Multivariate geostatistical tools for time series modeling and prediction

S. De Iaco[1], S. Maggio[1], M. Palma[1,*] and D. Pellegrino[1]

[1] *University of Salento, Via per Monteroni, Complesso Ecotekne, Lecce, Italy; sandra.deiaco@unisalento.it, sabrina.maggio@unisalento.it, monica.palma@unisalento.it, daniela.pellegrino@unisalento.it*
[*] *Corresponding author*

Abstract. *Modeling and prediction multivariate geostatistical techniques can be successfully applied to study the temporal behaviour of several correlated time series. In particular, in the time domain, by using variogram-based tools the analyst can easily a) identify trend and periodicity which characterize each time series, b) fit a properly Multivariate Linear Temporal (MLT) model to multiple correlated time series, c) predict the variable of interest (primary variable) at some time points after the last available observation, by taking into account the fitted model as well as the auxiliary information coming from the secondary variables. In this paper the convenience of performing a complete analysis of multiple correlated time series on the basis of geostatistical tools is illustrated through a case study concerning three environmental variables. As regards the computational aspects, a new version of the GSLib Cokb3d routine has been implemented for prediction purposes.*

Keywords. *Multivariate linear temporal model; Temporal cross-variogram; Temporal cokriging.*

1 Introduction

In time series analysis, the methodology developed by Box and Jenkins (1976) is commonly applied to detect the most suitable model which reasonable might describe the temporal evolution of the analyzed process. Then, the model is used in the prediction stage. On the basis of the Box-Jenkins approach, the auto-correlation and the partial auto-correlation functions (ACF and PACF, respectively), as well as the cross-correlation function (CCF) have a crucial role in the modeling selection, indeed through the visual inspection of the sample ACF, PACF and CCF, the most appropriate model for the process under study can be identified. In the multivariate context, several approaches have been proposed in order to model the joint relationships between multiple time series. Among the different types of models (Reinsel, 2003), the most common are Vector AutoRegressive, AutoRegressive-MovingAverage or AutoRegressive-Integreted-MovingAverage models in the presence of exogenous variables, the models based on a transferring function and the co-integrated models (mainly used in the economic field). However, for the analysis of multiple correlated time series, multivariate geostatistics could also be a very useful approach, nevertheless it is widely applied to investigate, through the matrix variogram, the spatial direct and cross-correlation which characterize the variables of interest and make predictions at unsampled locations of the spatial phenomena.
In this paper the use of variogram-based multivariate geostatistical techniques have been enlarged to analyze multiple time series, in order to identify trends and periodicity exhibited by the data, model the temporal evolution of the variables and make temporal predictions for the primary variable using the auxiliary information coming from the secondary available variables. The computational aspects have

been tackled by implemented a new version of the GSLib *Cokb3d* routine (Deutsch and Journel, 1998) which allows the analyst to use the fitted model in the cokriging system and define appropriate temporal search neighborhoods for prediction purposes.

2 Variogram-based modeling and prediction for multiple time series

In time series analysis, the measurements of $p \geq 2$ correlated variables, at different time points or intervals, can be considered as a finite realization of a real-valued Multivariate Random Process (MRP) $\{Z(t), t \in T \subseteq \mathbb{R}\}$, with $\mathbf{Z}(t) = [Z_1(t), Z_2(t), \ldots, Z_p(t)]^T$. Under second-order stationarity, the mean vector of $\mathbf{Z}$ exists and does not depend on t, and the $(p \times p)$ variogram matrix $\boldsymbol{\Gamma}$ defined for two MRP, $\mathbf{Z}(t)$ and $\mathbf{Z}(t')$, exists and depends on the temporal separation h, i.e.:

$$\boldsymbol{\Gamma}[\mathbf{Z}(t), \mathbf{Z}(t')] = E\left\{[\mathbf{Z}(t) - \mathbf{Z}(t')][\mathbf{Z}(t) - \mathbf{Z}(t')]^T\right\} = \boldsymbol{\Gamma}(h) = [\gamma_{ij}(h)],$$

where $h = (t - t')$ and $\gamma_{ij}(h),\ i, j = 1, \ldots, p,$ are the cross-variogram (if $i \neq j$) between the random variables $Z_i(t)$ and $Z_j(t+h)$ and the direct variogram (if $i = j$) of the i-th random variable. In the multivariate context, the empirical temporal variogram matrix can be modelled through the most used model in the spatial multivariate analysis, namely the *Linear Coregionalization Model* (Wackernagel, 2003). In this case, a Multivariate Linear Temporal (MLT) model $\boldsymbol{\Gamma}(h) = \sum_{l=1}^{L} \mathbf{B}_l\, g_l(h)$, can be developed, where $\mathbf{B}_l = [b_{ij}^l], i, j = 1, \ldots, p$, are $(p \times p)$ positive-definite matrices and $g_l(h)$, $l = 1, \ldots, L$, are basic temporal variograms identified at $L \geq 2$ temporal variability scales. Before modeling the temporal direct and cross-correlation among the variables, the direct and cross-variograms are estimated as follows:

$$\widehat{\gamma}_{ii}(r) = \frac{1}{2\,|N_i(r)|} \sum_{N_i(r)} [Z(t+h) - Z(t)]^2; \quad \widehat{\gamma}_{ij}(r) = \frac{1}{2|N_{ij}(r)|} \sum_{N_{ij}(r)} \left[(Z_i(t+h) - Z_i(t)) \cdot (Z_j(t+h) - Z_j(t))\right],$$

where $N_i(r) = \{t, t+h \in H_i, i = 1, \ldots, p,$ such that $|r-h| < \delta\}$, $|N_i(r)|$ is the cardinality of this last set, $N_{ij}(r) = \{t, t+h \in (H_i \cap H_j), i, j = 1, \ldots, p,\ i \neq j$ such that $|r-h| < \delta\}$, and $|N_{ij}(h)|$ is its cardinality, r is the temporal lag, δ is the tolerance and H_i is the set of the measurements for the i-th time series, $i = 1, \ldots, p$. As pointed out in De Iaco et al. (2013), the variogram could be efficiently applied in time series analysis (Haslett, 1997), since it can describe a wide class of stochastic processes (the class of intrinsic stochastic processes), and also its estimation does not require the knowledge of the expected value of the associated stochastic process. Moreover, the variogram is a useful tool to identify trend and periodicity exhibited by data and to make temporal predictions for the variable of interest. For a second-order stationary MRP $\mathbf{Z}$, a linear prediction of the time series under study at an unsampled time point $t \in T$, can be obtained by using the well-known cokriging predictor (Wackernagel, 2003). In this case, the temporal cokriging predictor is expressed as: $\widehat{\mathbf{Z}}(t) = \sum_{\alpha=1}^{N} \boldsymbol{\Lambda}_\alpha(t)\mathbf{Z}(t_\alpha)$, where $t_\alpha \in T$, $\alpha = 1, \ldots, N$, are the sampled points and $\boldsymbol{\Lambda}_\alpha(t)$, $\alpha = 1, \ldots, N$, are the $(p \times p)$ matrices of the weights which are determined so that the above temporal predictor is unbiased and efficient (Journel and Huijbregts, 1981). The ordinary cokriging requires only the knowledge of the model for the matrix variogram and it is used when the expected value of the process is constant and unknown.

3 A case study

Two atmospheric variables, i.e. daily Temperature ($^\circ C$) and daily Wind Speed (m/sec), as well as PM_{10} daily concentrations ($\mu g/m^3$), measured from 2010 to 2013, at one survey station belonging to the environmental network of the Apulian Protection Agency and located in Brindisi district (South of Italy),

have been analyzed by multivariate geostatistical tools. The survey station, called "Torchiarolo", is very close to the thermoelectric power station "Enel-Federico II", and all the surrounding area is considered being at high risk of air pollution, especially during the winter and during long period of low rainfall. PM_{10} is strongly influenced by meteorological conditions. In particular, the horizontal transport, dispersion and resuspension of PM_{10} are mainly determined by Wind speed: low values of this meteorological variable are related to high PM_{10} concentrations (Harrison et al., 1997; Sayegh et al., 2014). Moreover, temperature is considered as one of the strongest predictors of PM_{10} concentrations. High values of this air pollutant are measured in winter, specially when the difference between maximum and minimum daily temperature is large (Perez et al., 2002). In the following sections, the advantages and the flexibility of the multivariate geostatistical techniques to analyze the times series under study will be pointed out.

3.1 Exploratory analysis and modeling

Exploratory data analysis has clearly highlighted that: a) PM_{10} daily concentrations present an annual periodicity at 12 months, b) Temperature and Wind Speed are characterized by opposite seasonal behaviors: in winter time, Temperature decreases, while Wind Speed increases; on the other hand, in summer time Temperature increases and Wind Speed decreases, c) over the four-year span (from 2010 to 2013), the PM_{10} daily values have exceeded 243 times the threshold value (50 $\mu g/m^3$) fixed by the national law for the human health protection; in particular, during the summer, PM_{10} does not exceed this limit value, instead, in winter time changes in the lower layer of the troposphere determine PM_{10} stagnation and consequently high concentrations of PM_{10}. The sample direct temporal variograms for the analyzed time series highlight the presence of periodicity for all variables (Fig. 1). These periodic components

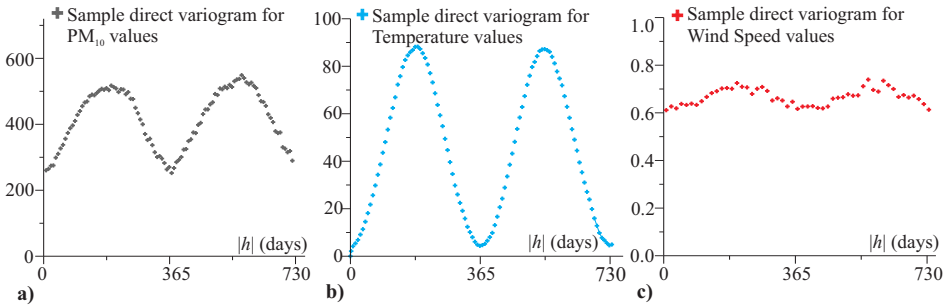

Figure 1: Sample temporal direct variograms of a) PM_{10}, b) Temperature and c) Wind Speed daily averages.

have been factored out from the observed data through moving average and monthly averages techniques. Hence, the residuals have been considered as a realization of a second-order stationary MRP $\mathbf{Z}(t) = [Z_1(t), Z_2(t), Z_3(t)]^T$, with $t \in T \subseteq \mathbb{R}$, and have been used in the following steps of the analysis. After computed the sample direct and cross temporal variograms of the residuals, two different scales of temporal variability have been detected through the visual inspection of the sample variograms. Hence, the following MLT model has been fitted to the sample matrix variogram:

$$\boldsymbol{\Gamma}(h) = \mathbf{B}_1\, g_1(h) + \mathbf{B}_2\, g_2(h), \tag{1}$$

where g_1 is the short-scale temporal component described by an exponential model (Cressie, 1993) with unit sill and range equal to 30 days, g_2 is the long-scale temporal component described by an exponential model with unit sill and range equal to 365 days and the positive-definite matrices $\mathbf{B}_l, l = 1,2$, are:

$$\mathbf{B}_1 = \begin{bmatrix} 230 & 3.07 & -3.4 \\ 3.07 & 4.9 & -0.025 \\ -3.4 & -0.025 & 0.58 \end{bmatrix}, \qquad \mathbf{B}_2 = \begin{bmatrix} 30 & 1.2 & -1.1 \\ 1.2 & 0.37 & -0.016 \\ -1.1 & -0.016 & 0.073 \end{bmatrix}. \tag{2}$$

At this point, it is convenient to check if the fitted model (1) can be considered suitable to make predictions of the primary variable, thus a validation procedure has been properly performed.

3.2 Model validation and temporal prediction

The goodness of model (1) has been checked through the cross-validation technique. In this stage of the analysis a modified version of the GSLib program *Cokb3D* (Deutsch and Journel, 1998), named *T-Cok*, has been implemented and used to compute temporal predictions of PM_{10} on the basis of a) the auxiliary variables, b) the fitted MLT model and c) a properly defined neighborhood, i.e. a subset of time data which can be considered in the cokriging system. Hence the cross-validation has been performed and the correlation between PM_{10} residuals and estimated ones has been measured. The high values of the linear correlation coefficient (0.780) has confirmed the goodness of the fitted MLT model, which can be used to predict PM_{10} daily concentrations in time points after the last available data. In particular, PM_{10} residuals have been predicted for six time points (1-6 January 2014), by using the new GSLib routine *T-Cok*. The deseasonalized PM_{10} observations, the residuals of the auxiliary variables and the model (1) are the input information for the *T-Cok* routine. Then, the diurnal component has been added to the predicted PM_{10} residuals in order to obtain predictions of PM_{10} daily concentrations. By comparing PM_{10} daily concentrations measured from the 1st to the 6th of January 2014 and the predicted ones, it is worth highlighting that the behavior of the predicted values is quite similar to the true PM_{10} daily concentrations; moreover, as it is for the true values recorded in the period 1-6 January 2014, some predicted values are greater than the limit value of 50 $\mu g/m^3$ and it can represent a hazardous condition for air quality and human health.

4 Conclusions

In this paper, the time series of PM_{10} daily concentrations and two meteorological variables (Temperature and Wind Speed), correlated with the pollutant under study, were analyzed through multivariate geostatistical techniques. The importance and the advantages of using variogram-based procedures were pointed out during both modeling and prediction stages. The scientific community should consider the flexibility of the geostatistical tools for the analysis of time series and more theoretical and computational efforts should be made in order to extend the variogram-based techniques in the time domain.

References

Box, G. E. P., Jenkins, G. M. (1976). *Time series analysis: forecasting and control*. Holden Day. San Francisco.

Cressie, N. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Mathematical Statistics. New York.

De Iaco, S., Palma, M., Posa, D. (2013). *Geostatistics and the Role of Variogram in Time Series Analysis: A Critical Review.* In: Montrone S., Perchinunno P. (eds) Statistical Methods for Spatial Planning and Monitoring. Contributions to Statistics. Springer, 47–75.

Deutsch, C. V., Journel, A. G. (1998). *GSLib: Geostatistical Software Library and User's Guide*. Oxford University Press. New York.

Haslett, J. (1997). On the sample variogram and sample autocovariance for non-stationary time series. *Statistician* **46**(4) 475–485.

Harrison, R. M., Deacon, A. R., Jones, M. R., Appleby, R. S. (1997). Sources and processes affecting concentrations of PM10 and PM2.5 particulate matter in Birmingham (U.K.). *Atmospheric Environment* **31** 4103–4117.

Journel, A. G., Huijbregts C.J. (1981). *Mining Geostatistics*. Academic Press. London.

Perez, P., Reyes, J. (2002). Prediction of maximum of 24-h average of PM10 concentrations 30 h in advance in Santiago, Chile. *Atmospheric Environment* **36**(28) 4555–4561.

Reinsel, G. C. (2003). *Elements of Multivariate Time Series Analysis*. Springer Science and Business Media.

Sayegh, A. S., Munir, S., Habeebullah, T. M. (2014). Comparing the performance of statistical models for predicting PM10 concentrations. *Aerosol Air Quality Research* **14**(3) 653–665.

Wackernagel, H. (2003). *Multivariate geostatistics-an introduction with applications*. 3rd edn. Springer. Berlin.