

Environmental misaligned data via HLM

Modelli lineari gerarchici per dati ambientali non allineati

Anna Lisa Bondì and Antonella Plaia

Dipartimento di Scienze Statistiche e Matematiche “S. Vianelli”, Università di Palermo
e-mail: bondi@dssm.unipa.it

Riassunto: Nell’analisi dei dati ambientali ci si trova spesso nella condizione di dover considerare in un modello dati spazialmente non allineati. Può capitare, per esempio, di volere studiare la relazione tra inquinamento ambientale e variabili meteorologiche non disponendo però di dati rilevati negli stessi siti. Il presente lavoro intende proporre una soluzione al problema nell’ambito dei modelli lineari gerarchici. In particolare, facendo riferimento ai dati del PM_{10} e meteorologici relativi all’area metropolitana di Palermo nell’anno 2003, vengono considerati il metodo del riallineamento punto a punto a partire da quattro stazioni meteo, quello basato su quattro stazioni meteo e quello ottenuto considerando una sola stazione meteo.

Keywords: spatially misaligned data, hierarchical linear models, PM_{10} , weather variables.

1. Introduction

Very often, while modelling pollution level variability as a function of some covariates, a researcher has to cope with the “change of support problem (COSP)”. The problem arises because covariates (like meteorological variables, for example) are not usually measured in the same sites where pollution is measured (Gelfand *et al.* (2003)). Different solutions, where the problem has been considered, have been proposed in literature; for example, Li Zhu *et al.* (2003) have used a hierarchical Bayesian spatial approach. Here we want to model PM_{10} variability as function of meteorological variables, but while PM_{10} data are registered by 8 monitoring stations, only 4 meteorological sites register information on temperature, wind direction, wind speed, humidity, solar radiations, precipitation. Data are spatially misaligned because weather data are collected in different points from those at which pollutant data are collected, but also because we have a different number of weather and pollutant monitoring sites. Considering daily average for all variables, according to LEE *et al.* (2003) we will use logarithmic transformation of PM_{10} , in order to make the frequency distribution of PM_{10} approximately symmetric. In the next section we will propose and compare three solutions to realign our data in the context of Hierarchical Linear Model (HLM), a natural approach to model these data because of its ability to consider PM_{10} and meteorology at different levels of the hierarchy.

2. Hierarchical linear models for spatially misaligned data

Let Y be a vector of responses observed at locations $s_1, \dots, s_n$ (PM_{10} in 8 monitoring sites), X a vector of observed predictors at sites $s_1, \dots, s_{n'}$ and X' a vector of unobserved predictors at sites $s_{n'+1}, \dots, s_n$. We want to realign PM_{10} and meteorological variables

(humidity, wind speed, temperature, solar radiation, wind direction, daily precipitation and delayed daily precipitation), that is to estimate $X_{s'}$ where not observed. The first method we will consider (Method 1) use a ‘*h point space moving average*’ technique, on the h observed sites nearest to the site where we want to estimate X . So we will have a global trend surface smoothing of meteorological information on the map. Taking into account the distance among sites, we consider a weighted moving average (here a weighted average because we have only 4 meteorological stations) with normalized weights given by reciprocal of Euclidean distance between observed site and unobserved sites. That is, let s_0 be the unobserved site and d_{i0} the distance between s_0 and s_j , $j=1, 2, 3, 4$

$$\widehat{X}_0 = \frac{\sum_{i=1}^4 \frac{1}{d_{i0}} x_i}{\sum_{i=1}^4 \frac{1}{d_{i0}}} \quad (1)$$

The second method (Method 2) is obtained replacing X in an unobserved site by X observed in the nearest site. The last method (Method 3), frequently used in recent studies on the effects of weather variables on PM_{10} concentration (Lee *et al.* (2001), Yun *et al.* (2003) e Paez *et al.* (2003)), considers the local climatological data recorded at one weather station (hopefully inside the metropolitan area where PM_{10} level concentration is analyzed). Therefore we will replace all the X (both observed and unobserved) with the data recorded in a single site. Denoting $\log(PM_{10})$ by Y_{ij} , where i stands for i -th day of year and j stands for the j -th monitoring sites, we consider a hierarchical linear model:

$$Y_{ij} = \beta_{0j} + \sum_k \beta_{kj} X_{kij} + R_{ij} \quad (2)$$

$i \in \{1, 2, \dots, 365\}, j \in \{1, 2, \dots, 8\}, k \in \{1, 2, \dots, 8\}$, where

$$\beta_{rj} = \gamma_{r0} + U_{rj} \quad (3)$$

with $r = 0, 1, \dots, k$. Substitution of these equations in (2) gives:

$$Y_{ij} = \gamma_{00} + \sum_k \gamma_{k0} X_{kij} + U_{0j} + \sum_k U_{kj} X_{kij} + R_{ij} \quad (4)$$

with the following assumption on Level-1 R_{ij} and Level-2 residuals U_{rj} :

$$var(R_{ij}) = \sigma^2, var(U_{rj}) = \tau_{rr}^2, cov(U_{rj}, U_{r'j}) = \tau_{rr'} \quad (5)$$

In this model we distinguish a “fixed part”, $\gamma_{00} + \sum_k \gamma_{k0} X_{kij}$, and a “random part”, $U_{0j} + \sum_k U_{kj} X_{kij} + R_{ij}$. The first part is a linear combination of the independent variables (humidity, wind speed, temperature, solar radiations, wind direction, daily precipitation, delayed daily precipitation and day); the second one reflects the unexplained variation R_{ij} inside a site as well as the unexplained variation U_{rj} between the sites. The second type of residuals U_{rj} characterizes HLM with respect to general linear model. The Level-1 residuals R_{ij} has usually a normal distribution with constant variance σ^2

and is independent on the Level-2 random coefficients U. Considering a first order autoregressive residual component AR(1) follow that

$$R_{ij} = \rho R_{i-1,j} + \eta_{ij} \quad (6)$$

where $\eta_{ij} \cong iidN(0, \sigma^2)$ and ρ is the correlation coefficient. All residuals have population mean 0, given the values of all the explanatory variables. The vector $(U_{0j}, U_{1j}, \dots, U_{kj})$ has a multivariate normal distribution with a constant covariance matrix. All these assumptions imply that units at Level-1 in the same Level-2 unit, are correlated. So using a first order autoregressive residual component we fit the model defined in equation (4) considering the three methods described above.

Table 1: Comparison of three methods (Model 4)

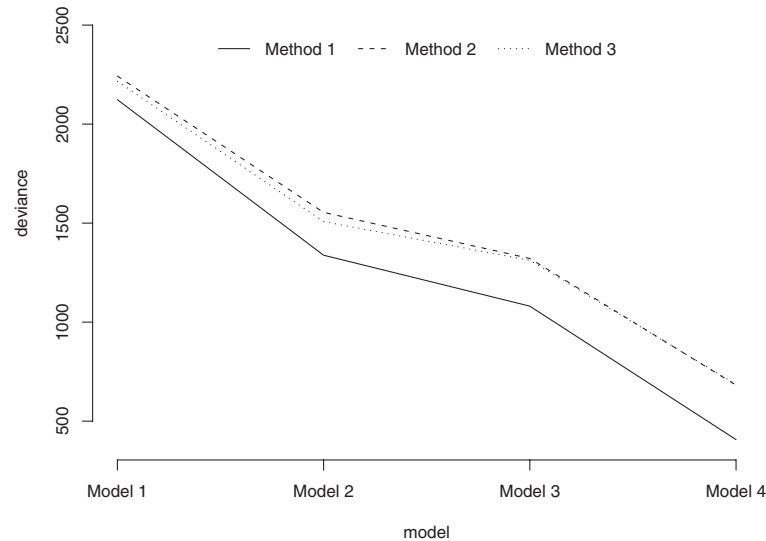
Fixed Effects				Random Effects			
	Method 1	Method 2	Method 3		Method 1	Method 2	Method 3
γ_{00}	3.461	3.298	3.619	τ_{00}^2	0.449	0.411	0.431
γ_{10}		0.003		τ_{22}^2	0.077	0.111	
γ_{20}	-0.133	-0.091	-0.186	τ_{33}^2	0.012	0.013	0.013
γ_{30}	0.021	0.0201	0.021	τ_{44}^2	0.000002		
γ_{40}	-0.0005	-0.0004	-0.0006	τ_{55}^2		0.000004	0.000008
γ_{50}	-0.077	-0.0007	-0.001	τ_{02}	-0.574	-0.411	
γ_{60}	-0.039	-0.059	-0.047	τ_{03}	-0.943	-0.861	-0.961
γ_{70}	-0.031	-0.041	-0.038	τ_{04}	-0.006		
γ_{80}	-0.123	-0.117	-0.122	τ_{05}		0.001	0.003
				τ_{23}	0.284	-0.048	
				τ_{24}	-0.001		
				τ_{25}		-0.001	
				τ_{34}	0.008		
				τ_{35}		-0.001	-0.003
				ρ	0.537	0.515	0.515
				σ^2	0.305	0.313	0.316

Deviance	Method 1: 406.7092	Method 2: 681.6118	Method 3: 685.7374
Df	20	21	16

1=humidity	2=wind speed	3=temperature	4=solar radiations
5=wind direction	6=daily precipitation	7=delayed daily precipitation	8=day

Table 1 shows both the fixed and random part estimates of the model fitted using a step-to-step procedure (Bondi (2005)). Actually the Table 1 shows only the final selected model (say Model 4 defined in equation (4) with a first order autoregressive residual component AR(1)), while Figure 1 shows the deviance functions for the four ‘best’ hierarchical linear models (from Model 1 to Model 4). For every Method, Model 1 includes only fixed and random intercept, Model 2 includes significative fixed effects and random intercept, Model 3 includes significative fixed effects, random intercept and random slopes and finally Model 4 is equal to Model 3 but includes a first order autoregressive residual component AR(1). In a detailed description of Method 1, Model 1 includes fixed and random intercept, Model 2 includes significative fixed effects (intercept, wind speed, temperature, solar radiation, wind direction, daily precipitation, delayed daily precipitation and day) and random intercept, Model 3 includes significative fixed effects, the same as Model 2, random intercept and random slopes (wind speed, temperature and solar radiation), Model 4 is equal to Model 3 but includes a first order autoregressive residual component AR(1).

Figure 1: Plot of deviances: three methods.



3. Results and concluding remarks

As expected the first method, based on ‘*h point space weighted average*’, shows a deviance function lower than the others, confirming the necessity to use site specific weather information. We suppose that observed site specific weather information could produce a better fit of the model, but, when unavailable, a realignment of data based on weighted average can help to improve model performance.

References

- Bondì A.L. (2005) *Modelli Lineari Gerarchici per lo studio degli effetti delle variabili meteorologiche sull'inquinamento ambientale*, Dipartimento di Scienze Statistiche e Matematiche “S. Vianelli”, Università di Palermo, Tesi di dottorato.
- Daniels M.J., Lee Y., Kaiser M. (2001) Assessing sources variability in measurement of ambient particulate matter, *Environmetrics*, 12, 547-558.
- Lee Y.D., Yun S., Lee Y. (2003) Analyzing weather effects on airborne particulate matter with HGLM, *Environmetrics*, 14, 687-697.
- Li Zhu B. P., Carlin P.E., Gelfand A.E. (2003) Hierarchical regression with misaligned spatial data: relating ambient ozone and pediatric asthma ER visits in Atlanta, *Environmetrics*, 14, 537-557.
- Paez M.S., Gamerman D. (2003) Study of the space-time effects in the concentration of airborne pollutants in the Metropolitan Region of Rio de Janeiro, *Environmetrics*, 14, 387-408.