Hierarchical Space-time Modelling of PM_{10} Pollution in the Emilia-Romagna Region[1]

Un modello gerarchico per lo studio dell'inquinamento da PM_{10} in Emilia-Romagna

Fedele Pasquale Greco
Dipartimento di Scienze Statistiche, Università di Bologna, greco@stat.unibo.it

Carlo Trivisano
Dipartimento di Scienze Statistiche, Università di Bologna, trivi@stat.unibo.it

Riassunto: In questo lavoro si propone un modello gerarchico per lo studio della distribuzione spazio-temporale dell'inquinamento da PM_{10} in Emilia-Romagna. L'obiettivo è quello di fornire una prima caratterizzazione della variabilità spaziale e temporale delle concentrazioni e di misurarne la dipendenza dalle principali grandezze meteorologiche. I risultati mostrano come la variabilità temporale sia largamente dominante rispetto all'eterogeneità spaziale ed alla variabilità non spiegata.

Keywords: hierarchical models, particulate matter, dynamic linear models, spatial modelling.

1. Introduction

The analysis of the dynamics of airborne particulate matter (PM) concentrations is a central issue in environmental monitoring. In fact, several epidemiological studies have shown association between daily levels of PM and adverse health effects (see Pope *et al.*, 1995, for a summary).
In recent years a number of papers has been devoted to spatio-temporal modelling of PM data recorded in monitoring networks (Park *et al.*,2004; Smith *et al.*, 2003; Shaddick and Wakefield, 2002). Besides the understanding of the observed processes dynamics, spatio-temporal modelling of PM concentrations can be useful to produce exposure variables useful in ecological risk models by: a) cleaning observed time series from confounding effects and measurement errors; b) adjusting observed time series for missing data; c) estimating exposure variables for sites where data are not available.
In this paper we propose a hierarchical model for daily mean concentrations of PM_{10} measured in 12 monitoring sites located in the main cities of the Emilia-Romagna Region from January 1st 2000 to December 31st 2002. Data are characterized by a considerable presence of missing values. A number of meteorological variables are available in each monitoring site. The proposed model explicitly takes into account the spatial relationship among data collected in each monitoring site, the temporal structure of the observed time series, and the relationships between PM_{10} and meteorological variables. The main aim of our model is to identify the different sources of variability of

[1] The research leading to this paper has been partially funded by a 2004 grant (Sector 13: Economics and Statistics, protocol n. 2004137478 001 for Research Project of National Interest by MIUR)

observed data (spatial variability, temporal variability, variability due to dependence on meteorological conditions) and to provide exposure measures in unmonitored sites.
As regards inference, we adopt a fully Bayesian approach. Posterior summaries of model parameters are obtained by means of Gibbs sampling routines, as they are implemented in the WinBUGS software. In the Bayesian context missing values can be treated as parameters. Inference on the parameters of interest is then performed via averaging over the missing values distribution. This approach to dealing with missing values can be easily implemented in WinBUGS.

2. The data set

The analysed data set comprises time series of PM_{10} daily means ($\mu g/m^3$) collected at 12 monitoring sites within the Emilia-Romagna Region from January 1st 2000 to December 31st 2002. At least one monitoring site is available for each of the 9 provinces of the Region. Percentage of missing values varies from 7% to 40% in the 12 time series. The monitoring sites have to be distinguished according to their collocation: 5 of them are located in background urban areas such as parks (Type A) while the remaining 7 are located in zones with high population density or high traffic density (Type B and C). PM_{10} levels are in average lower in Type A monitoring sites while the time series seasonality is very similar regardless of the Type. A strong correlation is observed among site measurements, ranging from 0.86 for nearest sites to 0.6 for those further away.
Meteorological variables for each site are obtained from the mass-consistent model CALMET, implemented by the Regional Meteorological Service. Such model provides estimates on a regular grid of $10km \times 10km$ for daily mean temperature, daily mean mixing height (*MH*) and daily mean wind speed (*WS*). Temperature is highly correlated among monitoring sites (the correlation between time series is always greater than .98). Moreover temperature and *MH* show the same seasonal trend in each site and are highly correlated. In order to avoid collinearity, we choose to include in the model only *MH* because of its greater spatial variability. *MH* and *WS* variables have been centred and divided by their range in order to speed up convergence of the Monte Carlo Markov Chain algorithm used for parameters estimation.

3. The hierarchical model

Let Y_{ts}, MH_{ts}, WS_{ts} denote respectively the log-PM_{10} concentration, the mixing height and the wind speed at spatial location s on day t and let (X_{1s}, X_{2s}) be the site s coordinates. We assume that:

$$Y_{ts} \mid \mu_{ts}, \sigma_s^2 \sim N\left(\mu_{ts}, \sigma_s^2\right) \tag{1}$$

$$\mu_{ts} = \alpha Z_s + \beta_1 X_{1s} + \beta_2 X_{2s} + \beta_3 MH_{ts} + \beta_4 WS_{ts} + \theta_t + \varepsilon_{ts} \tag{2}$$

In this model σ_s^2 represents the residual variance in the site s, also defined as the measurement error variance (Shaddick and Wakefield, 2002). The variable Z is defined

as follows: Z_s=1 if the site s is of Type A while Z_s=-1 otherwise; hence the parameter α measures the effect of the monitoring site Type on the average log-PM_{10} levels.
Parameters β_1 and β_2 capture the large scale spatial trend while coefficients β_3 and β_4 capture the dependence of log-PM10 values on the considered meteorological variables. With regard to θ_t we assume that:

$$\theta_t = \theta_{t-1} + \omega_t, \quad \omega_t \sim N\left(0, \sigma_\theta^2\right) \tag{3}$$

This represents a first-order smoothing non-stationary temporal model. In terms of Dynamic Linear Models, equation (2) is known as the observation equation, equation (3) is the system equation and θ_t is the state.
The terms ε_{ts} represents spatially correlated random effects. We assume that at each time t, the random effects $\boldsymbol{\varepsilon}_t = \left(\varepsilon_{t1}, \varepsilon_{t2}, ..., \varepsilon_{tS}\right)$ arise from a multivariate normal distribution with mean vector 0 and $S \times S$ correlation matrix $\boldsymbol{\Sigma}$:

$$\boldsymbol{\varepsilon}_t \sim MVN\left(\boldsymbol{\theta}_s, \sigma_\varepsilon^2 \boldsymbol{\Sigma}\right) \tag{3}$$

The underlying assumption is that the spatial correlation among random effects does not depend on time, that is spatial and temporal processes are separable. A zero-mean constraint for the random effects at each time t has to be used for model identifiability. The parameter σ_ε^2 plays the role of the between site variance. The *ss'* entry of the correlation matrix represents the correlation between site s and s' and is specified as follows:

$$\Sigma_{ss'} = \exp\left(-\phi d_{ss'}\right) \tag{4}$$

Therefore, the correlation is a decreasing function of the distance $d_{ss'}$. The parameter $\phi > 0$ describes the decay of correlation with distance. Model hierarchy is completed by prior specification for the hyparameters. A normal prior $N(0,1000)$ is assumed for the regression coefficients β_i, i=1,…,4. For the variance parameters σ_s^2, σ_θ^2 and σ_ε^2, small parameters inverse Gamma ($IG(0.01,0.01)$) have been specified. A uniform distribution $U(0,2)$ is assumed for ϕ: this turns out in a prior belief for the spatial correlation ranging from .13 to 1 at a distance of 1 *km* and from 0 to 1 at the maximum distance of 250 *km*.

4. Main results and discussion

In Table 1 posterior distributions of model parameters are summarized. The posterior means of parameters β_1 and β_2 indicate a decreasing spatial trend in the *N-S* and *W-E* directions. The effect of the monitoring site Type, as measured by α, has the expected sign. A negative relationship has been estimated between the considered meteorological variables and the level of PM_{10} concentrations. In the original scale of the

meteorological variables, when *MH* increases 100 *m*, a decrease of 0.02 in PM_{10} concentrations (in the log scale) is estimated and when and *WS* increases 1 *m/s*, a decrease of 0.04 in PM_{10} concentrations (in the log scale) is estimated.

Table 1: *Summaries of model parameters posterior distributions*

	Posterior mean	Posterior st. dev.	Posterior median	95% credibility interval			Posterior mean	Posterior st. dev.	Posterior median	95% credibility interval	
α	-0.1181	0.0041	-0.1180	-0.1266	-0.1103	σ_3^2	0.0872	0.0082	0.0715	0.0869	0.1040
β_1	-0.0018	0.0001	-0.0020	-0.0018	-0.0015	σ_4^2	0.1297	0.0083	0.1139	0.1293	0.1461
β_2	-0.0055	0.0003	-0.0061	-0.0055	-0.0050	σ_5^2	0.0502	0.0050	0.0413	0.0500	0.0610
β_3	-0.2079	0.0315	-0.2740	-0.2090	-0.1463	σ_6^2	0.0452	0.0055	0.0358	0.0449	0.0569
β_4	-0.1871	0.0319	-0.2482	-0.1878	-0.1247	σ_7^2	0.1014	0.0087	0.0857	0.1010	0.1202
ϕ	0.0362	0.0032	0.0312	0.0357	0.0443	σ_8^2	0.1033	0.0062	0.0921	0.1030	0.1164
σ_ε^2	0.0622	0.0029	0.0564	0.0622	0.0681	σ_9^2	0.0392	0.0035	0.0324	0.0392	0.0460
σ_θ^2	0.0615	0.0033	0.0561	0.0614	0.0684	σ_{10}^2	0.0296	0.0041	0.0223	0.0297	0.0383
σ_1^2	0.1489	0.0119	0.1266	0.1489	0.1723	σ_{11}^2	0.0137	0.0027	0.0085	0.0135	0.0196
σ_2^2	0.0881	0.0078	0.0741	0.0877	0.1042	σ_{12}^2	0.0490	0.0066	0.0388	0.0480	0.0648

The posterior mean of parameter ϕ shows that the spatial correlation decreases to zero at a distance of approximately 70 *km*. By the way the contribution of the spatial dependence on the overall variability is ignorable with respect to the contribution of the temporal variability. Such contribution cannot be measured by the conditional variance σ_θ^2. An approximate measure of the temporal variability is given by the posterior variance $V(\boldsymbol{\theta} \mid \boldsymbol{y})$ (Shaddick and Wakefield, 2002) that in the estimated model results equal to 0.22 and is the greater source of variability when compared with spatial variability. The addition in the model of meteorological variables accounts for a negligible reduction (about 2.5%) in the residual variances σ_s^2. The residual variances are quite high in 4 sites (namely site 1, 4, 7, 8) because of a massive presence of outliers. Actually, an analysis of model adequacy by means of posterior predictive checks (not reported) shows some inadequacy of the estimated model in fitting data collected at these four sites in the tails of their distributions. Nevertheless, the overall adequacy of the model can be considered satisfactory.
Finally, we remark that observed time series can be broadly thought as replications of the same temporal process, with a feeble large scale spatial trend and a spatial correlation that vanish at a distance of 70 *km*.

References

Park E. S., Guttorp P., Kim H. (2004) Location major PM_{10} source areas in Seul using multivariate receptor modelling, *Environmental and Ecological Statistics*, 11, 9-19.

Pope C.I., Dockery D., Schwarts J. (1995) Review of epidemiological evidence of health effects of particulate air pollution. *Inhaln. Toxicol.*, 7, 1-18.

Shaddick G., Wakefield J.C. (2002) Modelling multiple pollutant data at multiple sites, *Applied Statistics*, 51, 351-372.

Smith R.L., Kolenikov S., Cox L.H. (2003) Spatio-temporal modelling of PM2.5 data whit missing values, *J. Geophys. Res.*, 108(D24).