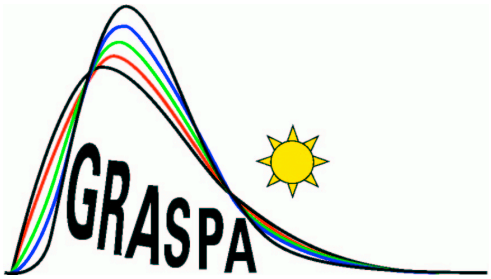
ISSN: 2037-7738
GRASPA Working Papers
[online]

Web Working Papers
by
The Italian Group of Environmental Statistics



Gruppo di Ricerca per le Applicazioni della Statistica
ai Problemi Ambientali

www.graspa.org

The dynamic coregionalization model with application to air quality remote sensing

Alessandro Fassò, Francesco Finazzi

GRASPA Working paper n.42, December 2010

Proceedings of the 25th International Workshop on Statistical Modelling

July 5-9, 2010

Glasgow

Adrian W. Bowman
(editor)

Proceedings of the 25th International Workshop on Statistical Modelling.
Glasgow, July 5-9, 2010
Adrian W Bowman, editor
Glasgow 2010.

Editor:
Adrian Bowman
Department of Statistics
The University of Glasgow
Glasgow G12 8QQ Scotland, UK
adrian@stats.gla.ac.uk

Printed by The University of Glasgow Print Unit

The dynamic coregionalization model with application to air quality remote sensing

Alessandro Fassò[1], Francesco Finazzi[1]

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

Abstract: In this paper, we discuss the dynamic coregionalization model and its capability for model selection inference and interpretation in relation to spatio-temporal dynamic calibration and mapping of daily concentration of airborne particulate matter. To do this, we consider the problem of joint modelling ground level concentration data and satellite measurements of aerosol optical thickness (AOT), which are rarely available. The maximum likelihood estimation for the large data set related to the "padano-veneto" region, North Italy, with missing data is covered by the stable EM algorithm and implemented on a small size computer cluster.

Keywords: EM algorithm; maximum likelihood estimation; multivariate spatio-temporal missing data; particulate matters; aerosol optical thickness.

1 Introduction

The increasing availability of large datasets on multivariate spatio-temporal data parallels the need for statistical models which are flexible enough for covering the underlying complexity and can be estimated by means of well founded inferential techniques. The dynamic coregionalization model, recently proposed by Fassò et al. (2009), has these advantages as it allows modelling of complex multivariate spatio-temporal dynamics and performing maximum likelihood parameter estimation by means of the EM algorithm. Moreover, it naturally covers large amounts of "structural" missing data. This is particularly important for remote sensing applications where, under cloud conditions, the satellite data are missing.

2 Dynamic coregionalization model

We consider multivariate data which are cross-correlated at each point in geographical space, say D, and discrete time $t = 1, 2, ..., T$. Each variable is allowed to have a different spatial correlation and/or serial correlation over time. This is different to standard application of the coregionalization model to spatio-temporal data where it is commonly considered continuous

time, see e.g. De Iaco et al. (2005). In other words, we suppose that at time t, the observed data follow the equation

$$Y_t = X_t\beta + KZ_t + \bar{W}_t + \varepsilon_t \tag{1}$$

Ignoring for a while missing data, the observed Y_t is a $N-$dimensional vector containing the maps related to the q observed variables. Namely $Y_t = (Y_1(S_1,t),...,Y_q(S_q,t)) = (Y_i(s_{i,j},t)))'_{j=1,...,n_i,i=1,...,q}$ so that each variable Y_i is observed at sites $S_i = \{s_{i,1},...,s_{i,n_i}\}$ and $N = n_1 + ... + n_q$.
The first term of the RHS of equation (1), X, is given by a set of known covariates. The second term, Z, covers for the time dynamics being a stable multivariate Markov process in the form $Z_t = GZ_{t-1} + \eta_t$, $\eta_t \sim N(0,\Sigma_\eta)$. The Gaussian error ε is a white noise process with diagonal variance-covariance matrix Γ_0 whose elements are $\sigma^2_{\varepsilon,i}$, $i = 1,...,q$.
Finally, the third term of RHS of equation (1) is a zero mean $q-$dimensional Gaussian process $\bar{W}(s,t) = \left(\bar{W}_1,...,\bar{W}_q\right)$ 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 $W_p(s,t) = (W_{p,1},...,W_{p,q})$ is white noise in time but correlated over space with a $q \times q$ covariance and cross-covariance matrix function given by $\Gamma_p(h,\theta_p) = (\mathrm{cov}(W_{p,i}(s),W_{p,j}(s')))_{i,j=1,...,q} = V_p\rho_p(h,\theta_p)$ where $h = \|s - s'\|$ is the Euclidean distance. 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, characterized by the parameter vector θ_p. In the sequel, the exponential correlation function is considered, namely $\rho_p(h,\theta_p) = \exp(-h/\theta_p)$. In addition, the processes W_p and $W_{p'}$ are uncorrelated so that the multivariate $q \times q$ covariance matrix for W is given by $\Gamma_{\bar{W}}(h,\theta_1,...,\theta_c) = \sum_{p=1}^{c}\Gamma_p(h,\theta_p) = \sum_{p=1}^{c} V_p\rho_p(h,\theta_p)$.
The model parameters are collected in the vector Ψ, which ignores duplications, namely $\Psi = vec^*(\beta,\Gamma_0;G,\Sigma_\eta,;V_1,\theta_1,...,V_c,\theta_c) = (\Psi_Y,\Psi_Z,\Psi_W)$.

3 Estimation and inference

Due to the Markovian assumption and to the space-time separability property of the model, the complete-data log-likelihood function takes the nice additive form

$$l(\Psi;Y,Z,W) = l(\Psi_Y;Y \mid Z,W) + l(\Psi_Z;Z) + l(\Psi_W;W) \tag{2}$$

However, (2) is not easily handled since Z is latent and Y is partially missing. The problem is overcome by considering the EM algorithm, already used by Fassò and Cameletti (2010) for univariate spatio-temporal environmental data and extended by Fassò et al. (2009) for the dynamic coregionalization model with missing data.
At the E-step of the algorithm, denoting by $Y^{(1)}$ the subset of actual observations, the expectation of the complete data log-likelihood under the parameter $\Psi^{(k)}$, conditionally to the observed data $Y^{(1)}$, is computed thanks to the iterated expectation theorem, that is

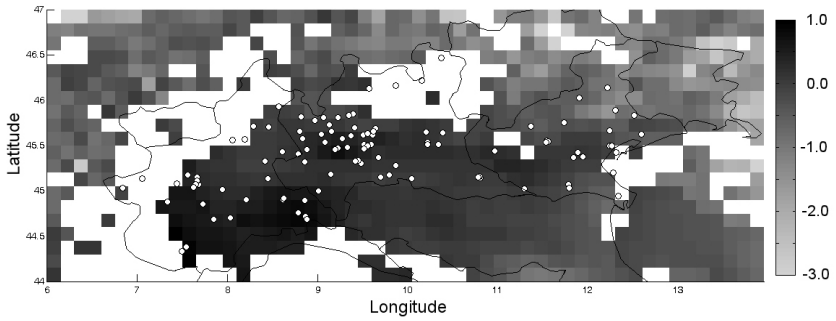


FIGURE 1. AOT standardized data (July 14th 2006) and ground level PM_{10} monitoring network sites (white circles).

$$Q\left(\Psi, \Psi^{(k)}\right) = E_{\Psi^{(k)}}\left[E_{\Psi^{(k)}}\left[l(\Psi; Y, Z, W) \mid Y, Z, W\right] \mid Y^{(1)}\right]$$

At the M-step, $Q\left(\Psi, \Psi^{(k)}\right)$ is maximized with respect to Ψ and $\Psi^{(k+1)}$ is chosen so that $\Psi^{(k+1)} = \arg\max Q\left(\Psi, \Psi^{(k)}\right)$. The solution of the maximization problem gives rise to quasi closed-form formulas for the update of Ψ, reported in detail in Fassò et al. (2009).
Since the EM algorithm is only guaranteed to converge to local maxima of the likelihood function, the whole estimation procedure is based on a set of EM estimation runs, each one characterized by different initial values for the parameter vector. Initial values are first evaluated through an estimation procedure based on the method of moments, detailed in Fassò and Finazzi (2010), and then locally perturbed by means of a random noise.
As the solution of the estimation procedure, the parameter vector $\hat{\Psi}$ is considered that gives rise to the maximum marginal log-likelihood $l_{Y^{(1)}}(\hat{\Psi})$. The role of the marginal log-likelihood, evaluated through the Kalman filter approach reported in Fassò et al. (2009), is important for computing likelihood-ratio tests and comparing nested models. Similarly, the estimated parameter vector $\hat{\Psi}$ is completed with standard deviations obtained by explicit recursive formulas for the Hessian matrix of the same marginal likelihood.

4 The case study

We consider ground-level data on concentration of airborne particulate matters PM_{10}, coming from $n_2 = 107$ monitoring stations. Although each station provides direct and reliable measures of the PM_{10} concentration, they have irregular spatial patterns. For this reason, a second variable is

model	M_0	M_1	M_2	M_3
$l_{Y^{(1)}}(\hat{\Psi})$	11271	21682	22543	22830
$Bias$	-0.0136	-0.0026	-0.0026	-0.0026
MSE	0.5004	0.2218	0.2214	0.2213

TABLE 1. Marginal log-likelihood and cross-validation results for models with $c = 0, ..., 3$ coregionalization components.

considered, namely the Aerosol Optical Thickness (AOT), which is known to be related with the particulate matters concentration and is useful to improve mapping capability of the PM_{10} concentration over the area of interest, see e.g. Koelemeijer et al. (2006).
AOT data are collected by the Terra and Aqua NASA satellites by means of the MODIS instrument (Moderate Resolution Imaging Spectroradiometer) and are provided with a spatial resolution of 10×10 km at nadir. The data set considered here covers the Italian region known as the padano-veneto area, bounded by a box of coordinates 44°N-6°E, 47°N-14°E, giving a daily data vector of 1134=54x21 elements, and the time period between March 2006 and September 2006 (see Figure 1). The daily average missing data rate for the AOT variable is 73% while it is 3% for the PM_{10}.
In order to improve calibration capability, several covariates are considered, including mixing height, accumulation of rain precipitation, land elevation, longitude of the site and percentage of urban area. PM_{10} concentrations and AOT measures are first log-transformed and then standardized, giving all variables with unit variance. Standardization is also applied to each covariate separately.

4.1 Model estimation and selection

In order to evaluate the role of the latent spatial variable W, models M_0, M_1, M_2 and M_3 are considered, with $c = 0, ..., 3$ coregionalization components respectively. It is worthwhile to note that, without coregionalization components, the spatial correlation function between sites is not directly modelled, though *lato sensu* a quota of the spatial correlation is covered by the covariates.
Models are estimated by means off the estimation procedure described in the previous section and compared by implementing likelihood-ratio tests between nested models. In order to evaluate the spatial prediction capability of each model, the leave-one-out crossvalidation method is applied over the PM_{10} sites S_2. Prediction bias and MSE are evaluated at each site $s_{2,j} \in S_2$. To do this, we estimate the model considering all data except the PM_{10} concentrations collected at $s_{2,j}$. The estimated model is then used to predict the PM_{10} concentration at $s_{2,j}$ for each day. Map average bias and map average MSE are reported in Table 1.

	$\hat{\beta}^{AOT}_{const}$	$\hat{\beta}^{AOT}_{MH}$	$\hat{\beta}^{AOT}_{Ele}$	$\hat{\beta}^{AOT}_{Urb}$	$\hat{\beta}^{AOT}_{Rain}$	$\hat{\beta}^{AOT}_{Long}$
value	−0.360	−0.143	−0.292	0.020	0.115	−0.005
std	0.140	0.009	0.006	0.002	0.010	0.001
	$\hat{\beta}^{PM}_{const}$	$\hat{\beta}^{PM}_{MH}$	$\hat{\beta}^{PM}_{Ele}$	$\hat{\beta}^{PM}_{Urb}$	$\hat{\beta}^{PM}_{Rain}$	$\hat{\beta}^{PM}_{Long}$
value	−0.097	−0.065	−0.133	0.106	−0.030	−0.133
std	0.136	0.010	0.006	0.004	0.011	0.005
	$\hat{\sigma}^2_{\varepsilon,AOT}$	$\hat{\sigma}^2_{\varepsilon,PM}$	$\hat{g}$	$\hat{\sigma}^2_{\eta}$		
value	0.041	0.192	0.880	0.084		
std	0.001	0.002	0.029	0.012		
	$\hat{v}^{AOT}_1$	$\hat{v}^{PM}_1$	$\hat{v}^{AOT,PM}_1$	$\hat{\theta}_1$		
value	0.923	0.367	0.177	162.194		
std	0.003	0.015	0.015	6.521		

TABLE 2. Estimated parameters and standard deviations for model M_1

4.2 Model interpretation

The map average MSEs of Table 1 suggest an improvement of the model predictive performance when the coregionalization variable $\bar{W}$ is considered. Indeed, the percentage of explained variance increases from 50 to 78 moving from M_0 to M_1. On the other hand, the performance is not significantly different when the number of coregionalization components is increased from one to either two or three.

If map prediction of the PM_{10} concentration is of concern, the more parsimonious M_1 model should be preferred, despite the likelihood-ratio test favouring, in this case, the unrestricted models with a near zero p-value.

Table 2 reports the estimated parameters along with their standard deviations. The $\hat{\beta}$ coefficients are directly comparable with each other since each covariate is standardized with respect to each variable. Note the opposite signs of $\hat{\beta}_{Rain}$, which is negative for PM, as precipitation usually reduces ground level concentrations, but is positive for AOT, due to the optical effect of those rare rainy days with non missing AOT data. Note also the difference in $\hat{\beta}_{Urb}$, which is related to the lower spatial resolution of AOT, so that single AOT pixels can include both urban and rural areas. Finally, the positive values of $\hat{\beta}$ suggest an east-west trend on the average PM_{10} concentration. In fact, the eastern side of the region considered is less urbanized and it is open to the winds from the Adriatic sea, while the western side is closed in by the Alps and is characterized by deficient air circulation. This aspect is confirmed by the positive sign of $\hat{g}$, related to the temporal dynamics. Net of the effect of covariates and time dynamics, the estimated cross-correlation between AOT and PM_{10}, based on matrix $\hat{V}_1$, is 0.30, which is consistent with marginal correlation of PM and AOT. The latent variable is also characterized by a persistent spatial correlation, described by the exponential correlation function with parameter $\hat{\theta}_1 \cong 162$

km.

5 Conclusions

We discussed the use of the dynamic coregionalization model in the framework of large dataset linear modelling for multivariate air quality data. Despite the large size of the data, the complex structure of the model and high rate of missing data, it is seen that this approach can be implemented with relatively standard computing facilities. Moreover, it covers in a natural way all inferential tools, including maximum likelihood estimation, classical likelihood inference, such as likelihood ratio tests, confidence intervals and crossvalidation. Of course, due to the large number of degrees of freedom, p-values have to be interpreted *cum grano salis*. Extensions to spatial nonstationarity and/or nonseparability can be naturally based on this model using the loess semiparametric approach of Bodnar and Schmid (2009) or the transformation based approach of Bruno et al. (2008).

References

Bruno F., Guttorp P., Sampson P.D., Cocchi D. (2008) A simple non-separable, non-stationary spatiotemporal model for ozone *Environmental and Ecological Statistics*. **16**, 515-529.

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.

Fassò, A., Cameletti, M. (2010). A unified statistical approach for simulation, modelling, analysis and mapping of environmental data. *Simulation*. **86**, 3, 139154.

Fassò, A., Finazzi, F. (2010). Statistical mapping of air quality by remote sensing: uncertainty and sensitivity to missing data. Submitted.

Fassò, A., Finazzi, F., D'Ariano, C. (2009) . Integrating satellite and ground level data for air quality monitoring and dynamical mapping. GRASPA Working Paper n.34. (www.graspa.org).

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**.

Bodnar, O., Schmid W. (2009). Nonlinear locally weighted kriging prediction for spatio-temporal environmental processes *Environmetrics*. DOI: 10.1002/env.1005