Statistical mapping of air quality by remote sensing

Uncertainty and sensitivity to missing data

Alessandro Fassò* and **Francesco Finazzi**
University of Bergamo – Dept. of Information Technology and Mathematical Methods
Bergamo – Italy
*alessandro.fasso@unibg.it

Abstract—**In this paper we consider the multivariate dynamic coregionalization model which has recently been introduced in environmental spatio-temporal statistics. The main modelling objective here is dynamic mapping of airborne particulate matters (PM_{10}) by merging measurements from an irregularly spaced ground level monitoring network and regularly spaced satellite measurements of aerosol optical thickness (AOT). Due to the fact that AOT measurements are not available under cloudy conditions, we have to manage a large amount of missing data. In principle this task is naturally handled by the state space representation of the model and the maximum likelihood estimation through the EM algorithm. After discussing the uncertainty sources of this model, we check the model sensitivity to missingness in the case of the "padano-veneto" region, North Italy, including the Alps. To do this, an extensive simulation campaign is performed with missing rate ranging from 0% to 90% and showing the reliability of the method for the case under study.**

Keywords: dynamic coregionalization model; aerosol optical thickness; multivariate spatio-temporal modeling; EM algorithm

I. INTRODUCTION

The recently introduced dynamic coregionalization model (DCM) is a statistical technique for multivariate spatio-temporal data which allows coverage of covariates, spatial correlation, temporal correlation and crosscorrelation among dynamic fields with missing data. Moreover, their parameters can easily be estimated even in large datasets by the maximum likelihood method thanks to a stable EM algorithm (see Fassò et al., 2009a and b).

Satellite measurement of aerosol optical thickness (AOT) has been proposed by various authors for assessing airborne particulate matter (PM_{10}) because of the positive correlation with ground level PM_{10}. Since these positive correlations (Wang, 2003) are not very high and AOT measurements are not available under cloudy conditions, the DCM is used as a calibration model able to work in absence of the primary information on AOT and to produce dynamic maps.

In order to understand the various sources of uncertainty which may be important in doing this, we consider the sensitivity of the mapping error to uncertainty in general and, in particular, to the rate of missing AOT data.

II. MODEL SETUP

Suppose that, at time $t = 1,2,...,T$, we have q different spatial variables $y_i(s,t)$ observed at sites $s \in S_i = \{s_{i,1},...,s_{i,n}\}$, $i = 1,...,q$. This gives the n_i-dimensional vectors Y_i and the overall N-dimensional vector $Y_t = (Y_{1,t},...,Y_{q,t})$, where $N = \sum_{i=1}^{q} n_i$. For example, in the present remote sensing application, we have $q=2$ variables, the ground level PM_{10} concentration and the AOT measurement, respectively. According to DCM, Y_t is generated by

$$Y_t = X_t\beta + Z_t + W_t + \varepsilon_t \qquad (1)$$

where X_t is given by a set of known covariates, including land features and meteorology. The second term, Z_t, is an unobserved trend with Markovian dynamics, given by $Z_t = gZ_{t-1} + \eta_t$, and Gaussian innovations η_t with variance σ_η^2. The third component, W_t, is given by an unobserved N-dimensional linear coregionalization model of order c, that is a zero-mean q-dimensional Gaussian process $W(s,t)$, which is white noise in time but has a $q \times q$ spatial covariance matrix generating function given by

$$\Gamma_W(h,\theta_1,...,\theta_c) = \sum_{p=1}^{c} V_p \rho_p(h,\theta_p)$$

where $h = \|s - s'\|$ is the Euclidean distance between two sites. All the matrices V_p are positive semi-definite and $\rho_p(h,\theta_p)$ are spatial correlation functions characterized by the parameter vector θ_p. In the sequel, the exponential model is considered, which is given by $\rho_p(h,\theta) = \exp(-h/\theta)$. Finally, the Gaussian error ε is a white noise process with variance $\sigma_{\varepsilon,i}^2$ for each variable y_i. The unknown model parameters are collected in the parameter set $\Psi = \{\beta, \sigma_\varepsilon^2; g, \sigma_\eta^2; V_1, \theta_1, ..., V_c, \theta_c\}$ and its estimate $\hat{\Psi}$ is obtained by the maximum likelihood method thanks to a version of the celebrated EM algorithm.

III. UNCERTAINTY ASSESSMENT

Here we consider the effect of the various sources of uncertainty in mapping PM_{10} ground level concentration. To do this, we first introduce the mapping algorithm which is based on plugging-in the estimated parameter set, say $\hat{\Psi}$, in the above DCM and taking conditional expectation. That is

$$\hat{y}_i(s,t) = X_i(s,t)\hat{\beta} + \hat{Z}(t) + \hat{W}_i(s,t) \quad (2)$$

where $\hat{Z}(t) = E_{\hat{\Psi}}\left(Z_t|Y_1,...,Y_T\right)$ is the Kalman smoother output discussed in detail by Fassò et al. (2009a) and $\hat{W}(s,t) = E_{\hat{\Psi}}\left(W(s,t)|Y_1,...,Y_T\right)$ is the kriging-like estimate of the spatial latent variable discussed in detail by Fassò and Finazzi (2010a).

Of course, assuming the model (1) as "the true data generating mechanism", the observation variance may be decomposed as follows

$$Var(y_i(s,t)) = Var\left(Z(t)\right) + \sum_{p=1}^{c} V_p(i,i)\rho_p + \sigma^2_{\varepsilon,i} \quad (3)$$

where the irreducible part is given by $\sigma^2_{\varepsilon,i}$. Nevertheless, Fassò and Cameletti (2010), considering the scalar version of DCM, proposed a prediction uncertainty decomposition which takes into account also the estimation uncertainty. This gives

$$E\left[\hat{y}_i(s,t) - y_i(s,t)\right]^2 \approx X_i(s,t)Var\left(\hat{\beta}\right)X_i(s,t)' + \\ + Var_{\hat{\Psi}}\left(\hat{Z}(t)\right) + Var_{\hat{\Psi}}\left(\hat{W}_i(s,t)\right) + \sigma^2_{\varepsilon,i} \quad (4)$$

This formula is approximated because it ignores the dependence among the components of $\hat{y}(s,t)$ and also ignores the influence of estimation uncertainty of $\hat{Z}$ and $\hat{W}$. Despite this, the same authors noticed that, for PM scalar modeling, both the estimation uncertainty $Var\left(\hat{\beta}\right)$ and the Kalman smoother error $Var\left(\hat{Z}(t)\right)$ are relatively small. Following this line, in the sequel of this short paper, we focus on two types of uncertainty.

The first type of uncertainty is related to the "modeled mapping error", which is an output of DCM and is given by the plug-in prediction variance $A_i(s,t) + \hat{P}(t)$, where $A_i(s,t) = Var_{\hat{\Psi}}\left(W_i(s,t)|Y_1,...,Y_T\right)$ and $\hat{P}(t) = Var_{\hat{\Psi}}\left(Z_t|Y_1,...,Y_T\right)$ is also given by the Kalman smoother. The prediction variance gives the local uncertainty for every estimate, resulting in an uncertainty map for every day. Fassò and Finazzi (2010a) give details for computing each term of the right-hand side of (4).

The second one is also able to cover for modeling errors and is given by comparing modeled values of PM_{10} concentrations with the ground level measures using the computer intensive crossvalidation technique, discussed in the following section V. In doing this, we will assess the "true bias" $E(y-\hat{y})$ and the "true mean square error" $E(y-\hat{y})^2$, in principle for every map pixel and every day or averaged in space and time.

IV. Model Fitting on Real Data

A. Data Structure

Data of PM_{10} concentrations and AOT measurements over the "padano-veneto" Italian region are considered. The region is about 600x330 km in size and, surrounded by the Alps and by the Apennines, is characterized by low air circulation. The time period considered goes from March to September 2006, including both spring and summer seasons.

The PM_{10} concentration is collected daily by a ground-level network of 107 monitoring stations scattered over the region. On the other hand, AOT data are collected by NASA Terra and Aqua satellites and provided daily over a 0.15x0.15° resolution grid.

The AOT missing-data rate is quite high, with a daily average of 73%. For this reason, naïve interpolation techniques are not suitable for missing data imputation, either over space or in time. The PM_{10} missing data rate, on the other hand, is much lower with a daily average of 3%.

Besides the relationship between PM_{10} and AOT, the PM_{10} concentration can be as well related to other variables, either meteorological, anthropological or morphological of the land. The set X of covariates considered in this work is composed of: mixing height, accumulation of rain precipitation, land elevation, longitude and percentage of urban area. The covariates are assumed to be known for each point in space and time without errors.

B. Model Estimation

The model considered here is a specialization of (1) of section II with $c=1$ coregionalization component. The choice of c is based on preliminary results obtained in Fassò and Finazzi (2010b).

PM_{10} concentrations and AOT measurements are first log-transformed, in order to reduce the skewness of the respective distributions, and then standardized giving all the y_i variables with unit variance. Standardization is also performed on each covariate.

The model parameter set is estimated by means of the EM algorithm, extensively used in spatial modeling when latent variables are considered (see for example Zhang, 2007). It should be noted that the estimation process calls for repeatedly solving large linear systems, with the number of equations proportional to the data size. In this particular case, where 107 PM_{10} concentrations and 54x21 AOT measurements are considered daily, the system matrix is 1241x1241 in size. For this reason, all the estimation algorithms have been implemented on a medium size computer cluster of three nodes and a total of 24 CPU cores.

The estimated $\hat{\beta}$ coefficients related to the covariate matrix X are reported in Table I along with their standard deviations.

TABLE I. Estimated Covariate $\hat{\beta}$ Coefficients

	const	Mixing height	Elevat.	%Urban	Rain	Long.
AOT	-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
PM_{10}	-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

As far as concern the latent temporal process Z, the estimated dynamic equation is $Z_t = 0.88 \cdot Z_{t-1} + \eta_t$, with $\eta_t \sim N(0, 0.084)$. The coregionalization matrix is equal to $\hat{V}_1 = \begin{bmatrix} 0.923 & 0.177 \\ 0.177 & 0.367 \end{bmatrix}$, while the parameter of the exponential correlation function is $\hat{\theta}_1 = 162.194$ km, suggesting a strong spatial correlation of the spatial latent variable W. The crosscorrelation between the AOT component of W and the PM_{10} component, evaluated from the matrix V_1, is around 0.30, which is consistent with preliminary analysis (Hoff and Sundar, 2009). Finally, the variances $\hat{\sigma}^2_{\varepsilon,i}$ of the Gaussian error ε are 0.041 and 0.192 for the AOT and the PM_{10} variable respectively. All the standard deviations related to the space-time dynamics parameters are less then 4% of their respective parameter value.

C. *Dynamical Mapping*

By considering the estimated parameters of the previous paragraph, the PM_{10} concentration is dynamically mapped on an 0.10°x0.10° resolution grid over the entire region by means of the plug-in approach described in section III.

For example, the result of the dynamical mapping procedure is reported in Fig.1, where the kriged PM_{10} concentration $\hat{y}_{PM}$ and its prediction variance $A_i(s,t) + \hat{P}(t)$ are depicted for one particular day. As expected, the prediction variance pattern mimes the spatial position of the PM_{10} monitoring stations which are depicted by the circles in Fig. 3.

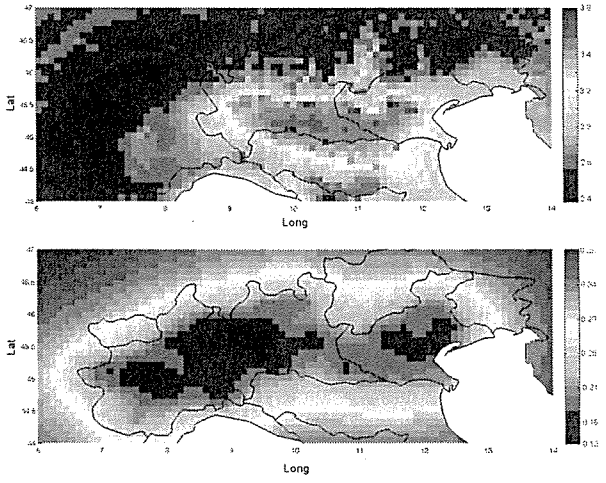

Figure 1. Kriged PM_{10} concentration (top) and prediction variance (bottom), April 2nd 2006.

In fact, the prediction variance, showed in the bottom part of Fig. 1, is higher over the areas of the region where no monitoring stations are available. The overall map and time average prediction variance is equal to 0.268, with a daily average increasing from 0.253 to 0.280 depending on the daily missing-data rate, as displayed in Fig. 2. The overall prediction variance restricted to the PM10 sites, instead, is equal to 0.139.

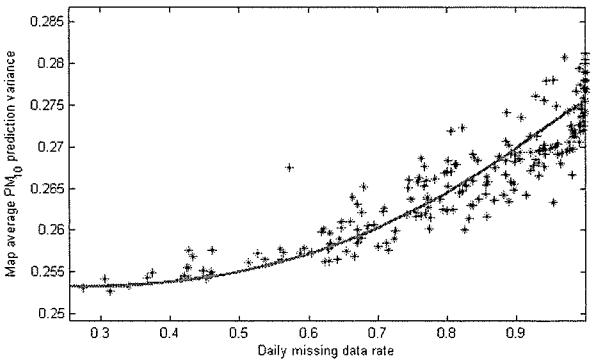

Figure 2. PM_{10} prediction variance with respect to daily missing-data rate.

D. *Cross-Validation*

The predictive performance of the estimated model is evaluated by means of the leave-one-out crossvalidation method. At each of the 107 ground-level monitoring stations, let's say s_j, the PM_{10} concentration is predicted using PM_{10} and AOT data at all sites except s_j. Using the observed PM_{10} concentration $y_{PM}(s_j,t)$ and its prediction $\hat{y}_{PM}(s_j,t)$, we evaluate at each site s_j the bias, namely $Bias_j = \sum_{t=1}^{T}\left[y_{PM}(s_j,t) - \hat{y}_{PM}(s_j,t)\right]/T$, and the mean squared error $MSE_j = \sum_{t=1}^{T}\left[y_{PM}(s_j,t) - \hat{y}_{PM}(s_j,t)\right]^2/T$.

The spatial distribution of the MSE is depicted in Fig. 3. The average value of the bias and MSE over the whole set of sites is 0.008 and 0.263, respectively.

V. Sensitivity to Missing Data

The estimation results presented in the previous section are based on real AOT observations characterized by an average missing-data rate of 73%. One naturally wonders if the results are reliable, namely if the estimation process is robust with respect to high missing-data rates.

In order to shed light on this aspect, an extensive simulation study has been implemented involving, on the one hand, simulation of data from the estimated model, on the other, simulation of the spatial pattern of the missing data.

A. *Data Simulation*

By considering the same data structure of the case study of section IV.A, (monitoring network sites, AOT grid and temporal frame), and the estimated model of section IV.B, PM_{10} concentrations $\tilde{y}_{PM}(s,t)$ and AOT measures $\tilde{y}_{AOT}(s,t)$ are simulated with an AOT missing-data rate α ranging between 0% to 90%. For each α considered, the parameter set Ψ is estimated and the leave-one-out crossvalidation method applied. As in section IV.D, the average value of bias and MSE over the PM_{10} sites will be evaluated for each α.

B. *Missing Data Pattern Simulation*

Lack of AOT data is usually due to either clouds or snow on the ground. Missing data positions are then spatially correlated, in the sense that they appear as large patches over the region. To emulate the missing data pattern, a random sample is drawn for each time t from a multivariate Normal distribution $U \sim N(0, \Sigma_m)$ over the AOT sites.

In order to impose a strong spatial correlation, the variance-covariance matrix Σ_m is based on the exponential correlation function with parameter $\theta = 100\,\text{km}$. If α is the missing data rate to be simulated and $u(s,t)$ is the value drawn from U at site s and time t, the AOT measure $\tilde{y}_{AOT}(s,t)$ is forced to be missing if $|u(s,t)| \geq \Phi^{-1}(1-\alpha/2)$, where Φ^{-1} is the inverse of the standard normal cumulative distribution function.

C. *Results*

Following the simulation procedure of the previous paragraphs, we obtained average bias and MSE values equal to 0.021 and 0.241 respectively, independently to the missing-data rate α. This is not surprising since missingness affects AOT data only while the crossvalidation is on PM_{10} concentrations.

The MSE value obtained here should be compared to the MSE value resulting from the crossvalidation procedure of section IV.D, namely 0.263. The small difference between the two value can be ascribed to model miss-specification.

To go further, we applied the same simulation procedure considering missingness over both AOT and PM_{10} variables. bias and MSE average values are reported in Table II, from which it can be noted that the MSE only increases by 26% when the missing-data rate moves from 0% to 80%. Indeed, a critical point between $\alpha = 0.8$ and $\alpha = 0.9$, seems to exist where the MSE rapidly increases.

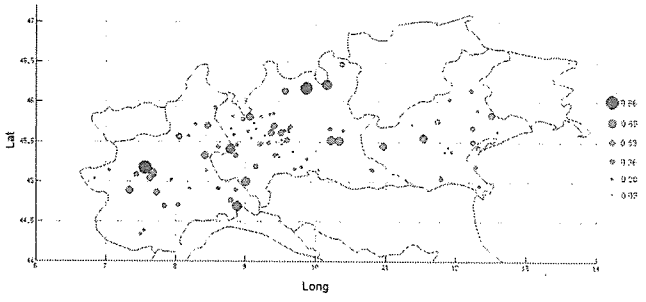

Figure 3. Spatial distribution of the mean-squared error over the PM_{10} sites.

TABLE II. BIAS AND MEAN SQUARED ERROR WITH RESPECT TO THE AOT AND PM_{10} MISSING-DATA RATE

α	0.0	0.2	0.4	0.5	0.6	0.7	0.8	0.9
bias	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.03
MSE	0.49	0.49	0.50	0.50	0.51	0.55	0.62	0.88

VI. DISCUSSION AND CONCLUSIONS

The dynamic coregionalization model can be used as a flexible mapping tool able to mix static and dynamic measurements at ground level and remote sensing and to work satisfactorily under bad weather conditions.

Moreover, it allows to assess the various sources of uncertainty on a common ground tanks to an approximate decomposition of the spatial prediction variance which can be considered locally in space and time or globally in space and/or time.

The sensitivity analysis of sections V and II.C shows that the relation between mapping precision and missing data has two faces at least for the region and seasons covered by this study. On the one hand, considering the daily mapping behavior, Fig. 2 shows that the average mapping uncertainty is higher for those days with bad weather and little satellite information, the increase being about 10% for the data considered. Moreover simulating missing data for both ground level and remote sensing data gives Table II which shows that mapping precision deteriorates at high missing rates.

On the other hand restricting the missing pattern to AOT only, which is appropriate in this study, it results that the dynamic mapping method based on DCM is quite robust with respect to remote sensing data availability as the global mapping mean square error results to be insensitive to the rate of AOT missing data.

REFERENCES

Fassò A, and Cameletti M. (2010). A unified statistical approach for simulation, modeling, analysis and mapping of environmental data. Simulation: *Transactions of the Society for Modeling and Simulation International*. 86(3), 139–154.

Fassò A, and Finazzi F. (2010a). Air quality mapping using the dynamic coregionalization model. Accepted in *Proceedings of 45th Scientific Meeting of the Italian Statistical Society*. Padua, June 16 -18, 2010.

Fassò A, and Finazzi F. (2010b). The dynamic coregionalization model with application to air quality remote sensing. *Submitted*.

Fassò A, Finazzi F. and D'Ariano C. (2009a). *Maximum likelihood estimation of the dynamic coregionalization model with heterotopic data*. Working Paper n.15/MS - 2009, Dept. IT and Math Meth. University of Bergamo, www.unibg.it.

Fassò A, Finazzi F, and 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.

Hoff R.M, and Sundar A.C. (2009). Remote sensing of particulate pollution from space: have we reached the promised land? *Air and Waste Manage. Assoc.* 59, 645-675.

Wang J, and 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.