Bayesian Modelling for Estimating Adverse Health Effects of Exposure to Multiple Air Pollutants in a Time Series Framework

Monica Pirani[1,4], Georgios Papageorgiou[2], Nicky Best[2],
Richard W. Atkinson[3], and Gary W. Fuller[1]
[1] King's College London, London, UK
[2] Imperial College London, London, UK
[3] St. George's, University of London, London, UK
[4] Corresponding author: Monica Pirani, e-mail: monica.pirani@kcl.ac.uk

Abstract

Polluted air contains a complex mixture of particles with a range of physical and chemical properties, gases, bioareosols and toxic substances. Estimation of how simultaneous exposure to multi-pollutants affects the risk of adverse health response represents a challenge for scientific research and air quality management. In this work we consider the problem of modelling these multi-pollutant systems within the framework of time series studies. We propose a Bayesian approach, using a Dirichlet process mixture, defined by a stick-breaking construction, to cluster time points with similar multi-pollutant and response profiles. Inference is carried out via Markov Chain Monte Carlo methods. The applicability of our approach is evaluated in a real data set which comprises daily time series of a range of air pollutants, meteorological variables and cardiovascular mortality counts for London (UK) during the years 2002-2005.

Keywords: Ambient air pollution, Bayesian model-based clustering, Cardiovascular mortality, Dirichlet process mixture model.

1. Introduction

The current scientific evidence of association between air pollution and short-term adverse endpoints, derives largely from observational ecological time series studies (e.g., Bell et al. 2004; HEI 2010; and references therein). Since the early 1990s this evidence has been playing an important role in setting standards for acceptable levels of ambient pollution. Within this study design, the quantification of impact of polluted air on the public health has been historically undertaken through a single pollutant approach, using regression-based techniques (mainly generalized linear and additive models) to examine the association between a pollutant and a health endpoint, where the co-pollutants have been treated as modifying or confounding factors. This reliance on single pollutant results is due, in part, to measurement and source complexities (such as the correlation between pollutants) which have limited the development of statistically robust multi-pollutant models, and in part to the regulatory strategies of air quality management which have addressed a single pollutant at time (Dominici et al. 2003).
Air pollution exists, however, as a complex mixture of particles with a range of physical and chemical properties, gases, bioareosols and myriad of volatile organic compounds mixed with biological material. This airborne mixture reacts dynamically with sunlight and is modified by meteorological conditions.
As alternative technique to classical regression models, we propose to analyse the impact of these pollutant systems on population health using a Bayesian nonparametric

mixture model that simultaneously clusters days according to their multiple pollutant profiles (i.e. a sequence of measurements for each predictor on the day) and investigates the relationship between these clusters and adverse health outcomes. Our approach builds on recent work of Molitor et al. (2010). Generally speaking, model-based clustering methods are based on the idea that the data are clustered using some assumed mixture modeling structures (Fraley and Raftery 2002; McLachlan and Baek 2010). Recently, Frühwirth-Schnatter and Kaufmann (2008) showed that model-based clustering based on finite mixture models (McLachlan and Peel 2000) can be extended to time series in a quite natural way. In contrast to finite mixture model, a nonparametric approach to mixture model avoids the constraint to a pre-determined number of clusters by assuming potentially infinite mixtures, described by unknown probability distributions, which are drawn themselves from prior distributions. The Dirichlet process (Ferguson 1973), is one such prior for unknown mixing probability distributions and was used by Molitor et al. (2010) to perform Bayesian clustering in an approach termed *profile regression*.
We extended this technique to analyse data that are time-evolving. To asses the applicability of our methodology in air pollution epidemiology, we used a database from Atkinson et al. (2010) and investigated the effects of different particle metrics and meteorological variables on cardiovascular mortality in London for the years 2002-2005.

2. Profile-based Bayesian model

Denote, for day t, $t = 1, \dots, T$, a covariate profile of pollutants and meteorological variables as $\mathbf{x}_t = (x_t^1, x_t^2, \dots, x_t^C)'$, where each covariate, x_t^c, $c = 1, \dots, C$, within each profile, denotes a measure of exposure for covariate c on day t, and let y_t denote a count number for health events on day t. A latent allocation variable z_t is introduced to indicate which cluster the day t belongs, i.e. $z_t = h$. It is also assumed that π_h indicates the probability of assignment to cluster h, that is $p(z_t = h) = \pi_h$, where the weights π_h are nonnegative and sum to one.
We assume the T days characterised by a pattern of exposure values and linked to the outcome, arise from a potentially infinite number of clusters, generated by a mixture of underlying probability distributions. In our approach the components are represented by parametric exponential family densities that are combined in a mixture model

$$f_{\mathbf{\Theta}}(y_t, \mathbf{x}_{t-l}) = \sum_{h=1}^{\infty} \pi_h f_{\mathbf{\Theta}}(y_t, \mathbf{x}_{t-l} | z_t = h) \tag{1}$$

where l represents a lag of l days and $\mathbf{\Theta}$ is used to denote collectively the model parameters. The weights, π_h, are generated using a stick-breaking procedure (Sethuraman 1994), based on i.i.d. Beta distributions, Beta$(1, \alpha)$, draws, $\{V_1 : i = 1, 2, \dots\}$. More precisely, the first weight is equal to V_1 and, for $h \geq 2$ the $h-$th weight is given by $V_h \prod_{i=1}^{h-1}(1 - V_i)$. α represents the concentration parameter that controls the clustering. We assumed that y_t has a Poisson distribution

$$y_t \sim \text{Poisson}(\mu_t) \tag{2}$$

where $\mu_1 = \mu_2 = \mu_h$ if $y_1, y_2 \in h$-th cluster. Conditionally to the membership in the $h-$th component of the mixture, the model for the mortality events is

$$\ln(\mu_t) = \mu_h + \ln(\text{offset}) + \epsilon_t \tag{3}$$

with $\epsilon_t \sim (0, \sigma^2)$. Finally, we assume for daily time series of pollutants and meteoro-

logical variables a mixture of Normals

$$\mathbf{x}_t | z_t = h \sim \text{Normal}(\mathbf{m}_h, \mathbf{\Sigma}_h) \tag{4}$$

where $\mathbf{m}_h = (m_h^1, \ldots, m_h^C)$ and $\mathbf{\Sigma}_h$ are the mean and covariance matrix of the Normal distribution for cluster h. Therefore, the days with similar risk and covariate profile are clustered together.

The computations are performed via Markov Chain Monte Carlo methods (MCMC), using an algorithm implemented in an open source C++ code, wrapped in the R package PReMiuM (Liverani et al. 2013).

The output coming from this modelling approach is very rich, as it allows the number of clusters to change from iteration to iteration of MCMC sampler. Thus, the challenge is to find the best way in which the algorithm groups daily profiles into clusters and then process this optimal partition using Bayesian model-averaging techniques. As in Molitor et al. (2010) and Liverani et al. (2013), the starting point is to construct, at each iteration of the sampler, a score matrix with $(i, j)-$th element of the matrix set equal to 1 if day i and day j belong to the same cluster and 0 otherwise. The end of this process leads to a probability matrix, $\mathbf{S}$, formed by averaging the score matrices obtained at each iteration, thus element S_{ij} denotes the probability that day i and j are assigned to the same cluster. The task is to find a partition that best represents $\mathbf{S}$. This is achieved using the partitioning around medoids (Molitor et al. 2010) on the dissimilarity matrix $1 - \mathbf{S}$. Once the optimal partitioning is defined, a model averaging approach is adopted to evaluate the uncertainty related to the characteristics of the groups found. This involves running through the MCMC output, obtaining, at each iteration, an average value for the model parameters across all days in a certain group. This model averaging approach produces narrower credible intervals for cluster parameters when the clustering is consistent from iteration to iteration.

A sensitivity analysis, setting different start points on the number of clusters was performed. The diagnosis showed consistency of the results.

3. Data

We selected a subset of four years data (2002-2005) from the database used by Atkinson et al. (2010) that comprises mortality counts for cardiovascular diseases, weather, and air pollutant concentrations for London (UK). For additional details, see Atkinson et al. (2010), but briefly this comprised:

Pollutant data. Particle metrics included: particle number concentration (PNC), inorganic anions (chloride (CI^-), nitrate (NO_3^-) and sulphate (SO_4^{2-})), black smoke (BS) and gravimetric measurements of particulate matter (PM) such as PM_{10}, $PM_{2.5}$, PM coarse fraction (i.e., $PM_{10-2.5}$ obtained by subtraction). We also included apportioned PM into primary and nonprimary sources, obtained using an apportionment model (Fuller and Green 2006), giving modelled primary PM_{10} (PPM_{10}), nonprimary PM_{10} ($NPPM_{10}$), nonprimary $PM_{2.5}$ ($NPPM_{2.5}$), and nonprimary PM coarse fraction. With the exception of blacke smoke, these daily concentrations were obtained from a single background monitoring station in central London. Black smoke was an average across several urban and suburban stations.

Meteorological data. Daily average temperature and dew point temperature (as proxy of humidity), measured in central London were obtained from the British Atmospheric Data Centre.

Mortality data. Counts mortality data for cardiovascular diseases were obtained from the Office for National Statistics and coded using the International Classification of Diseases, 10th Revision (ICD$-$10: Chapter I).

4. Results

We present the preliminary results obtained using pollution data and meteorological variables one day lagged (i.e., $l = 1$). Table 1 contains the summary statistics for the cardiovascular mortality, environmental and meteorological data in London during the study period.

		Range		Percentiles		
Variable	% Missing	Min	Max	25th	50th	75th
Deaths	0	27	116	47	53	60
PNC (n/cm^3)	33.06	5543.00	52443.00	14277.00	18944.50	24582.00
NO_3^- ($\mu g/m^3$)	22.52	0.03	30.89	1.35	2.44	4.48
CI^- ($\mu g/m^3$)	22.31	0.01	9.06	0.25	0.88	1.98
SO_4^{2-} ($\mu g/m^3$)	22.18	0.23	20.63	1.51	2.25	3.89
BS ($\mu g/m^3$)	0	1.40	31.33	4.00	5.40	7.06
PM_{10} ($\mu g/m^3$)	15.95	5.00	119.00	17.00	23.00	32.00
$PM_{2.5}$ ($\mu g/m^3$)	9.99	1.00	104.00	11.00	15.00	22.00
Coarse ($\mu g/m^3$)	18.41	-5.00	33.00	5.00	7.00	10.00
PPM_{10} ($\mu g/m^3$)	0	0.80	39.10	2.50	3.70	5.60
$NPPM_{10}$ ($\mu g/m^3$)	0	-1.00	61.00	7.00	9.90	14.20
$NPPM_{2.5}$ ($\mu g/m^3$)	0	-3.50	32.60	2.30	4.00	7.20
NP coarse ($\mu g/m^3$)	0	-1.00	42.20	4.00	5.60	7.40
Temperature (°C)	0	-0.10	29.30	8.60	12.20	16.70
Dew point temp. (°C)	0	-6.50	18.00	3.50	7.00	10.10

Table 1: Descriptive statistics of daily cardiovascular mortality, pollutants and meteorological variables. London, 2002-2005.

The analysis of correlations between pairs of pollutants, showed different degrees of interdependence (data not showed). The results in terms of optimal clustering produced the compartment of days in four main clusters. Figure 1 displays graphically the posterior mean estimates, with the 95% credible intervals (CI) by cluster, sorted by increasing mortality risk. The first two clusters showed low risk for cardiovascular mortality, with 95% CI including 1. The last two clusters, related to 129 and 6 days respectively, presented risk for cardiovascular mortality respectively equal to 1.02 (95% CI: 1.00-1.06) and 1.59 (95% CI: 1.38-1.82). The first of these two high risk clusters is mainly related to high levels of PNC, black smoke, primary PM and cold days (low levels of temperature and dew point temperature). This suggested a cluster based on primary pollutant emissions. The second cluster is characterised by high concentrations of PM_{10}, $PM_{2.5}$, modelled nonprimary PM, with high temperature and dew point temperature. It is related to a summer heat-wave event that London experienced in the year 2003.

5. Discussion

In this paper, we have proposed a new Bayesian clustering approach to investigate a combined short-term effect of multiple pollutants on mortality, using a mixture model to cluster time points based on their multi-pollutant profiles. The preliminary results show consistency with the findings in Atkinson et al. (2010), where using an univariate Poisson log-linear time series model, the authors found a positive association between PNC and mortality for cardiovascular diseases lagged 1-day. In comparison to classical regression models, however, Bayesian profile regression represents a technique that overcomes the issue of multicollinearity, allowing an estimate of the simultaneous effect of multiple pollutants.

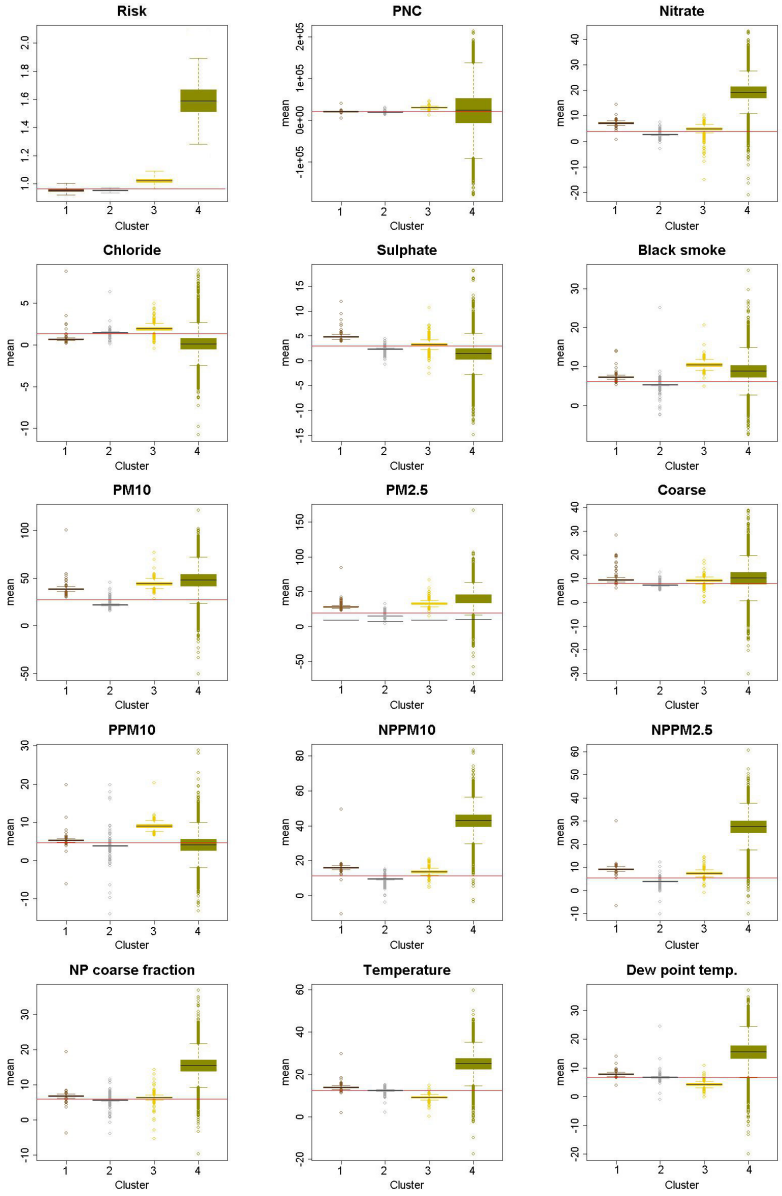

Figure 1: Box plots showing the posterior estimates: risk for cardiovascular mortality with 95% credible intervals (CI), and posterior mean with 95% CI of each covariate for the 4 clusters that form the optimal clustering.

The idea of using clustering technique to identify groups of days with distinct profiles of pollutants has also been recently presented by Austin et al. (2012). The authors used not model-based clustering techniques (*k*-means partitioning and hierarchical clustering) for clustering daily measurements of gaseous and particulate. Our proposed model also generates a covariate multi-pollutant profile for each day, however, it presents a number of advantages in comparison to traditional clustering techniques. Firstly, our approach is model-based, thus the probability distribution of the data is approximated by a statistical model. Moreover, it discovers clusters in data and does not require the user to know the number of clusters a priori. Finally, it provides a way to handle missing data.

Our use of the Dirichlet process mixture to model time series data is still ongoing, particularly with respect to: (i) the combination of the Dirichlet process mixture with smooth functions to account for aspects associated with time variation such as trend

and seasonality; (ii) the production of predicted values (using the probability of allocation instead of performing a random allocation); (iii) the understanding of the different harmful effect of pollutants, using technique such as variable selection on the covariates.

Acknowledgements

We wish to thank Silvia Liverani for her help with the R code and for her comments on model. The study is supported by the MRC-HPA Centre for Environment and Health.

References

Atkinson, R. W., Fuller, G. W., Anderson, R. H., Harrison, R. M., and Armstrong, B. (2010), "Urban ambient particle metrics and health: a time-series analysis," *Epidemiology*, 21, 501–511.

Austin, E., Coull, B., Thomas, D., and Koutrakis, P. (2012), "A framework for identifying distinct multipollutant profiles in air pollution data," *Environment International*, 45, 112–121.

Bell, M. L., Samet, J. M., and Dominici, F. (2004), "Time-series studies of particulate matter," *Annual Review of Public Health*, 25, 247–280.

Dominici, F., Sheppard, L., and Clyde, M. (2003), "Health effects of air pollution: a statistical review," *International Statistical Review*, 71, 243–276.

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

Fraley, C. and Raftery, A. E. (2002), "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, 97, 611–631.

Frühwirth-Schnatter, S. and Kaufmann, S. (2008), "Model-based clustering of multiple time series," *Journal of Business & Economic Statistics*, 26, 78–89.

Fuller, G. W. and Green, D. (2006), "Evidence for increasing primary PM_{10} in London," *Atmosferic Environment*, 40, 6134–6145.

HEI (2010), "Traffic-related air pollution. A critical review of the literature on emission, exposure, and health effects," Tech. Rep. 17, Health Effect Institute Special Report, Boston.

Liverani, S., Hastie, D., Papathomas, M., and Richardson, S. (2013), "PReMiuM: an R package for profile regression mixture models using Dirichlet processes," *Submitted*, avalable at http://uk.arxiv.org.

McLachlan, G. J. and Baek, J. (2010), *Clustering of high-dimensional data via finite mixture models*, Springer-Verlag, pp. 33–44.

McLachlan, G. J. and Peel, D. (2000), *Finite Mixture Models*, John Wiley & Sons, Inc.

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010), "Bayesian profile regression with an application to the National Survey of Children's Health," *Biostatistics*, 11, 484–498.

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