A spatio-temporal model for particulate matter monitored from a heterogeneous network

Sujit K. Sahu[1], Orietta Nicolis[2]

[1] School of Mathematics, University of Southampton, UK
[2] Dept. IGI, University of Bergamo, Italy

Abstract: Modeling atmospheric particulate matter (PM) is an important problem since PM has been linked to many respiratory and cardiovascular diseases. An even more interesting and challenging task is to model PM concentrations monitored on a heterogeneous network over different seasons. In this paper we analyze a data set of daily PM_{10} concentrations in north Italy for four months of 2003. The data set contained observations from two PM_{10} monitoring networks one measuring Low Volume sampler Gravimetric (LVG) and the other a tapered element oscillating microbalance (TEOM). We develop a flexible hierarchical Bayesian spatio-temporal model which includes seasonal (winter and summer) effects. The fully Bayesian model is implemented, using MCMC techniques, which enables full inference with regard to process unknowns, calibration, validation and predictions in time and space.

Keywords: Hierarchical model; MCMC; separable spatio-temporal process.

1 Introduction

The reference sampling and measurement of PM_{10} are based on the gravimetric method known as LVG (*Low Volume sampler Gravimetric*). However, in our study region of North Italy the local authority, due to technical, administrative and historical reasons, maintain a dense network of automatic monitors based on a *tapered element oscillating microbalance* (TEOM). These monitors are known to underestimate the true PM_{10} levels given by the reference LVG method. Moreover, there is heterogeneity between the two sets of measurements. Thus there is an urgent need to to correct the TEOM measurements so that those are comparable with the LVG measurements.

The primary aim of this paper is to develop joint space-time modeling of data provided by above two heterogeneous instruments, LVG and TEOM. Except for one station, the LVG and TEOM were observed at completely different sites. This gives rise to a problem of spatial mis-alignment. Our modelling strategy avoids that by incorporating a space-time process common to both types of measurements. This induces correlation between the

two measurements in space and time. The model also include seasonality effects often found in PM_{10} data. The full Bayesian model, implemented using MCMC, enables: calibration for the station where both were measured, validation at a number of sites, and spatial interpolation and forecasting of PM_{10} at future time points. The structure of the paper is as follows. In Section 2, we provide a description of the data. Section 3 develops the Bayesian spatio-temporal model that accounts for monitor type, seasonality and random effects. Model based data analysis are presented in Section 4.

2 The data set

We consider the PM_{10} daily concentrations for $T = 120$ days in the period February 1 to May 31, 2003. The study region covers approximately an area of 400 kilometer by 200 kilometer grid and the monitors are located in the main city centers and along the main roads of three Italian regions: Piemonte, Lombardia and Emilia Romagna. The monitoring network is characterized by instrument heterogeneity: some regions have many TEOM sites and few LVG sites whilst the opposite holds for others. We analyze data from $n = 54$ stations composed of $n_1 = 34$ LVG monitors and $n_2 = 20$ TEOM monitors. There are less than 10% missing data. In one monitoring station, Consolata, in the Piemonte region we have both LVG and TEOM readings, but we consider only the TEOM monitors for modeling. The LVG measurements at Consolata are used for calibration purposes. We use data from 8 stations for validation of the model.

3 The PM_{10} Model

Let $Z_G(\mathbf{s}, t)$ denote the logarithm of the observed LVG measurement at a location $\mathbf{s}$ and at time t. Recall that we have LVG measurements from $n_1 = 34$ stations $\mathbf{s}_1, \ldots, \mathbf{s}_{n_1}$ at $T = 120$ days from February 1 to May 31, 2003. Let $Z_T(\mathbf{s}, t)$ denote the logarithm of the observed TEOM measurement at a location $\mathbf{s}$ and at time t. There are TEOM data from $n_2 = 20$ stations $\mathbf{s}_{n_1+1}, \ldots, \mathbf{s}_{n_1+n_2}$ at each of T days.
First, we assume the hierarchical models:

$$Z_G(\mathbf{s}_i, t) = Y_G(\mathbf{s}_i, t) + \epsilon_G(\mathbf{s}_i, t),\ i = 1, \ldots, n_1,\ t = 1, \ldots, T \qquad (1)$$

$$Z_T(\mathbf{s}_i, t) = Y_T(\mathbf{s}_i, t) + \epsilon_T(\mathbf{s}_i, t),,\ i = n_1 + 1, \ldots, n_1 + n_2,\ t = 1, \ldots, T, \quad (2)$$

where $Y_G(\mathbf{s}, t)$ and $Y_T(\mathbf{s}, t)$ are true space-time processes for LVG and TEOM measurements respectively; $\epsilon_G(\mathbf{s}_i, t)$ and $\epsilon_T(\mathbf{s}_i, t)$ are independent white noise processes assumed to follow $N(0, \sigma_G^2)$ and $N(0, \sigma_T^2)$ respectively. We suppose that the spatio-temporal processes $Y_G(\mathbf{s}, t)$ and $Y_T(\mathbf{s}, t)$ have different mean structures but are governed by a single un-observed spatio-temporal process $u(\mathbf{s}, t)$. This zero mean spatio-temporal process introduces

dependence between the Y_G and Y_T processes and these in turn influence dependencies between the observation processes $Z_G(\mathbf{s},t)$ and $Z_T(\mathbf{s},t)$. Thus we assume that:

$$Y_G(\mathbf{s}_i,t) = \mu_G(\mathbf{s}_i,t) + u(\mathbf{s}_i,t),\ i=1,\ldots,n_1,\ t=1,\ldots,T \quad (3)$$

$$Y_T(\mathbf{s}_i,t) = \mu_T(\mathbf{s}_i,t) + u(\mathbf{s}_i,t),,\ i=n_1+1,\ldots,n_1+n_2,\ t=1,\ldots,T. \quad (4)$$

We model the means $\mu_G(\mathbf{s},t)$ and $\mu_T(\mathbf{s},t)$ as linear functions of seasonality (coded by one for winter and zero for summer) and the longitude and latitude of the location $\mathbf{s}$. Let $\mathbf{x}(\mathbf{s},t) = (1, w_t, Lon(\mathbf{s}), Lat(\mathbf{s}))'$ where $w_t = 1$ if t falls in the two winter months, February and March, and 0 otherwise, and $Lon(\mathbf{s})$ and $Lat(\mathbf{s})$ are the longitude and the latitude of the location $\mathbf{s}$ respectively. Thus we assume that: $\mu_G(\mathbf{s},t) = \mathbf{x}(\mathbf{s},t)'\boldsymbol{\beta}_G$, and $\mu_T(\mathbf{s},t) = \mathbf{x}(\mathbf{s},t)'\boldsymbol{\beta}_T$ where
$\boldsymbol{\beta}_G = (\beta_G(1), \beta_G(2), \beta_G(3), \beta_G(4))'$ and $\boldsymbol{\beta}_T = (\beta_T(1), \beta_T(2), \beta_T(3), \beta_T(4))'$. Let Σ_s and Σ_t denote the spatial and temporal correlation matrices of the $u(\mathbf{s},t)$ process. That is, for $i,j = 1,\ldots,n$ and $k,l = 1,\ldots,T$, we have $(\Sigma_s)_{ij} = \rho_s(\mathbf{s}_i - \mathbf{s}_j;\phi_s), (\Sigma_t)_{kl} = \rho_t(k-l;\phi_t)$. We assume a separable covariance structure, see e.g. Mardia and Goodall (1993), for the unobserved $u(\mathbf{s},t)$ process. The prior specification is given by: $\mathbf{u} \sim N\left(\mathbf{0}, \sigma_u^2, \Sigma_s \otimes \Sigma_t\right)$ where $\otimes$ denotes the Kronecker product. Denote the unknown parameters by $\boldsymbol{\theta} = (\boldsymbol{\beta}_G', \boldsymbol{\beta}_T', \sigma_G^2, \sigma_T^2)'$. We assume that, a priori, the $\beta_G \sim N(0, A^2 I)$ where I denote the identity matrix and A^2 is a large positive constant. Similarly, we assume $\beta_T \sim N(0, A^2 I)$. For the three variance parameters σ_G^2, σ_T^2 and σ_u^2 we assume independent proper inverse gamma prior distributions, $IG(a,b)$ (with $a > 1$, hence, mean $b/(a-1)$) to avoid having an improper posterior distribution. In our numerical example we set $a = 2$ and $b = 1$ so that the resulting prior distribution has mean 1 and infinite variance.

For calibration purposes we want to predict LVG (or TEOM) at one of the sampled locations, $\mathbf{s}_{n_1}, \ldots, \mathbf{s}_n$ (or $\mathbf{s}_1, \ldots, \mathbf{s}_{n_1}$). To predict the LVG at a location $\mathbf{s}$ at a time t we see from (1) that $Z_G(\mathbf{s},t)$ as a Normal distribution with mean $\mu_G(\mathbf{s},t) + u(\mathbf{s},t)$ and variance σ_G^2. The predictions are obtained using the distribution

$$\pi\left(Z_G(\mathbf{s},t)|\mathbf{z}\right) = \int \pi\left(Z_G(\mathbf{s},t)|\boldsymbol{\theta},\mathbf{U}\right)\pi(\boldsymbol{\theta},\mathbf{U}|\mathbf{z})\, d\mathbf{U}\, d\boldsymbol{\theta}.$$

We perform this integration using the draws from the posterior distribution $\pi(\boldsymbol{\theta},\mathbf{U}|\mathbf{z})$. If, instead we want to predict TEOM at one of the LVG sites $\mathbf{s}_1, \ldots, \mathbf{s}_{n_1}$, we use the above methodology with obvious modifications; in particular we simulate a new $z_T(\mathbf{s},t)$ from $Z_T(\mathbf{s},t) \sim N\left(\mu_T(\mathbf{s},t) + u(\mathbf{s},t), \sigma_G^2\right)$ at each MCMC iteration. We then exponentiate the realizations to obtain the values in original scale.

It is of interest to provide the gravimetric concentrations $Z_G(\mathbf{s}',t')$ for a new location $\mathbf{s}'$ and time t'. We note that from (1) and (3)

$$Z_G(\mathbf{s}',t') \sim N\left(\mu_G(\mathbf{s}',t') + u(\mathbf{s}',t'), \sigma_G^2\right). \quad (5)$$

In order to simulate from this distribution we construct $\mu_G(\mathbf{s}', t')$ using the longitude and latitude of $\mathbf{s}'$ and seasonality of t'. We need the conditional distribution of $u(\mathbf{s}', t')$ given $\mathbf{U}$ at the n observed locations and at T time points.

4 Estimation, validation, prediction and calibration

The decay parameters $\mathbf{\Phi} = (\phi_s, \phi_t)$ are selected by a validation criterion, using the MSE on the 8 validation sites. The optimal values of ϕ_s and ϕ_t are found to be 0.02 and 0.75, respectively. See Sahu *et al.* (2006) for more detailed discussions regarding the choice of $\boldsymbol{\phi}$.
The following table shows the parameter estimates, their posterior standard deviations and the associated 95% intervals.

	mean	sd		mean	sd
$\beta_G(1)$	–6.261	1.505	$\beta_T(3)$	-0.127	0.014
$\beta_G(2)$	0.619	0.040	$\beta_T(4)$	0.061	0.030
$\beta_G(3)$	0.053	0.013	σ_G^2	0.025	0.0012
$\beta_G(4)$	0.207	0.032	σ_T^2	0.0024	0.0005
$\beta_T(1)$	1.891	1.426	σ_U^2	0.145	0.004
$\beta_T(2)$	0.297	0.041			

We considered calibration problem for the station Consolata. Using the calibration methods we have predicted the LVG measurements and their 95% prediction intervals. The prediction intervals contain 78% of the actual observations. Considering the validation of the model, except for one site, the percentage of observations inside the 95% prediction interval, that is the coverage, is about 95%. From a comparative analysis, the obtained results seem better, in terms of MSE, than the calibration results of Fasso and Nicolis (2005).

References

Fasso, A. and Nicolis, O. (2005). Space-Time Integration of Heterogeneous Networks in Air Quality Monitoring. *Statistics and Environment*, SIS Invited Papers, CLEUP, 265-276.

Mardia, K. V. and Goodall, C. (1993). Spatial-temporal analysis of multivariate environmental monitoring data. *Multivariate Environmental Statistics*. Eds. G. P. Patil and C. R. Rao, Elsevier, 347-386.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2006). Spatio-temporal modeling of fine particulate matter. *Journal of Agricultural Biological and Environmental Statistics* **11**, 61-86.