

Air quality mapping using the dynamic coregionalization model

Alessandro Fassò and Francesco Finazzi

Abstract In this paper we study the ability of the multivariate dynamic coregionalization model for describing and mapping air quality in the "padano-veneto" region, North Italy. In particular, we consider spatio-temporal dynamic mapping of average daily concentration of airborne particulate matter. To do this, we combine air quality information of ground level particulate concentration measured by monitoring networks and satellite measurements of aerosol optical thickness (AOT). Although these data are given over a regular grid over space, they are missing under cloud, ice and snow conditions. Thanks to the EM algorithm which naturally covers the missing data, the model parameters are suitably estimated by maximum likelihood. Inference on the model, including parameter confidence intervals and likelihood ratio tests, is also suitably performed. Moreover, Kriging-like formulas are obtained and applied for dynamic mapping in presence of missing data.

Key words: EM algorithm; maximum likelihood estimation; multivariate spatio-temporal missing data; model selection; particulate matters; aerosol optical thickness.

1 Introduction

The multivariate spatio-temporal model, named dynamic coregionalization model, has recently been introduced by Fassò et al. (2009a and b). It is particularly suitable for environmental applications because it allows for maximum likelihood estimation and likelihood based inference even in rather large data sets.

Here we consider ground level monitoring network data on the concentration of airborne particulate matters PM_{10} . Although each station provides direct and reliable measurements of the PM_{10} concentration, they have an irregular spatial pattern. For

Dept. of IT and Mathematical Methods, University of Bergamo, Via Marconi 5, 24044 Dalmine BG, Italy. alessandro.fasso@unibg.it

this reason, a second variable is considered, namely the aerosol optical thickness (AOT), which is known to be related with the particulate matter concentration and is useful to improve mapping capability of the PM_{10} concentration over the area of interest, see e.g. Kasstele et al. (2006) and Koelemeijer et al. (2006).

Since the AOT data are given by remote sensing by means of the spectroradiometer known as MODIS, their quality is affected by weather and ground conditions, so in certain areas missing data are the majority. It follows that a stochastic model capable of using noisy data, when present, and to working well even in their absence, is needed. The above-mentioned dynamic coregionalization model is appropriate for this task because is a spatio-temporal state space model which "naturally" allows for missing data at both the parameter estimation stage and the missing data filling and mapping stage.

The paper is organized as follows. the dynamic coregionalization model is introduced in Section 2 and in Section 3 it is used to derive the spatial estimator which is a dynamic Kriging estimator and gives the dynamic mapping algorithm.

In Section 4 we discuss the application data concerning the "padano-veneto" area, North Italy. Using these data, in Section 5 the initial values of the EM algorithm are selected for the model fitting presented in Section 6, where mapping results are obtained using the plug-in approach.

2 The statistical model

In this paper we consider a special case of the model introduced by Fassò et al. (2009a and b). To do this, ignoring the missing data problem for a while, we denote the observed q -dimensional data at each point s in geographical space D and discrete time $t = 1, 2, \dots, T$ by $Y(s, t)$, and the multivariate map at time t by the N -dimensional vector $Y_t = (Y_1(s_{1,1}, t), \dots, Y_1(s_{1,n_1}, t), \dots, Y_q(s_{q,1}, t), \dots, Y_q(s_{q,n_q}, t))$. We suppose that Y_t is generated by the following equation

$$Y_t = X_t \beta + 1Z_t + \bar{W}_t + \varepsilon_t \quad (1)$$

where each variable Y_i , $i = 1, \dots, q$, is observed at sites $S_i = \{s_{i,1}, \dots, s_{i,n_i}\}$ and $N = n_1 + \dots + n_q$. In the sequel, we consider the completely heterotopic case for which different variables, Y_i and Y_j , are never observed in the same location so that S_i and S_j are disjointed sets.

The first term of the right hand side of equation (1), X_t , is given by a set of known covariates, including land features and meteorology. The second term, Z_t , is a common trend and covers for time dynamics. It is a stable Markov process given by $Z_t = gZ_{t-1} + \eta_t$, where η_t is a white noise with variance σ_η^2 . The vector of ones 1 allows Z_t to be conformable with respect to Y_t .

The third component is a linear coregionalization model. This means that the zero mean q -dimensional Gaussian process $\bar{W}(s, t) = (\bar{W}_1(s, t), \dots, \bar{W}_q(s, t))$ is defined by the so-called linear coregionalization model with c components, namely

$\bar{W}(s, t) = \sum_{p=1}^c W_p(s, t)$, where $W_p(s, t) = (W_{p,1}(s, t), \dots, W_{p,q}(s, t))$ is white noise in time but has a $q \times q$ spatial covariance and cross-covariance matrix function given by $\Gamma_p(h, \theta_p) = (\text{Cov}(W_{p,i}(s, t), W_{p,j}(s', t)))_{i,j=1,\dots,q} = V_p \rho_p(h, \theta_p)$, where $h = \|s - s'\|$ is the Euclidean distance between two sites $s, s' \in D$. For each $p = 1, \dots, c$, the $q \times q$ dimensional matrix V_p is positive semi-definite and $\rho_p(h, \theta_p)$ is a valid correlation function, characterized by the parameter vector θ_p . In particular, the exponential model is considered for the application of Section 6. In addition, the processes $W_p(s, t)$ and $W_{p'}(s, t)$ are uncorrelated for $p \neq p'$, so that the multivariate $q \times q$ covariance matrix for $\bar{W}(s, t)$ is given by $\Gamma_{\bar{W}}(h, \theta_1, \dots, \theta_c) = \sum_{p=1}^c V_p \rho_p(h, \theta_p)$. Finally, the Gaussian error ε is a white noise process with diagonal variance-covariance matrix. The parameters to be estimated are collected in the vector $\Psi = \text{vec}^*(\beta; g, \sigma_\eta^2; V_1, \theta_1, \dots, V_c, \theta_c)$, where vec^* is the vectorizing operator which ignores duplicate elements.

3 Dynamic mapping

The aim of dynamic mapping is to predict the variables $Y_i(S_0, t)$, $i = 1, \dots, q$, for each $t = 1, 2, \dots, T$ over the set of sites S_0 , with $S_0 \cap S_t = \emptyset$. The dynamic coregionalization model introduced by Fassò et al. (2009a and b) provides a natural way to solve the mapping problem since $Y_i(S_0, t)$ can be treated as missing data.

If $\hat{\Psi}$ is the vector parameter given by the EM algorithm, then $Y_i(S_0, t)$ is estimated by means of a plug-in approach as $Y_i(S_0, t) = X(S_0, t)\hat{\beta} + 1Z_t^T + \bar{W}_i(S_0, t)$, where $\hat{\beta} \in \hat{\Psi}$, Z_t^T is the Kalman smoother output and $\bar{W}_i(S_0, t)$ is selected from

$$\bar{W}(S_0 \cup S, t) = (\Sigma_{\bar{W}} L_t') (L_t \Sigma L_t')^{-1} \left[L_t \left(Y(S_0 \cup S, t) - X(S_0 \cup S, t)\hat{\beta} - 1Z_t^T \right) \right]$$

where $\Sigma_{\bar{W}} = \Sigma_{\bar{W}}(S_0 \cup S, \hat{\Psi})$ is the spatial covariance matrix of the process $\bar{W}$ extended over the sites in $S_0 \cup S$, with $S = \bigcup_{i=1}^q S_i$, $\Sigma = \Sigma_{\bar{W}} + \Sigma_\varepsilon$, Σ_ε is the variance-covariance matrix of the Gaussian error ε_t and L_t is the selection matrix of the non-missing data at time t . The vector $\bar{W}_i(S_0, t)$ is obtained from $\bar{W}(S_0 \cup S, t)$ by only considering the sites in S_0 .

It should be noted that, even if the matrix $\Sigma_{\bar{W}_i}$ can be very large, the $L_t \Sigma L_t'$ matrix to be inverted has the dimension of the non-missing data at time t , which is independent from $|S_0|$.

The prediction variance-covariance matrix of $Y_i(S_0, t)$ is given by $\Sigma_{Y_i}(S_0, t) = 1P_t^T 1' + \bar{A}_i(S_0, t)$, where P_t^T is the estimation variance of Z_t given by the Kalman smoother and $\bar{A}_i(S_0, t)$ is selected from

$$\begin{aligned} \bar{A}(S_0 \cup S, t) = & (\Sigma_{\bar{W}} L_t') (L_t \Sigma L_t')^{-1} [L_t (1P_t^T 1') L_t'] (L_t \Sigma L_t')^{-1} (L_t \Sigma_{\bar{W}}) + \\ & + \left[\Sigma_{\bar{W}} - (\Sigma_{\bar{W}} L_t') (L_t \Sigma L_t')^{-1} (L_t \Sigma_{\bar{W}}) \right] \end{aligned}$$

4 The case study

Here we consider daily air quality data, from March to September 2006, for the Italian region known as the "padano-veneto" area, sized about $600km \times 330km$, surrounded by the Alps on the northern and western sides, by the Apennines to the south and by a plain descending down towards the Adriatic sea to the east. The overall area is subject to a heavy anthropic load as it is characterized by the presence of large, densely populated urban centres and metropolitan areas with a busy motorway network.

The ground level data on particulate matters PM_{10} have been collected by a ground-level network composed of $n_2 = 107$ monitoring stations.

AOT data are daily collected by NASA Terra and Aqua satellites and provided over a $n_1 = 54 \times 21$ regularly spaced grid. These data are affected by structural missing data as cloud and ice conditions prevent remote sensing. For the seven months in hand, the average missing data rate is 73%, hence the state space approach to missing data is called for.

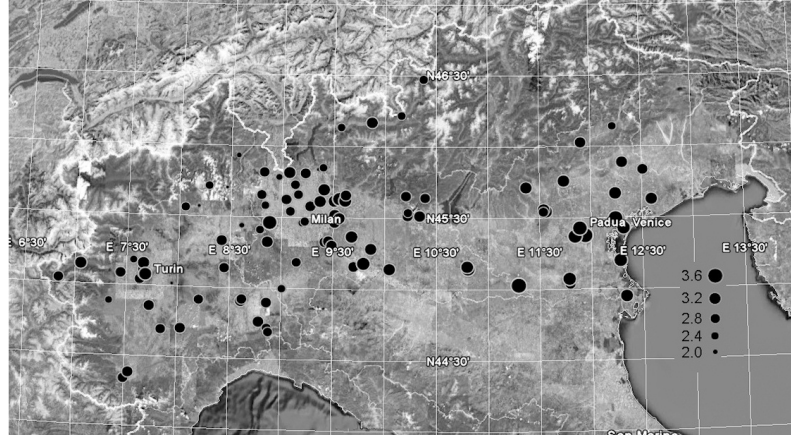


Fig. 1 Log-scale ground-level PM_{10} concentrations, 16th July 2006.

In order to improve calibration capability, several covariates are considered, including mixing height, accumulation of rain precipitation, land elevation and percentage of urban area.

5 EM estimation and initial values assessment

The EM algorithm is considered, which has been used by Fassò and Cameletti (2009a and b) for univariate spatio-temporal environmental data and by Zhang

Table 1 Estimated parameters, standard deviations and confidence intervals

	$c = 1$ ($\log L = 21584$)				$c = 2$ ($\log L = 22439$)			
	$\hat{\Psi}$	std	95% CI		$\hat{\Psi}$	std	95% CI	
$\beta_{AOT,const}$	-0.360	0.141	-0.636	-0.085	-0.359	0.153	-0.659	-0.059
$\beta_{AOT,MH}$	-0.143	0.009	-0.160	-0.127	-0.146	0.009	-0.162	-0.129
$\beta_{AOT,Ele}$	-0.309	0.006	-0.321	-0.297	-0.217	0.006	-0.230	-0.205
$\beta_{AOT,Urb}$	0.020	0.002	0.017	0.023	0.019	0.002	0.016	0.022
$\beta_{AOT,Rain}$	0.118	0.010	0.099	0.138	0.101	0.010	0.082	0.121
$\beta_{PM,const}$	-0.085	0.138	-0.355	0.186	-0.162	0.150	-0.457	0.133
$\beta_{PM,MH}$	-0.061	0.011	-0.082	-0.039	-0.056	0.011	-0.077	-0.034
$\beta_{PM,Ele}$	-0.118	0.005	-0.129	-0.108	-0.115	0.005	-0.125	-0.104
$\beta_{PM,Urb}$	0.102	0.004	0.094	0.109	0.100	0.004	0.092	0.108
$\beta_{PM,Rain}$	-0.028	0.010	-0.047	-0.010	-0.030	0.010	-0.049	-0.011
$\sigma_{\varepsilon,AOT}^2$	0.042	0.001	0.040	0.044	0.028	0.002	0.026	0.031
$\sigma_{\varepsilon,PM}^2$	0.192	0.002	0.188	0.195	0.189	0.002	0.184	0.194
g	0.883	0.029	0.826	0.939	0.894	0.031	0.834	0.954
σ_{η}^2	0.081	0.012	0.058	0.104	0.072	0.014	0.045	0.099
v_1^{AOT}	0.952	0.031	0.891	1.013	0.825	0.092	0.644	1.005
v_1^{PM}	0.389	0.015	0.360	0.418	0.343	0.038	0.268	0.418
$v_1^{AOT,PM}$	0.181	0.015	0.151	0.210	0.173	0.034	0.107	0.239
θ_1	168.051	6.212	155.881	180.212	238.585	34.006	171.944	305.239
v_2^{AOT}					0.165	0.084	0.001	0.329
v_2^{PM}					0.066	0.034	0.000	0.132
$v_2^{AOT,PM}$					0.031	0.020	-0.007	0.069
θ_2					55.37	15.44	25.10	85.64

(2007) for the static linear coregionalization model. These results have been extended by Fassò et al. (2009a and b) for the dynamic coregionalization model with missing values.

The application of the case study is an example that, despite the large amount of data involved, the above EM algorithm gives reliable results with an acceptable amount of computing time.

The initial values $\hat{\Psi}^{(0)}$ for initializing the EM algorithm are obtained as follows. The β coefficients are computed using the least squares on the overall data matrix. Then, using the univariate residuals $y(s, t) = Y(s, t) - X(s, t)\hat{\beta}^{(0)}$ for the ground-level monitoring data, we get estimates of the Markov inertia g by taking the spatial median of the "local moment estimator" based on the first order autoregression properties, that is $\hat{g}^{(0)}(s) = \frac{r_{y(s, \cdot)}(2)}{r_{y(s, \cdot)}(1)}$, where $r_{y(s, \cdot)}(h) = \frac{c_{y(s, \cdot)}(h)}{c_{y(s, \cdot)}(0)}$ and $c_{y(s, \cdot)}(h)$ are the empirical autocorrelation and autocovariance at lag h of $y(s, 1), \dots, y(s, T)$, respectively. Similarly, we have the initial value for σ_{η}^2 from the spatial median of $\hat{\sigma}_{\eta}^{(0)2}(s) = c_{y(s, \cdot)}(1) \frac{1 - \hat{g}^{(0)}(s)^2}{\hat{g}^{(0)}(s)}$. Moreover, the geostatistical parameters, namely V_p, θ_p and σ_{ε}^2 , are obtained by the temporal median of the non-linear least square estimates of an exponential bivariate covariogram model with nugget using the spatial bivariate data $y(\cdot, t)$.

6 Estimation and mapping results

This section reports the estimation and mapping results for the case study data previously described. It should be noted that both tasks are computer intensive since they involve the inversion of large matrices with dimension of order $N \times N$. All the results reported here are the output of algorithms implemented on a small size computer cluster of three nodes and a total of 24 CPU cores.

6.1 Estimation

Two models are considered, with number of coregionalization components equal to $c = 1$ and $c = 2$ respectively. Indeed, c is not easily estimated from the empirical variograms, especially in the case of high rates of missing data. For this reason, the number of coregionalization components is chosen a posteriori by means of the likelihood-ratio test.

Table 1 reports, for both models, the estimated parameters and their standard deviation obtained by evaluating an approximate Hessian matrix as described in Fassò et al. (2009b). Following Fassò and Cameletti (2009b), each model has been estimated five times by randomly perturbing the initial values $\hat{\Psi}^{(0)}$. As result of the estimation procedure, the parameter vector that gives rise to the highest marginal likelihood, as defined in Fassò et al. (2009b), is considered. This practice allows the effect of the local-minima convergence typical of the EM algorithm to be reduced.

By analysing the results of Table 1, it can be observed that the values of the β coefficients related to the set of covariates, are quite the same for the two models, as expected. This is also true for the temporal dynamic parameters g and σ_η^2 . The positive value of g reflects the temporal persistence of particulate matters.

On the contrary, significant differences can be noted for the geostatistical parameters V_p , θ_p and σ_ε^2 . While the coregionalization model based on one component identifies a large-scale spatial process ($\theta_1 \simeq 170$ km), the two-components model is able to discriminate a large-scale process ($\theta_1 \simeq 240$ km) and a medium-scale process ($\theta_2 \simeq 55$ km) characterized by a lower variance. Moreover, the unexplained variance σ_ε^2 is lower for the model with $c = 2$. Also note that, for this model, the cross-correlation between PM_{10} and AOT is 0.32 for the first coregionalization component and 0.30 for the second. However, the rather large confidence intervals on the element of the coregionalization matrix V_2 suggest to implement further analysis on the spatial structure of the second component.

By considering the two models as nested, the hypothesis for θ_2 and V_2 to be zero can be tested by means of the likelihood-ratio test as mentioned before. The test, based on a χ^2 distribution with 4 degrees of freedom, favours the unrestricted model ($c = 2$) with a p-value close to zero.

6.2 Dynamical mapping

To assess air quality for the "padano-veneto" region, the PM_{10} concentration is dynamically mapped on a $0.05^\circ \times 0.05^\circ$ resolution grid over the area of interest for each day of the seven months. As an example, the 16th of July 2006 is considered here. The pictures in Fig.1 shows, on log-scale, the PM_{10} concentrations collected from the ground-level network. For the day considered, the AOT missing data rate is 25%.

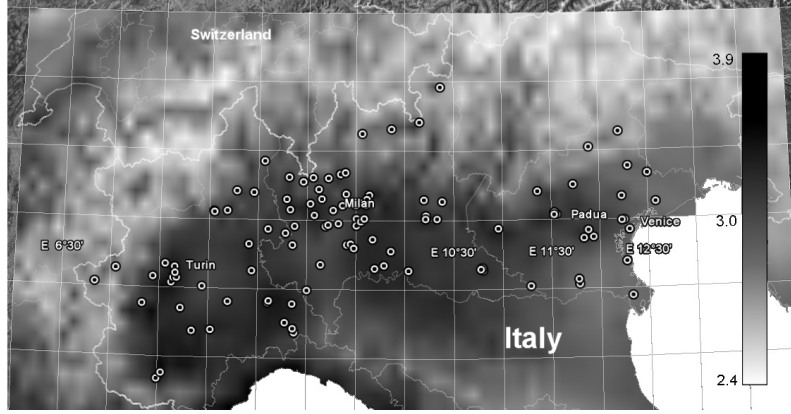


Fig. 2 Mapped log-scale PM_{10} concentrations - 16th July 2006

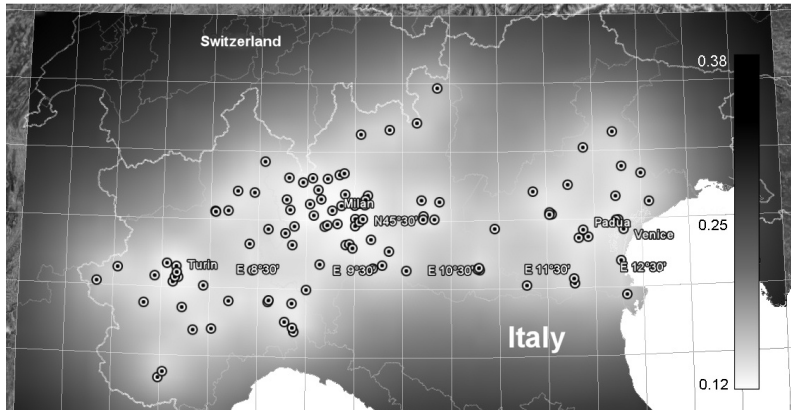


Fig. 3 Mapped PM_{10} concentration prediction variance - 16th July 2006

The mapped log-scale PM_{10} concentration and its prediction variance are depicted in Fig.2 and Fig.3, respectively. It is worthwhile to note that the spatial pattern

of the prediction variance emulates the spatial structure of the ground-level monitoring network. Indeed, the prediction variance is higher near the borders of the geographical space considered. The effect of the *AOT* missingness on the prediction variance can be noted from the graph in Fig.4. As expected, the lower the missing data rate the lower the prediction variance. Moreover, note that the prediction variance has an upper bound of about 0.28, irrespective of the *AOT* missingness.

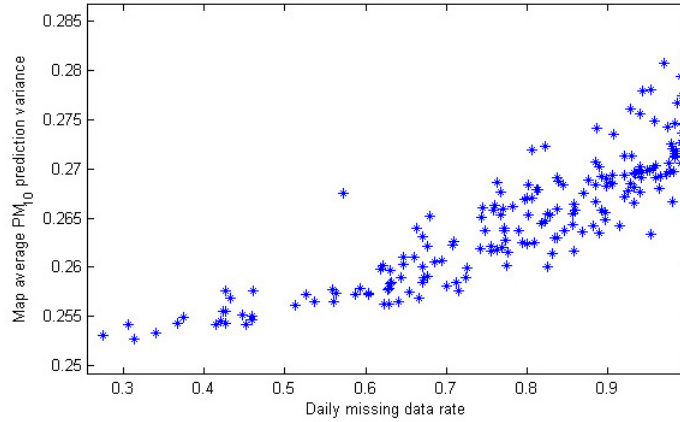


Fig. 4 Map average PM_{10} prediction variance with respect to daily *AOT* missing data rate

References

1. 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*, doi:10.1177/0037549709102150.
2. 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.
3. Fassò, A., Finazzi, F., D'Ariano, C. (2009a). Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data. WP n.15/MS - 2009, Dept. IT and Mathematical Methods, University of Bergamo (www.unibg.it).
4. Fassò, A., Finazzi, F., D'Ariano, C. (2009b). Integrating satellite and ground level data for air quality monitoring and dynamical mapping. GRASPA Working Paper n.34 (www.graspa.org).
5. Kasstele, J., Van de Koelemeijer, R.B.A., Dekkers, A.L.M., Schaap, M., Homan, C.D., 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**, 183 - 194.
6. Koelemeijer, R.B.A., Homan, C.D., Matthijsen, J. (2006). Comparison of spatial and temporal variations of aerosol optical thickness and particulate matter over Europe. *Atmospheric Environment*, **40**.
7. Zhang, H. (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics*, **18**, 125-139.