that they sum to the unity. For this representation, the base measure F_0 defines the mean of the process, since $E(F) = F_0$ and it can be considered as the prior guess (Antoniak, 1974); while the α parameter can be understood as an inverse variance (Teh, 2010), so the larger α is, the smaller the variance (broadly speaking, α controls the number of components of the mixture).

Although flexible, this mixture modelling approach does not explicitly exploits the order information contained in the data, and assumes that the observations are exchangeable. This assumption usually leads to a easy tractable model form for the posterior computation, however, it may degrade the clustering performance in time-series studies as the temporal order of the data is not accounted for. The same problem applies for data characterised by spatial dependence. One solution suggested in literature to address this issue, is given by a DDP (MacEachern, 1999), which general formulation allows the weights and/or the locations in the stick-breaking construction of the DP in (1) to depend on covariates.

We consider wind speed and direction as our main covariates. The covariates influence only the first $K-1$ mixing weights as the K^{th} is set to take on the remaining probability up to 1. Therefore, the structure in (1) is modified as follows:

$$
\begin{aligned}
F_t &= \sum_{k=1}^{\infty} w_{k,t} \delta_{\boldsymbol{\theta}_k} \\
w_{k,t} &= v_{k,t} \prod_{l<k} (1 - v_{l,t}), \quad k = 2, \dots, K \\
v_{k,t} =& w_{k,t} \cdot \eta_{k,t}, \quad w_{1,t} = v_{1,t} \quad v_{K,t} = 1 \\
\eta_{k,t} &\sim \text{Beta}(1, \alpha), \quad k = 1, \dots, K-1 \\
\boldsymbol{\theta}_k &\overset{iid}{\sim} F_0
\end{aligned}
\tag{2}
$$

where $w_{k,t}$ is a kernel that incorporates weights dependent on wind speed and wind direction. When $w_{k,t} = 1$ the model becomes a standard time varying DP without time dependency. For computational feasibility, we implement the model with finite approximation to the infinite stick-breaking process (2), that is $k = 1, 2, \dots, K$.

2.2.2 Characterization of particle source profiles via wind Kernel Stick-Breaking Dirichlet Process(KSBDP)

Pollutant model: A Gaussian Mixture model

Under this setting for each cluster k, the cluster specific parameters are given by Θ_k. We consider X_{pt} as a measurement of component p ($p = 1, \dots, P$) on day $t (t = 1, \dots, n)$. We assumed a multivariate normal distribution for the P covariates:

$$
\begin{aligned}
p(X_t | z_t, \Theta_k, \Theta_0) = f(X_t | m_{z_t}, \Theta_k) = (2\pi)^{-\frac{P}{2}} |\Sigma_{z_t}|^{-\frac{1}{2}} exp\big[-\frac{1}{2}(X_t - m_{z_t})^T \Sigma_k^{-1} (X_t - m_{z_t})\big], \\
z_t \sim Discrete(w_{t,k})
\end{aligned}
\tag{3}
$$

where $X_t = (X_{t1}, \dots, X_{tP})^T$ represents a daily covariate profile of air pollutants, the cluster specific parameters are given by $\Theta_k = (m_k, \Sigma_k)$, where m_k is a mean vector and Σ_k is a covariance matrix. There are no additional global parameters Θ_0. We choose $m_k \sim N(m_0, \Sigma_0)$ and $\Sigma_k \sim InvWishart(R_0, \kappa_0)$ (for each k) for our prior.

Health model: Count response

Let y_t be the observed number of health outcome for the day t, and is is Poisson distributed with mean λ_t. Following (Pirani et al., 2015) , we define $u_t = (u_{t1}, \dots, u_{tH})^T$ to be a B-spline basis matrix for natural cubic splines of temperature, relative humidity and calendar time, we have

$$
\begin{aligned}
&y_t \sim Pois(\lambda_t), \quad \lambda_t = E_t exp\{\mu_t\} \\
&\mu_t = \beta_0 + \phi_k + \sum_{h=1}^{H} f_h(u_{th}, df_h) + Dummy.DayWeek
\end{aligned}
\tag{4}
$$

The parameter ϕ_k represents the log relative risk for the outcome of interest associated with the kth cluster, were each cluster includes days with similar multipollutant profile. The functions $f(., dfh)$ denote smooth functions of confounding factors, with smoothing parameters dfh. We use this kernel to incorporate wind speed in m/s ($\{ws_t\}$) and wind direction in degrees clockwise from north ($\{wd_t\}$) both aggregated at the same daily time steps as the observed data. Explicitly,

$$
w_{k,t} \quad = \quad \exp\left(\frac{\left(\nu_{1,k} - ws_t\right)^2}{2 \cdot h_{1,k}}\right) \cdot \exp\left(\frac{\sin\left((\nu_{2,k} - wd_t) \cdot \pi/360\right)^2}{2 \cdot h_{2,k}}\right), \tag{5}
$$

where $\nu_{1,k}$ and $\nu_{2,k}$ are the kernel modes (knots) for the two wind-related covariates, and $h_{1,k}$, $h_{2,k}$ the corresponding bandwidths, which control the spread for source k both in angular and speed direction.

3 Preliminary Results

Six cluster solutions are derived from the SBDP model, that represent corresponding aerosol types (see Fig.1). Our preliminary results are in line with Beddows et al. (2015), which used positive matrix factorization for apportioning PM in London. The study of Beddows et al. (2015), however, failed to resolve a number of diffuse urban sources including some aspects linked to fireworks.

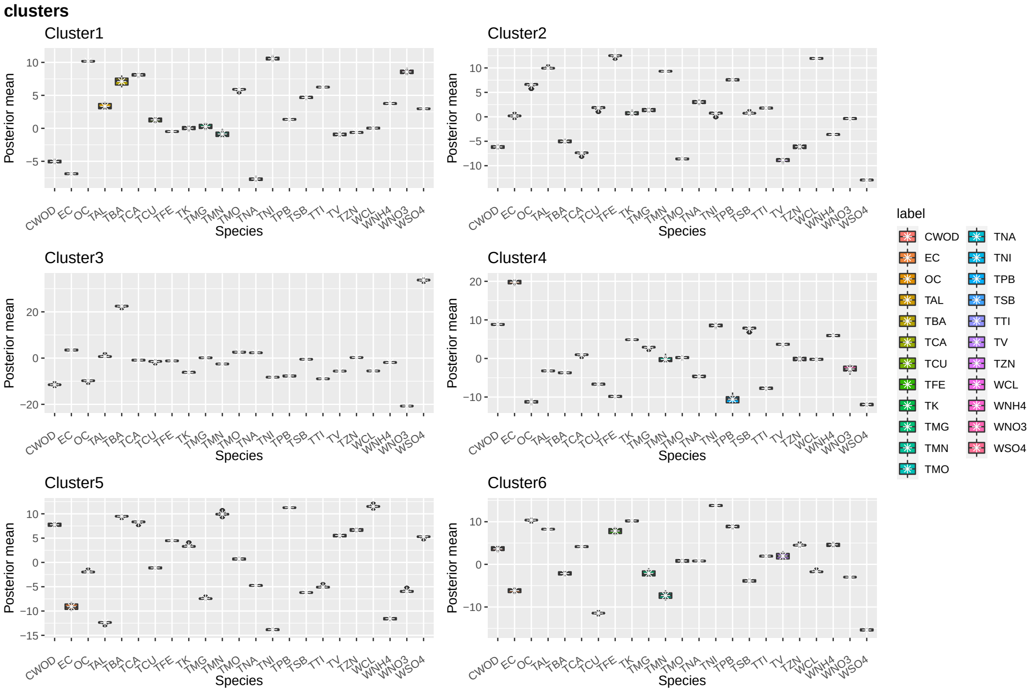

Figure 1: Clusters outputted from the kernel stick-breaking process run on PM_{10} components

Our analysis on PM compositional data using the SBDP shows a better characterization of this diffuse urban profile defined by contributions from both wood-smoke and road traffic, since these are ground-level sources which are affected by meteorology. The link with health outcomes is currently on going.

REFERENCES

JA Adame, A Notario, F Villanueva, and J Albaladejo. Application of cluster analysis to surface ozone, no2 and so2 daily patterns in an industrial area in central-southern spain measured with a doas system. *Science of the Total Environment*, 429:281–291, 2012.

Wedad Alahamade, Iain Lake, Claire E Reeves, and Beatriz De La Iglesia. Clustering imputation for air pollution data. In *International Conference on Hybrid Artificial Intelligence Systems*, pp. 585–597. Springer, 2020.

Charles E Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pp. 1152–1174, 1974.

Elena Austin, Brent Coull, Dylan Thomas, and Petros Koutrakis. A framework for identifying distinct multipollutant profiles in air pollution data. *Environment international*, 45:112–121, 2012.

D. C. S. Beddows, R. M. Harrison, D. C. Green, and G. W. Fuller. Receptor modelling of both particle composition and size distribution from a background site in london, uk. *Atmospheric Chemistry and Physics*, 15(17):10107–10125, 2015.

Colin Bellinger, Mohomed Shazan Mohomed Jabbar, Osmar Zaïane, and Alvaro Osornio-Vargas. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health*, 17(1):1–19, 2017.

Dimitrios Bousiotis, Ajit Singh, Molly Haugen, David Beddows, Sebastián Diez, Killian L Murphy, Pete M Edwards, Adam Boies, Roy M Harrison, and Francis D Pope. Assessing the sources of particles at an urban background site using both regulatory instruments and low-cost sensors–a comparative study. *Atmospheric Measurement Techniques*, 14(6):4139–4155, 2021.

David C. Carslaw and Karl Ropkins. openair - an r package for air quality data analysis. *Environmental Modelling & Software*, 27–28(0):52–61, 2012.

Mei Chen, Pengfei Wang, Qiang Chen, Jiadong Wu, and Xiaoyun Chen. A clustering algorithm for sample data based on environmental pollution characteristics. *Atmospheric Environment*, 107: 194–203, 2015.

L Dai, A Zanobetti, P Koutrakis, and J Schwartz. Associations of fine particulate matter species with mortality in the united states: A multicity time-series analysis. *Environmental Health Perspectives*, 122:837–842, 2014.

David B. Dunson and Ju-Hyun Park. Kernel stick-breaking processes. *Biometrika*, 95:307–323, 2008.

T. Ferguson. A Bayesian analysis of some non-parametric problems. *The Annals of Statistics*, 1: 209–230, 1973.

J. Krall and H. Chang. *Statistical methods for source apportionment*, pp. 523–546. Chapman and Hall/CRC, 2019.

S. N. MacEachern. *Dependent nonparametric processes*. American Statistical Association, Alexandria, VA, 1999.

S. N. MacEachern. *Dependent Dirichlet processes Technical Report*. Department of Statistics, The Ohio State University, 2000.

Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Eun Sug Park, Philip K Hopke, Man-Suk Oh, Elaine Symanski, Daikwon Han, and Clifford H Spiegelman. Assessment of source-specific health effects associated with an unknown number of major sources of multiple air pollutants: a unified bayesian approach. *Biostatistics*, 15(3): 484–497, 2014.

M. Pirani, N. Best, M. Blangiardo, S. Liverani, R. W. Atkinson, and G. W. Fuller. Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environment International*, 79:56–64, 2015.

Fernand A Quintana, Peter Mueller, Alejandro Jara, and Steven N MacEachern. The dependent dirichlet process and related models. *Statistical Science*, 2020.

B. J. Reich and M. Fuentes. A multivariate semiparametric bayesian spatial modeling framework for hurricane surface wind fields. *Annals of Applied Statistics*, 1:249–264, 2007.

E Samoli, R Atkinson, A Analitis, G Fuller, D Green, I Mudway, and H et al. Anderson. Associations of short-term exposure to traffic-related air pollution with cardiovascular and respiratory hospital admissions in london, uk. *Occupational and Environmental Medicine*, 75(3):300–307, 2016.

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

Y. W. Teh. Dirichlet processes. In *Encyclopedia of Machine Learning*. Springer, 2010.

M. Viana, T.A.J. Kuhlbusch, X. Querol, A. Alastuey, R.M. Harrison, P.K. Hopke, W. Winiwarter, M. Vallius, S. Szidat, A.S.H. Prevot, C. Hueglin, H. Bloemen, P. Wahlin, R. Vecchi, A.I. Miranda, A. Kasper-Giebl, W. Maenhaut, and R. Hitzenberger. Source apportionment of particulate matter in Europe: a review of methods and results. *Journal of Aerosol Science*, 39:827–849, 2008.

Antonella Zanobetti, Elena Austin, Brent A Coull, Joel Schwartz, and Petros Koutrakis. Health effects of multi-pollutant profiles. *Environment international*, 71:13–19, 2014.