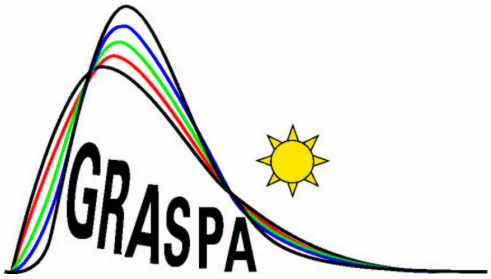
Web Working Papers
by
The Italian Group of Environmental Statistics



Gruppo di Ricerca per le Applicazioni della Statistica
ai Problemi Ambientali

www.graspa.org

Integrating satellite and ground level data for air quality monitoring and dynamical mapping

Alessandro Fassò, Francesco Finazzi, Cinzia D'Ariano

GRASPA Working paper n.34, November 2009

Integrating satellite and ground level data for air quality monitoring and dynamical mapping

Alessandro Fassò, Francesco Finazzi, Cinzia D'Ariano
DIIMM, University of Bergamo
alessandro.fasso@unibg.it

20/11/2009

Abstract

Aerosol optical thickness (AOT) is a satellite measurement useful for assessing airborne particulate matter concentration. Although it has nice properties for air quality mapping due to regularity in space and time, it is known to have large uncertainty and structural missing due to cloud conditions.

In this paper, we discuss the multivariate spatio-temporal calibration model named dynamical coregionalization model, which allows us to merge AOT and the data coming form ground level networks even in presence of extensive missing data. Thanks to this, we can define a spatio-temporal dynamical calibration algorithm for high quality mapping of particulate matter concentration, which gives good results even in areas without ground level instruments.

In particular in order to perform maximum likelihood estimation, formulas for the generalized EM algorithm and of the observed information matrix are given.

The proposed method is applied to the large data set related to "padano-veneto" region, North Italy. This includes particulate matter concentration and various spatio-temporal covariates related to meteorology and land properties. The relevant computational burden is covered by a medium size computer cluster.

1 Introduction

Air quality monitoring networks have been increasingly installed around Europe, mainly on a local basis. Satellite measurements give interesting data because of homogeneity over time and space. For example, aerosol optical thickness (AOT) may be used to get information on airborne particulate matters PM_{10} and $PM_{2.5}$. As a matter of fact, although AOT measurements are less precise than ground-level ones of particulate matters, various authors, with the aim of calibrating satellite data, have reported positive temporal correlations between satellite data and both PM_{10} and $PM_{2.5}$ and suggested their use as components of pollution models, see, for example, Kaastle et al., 2006, Koelemeijer et al., 2006, and Wang et al., 2003.

Along these lines, we discuss multivariate spatio-temporal modelling for merging ground level data, satellite data and a number of spatio-temporal covariates, which are used for improving calibration capability and dynamical mapping quality. Due to the fact that AOT availability is restricted to cloud-free conditions, the model has to cover with extensive missing data.

The approach to multivariate spatio temporal modelling based on the linear coregionalization model (e.g. Rouhani and Wackernagel, 1990) is widely used in applications, see, for example, De Iaco et al. (2005) which used a bivariate LCM model for spatio-temporal cokriging of isotopic temperature and humidity in northern Italy. Similarly, Jost et al. (2005) used the same cross-product model to analyse water storage in a forest ecosystem in Austria. Moreover, Liu and Koike (2007) used the spatio-temporal coregionalization model for partially heterotopic data on water quality in the Arike Sea, in Japan.

In this paper, we propose a frequentist hierarchical model which generalizes coregionalization of Zhang (2007) and missing data handling of Amisigo and Van de Giesen (2005). Moreover, the model can deal with entirely heterotopic datasets, for which AOT and PM concentration are observed at different locations. Maximum likelihood estimation is covered by the generalized EM algorithm extending Fassò and Cameletti (2009), which is completed by the observed Information matrix computed in closed form, generalizing Shumway and Stoffer (2006)

The rest of the paper is organized as follows. Section 2 introduces the multivariate spatio-temporal model based on coregionalization and autoregressive temporal dynamics. Section 4 discusses the data used for the application which are related to daily dynamics in year 2006 for the "padano-

veneto" region, North Italy. Section 5 discusses the estimation and results while the Appendix A gives technical details on both marginal likelihood and Information matrix computations.

2 Dynamical coregionalization model

We suppose that the $q-$dimensional data $Y(s,t) = (Y_1(s,t),...,Y_q(s,t))'$, observed at time $t = 0,...,T$ and at site $s \in D \subset \mathbb{R}^2$, is given by the following measurement equation

$$Y(s,t) = U(s,t) + \varepsilon(s,t) \tag{1}$$

where $\varepsilon(s,t)$ is the Gaussian instrumental error which is white noise in space and time with $q \times q$ covariance diagonal matrix

$$\Gamma_0 = diag\left(\sigma_1^2,...,\sigma_q^2\right) \tag{2}$$

The underlying true local $q-$dimensional $U(s,t)$ has the $j-th$ element given by

$$U_j(s,t) = X_j(s,t)\beta_j + \mathcal{K}_j' Z(t) + \bar{W}_j(s,t) \tag{3}$$

for $j = 1,...,q$, where $X_j(s,t)$ is the b_j-dimensional spatio-temporal field of known covariates for $Y_j(s,t)$, $Z(t)$ is the d-dimensional latent temporal state, which is constant in space and has Markovian temporal dynamics. In particular, the latent state at time t is defined as

$$Z(t) = GZ(t-1) + \eta(t)$$

where G is a $d \times d$ transition matrix with eigenvalues λ_i such that $|\lambda_i| < 1$, $\eta(t) \sim N_d(0, \Sigma_\eta)$ and $Z(0) \sim N_d(\mu_0, \Sigma_{Z_0})$. The $q \times d$ matrix $\mathcal{K} = (\mathcal{K}_1,...,\mathcal{K}_q)'$ is fixed in time and accounts for the weights of the d components of $Z(t)$ for the different components of Y.

The zero mean $q-$dimensional Gaussian process $\bar{W}(s,t) = \left(\bar{W}_1,...,\bar{W}_q\right)$ is the latent random spatial effect at time t and it is defined by the q-dimensional linear coregionalization model (LCM) with c components, namely

$$\bar{W}(s,t) = \sum_{p=1}^{c} W_p(s,t)$$

where $W_p = (W_{p,1}, ..., W_{p,q})$ is white noise in time but correlated over space with a $q \times q$ covariance and cross-covariance matrix function given by

$$\Gamma_p(h, \theta_p) = (\text{cov}(W_{p,i}(s), W_{p,j}(s')))_{i,j=1,...,q} = V_p \rho_p(h, \theta_p)$$

The above covariance and cross-covariance functions are assumed to be isotropic and $h = \|s - s'\|$ is the Euclidean distance between two sites $s, s' \in D$. For each $p = 1, ..., c$, V_p is a positive semi-definite $q \times q$ matrix and $\rho_p(h, \theta_p)$ is a valid correlation function, for example the Matern function, characterized by the parameter vector θ_p. In addition, the processes W_p are uncorrelated in the sense that, for any $i \neq j$

$$\text{cov}(W_i(s), W_j(s')) = 0, \ \forall s, s' \in D$$

The multivariate $q \times q$ covariance matrix for W is then

$$\Gamma_{\bar{W}}(h, \theta_1, ..., \theta_c) = \sum_{p=1}^{c} \Gamma_p(h, \theta_p) = \sum_{p=1}^{c} V_p \rho_p(h, \theta_p) \tag{4}$$

The model parameters are collected in the following vector

$$\Psi = vec^*(\beta, \Gamma_0; \mu_0; G, \Sigma_\eta, ; V_1, \theta_1, ..., V_c, \theta_c) = (\Psi_Y, \Psi_{Z_0}, \Psi_Z, \Psi_W) \tag{5}$$

where $\beta = \left(\beta_1', ..., \beta_q'\right)'$ is $b-$dimensional with $b = \sum_{j=1}^{q} b_j$ and the operator vec^* vectorizes all the unique parameters contained in the covariance matrices excluding zeros.

Note that the matrices $\mathcal{K}$ and Σ_{Z_0} are assumed to be known and do not take part in model parametrization and estimation.

2.1 Data structure

We assume that each variable is observed over different sets of sites. If $S_i = \{s_{i,1}, ..., s_{i,n_i}\}$ is the set of sites for the variable Y_i, $i = 1, ..., q$, three cases can be distinguished: the isotopic case, the partially heterotopic case and the entirely heterotopic case (Wackernagel, 1998). The isotopic case is characterized by the fact that each variable is observed at each site, so that $S_1 = ... = S_q$. Within the entirely heterotopic case, two different variables are never observed at the same site and $\bigcap_{i=1}^{q} S_i = \emptyset$. A less narrow case is the

partially heterotopic setting, in which only some variables share only some sites. In this work it is also assumed that the sets S_i do not change with time $t \in \mathcal{T} = \{1, ..., T\}$.

Now, let

$$Y_t = \big(Y_1(s_{1,1}, t), ..., Y_1(s_{1,n_1}, t), ..., Y_q(s_{q,1}, t), ..., Y_q(s_{q,n_q}, t)\big)' \tag{6}$$

be the $N = n_1 + ... + n_q$ dimensional observation vector at time t at the sampling sites $S = \{S_1, ..., S_q\}$ and let X_t be the $N \times b$ block diagonal matrix of known regressors at time t. Let $Y = (Y_1, ..., Y_T)$ be the $N \times T$ full data matrix, $Z = (Z_0, Z_1, ..., Z_T)$ be the $d \times (T + 1)$ matrix of the latent temporal process and $W_p = (W_{p,1}, W_{p,2}, ..., W_{p,T})$, $p = 1, ..., c$, be the $N \times T$ matrices of the c spatial latent processes. The collection of all processes is $W_t = (W_{1,t}, ..., W_{c,t})$ and their sum is denoted by $\bar{W}_t = \sum_{p=1}^{c} W_{p,t}$. Similar definitions hold for ε_t, W and $\bar{W}$.

Using these symbols, the multivariate geographical observations at time t follow the measurement equation

$$Y_t = X_t\beta + KZ_t + W_t + \varepsilon_t$$

where the $N \times d$ matrix K has rows $\mathcal{K}_j'$ matching those of Y_t as in equation (6).

The $N \times N$ symmetric matrix of distances between each pair of sampling locations in S is given by the block representation

$$H = \big(H^{i,j}\big)_{i,j=1,...,q}$$

with $n_i \times n_j$ dimensional blocks $H^{i,j}$. If $\rho_p(H, \theta_p)$ is the spatial correlation matrix for the W_p process observed at S, then

$$\Sigma_p(V_p, \theta_p, H) = \big(v_p^{i,j}\rho_p(H^{i,j}, \theta_p)\big)_{i,j=1,...,q} \tag{7}$$

is the $N \times N$ variance-covariance block matrix of the process W_p, $p = 1, ..., c$. In particular, the k, l-th element of the block $v_p^{i,j}\rho_p(H^{i,j}, \theta_p)$ is $\mathrm{cov}(W_p(s_{i,k}), W_p(s_{j,l}))$, $k = 1, ..., n_i$, $l = 1, ..., n_j$. It must be noted that, unless the isotopic setting is considered, only the diagonal blocks of Σ_p are square.

Moreover

$$\Sigma_{\bar{W}} = \sum_{p=1}^{c} \Sigma_p$$

and the $N \times N-$dimensional diagonal matrix Σ_0 is the variance covariance matrix of the measurement errors stacked conformably to Y_t as in equation (6) and based on variances of equation (2).

3 MLE with missing data

In this section we define the generalization of the EM algorithm of Fassò et al. (2009) for the presence of missing data by using the same notation and concepts. To do this, if, at time t, some observations in Y_t are missing, we define the observed data by

$$Y_t^{(1)} = L_t Y_t$$

where L_t is the corresponding selection matrix. In other words, $\left(Y_t^{(1)}; Y_t^{(2)}\right)$ is the permutation of the vector Y_t, where the first component $Y_t^{(1)}$ is observed while the second component $Y_t^{(2)}$ is missing. It is then defined the inverse permutation D_t, which reorders the vector in its original order, namely $Y_t = D_t \begin{pmatrix} Y_t^{(1)} \\ Y_t^{(2)} \end{pmatrix}$. Then the partitioned measurement equations, for $i = 1, 2$, become:

$$Y_t^{(i)} = X_t^{(i)}\beta + K^{(i)} Z_t + \bar{W}_t^{(i)} + \varepsilon_t^{(i)} \tag{8}$$

and the variance covariance matrix of the permuted errors is conformably partitioned, namely

$$Var \begin{bmatrix} \varepsilon_t^{(1)} \\ \varepsilon_t^{(2)} \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix} \tag{9}$$

In the following sections, we discuss the two-steps of the iterative EM algorithm, that is the socalled E-step and M-step.

3.1 E-step computations

In this phase, the expectation of the complete data log-likelihood under the parameter $\Psi^{(k)}$ conditionally to the observed data $Y^{(1)}$ can be computed thanks to the iterated expectation theorem, that is

$$\begin{aligned} Q\left(\Psi, \Psi^{(k)}\right) &= E_{\Psi^{(k)}}\left[-2 \log L_c(\Psi; Y, Z, W) \mid Y^{(1)}\right] \\ &= E_{\mathbf{\Psi}^{(k)}}\left[E_{\Psi^{(k)}}\left[-2 \log L_c(\Psi; Y, Z, W) \mid Y^{(1)}, Z, W\right] \mid Y^{(1)}\right] \end{aligned} \tag{10}$$

To compute the right hand side of 10, we need some expected values. First, let us observe that the purely spatial quantities

$$\begin{aligned} W_{p,t}^T &= E_{\Psi^{(k)}}(W_{p,t} \mid Y^{(1)}) \\ A_{p,t}^T &= Var_{\Psi^{(k)}}(W_{p,t} \mid Y^{(1)}) \end{aligned}$$

are computed in closed form in Appendix B.

Next, the Kalman smoother outputs $Z_t^T = E_{\Psi^{(k)}}(Z_t \mid Y^{(1)})$ and $P_t^T = Var_{\Psi^{(k)}}(Z_t \mid Y^{(1)})$ are defined as in Fassò et al. (2009), Appendix A, with standard modifications for missing values, see e.g. Durbin and Koopman (2001), p.92.

Now, using the techniques of Fassò et al. (2009), we get

$$\begin{aligned} & Q\left(\Psi, \Psi^{(k)}\right) \\ & = T \log|\Sigma_0| + \sum_{t=1}^{T} \operatorname{tr}\left\{(\Sigma_0)^{-1}\,\Omega_t^{(k)}\right\} \\ & + \log|\Sigma_{Z_0}| + \operatorname{tr}\left\{\Sigma_{Z_0}^{-1}\left[\left(Z_0^T - \mu_0\right)\left(Z_0^T - \mu_0\right)' + P_0^T\right]\right\} \\ & + T \log|\Sigma_\eta| + \operatorname{tr}\left[\Sigma_\eta^{-1}\left(S_{11} - S_{10}G' - GS_{10}' + GS_{00}G'\right)\right] \\ & + \sum_{p=1}^{c}\left[T \log|\Sigma_p| + \sum_{t=1}^{T} \operatorname{tr}\left\{(\Sigma_p)^{-1}\left[A_{p,t}^T + W_{p,t}^T\left(W_{p,t}^T\right)'\right]\right\}\right] \end{aligned} \tag{11}$$

In order to compute $\Omega_t^{(k)}$ in equation (11), let us consider

$$e_t = E_{\Psi^{(k)}}\left(Y_t - X_t\beta - KZ_t - \bar{W}_t \mid Y^{(1)}, Z_t, W_t\right) = D_t \begin{bmatrix} e_t^{(1)} \\ e_t^{(2)} \end{bmatrix}$$

which is given by

$$e_t = D_t \begin{bmatrix} Y_t^{(1)} - X_t^{(1)}\beta - K^{(1)}Z_t - \bar{W}_t^{(1)} \\ R_{21}R_{11}^{-1}\left(Y_t^{(1)} - X_t^{(1)}\beta - K^{(1)}Z_t - \bar{W}_t^{(1)}\right) \end{bmatrix}$$

Moreover, observe that

$$\Lambda_t = Var_{\Psi^{(k)}}\left[Y_t - X_t\beta - KZ_t - \bar{W}_t \mid Y^{(1)}, Z_t, W_t\right] = D_t \begin{bmatrix} 0 & 0 \\ 0 & R_{22} - R_{21}R_{11}^{-1}R_{12} \end{bmatrix} D_t'$$

Using these results, we have that

$$\begin{aligned} \Omega_t^{(k)} &= E_{\Psi^{(k)}}\left[e_t e_t' + \Lambda_t \mid Y^{(1)}\right] \\ &= D_t \begin{bmatrix} \Omega_t^{(11)} & \Omega_t^{(11)}R_{11}^{-1}R_{21} \\ R_{21}R_{11}^{-1}\left(\Omega_t^{(11)}\right)' & R_{21}R_{11}^{-1}\Omega_t^{(11)}R_{11}^{-1}R_{21} + \left(R_{22} - R_{21}R_{11}^{-1}R_{12}\right) \end{bmatrix} D_t' \end{aligned} \tag{12}$$

where

$$\Omega_t^{(11)} = \sum_{p=1}^{c} A_{p,t}^{(1),T} - K^{(1)} P_t^T \left(K^{(1)}\right)' + \bar{e}_t^{(1)} \left(\bar{e}_t^{(1)}\right)'$$
$$\bar{e}_t^{(1)} = Y_t^{(1)} - X_t^{(1)}\beta - K^{(1)} Z_t^T - \bar{W}_t^{(1),T}$$

and $R_{ij} = R_{ij}^{(k)}$ are the blocks of matrix (9) computed using the current parameter $\Psi^{(k)}$.

Finally, note that in equation (11)

$$S_{11}^{(k)} = \sum_{t=1}^{T} \left(Z_t^{T,(k)}\right) \left(Z_t^{T,(k)}\right)' + P_t^{T,(k)}$$
$$S_{10}^{(k)} = \sum_{t=1}^{T} Z_t^{T,(k)} Z_{t-1}^{T,(k)} + P_{t,t-1}^{T,(k)}$$
$$S_{00}^{(k)} = \sum_{t=1}^{T} \left(Z_{t-1}^{T,(k)}\right) \left(Z_{t-1}^{T,(k)}\right)' + P_{t-1}^{T,(k)}$$

3.2 M-step computations

The maximization of $Q\left(\Psi, \Psi^{(k)}\right)$ is now is perfomred mostly in closed-form avoiding large numerical optimization. To see this the solution of $\frac{\partial Q}{\partial \mathbf{\Psi}} = 0$ is obtained by partitioning the parameter vector $\Psi = \left\{\mathring{\Psi}, \breve{\Psi}\right\}$, where

$$\mathring{\Psi} = vec^* \left(\beta, \Gamma_0; G, \Sigma_\eta, \mu_0; V_1, ..., V_c\right)$$

and $\breve{\Psi} = (\theta_1, ...\theta_c)$. In particular, the closed forms for $\mathring{\Psi}$ are given by

$$V_p^{(k+1)} = \begin{bmatrix} \frac{\operatorname{tr}\left[R_p\left(H^{1,1},\theta_p^{(k)}\right)\bar{U}_p^{1,1,(k)}\right]}{\operatorname{tr}\left[R\left({}_pH^{1,1},\theta_p^{(k)}\right)^2\right]} & \cdots & \frac{\operatorname{tr}\left[R_p\left(H^{1,q},\theta_p^{(k)}\right)\bar{U}_p^{1,q,(k)}\right]}{\operatorname{tr}\left[R\left({}_pH^{1,q},\theta_p^{(k)}\right)^2\right]} \\ \vdots & \ddots & \vdots \\ \frac{\operatorname{tr}\left[R_p\left(H^{q,1},\theta_p^{(k)}\right)\bar{U}_p^{q,1,(k)}\right]}{\operatorname{tr}\left[R\left({}_pH^{q,1},\theta_p^{(k)}\right)^2\right]} & \cdots & \frac{\operatorname{tr}\left[R_p\left(H^{q,q},\theta_p^{(k)}\right)\bar{U}_p^{q,q,(k)}\right]}{\operatorname{tr}\left[R\left({}_pH^{q,q},\theta_p^{(k)}\right)^2\right]} \end{bmatrix}, \ p = 1, ..., c \tag{13}$$

$$\Gamma_0^{(k+1)} = \begin{bmatrix} \frac{1}{n_1}\operatorname{tr}\bar{\Omega}^{1,1,(k)} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \frac{1}{n_q}\operatorname{tr}\bar{\Omega}^{q,q,(k)} \end{bmatrix} \tag{14}$$

$$\beta^{(k+1)} = \left[\sum_{t=1}^{T} X_t' \left(\Sigma_0^{(k)}\right)^{-1} X_t\right]^{-1} \left[\sum_{t=1}^{T} X_t' \left(\Sigma_0^{(k)}\right)^{-1} \left(Y_t^{(k)} - KZ_t^{T,(k)} - \bar{W}_t^{T,(k)}\right)\right] \tag{15}$$

$$\mu_0^{(k+1)} = Z_0^{T,(k)} \tag{16}$$

$$G^{(k+1)} = S_{10}^{(k)} \left(S_{00}^{(k)}\right)^{-1} \tag{17}$$

$$\Sigma_\eta^{(k+1)} = \frac{1}{T}\left[S_{11}^{(k)} - S_{10}^{(k)}G' - G\left(S_{10}^{(k)}\right)' + GS_{00}^{(k)}G'\right] \tag{18}$$

where

$$\bar{U}_p^{(k)} = \begin{bmatrix} \bar{U}_p^{1,1,(k)} & \cdots & \bar{U}_p^{1,q,(k)} \\ \vdots & \ddots & \vdots \\ \bar{U}_p^{q,1,(k)} & \cdots & \bar{U}_p^{q,q,(k)} \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T} U_{p,t}^{(k)} = \frac{1}{T}\sum_{t=1}^{T}\left[A_{p,t}^{T,(k)} + W_{p,t}^{T,(k)}\left(W_{p,t}^{T,(k)}\right)'\right]$$

and

$$\bar{\Omega}^{(k)} = \begin{bmatrix} \bar{\Omega}^{1,1,(k)} & \cdots & \bar{\Omega}^{1,q,(k)} \\ \vdots & \ddots & \vdots \\ \bar{\Omega}^{q,1,(k)} & \cdots & \bar{\Omega}^{q,q,(k)} \end{bmatrix} = \frac{1}{T}\sum_{t=1}^{T}\left(D_t\Omega_t^{(k)}D_t'\right)$$

Moreover, the updates for $\breve{\Psi}$ are given by

$$\theta_p^{(k+1)} = \arg\min_{\theta_p}\left\{T\log\left|\Sigma_p^{(k)}\right| + \operatorname{tr}\left[\left(\Sigma_p^{(k)}\right)^{-1}\sum_{t=1}^{T} U_{p,t}^{(k)}\right]\right\}, \ p = 1, ..., c \tag{19}$$

which are computed using c one-dimensional numerical optimizations as a function of $\breve{\Psi}$ only.

4 Data of "Padano-veneto" region

This section provides a description of the variables and covariates involved in this work, with information about their spatio-temporal structure. In particular, the data set here considered covers the Italian region known as padano-veneto area, bounded by a box of coordinates $44°N - 6°E, 47°N - 14°E$, and the time period between March 2006 and September 2006.

4.1 Variables

The main variable of interest is the airborne particulate matter PM_{10}, the concentration of which is measured by the stations of a ground level monitoring network. Altough each station provides direct and reliable measures of the PM_{10} concentration, they are limited in number with respect to the area to be covered. For this reason, a second variable is considered, namely the Aerosol Optical Thickness (AOT). The AOT is known to be related with the particulate matters concentration and it is useful to improve mapping capability of the PM_{10} concentration over the area of interest.

4.1.1 Particulate matters PM10

The PM_{10} concentration is daily collected by a ground level monitoring network composed of $n_{PM} = 107$ gravimetric samplers sparsed over the padano-veneto area as depicted in Figure 1. Data are provided by the locale environmental agencies (ARPA) from regione Piemonte, regione Lombardia and regione Veneto. The average missing data rate for the time period considered is 6%, with specific missing data rates displayed in Figure 1. The image of Figure 2 represents the PM_{10} concentration measured at a specific day.

4.1.2 Aerosol Optical Thickness

Atmospheric aerosol concentrations are expressed in terms of Aerosol Optical Thickness, defined as the degree to which aerosols prevent the transmission of sun light to the Earth's surface. Since particulate matter PM_{10} is seen as a specific aerosol type, AOT data can be analyzed in order to assess ground level concentrations of PM_{10}. On the other hand, optical thickness includes the scattering of light by molecules in the atmosphere, which occurs even when no aerosols are present; for this reason, the correlation between AOT and PM_{10} concentration is corrupted by noise.

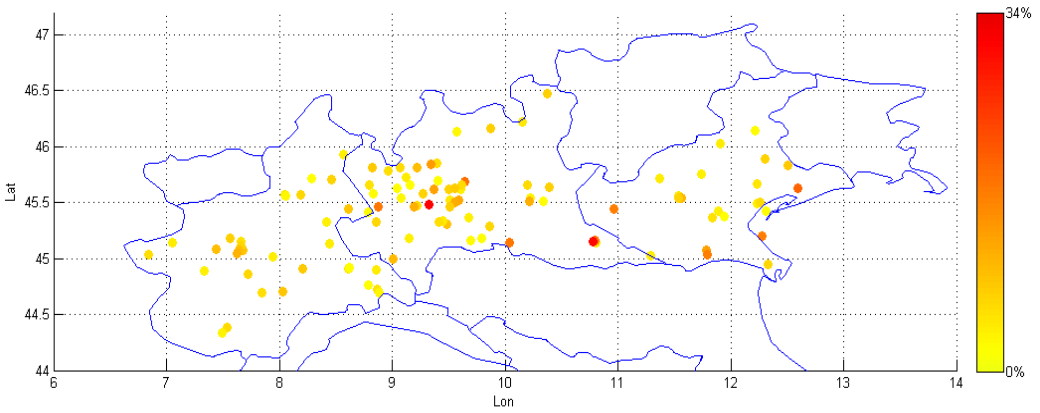


Figure 1: PM10 monitoring network and time-average missing data rates

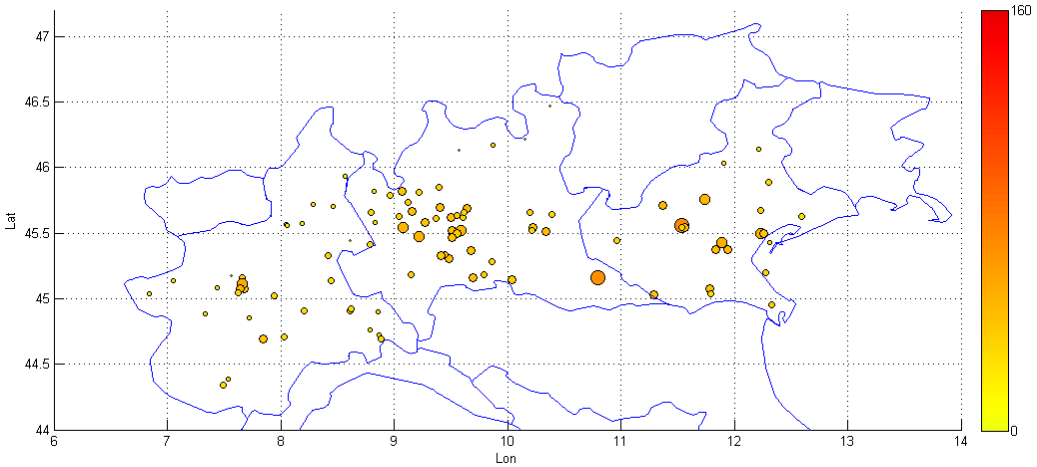


Figure 2: PM10 concentrations $[\mu g/m^3]$ for day 202.

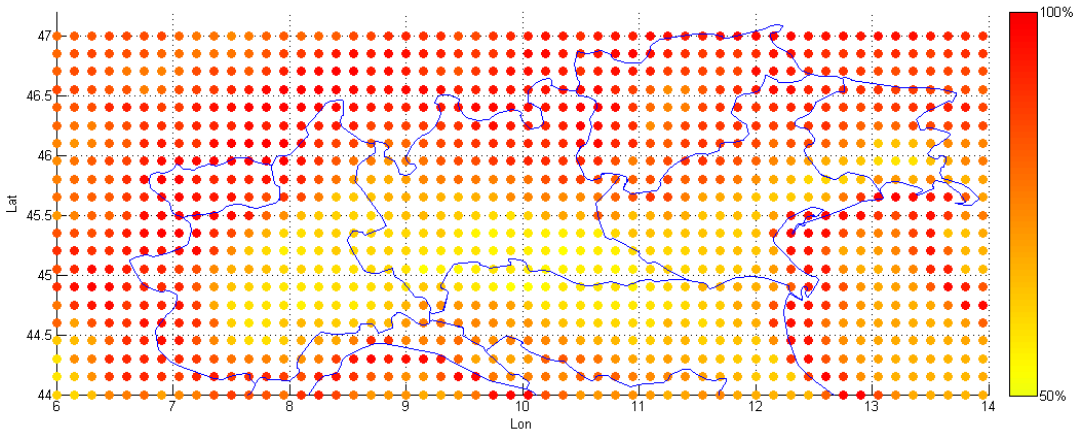


Figure 3: AOT sites at ground and time-average missing data rates

AOT data are collected by remote sensing. NASA satellites Terra and Aqua cover the entire Earth's surface every 1 to 2 days, acquiring data in 36 spectral bands by means of the MODIS instrument (Moderate Resolution Imaging Spectroradiometer). The Aerosol Optical Thickness is a product derived from MODIS observations and it is provided with a spatial resolution of 10×10 km at nadir. In particular, the *Optical_Depth_Land_And_Ocean* data field of the level 2 MOD04_L2 product from collection 5 provided by the NASA agency is here considered.

Since the spatial resolution is not constant over the area of interest (due to spherical geometry) and since the grid change every day (Terra and Aqua satellites repeat their orbits every 16 days), AOT data are re-located through interpolation over a fixed in time grid of $0.15 \times 0.15^{\circ}$ spatial resolution ($n_{AOT} = 1134$). The image of Figure 3 represents the AOT sites and their missing data rate with respect to the time period considered. The missing data rate is quite high since AOT availability is restricted to cloud-free and snow-free conditions (see Figure 4). The histogram of Figure 5 represents the daily missing data rate distribution for the area considered. It must be noted that the missing data rate is higher than 85% for half of the days considered, with an average rate of 79%.

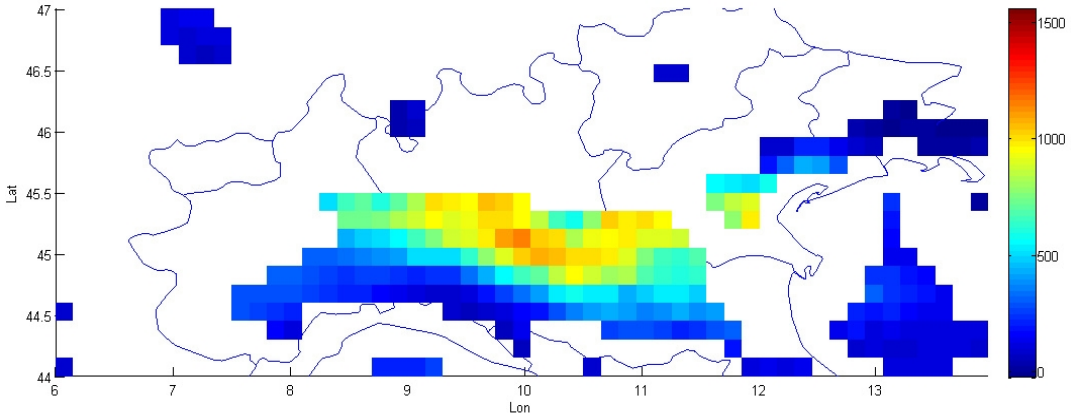


Figure 4: AOT data for day 202 (raw data scale)

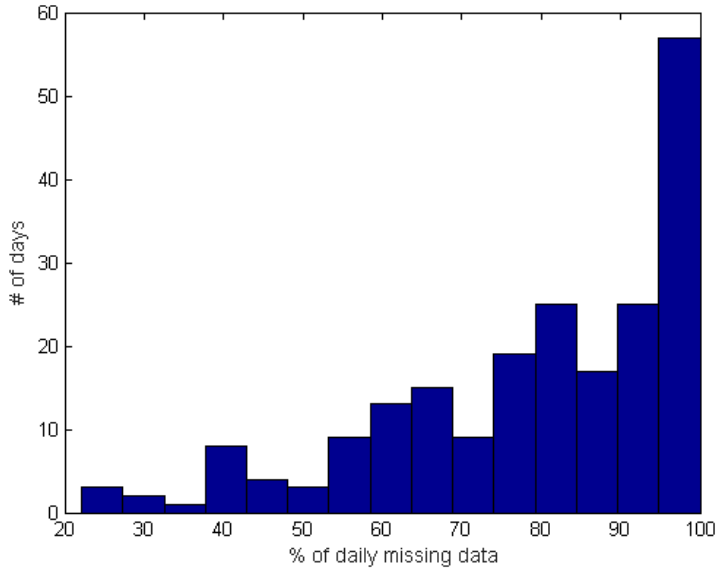


Figure 5: Daily missing data rate distribution

4.2 Covariates

In order to improve calibration capability, several covariates are considered, including mixing height, accumulation of rain precipitation, land elevation and percentage of urban area. Each covariate is described within the next paragraphs.

4.2.1 Mixing height

Mixing height is the height above ground level through which relatively vigorous vertical mixing occurs. It is the height at which smoke will lose its buoyancy and stop rising. Low mixing heights mean that the air is generally stagnant with very little vertical motion; pollutants are usually trapped near the ground surface. High mixing heights allow vertical mixing within a deep layer of the atmosphere and good dispersion of pollutants. For this reason, mixing height can be useful in explaining particulate matters concentration.

The mixing height data here considered are produced by the meteorological module of the EMCT modeling system of ARPA Piemonte (see Fassò and Cameletti 2009a for details) and are provided with daily temporal resolution and $0.1^\circ \times 0.2^\circ$ spatial resolution.

4.2.2 Rain precipitation

Also useful to explain particulate matters concentration is the accumulation of rain precipitation over the area of interest (Mossetti et al., 2005). In fact, persistent or heavy rain precipitations are usually followed by a reduction of airborne pollutants.

The Tropical Rainfall Measuring Mission (TRMM) is a joint U.S.-Japan satellite mission to monitor tropical and subtropical precipitation and to estimate its associated latent heating. The rainfall measuring instruments on the TRMM satellite include an electronically scanning radar operating at 13.8 GHz, a nine-channel passive microwave radiometer and a five-channel visible/infrared radiometer. The data-set considered is based on the 3B42 algorithm which is implemented to produce TRMM-adjusted merged-infrared precipitation and root-mean-square precipitation-error estimates. The final gridded, adjusted merged-IR precipitation (mm/hr) and RMS precipitation-error estimates have a 3-hour temporal resolution and a $0.25^\circ \times 0.25^\circ$ spatial resolution, here aggregated in order to obtain a daily temporal resolution.

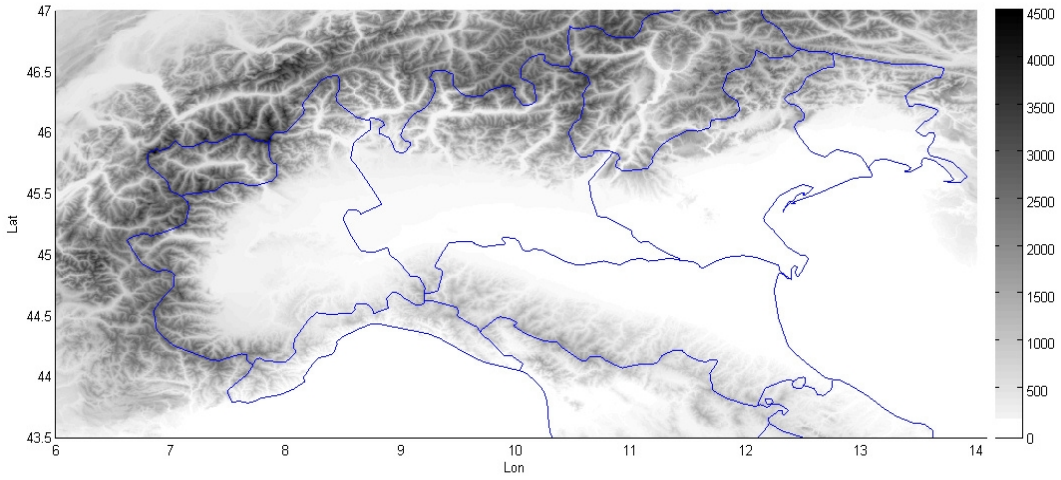


Figure 6: Land elevation $[m]$

It must be noted that spatial coverage extends from 50 degrees south to 50 degrees north latitude, which includes the area of interest.

4.2.3 Land elevation

Since particulate matter PM_{10} is heavier than other air compounds, it is inclined to occupy the layers of the atmosphere near the ground. Land elevation is then a candidate covariate for explaining the PM_{10} concentration. The elevation data here considered comes from the global digital elevation model called GTOPO30 and are characterized by a spatial resolution of 30" $\times$ 30". The image of Figure 6 represents the elevation data for the area of interest.

4.2.4 Percentage of urban area

It is well known that urban areas are a major source of particulate matters (Brizio et al., 2007). Indeed, the three main sources of pollutants are heating plants, industrial plants and vehicular traffic. In order to explain the PM_{10} concentration, the spatial distribution of urban areas over the region of interest can be considered.

The Modis product MOD12C1 provides the yearly dominant land cover type with spatial resolution $0.05° \times 0.05°$. In addition, it also provides a land

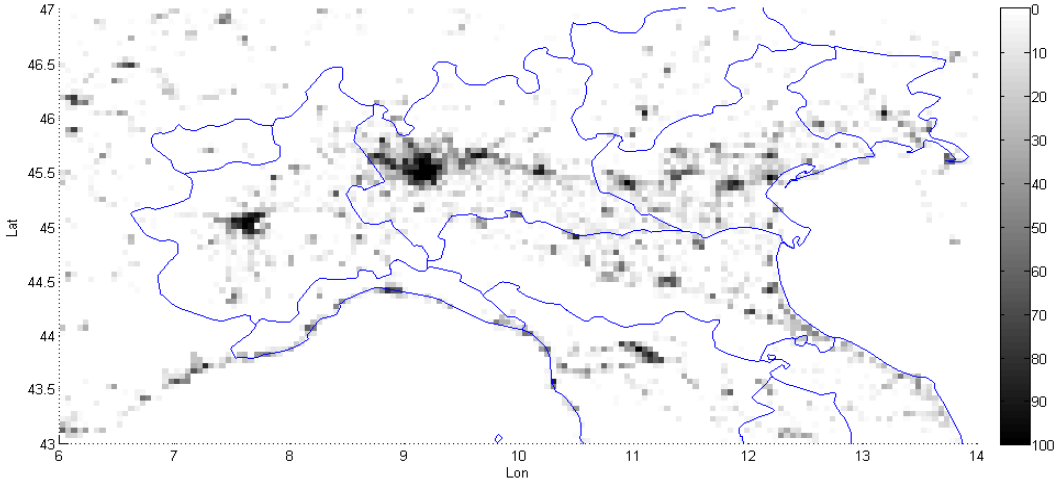


Figure 7: Percentage of urban area and built-up for the year 2006.

cover type assessment, percent distribution, and quality control information. In this work, it is considered the *Majority_Land_Cover_Type_1_Percent* layer of the 13-th class (urban and built-up) of the IGBP global classification scheme for the year 2006. In short, the covariate represents the percentage of urban area within each pixel of dimension $0.05° \times 0.05°$. The image of Figure 7 displays the urban areas distribution over the region of interest, from which the metropolitan areas of Milan and Turin stand out.

5 Estimation and results

The EM algorithm described in Section 3 is here adopted in order to estimate the vector parameter of the model of Section 2 by means of the data-set previously described.

5.1 Model and data settings

At the first stage of the model equations, $Y(s,t) = (Y_{AOT}(s,t), Y_{PM}(s,t))$ is a bi-variate vector accounting for the AOT and PM_{10} variables at site s and time t. The Gaussian instrumental error ε has then a 2×2 covariance

diagonal matrix $\Gamma_0 = \begin{bmatrix} \sigma^2_{AOT} & 0 \\ 0 & \sigma^2_{PM} \end{bmatrix}$. At the second stage, $X_j(s,t)$, is the vector of covariates at site s and time t for the $j-th$ variable, $j = 1, 2$, while β_j is the vector of parameters including a constant. The latent temporal state Z is univariate ($d = 1$) and the related transition matrix G is represented by a scalar g satisfying the condition $|g| < 1$. The 2×1 loading vector $\mathcal{K}$ is not estimated and its value is set to $\mathcal{K} = \begin{bmatrix} 1 & std(Y_{PM})/std(Y_{AOT}) \end{bmatrix}'$.

For what concerns the latent spatial process $\bar{W}$, it is considered a bivariate linear coregionalization model of one component ($c = 1$) with coregionalization weigthing matrix $V_1 = \begin{bmatrix} v_1^{AOT,AOT} & v_1^{AOT,PM} \\ v_1^{AOT,PM} & v_1^{PM,PM} \end{bmatrix}$. The correlation structure between different sites is based on the exponential model, that is $\rho(h, \theta_1) = \exp(-h/\theta_1)$.

In order to reduce data long tails, the logarithmic transformation for the AOT and PM_{10} variables is used. Moreover, all variables and covariates are standardized in order to avoid numerical instability.

5.2 EM estimation

Triggered with suitable starting values Ψ^0, the EM algorithm is used to estimate the vector parameter Ψ. The maximum iteration number of the algorithm is set to 250 while the exit condition is based on the convergence criterion $\frac{\left\|\Psi^{(k+1)} - \Psi^{(k)}\right\|}{\left\|\Psi^{(k)}\right\|} < 5 \cdot 10^{-4}$. It should be noted that the estimation procedure involves large variance-covariance matrices of dimension $N \times N$, with $N = n_{PM} + n_{AOT} = 1241$. For this reason, the estimation procedure is attained by a medium size cluster of three computers.

Table 1 reports the estimates together with the corresponding standard errors, based on the formulas in A.1.1-A.1.2, and the bounds of the 95% confidence intervals.

Examining the size of the confidence intervals reported in Table 1, it can be noted that all parameters except some β coefficients are characterized by a good level of estimation accuracy. By excluding the $\beta_{AOT,rain}$ coefficient , that seems to suggest that rain does not explain AOT variability, all the other β coefficients are characterized by the expected sign: positive for the percentage of urban area and negative for mixing height, elevation and rain.

The temporal coefficient g is within the stationarity range and its positive value confirms the temporal persistence of particulate matters even after

adjusting for the considered covariates. The parameter $\theta_1 \simeq 160$ suggests a strong spatial correlation of the latent variable W, with a cross-correlation of 0.31 (derived from the matrix V_1) between the W latent variable over the AOT sites and the W variable over the PM sites.

6 Conclusions

This paper introduces the dynamical coregionalization model that, thanks to stable expectation-maximization formulas provided in quasi-closed form, is suitable to be estimated by maximum likelihood. The model is shown to be capable of managing large numbers of missing data, entirely heterotopic data and high dimensionality, such as in the real application of integrating satellite aerosol optical thickness data with ground level PM_{10} data. Also provided are computationally affordable formulas for the observed information matrix related with the parameters set.

	$\hat{\Psi}$	std	$95\%CIbounds$	
$\beta_{AOT,const}$	-0.256	0.106	-0.465	-0.047
$\beta_{AOT,MH}$	-0.041	0.008	-0.058	-0.025
$\beta_{AOT,Rain}$	0.011	0.010	-0.086	0.031
$\beta_{AOT,Ele}$	-0.356	0.006	-0.368	-0.343
$\beta_{AOT,Urb}$	0.020	0.002	0.017	0.023
$\beta_{PM,const}$	-0.052	0.104	-0.257	0.152
$\beta_{PM,MH}$	-0.049	0.011	-0.071	-0.028
$\beta_{PM,Rain}$	-0.038	0.010	-0.057	-0.018
$\beta_{PM,Ele}$	-0.118	0.005	-0.129	-0.108
$\beta_{PM,Urb}$	0.101	0.004	0.093	0.109
σ^2_{AOT}	0.047	0.001	0.045	0.049
σ^2_{PM}	0.190	0.002	0.186	0.194
g	0.745	0.035	0.677	0.813
σ^2_η	0.183	0.017	0.150	0.216
$v_1^{AOT,AOT}$	0.890	0.030	0.831	0.950
$v_1^{AOT,PM}$	0.180	0.015	0.150	0.209
$v_1^{PM,PM}$	0.387	0.015	0.358	0.417
θ_1	163.855	6.364	151.383	176.328

Table 1: Estimated parameters, standard deviations and confidence intervals

A Marginal likelihood and Hessian

A.1 No missing data

We first consider the case of no missing data. Since the direct likelihood computation is exceedingly high dimensional with inversion of matrixes $NT \times NT$, we use here the computation based on the Kalman filter which is computationally affordable as it requires T inversions of $N \times N$ dimensional matrices.

To do this, we adapt the standard results of, for example, Shumway and Stoffer (2006), §6.3, and define

$$\bar{\varepsilon}_t = \bar{\varepsilon}\left(\Psi\right)_t = Y_t - X_t\beta - KZ_t^{t-1}$$

and

$$\Sigma_t = \Sigma\left(\Psi\right)_t = KP_t^{t-1}K' + \Sigma_0 + \Sigma_{\bar{W}}$$

where $Z_t^{t-1} = Z\left(\Psi\right)_t^{t-1}$ and $P_t^{t-1} = P\left(\Psi\right)_t^{t-1}$ are the Kalman filter outputs. Hence, the loglikelihood function, ignoring an additive constant, is given by

$$-2lnL_Y\left(\Psi\right) = \sum_{t=1}^{T} ln\left|\Sigma_t\right| + \sum_{t=1}^{T} \bar{\varepsilon}_t'\Sigma_t^{-1}\bar{\varepsilon}_t \tag{20}$$

A.1.1 Hessian

Using the observed information matrix denoted by $\hat{\mathfrak{I}}$, we have that the variance-covariance matrix of the estimates is given by

$$V\left(\hat{\Psi}\right) \cong \hat{\mathfrak{I}}^{-1}$$

where

$$\begin{aligned}\hat{\mathfrak{I}}_{i,j} \cong & \sum_{t=1}^{T} \left(\partial_i\bar{\varepsilon}_t'\right)\Sigma_t^{-1}\left(\partial_j\bar{\varepsilon}_t\right) + \\ & + \frac{1}{2}tr\left(\Sigma_t^{-1}\left(\partial_i\Sigma_t\right)\Sigma_t^{-1}\left(\partial_j\Sigma_t\right)\right) \\ & + \frac{1}{4}tr\left(\Sigma_t^{-1}\left(\partial_i\Sigma_t\right)\right)tr\left(\Sigma_t^{-1}\left(\partial_j\Sigma_t\right)\right)\end{aligned} \tag{21}$$

and the involved derivatives are given in the sequel.

A.1.2 Derivatives

Extending Shumway and Stoffer (2006) p. 408, we have the following recursions

1. $\partial_i \bar{\varepsilon}_t = -X_t \partial_i \beta - K \partial_i Z_t^{t-1}$
2. $\partial_i Z_t^{t-1} = (\partial_i G) Z_{t-1}^{t-2} + G \partial_i Z_{t-1}^{t-2} + (\partial_i J_{t-1}) \bar{\varepsilon}_{t-1} + J_{t-1} \partial_i \bar{\varepsilon}_{t-1}$
3. $\partial_i \Sigma_t = K \left(\partial_i P_t^{t-1} \right) K' + \sum_{p=0}^{c} \partial_i \Sigma_p$
4. $\partial_i J_t = \left((\partial_i G) P_t^{t-1} K' + G \left(\partial_i P_t^{t-1} \right) K' - J_t \left(\partial_i \Sigma_t \right) \right) \Sigma_t^{-1}$
5. $$\begin{aligned} \partial_i P_t^{t-1} = & (\partial_i G) (I - J_{t-1} K) P_{t-1}^{t-2} G' - G (\partial_i J_{t-1}) K P_{t-1}^{t-2} G' \\ & + G (I - J_{t-1} K) \left(\partial_i P_{t-1}^{t-2} \right) G' + G (I - J_{t-1} K) \left(P_{t-1}^{t-2} \right) (\partial_i G') \\ & + \partial_i \Sigma_\eta \end{aligned}$$

In order to compute the above recursions we need the closed form for the derivatives which are computed below, with $d = 1$, zero initial values and exponential spatial correlations. Hence, recalling definition of Ψ given by equation (5) and Γ_0 from equation (2), we have the following cases:

- $\partial_i \beta_j = 1$ if $\Psi_i = \beta_j$ and $\partial_i \beta_j = 0$ else;
- $\partial_i G = 1$ if $\Psi_i = G$ and $\partial_i G = 0$ else;
- $\partial_i \Sigma_\eta = 1$ if $\Psi_i = \Sigma_\eta$ and $\partial_i \Sigma_\eta = 0$ else;
- $\left(\partial_i \Sigma_0 \right)_{j,j} = 1$ if $\Psi_i = \sigma_{j'}^2$ and $\sum_{r=1}^{j'-1} n_r \leq j \leq \sum_{r=1}^{j'} n_r$ else $\left(\partial_i \Sigma_0 \right)_{j,m} = 0$;
- For $p = 1, ..., c$, let
  - $\left(\partial_i \Sigma_p \right)_{r,r'} = \left(\rho_p \left(H^{j,j'}, \theta_p \right) \right)_{s,s'}$ if $\Psi_i = v_p^{j,j'}$ and (r, r') correspond to (s, s'),
  - $\left(\partial_i \Sigma_p \right)_{r,r'} = \left(v_p^{j,j'} \frac{H^{j,j'}}{\theta_p^2} \circledast \rho_p \left(H^{j,j'}, \theta_p \right) \right)_{s,s'}$ if $\Psi_i = \theta_p$, (r, r') correspond to (s, s') and we used the Hadamar or elementwise product,
  - $\left(\partial_i \Sigma_p \right)_{r,r'} = 0$ else.

A.2 Missing data

As in section A.1, we get a computable formula, thanks to the Kalman filter approach. To do this we define

$$\bar{\varepsilon}_t^{(1)} = Y_t^{(1)} - X_t^{(1)}\beta - K^{(1)}Z_t^{t-1}$$

and

$$\Sigma_t^{(1)} = K^{(1)}P_t^{t-1}K^{(1)\prime} + \Sigma_0^{(1)} + \Sigma_{\bar{W}}^{(1)}$$

where the matrices with exponent (1) are obtained by the observed data selector operator.

Hence, as above, we get

$$-2lnL_{Y^{(1)}}(\Psi) = \sum_{t=1}^{T} ln\left|\Sigma_t^{(1)}\right| + \sum_{t=1}^{T} \bar{\varepsilon}_t^{(1)\prime}\left(\Sigma_t^{(1)}\right)^{-1}\bar{\varepsilon}_t^{(1)} \tag{22}$$

A.3 Proof of formula 5 paragraph A.1.2

From (6.20) of Shumway and Stoffer (2006) we have

$$P_t^{t-1} = GP_{t-1}^{t-1}G' + \Sigma_\eta$$

and, from (6.22), it follows that

$$P_t^{t-1} = G\left(I - J_{t-1}K\right)P_{t-1}^{t-2}G' + \Sigma_\eta$$

Hence formula 5 is obtained straightforwardly by derivation and using $\partial_i K = 0$.

B Estimation of the spatial latent components

For a fixed $t \in \{1, ..., T\}$ and for each $p = 1, ..., c$, $W_{p,t}$ and Y_t have a joint multivariate normal distribution and:

$$\text{cov}\left(W_{p,t}, Y_t\right) = \text{cov}(W_{p,t}, W_{p,t}) = \Sigma_p$$

If, at time t, some observations in Y_t are missing, $Y_t^{(1)}$ can be obtained by downsampling the variable Y_t: $Y_t^{(1)} = L_t Y_t$, where L_t is the corresponding selection matrix. In this case,

$$\operatorname{cov}\left(W_{p,t}, Y_t^{(1)}\right) = \Sigma_p L_t'$$

From the properties of the multivariate normal distribution, it follows that the conditional expectation and conditional covariance matrix of $W_{p,t}$ given the observed data $Y_t^{(1)}$ are:

$$\begin{aligned}
W_{p,t}^T &= E(W_{p,t} \mid Y^{(1)}) \\
&= E\left[E\left(W_{p,t} \mid Y^{(1)}, Z\right) \mid Y^{(1)}\right] \\
&= E\left[\hat{W}_{p,t} \mid Y^{(1)}\right] \\
&= E\left[\left(\Sigma_p L_t'\right)\left(L_t \Sigma L_t'\right)^{-1}\left[L_t\left(Y_t - X_t\beta - KZ_t\right)\right] \mid Y^{(1)}\right] \\
&= \left(\Sigma_p L_t'\right)\left(L_t \Sigma L_t'\right)^{-1}\left[L_t\left(Y_t - X_t\beta - KZ_t^T\right)\right]
\end{aligned}$$

where

$$\Sigma = \operatorname{Var}(Y_t|Z_t) = \sum_{p=0}^{c} \Sigma_p$$

and

$$\begin{aligned}
A_{p,t}^T &= \operatorname{Var}(W_{p,t} \mid Y^{(1)}) = \\
&= \operatorname{Var}\left[E\left(W_{p,t} \mid Y^{(1)}, Z\right) \mid Y^{(1)}\right] + E\left[\operatorname{Var}\left(W_{p,t} \mid Y^{(1)}, Z\right) |Y^{(1)}\right] \\
&= \operatorname{Var}\left[\left(\Sigma_p L_t'\right)\left(L_t \Sigma L_t'\right)^{-1}\left[L_t\left(Y_t - X_t\beta - KZ_t\right)\right] \mid Y^{(1)}\right] + \\
&+ \left[\Sigma_p - \left(\Sigma_p L_t'\right)\left(L_t \Sigma L_t'\right)^{-1}\left(L_t \Sigma_p\right)\right] \\
&= \left(\Sigma_p L_t'\right)\left(L_t \Sigma L_t'\right)^{-1}\left[L_t\left(KP_t^T K'\right) L_t'\right]\left(L_t \Sigma L_t'\right)^{-1}\left(L_t \Sigma_p\right) + \\
&+ \left[\Sigma_p - \left(\Sigma_p L_t'\right)\left(L_t \Sigma L_t'\right)^{-1}\left(L_t \Sigma_p\right)\right]
\end{aligned}$$

References

Amisigo, B. A. and Van De Giesen, N. C. (2005) Using a spatio-temporal dynamic state-space model with the EM algorithm to patch gaps in daily riverflow series, Hydrology and Earth System Sciences, 9

Brizio E, Genon G, Borsarelli S, PM emissions in an urban context, American Journal Of Environmental Sciences, pp. 166-174, 2007, Vol. 3 (3), ISSN: 1553-345X

De Iaco S, Palma M, Posa D. 2005. Modeling and prediction of multivariate space-time random fields. Computational Statistics and Data Analysis **48**: 525-547

Durbin J. and Koopman S.J. (2001) Time series analysis by state space methods. Oxford University Press

Fassò A, Cameletti M. 2009a. A unified statistical approach for simulation, modelling, analysis and mapping of environmental data. Simulation: Transactions of the Society for Modeling and Simulation International, accepted. OnlineFirst, doi:10.1177/0037549709102150

Fassò A, Cameletti M. 2009b. The EM algorithm in a distributed computing environment for modelling environmental space-time data. Environmental Modelling & Software **24**: 1027-1035

Fassò A, Finazzi F, D'Ariano C. (2009) Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. Submitted

Jost G.Heuvelink, G.B.M. Papritz A. (2005) Analysing the space–time distribution of soil water storage of a forest ecosystem using spatio-temporal kriging. Geoderma, 128. 258– 273

Kassteele, J. van de, Koelemeijer, R.B.A., Dekkers, A.L.M., Schaap, M. , Homan, C.D., and Stein, A. (2006). Statistical mapping of PM10 concentrations over Western Europe using secondary information from dispersion modeling and MODIS satellite observations, Stochastic environmental research and risk assessment , 21, p. 183 - 194

Koelemeijer, R.B.A., Homan, C.D., and Matthijsen, J. (2006).Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. Atmospheric Environment, 40

Liu C, Koike K. 2007. Extending Multivariate Space-Time Geostatistics for Environmental Data Analysis. Mathematical Geology **39**: 289–305

Lutkephol (1993) Introduction to multiple time series analysis. Springer. New York

Mossetti S, Angius S, Angelino E, Assessing the impact of particulate matter sources in the Milan urban area, International Journal of Environment and Pollution, Volume 24, Numbers 1-4, 18 July 2005 , pp. 247-259(13)

Rouhani S, Wackernagel H. 1990. Multivariate geostatistical approach to space-time data analysis: Water Resources Res. **36**: 585–591

Shumway R.H., Stoffer D.S. (2006) Time series analysis and its applications, with R examples. Springer. New York

Wang J, Christopher SA. 2003. Intercomparison between satellite-derived aerosol optical thickness and PM2.5 mass: Implications for air quality studies. Geophys. Res. Lett., 30, doi:10.1029/2003GL018174

Zhang H. 2007. Maximum-likelihood estimation for multivariate spatial linear coregionalization models. Environmetrics **18**: 125-139