ENVIRONMETRICS
Environmetrics 2001; **12**:11–23

Product-sum covariance for space-time modeling: an environmental application

L. De Cesare[1,2], D. E. Myers[3]* and D. Posa[2,4]

[1] *Facoltà di Economia, Università di Bari, Italy*
[2] *Istituto per Ricerche di Matematica Applicata (CNR), Bari, Italy*
[3] *Dept of Mathematics, Tucson AZ, U.S.A.*
[4] *Facoltà di Economia, Università di Lecce, Italy*

SUMMARY

In this paper a product-sum covariance for space-time modeling of nitrogen dioxide in the Milan district is introduced. Residuals have been generated for all stations after the removal of daily and seasonal trends, which are readily interpretable, in order to estimate and model the spatial-temporal variogram. The trend component and the residual variogram model have been used to predict the hourly averages of nitrogen dioxide for the first two days of January 1997. GSLIB programs were modified for sample variogram computations, cross-validation and kriging. Copyright © 2001 John Wiley & Sons, Ltd.

KEY WORDS: geostatistics; space-time correlation; time series

1. INTRODUCTION

A large number of environmental phenomena may be regarded as realizations of space-time random functions and geostatistics offers a variety of methods to model such processes as an extension of the existing spatial techniques into the space-time domain. Despite the straightforward appearance of this extension, there are a number of theoretical and practical problems that must be addressed prior to any successful application of geostatistical methods to space-time data: (a) qualitative differences between spatial and temporal processes; (b) the presence of spatial non-stationarity and temporal periodicity; (c) spatial and temporal scales non-comparable in a physical sense; (d) space-time data sets are usually composed of few spatial locations, each with a long time series: this kind of sampling yields information which is rich in time but poor in space.

In a previous paper (De Cesare *et al.*, 1996) space-time analysis on monthly averages of SO_2 measurements taken at 33 monitoring stations from January 1983 to December 1986 in the Milan district was described: in analyzing this data-set there were about the same number of spatial

*Correspondence to: D. E. Myers, Department of Mathematics, University of Arizona, Building #89, 617 N. Santa Rita, Tucson, AZ 85721, U.S.A.

Received 10 November 1999
Revised 23 February 2000

data locations as temporal data locations (48 time points). Moreover, the authors modeled a spatial variogram and a temporal variogram and essentially the sill of the spatial-temporal model was the product of the sills; starting from the covariance form it is only necessary that the coefficients be positive.

In this paper a product-sum covariance model has been introduced for predicting and modeling NO_2, a hazardous pollutant, in the Milan district in 1996; the measurements have been obtained from an irregular spatial network of monitoring stations (48 locations), at high temporal resolutions (hourly data), which represents a continuous-time and discrete-space sampling network: this means almost $48 \times 24 \times 366$ data for the above pollutant for one year.

Although there are a number of articles about space-time models (Myers, 1992; Rodriguez-Iturbe *et al.*, 1974; Rouhani and Hall, 1989) and analysis in the area of atmospheric pollution (Bilonick, 1985; Eynon & Switzer, 1983; Le and Petkau, 1988; Soares *et al.*, 1992), there are few examples in the literature such as the present case study in which a very exhaustive information in time (hourly measurements for one year) and moderate information in space is available.

The original data set consists of hourly average concentrations of five pollutants (SO_2, NO_2, PTS, CO and O_3) measured at 48 monitoring stations in 1996, however in this paper only the pollutant NO_2 will be considered: the main reason for this choice is that the concentration levels set by the Regional Law have been exceeded by nitrogen dioxide.

The air quality standards (limit and guidelines values) are given in the Premier's Decree of the 28th of March 1983 and Decree by the President of the Republic n.203 of the 24th of May 1988 and incorporate EEC directives regarding air quality.

Decree by the President of the Republic n.203 of the 24th of May 1988 requires the 98th percentile of the hourly average concentrations (recorded during the solar year) of nitrogen dioxide to be below 200 $\mu g/m^3$, while the Premier's Decree of the 28th of March 1983 states that this value must not be exceeded more than once a day.

Nitrogen dioxide (NO_2) is a secondary pollutant caused by the oxidation of NO in the air. The rise in temperature during the summer makes the consequences of photochemical smog even more serious: this explains why nitrogen dioxide, has notably different daily and seasonal trends than other pollutants.

The environmental decision-making process is preceded by data analysis, which is conditioned by a specific temporal and spatial scale. Air quality management decisions are different from regional to local scale and very local spatial frame, and from hourly or daily patterns, according to the different types of pollutant characteristics. Moreover, by utilizing the information in this data set, spatial structures, chronological trends and cycles will be detected.

The estimated spatial-temporal variogram model and the hourly averages of NO_2 from January 1996 to December 1996 have been used to predict, the hourly averages of NO_2 on the first day of January 1997, by ordinary kriging at the same monitoring stations available on the 1st of January. The predicted values have been compared to the true ones; of course, the January 1997 data were not used throughout the above analysis.

2. THE DATA SET

The data collection network, which was planned in relation to the standards and information provided by the Lombardy Region's Environmental Town Council, consists of 48 survey stations

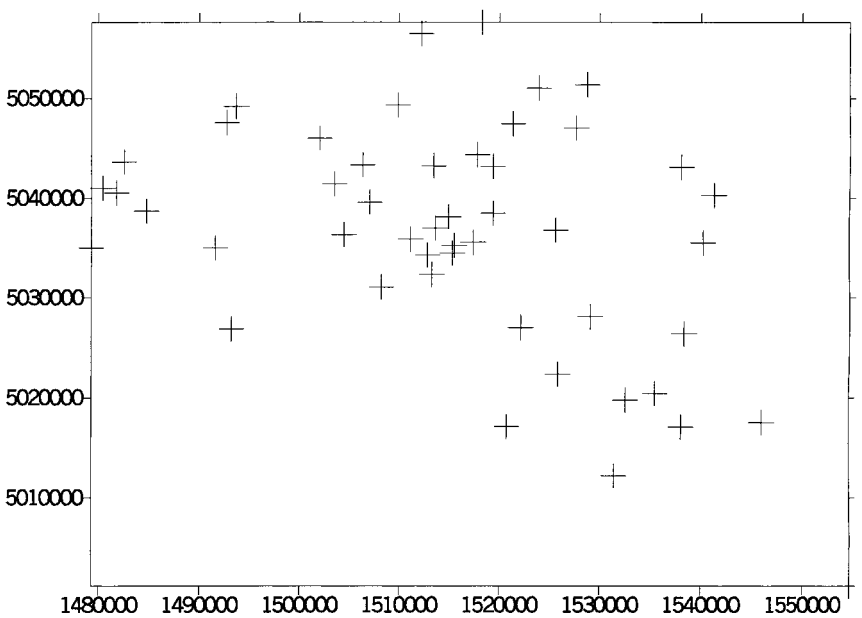

Figure 1. Network of Milan district. The coordinate system of the monitoring stations refers to the Italian national grid system (Gauss-Boaga).

for NO_2 (Figure 1). The coordinate system of the monitoring stations refers to the Italian national grid system (Gauss-Boaga), which is based on the Universal Transverse Mercator (UTM) projection.

The data set described throughout the paper consists of hourly averages of NO_2 measured from 1 January 1996 to 31 December 1996. The sampling points are not uniformly distributed; the central portion of the study area has a higher concentration of points. Figure 2 shows the histogram of the data set.

3. SPACE-TIME MODELING FOR NO_2

It is assumed that the observations are a realization of the spatial-temporal random field

$$(Z(s,t); \quad s \in D, \ t \in T) \tag{1}$$

where the domain $D \subseteq \mathbb{R}^2$ and $T \subseteq \mathbb{N}$, which allows Equation (1) to be viewed as a time series of spatial processes, each process occurring at equally spaced time points. In this case $T = \{1, 2, \ldots, 8784\}$.

Assuming that the first and second moments of Z exist, Z can be decomposed as

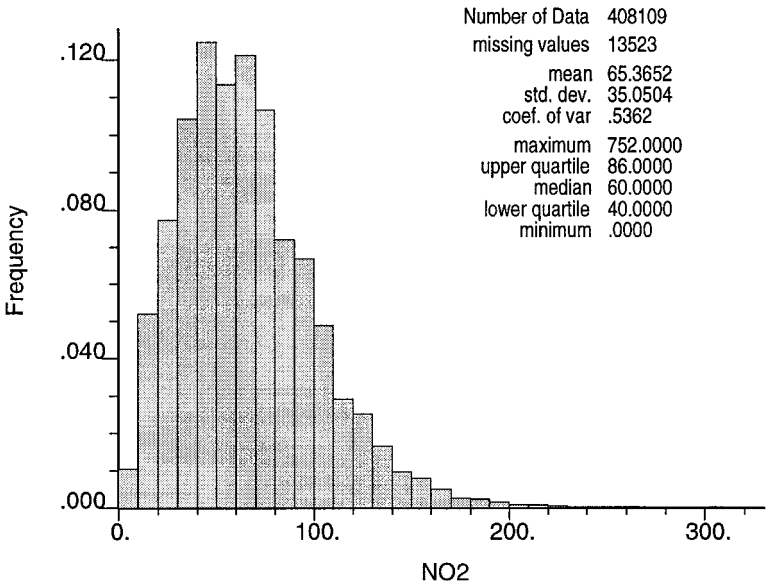

Figure 2. Histogram of the hourly averages of NO_2.

$$Z(s,t) = m(s,t) + Y(s,t) \tag{2}$$

where Y is a second order stationary stochastic process with

$$E(Y(s,t)) = 0$$

and $m(s,t)$ is the mean function of Z; the covariance function

$$C_{st}(h) = \mathrm{Cov}(Y(s+h_s, t+h_t), Y(s,t)) \tag{3}$$

where $h = (h_s, h_t) \in \mathbb{R}^2 \times \mathbb{R}$, $(s, s+h_s) \in D^2$ and $(t, t+h_t) \in T^2$ and the variogram

$$\gamma_{st}(h_s, h_t) = \frac{\mathrm{Var}(Y(s+h_s, t+h_t) - Y(s,t))}{2}$$

depend solely on the lag vector h, not on location or time.

3.1. Seasonal effect and missing values

Figure 3 shows the temporal variogram using the original data: note two periodic structures at 12 and 24 h. The first periodic structure could be due to more intense traffic during the morning and the evening (12 h time interval, see Figure 4).

In order to remove the seasonal effect, the trend $m(s,t)$ has been modelled in the following way: for each monitoring station, the whole year 1996 has been split in 12 months; for each of them the daily cycle (i.e. cycle with period $d = 24$) has been computed. That is:

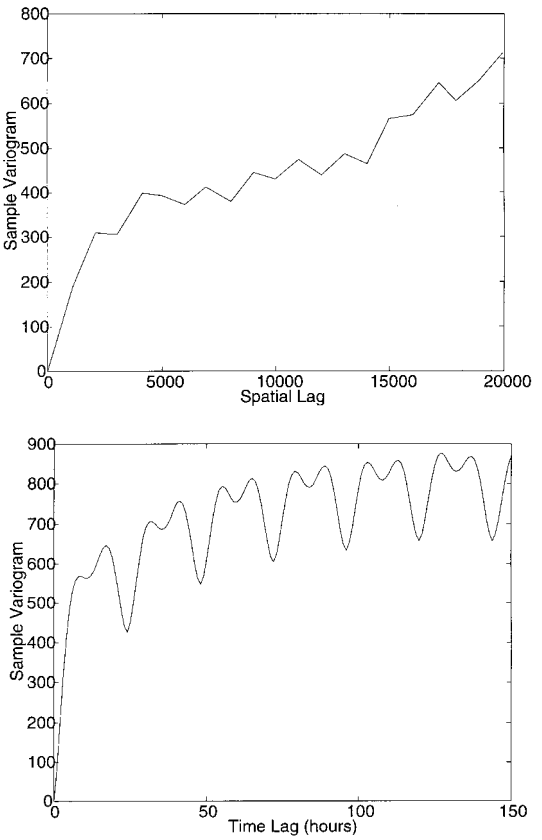

Figure 3. Sample variograms using the original data set.

$$m(s,t) = \alpha_i(s,t) + \mu_i$$

for

$$t \in T_i = \left[\frac{8784}{12}(i-1)+1, \frac{8784}{12} i\right] \cap \mathbb{N}, \quad i = 1,\ldots,12$$

where

i. $\alpha_i(s,t) = \alpha_i(s,t+d) \quad \forall s \in D \quad \forall t, t+d \in T_i$

ii. $\sum_{j=1}^{d} \alpha_i(s,j) = 0 \quad \forall s \in D$

$\alpha_i(s,t)$ is called *seasonal component* for the ith month and μ_i is a constant trend for the same month. Time series analysis has been carried out for each location by the standard technique of moving average estimating (Brockwell and Davis, 1987). If in the above time series less than five consecutive missing values are present, these missing values have been linearly interpolated. Residuals have been generated for all stations after removal of the seasonal component. Figure 4 shows the time series of the hourly averages and the corresponding seasonal component of a

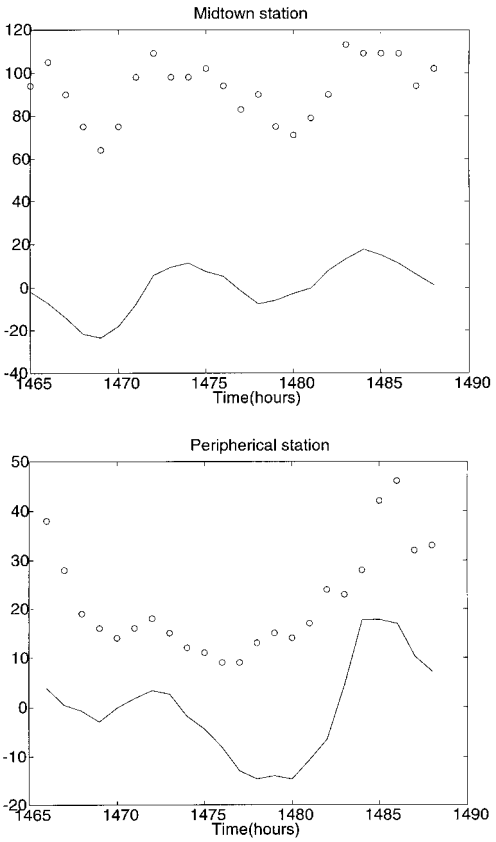

Figure 4. Time series of the hourly averages (dotted line) and the corresponding seasonal component (continuous line).

generic day (2 March) for two different stations. Figure 5 shows the spatial-temporal variogram surface for these residuals.

3.2. The product-sum model

In the present analysis, the following spatial-temporal covariance model

$$C_{st}(h_s, h_t) = k_1 C_s(h_s) C_t(h_t) + k_2 C_s(h_s) + k_3 C_t(h_t) \tag{4}$$

for the process Y has been used. The previous model could be easily written in terms of the spatial-temporal variogram:

$$\gamma_{st}(h_s, h_t) = (k_1 C_s(0) + k_3)\gamma_t(h_t) + (k_1 C_t(0) + k_2)\gamma_s(h_s) - k_1 \gamma_s(h_s)\gamma_t(h_t) \tag{5}$$

where γ_{st} is the spatiotemporal variogram, γ_t the temporal variogram and γ_s the spatial variogram.

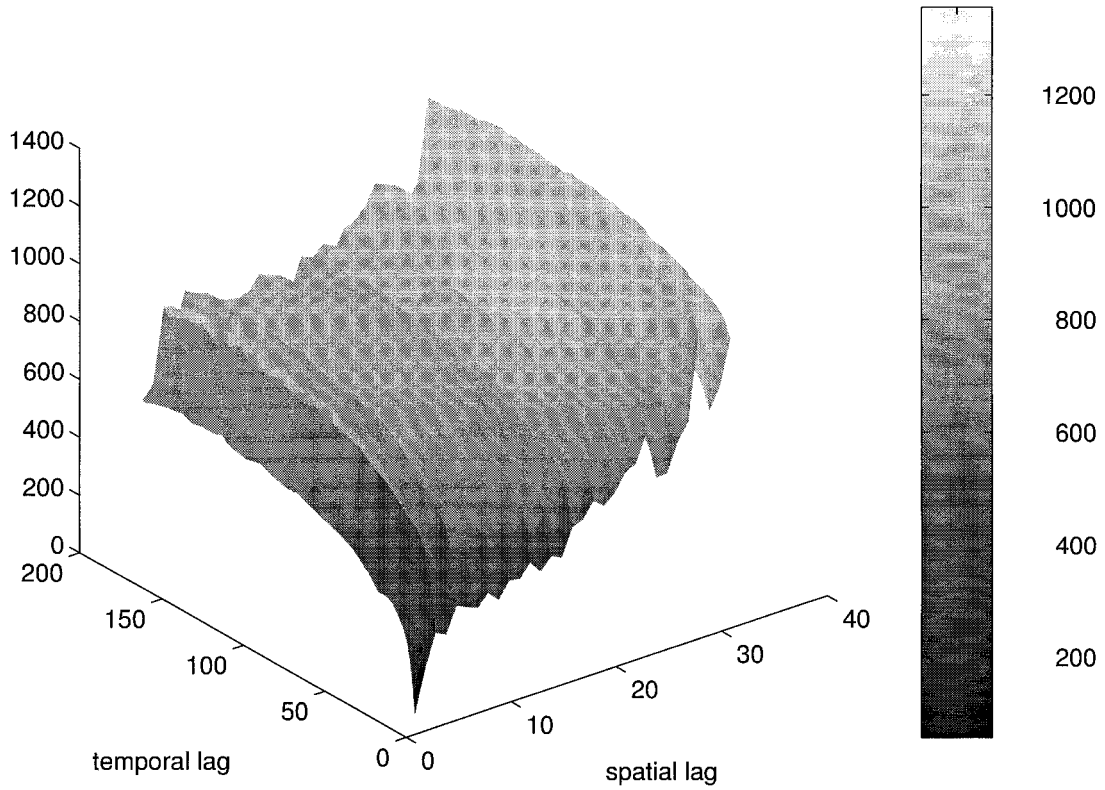

Figure 5. Sample spatial-temporal variogram surface using the residuals.

Note that $C_{st}(0,0)$ is the 'sill' of γ_{st}, $C_s(0)$ is the sill of γ_s and $C_t(0)$ is the sill of γ_t; moreover, by definition $\gamma_{st}(0,0) = \gamma_s(0) = \gamma_t(0) = 0$.

The following condition is implicit in the transformation from covariance form to variogram form:

$$k_1 C_s(0) C_t(0) + k_2 C_s(0) + k_3 C_t(0) = C_{st}(0,0) \tag{6}$$

Note also from Equation (5) that

$$\gamma_{st}(h_s, 0) = [k_2 + k_1 C_t(0)]\gamma_s(h_s) \tag{7}$$

$$\gamma_{st}(0, h_t) = [k_3 + k_1 C_s(0)]\gamma_t(h_t) \tag{8}$$

It is assumed that

$$k_2 + k_1 C_t(0) = 1$$

$$k_3 + k_1 C_s(0) = 1 \tag{9}$$

in order to estimate and model $\gamma_s(h_s)$ and $\gamma_t(h_t)$ by $\gamma_{st}(h_s, 0)$ and $\gamma_{st}(0, h_t)$, respectively.

Let $H = D \times T$ be the set of data location, then the standard moment estimator for the spatial variogram at the vector lag $r_s \in \mathbb{R}$, with spatial tolerance δ, is:

$$\hat{\gamma}_s(r_s) = \hat{\gamma}_{st}(r_s, 0) = \frac{1}{2|N_\delta(r_s)|} \sum [Z(s+h_s, t) - Z(s, t)]^2$$

where the summation is over the set

$$N_\delta(r_s) = \{((s+h_s, t),(s, t)) \in H^2 \text{ such that } \|r_s - h_s\| < \delta\},$$

and $|N_\delta(r_s)|$ is the cardinality of this set. Similarly, the standard moment estimator for the temporal variogram at lag $r_t \in N$ is:

$$\hat{\gamma}_t(r_t) = \hat{\gamma}_{st}(0, r_t) = \frac{1}{2|N(r_t)|} \sum [Z(s, t+r_t) - Z(s, t)]^2$$

where

$$N(r_t) = \{((s, t),(s, t+r_t)) \in H^2\}$$

Usually the spatial locations are not on a regular grid, while the temporal points are regularly spaced and hence it is not necessary to use temporal distance classes.

Figure 6 shows the sample spatial and temporal variograms of the residuals with the fitted models whose analytical expression is given below:

$$\gamma_s(h_s) = 220\text{Sph}\left(\frac{h_s}{2000}\right) + 450\left(1 - \exp\left(-\frac{h_s}{18000}\right)\right)$$

$$\gamma_t(h_t) = 280\text{Sph}\left(\frac{h_t}{12}\right) + 90\text{Sph}\left(\frac{h_t}{24}\right) + 250\text{Sph}\left(\frac{h_t}{96}\right) \tag{10}$$

In order to obtain a diagnostic check of the fit of the above variogram model vs. the raw variogram estimates, cross-validation (Cressie, 1993; Myers, 1991) has been used. Figure 7 shows the scatterplot of the residual values towards the predicted ones: the correlation coefficient is $\rho = 0.96$.

3.3. Spatial-temporal prediction

The hourly averages of NO_2 from January 1996 to December 1996 and the previous spatial-temporal variogram model have been used to predict the hourly averages of NO_2 on 1 January 1997, by ordinary kriging at the same monitoring stations available on 1 January. The predicted values have been compared to the true ones (Figure 8); of course, these last values have not been used throughout the above analysis. The correlation coefficient between true and predicted values is 0.67.

In order to evaluate the evolution of the hourly averages of NO_2 for 1 January 1997, the predicted values, at the same locations where the true values are available, were computed in the following way:

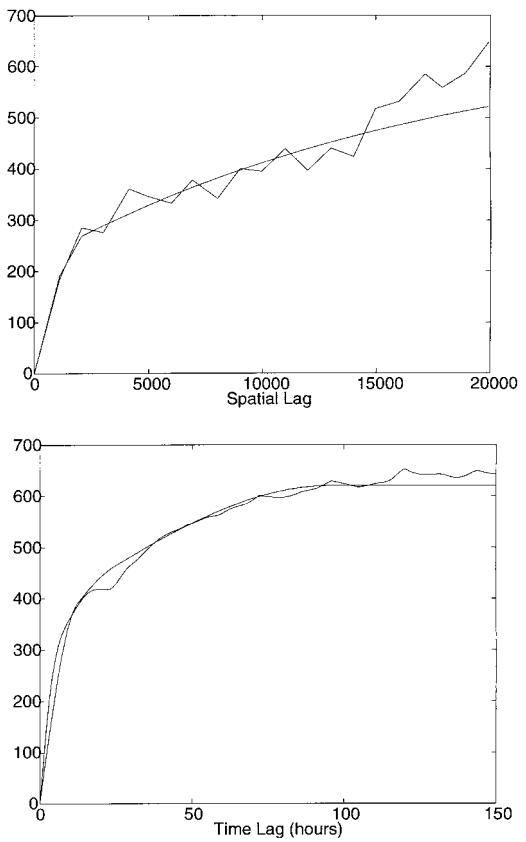

Figure 6. Sample spatial and temporal variograms using the deseasonalized data with the fitted models.

i. the residual variogram model given in Equation (5) was used to predict by ordinary kriging the residuals for 1 January 1997 at the same locations where the true values are available;
ii. the seasonal component corresponding to January 1996, which was previously estimated as discussed in Section 3.1 has been added to the predicted residuals in order to obtain the predicted values.

Of course, in the above modeling procedure, it is assumed that the trend component does not change from one year to the next.

Figures 9 and 10 show, respectively, the postplots of the predicted values and the true ones at a time interval of 4 h. The highest values are clearly localized in the central part of the graph, corresponding to the city of Milan and some outlying districts, while the peripheral parts of the district are characterized by the lowest values. Also note an overestimation of the true values in the first part of the day, maybe because the prediction corresponds to the New Year Day.

It is important to point out that NO_2 is caused by the oxidation of NO (nitric oxide) in the air: nitric oxide, generated by civil and industrial heating systems and thermoelectric power stations located in the Milan district, is mainly given off by motor vehicles. Peak nitrogen dioxide

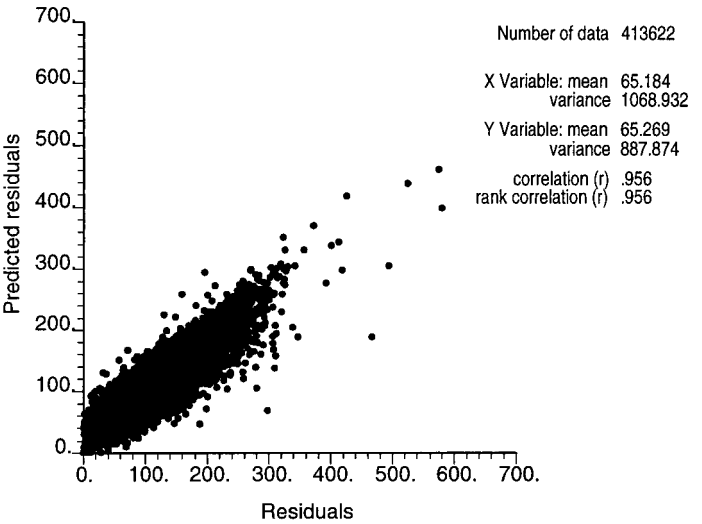

Figure 7. Scatterplot of the residual values vs. predicted ones obtained by cross-validation.

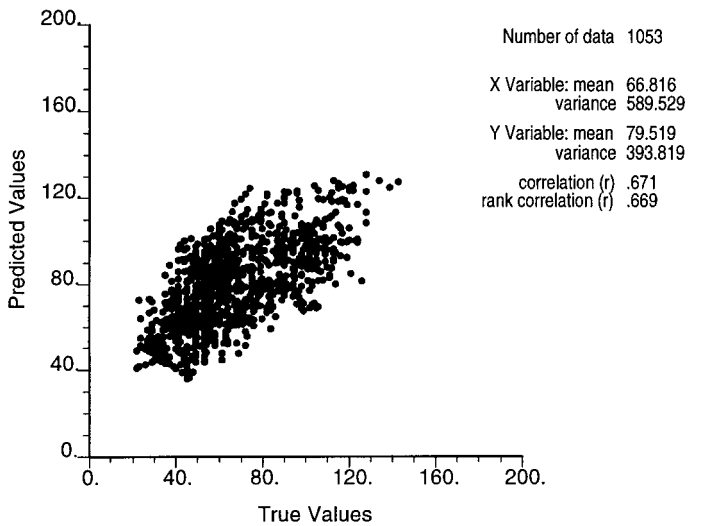

Figure 8. Scatterplot of the true values vs. predicted ones for the hourly averages for 1 January 1997.

readings generally tend to occur in the evening, when certain chemical reactions involving the photolysis of NO_2 and formation of O_3 no longer take place.

Different highlights and decisions can be taken from this kind of analysis, from technological investments in factories, to changes on the corridors in the entrance city and to limit the circulation of bus and trucks at specific streets.

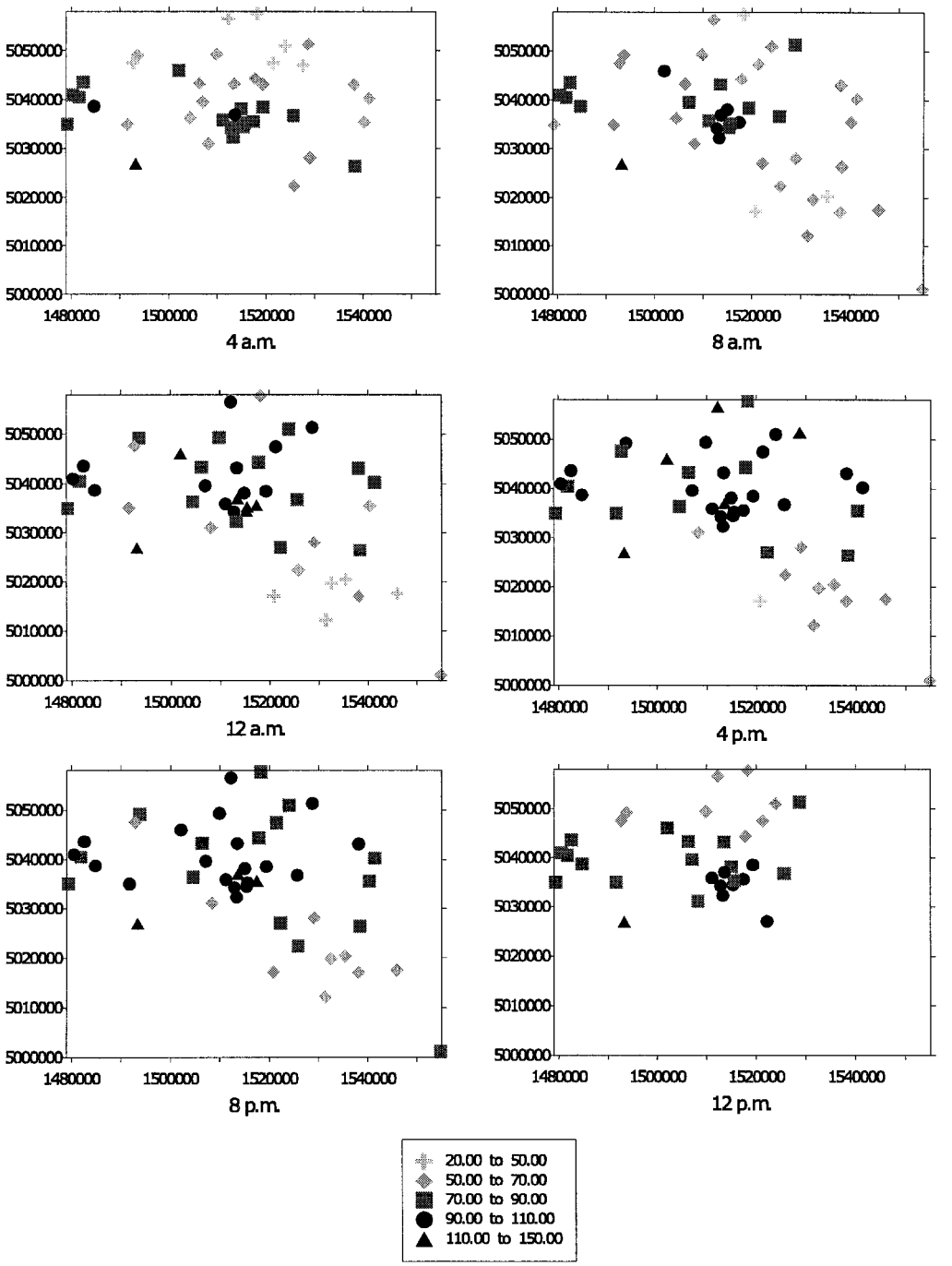

Figure 9. Postplots of the predicted values, at a time interval of 4 h, for 1 January 1997.

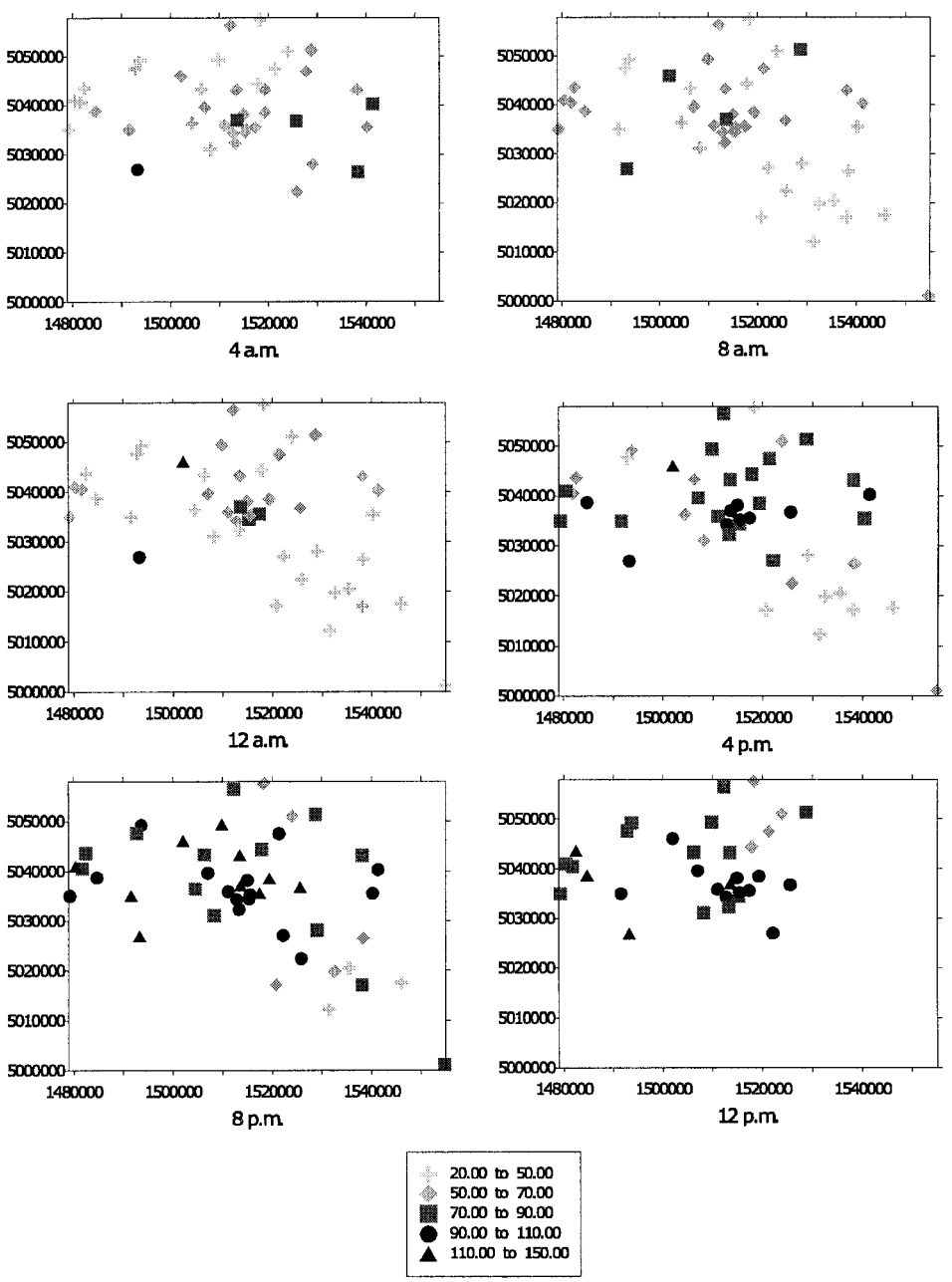

Figure 10. Postplots of the true values, at a time interval of 4 h, for 1 January 1997.

4. SOFTWARE

Standard geostatistical programs, such as in GSLIB and GEOEAS, do not allow for treating time as a separate coordinate (and not just as a 'dimension'). Without modification, only metric spatial-temporal covariances/variograms could be used. For that reason it was necessary to modify existing software. The 3-D variogram program in GSLIB (Deutsch and Journel, 1992) was modified in order to compute $\gamma_{st}(r_s, r_t)$, $\gamma_{st}(r_s, 0)$ and $\gamma_{st}(0, r_t)$. MATLAB was used for plotting these sample variograms. The 3-D cross-validation program in GSLIB was modified for spatial-temporal cross-validation. The 3-D kriging program in GSLIB was modified for spatial-temporal kriging. These modified programs will be contained in a manuscript under preparation.

REFERENCES

Bilonick RA. 1985. The space-time distribution of sulfate deposition in the northeastern United States. *Atmospheric Environment* **19**:1829–1845.

Brockwell PJ, Davis RA. 1987. *Time Series: Theory and Methods*, Springer-Verlag: New York.

Cressie NAC. 1993. *Statistics for Spatial Data*. John Wiley and Sons: New York.

De Cesare L, Myers D, Posa D. 1996. Spatial-temporal modeling of SO_2 in Milan District. In *Geostatistics Wollongong '96*, Vol. 2, Baafi EY, Schofield NA (eds). Kluwer Academic Publishers: Dordrecht; 1031–1042.

Deutsch CV, Journel AG. 1992. *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press: New York.

Eynon BP, Switzer P. 1983. The variability of rainfall acidity. *Canadian Journal of Statistics* **11**:11–24.

Le DN, Petkau AJ. 1988. The variability of rainfall acidity revisited. *Canadian Journal of Statistics* **16**:15–38.

Myers DE. 1991. On variogram estimation. *Proceedings of the First Inter. Conf. Stat. Comp.*, Cesme, Turkey, 30 March–2 April 1987, Vol. II, American Sciences Press, 261–281.

Myers DE. 1992. Spatial-temporal geostatistical modeling in hydrology. *Proceedings of the International Workshop*, Karlsruhe, Germany.

Rodriguez-Iturbe I, Meija JM. 1974. The design of rainfall networks in time and space. *Water Resources Research* **10**:713–728.

Rouhani S, Hall TJ. 1989. Space-time kriging of groundwater data. In *Geostatistics*, Vol. 2, Armstrong M (ed.). Kluwer Academic Publishers: Dordrecht; 639–651.

Soares A, Tavora J, Pinheiro L, Freitas C, Almeida J. 1992. Predicting probability maps of air pollution concentration: a case study on Barreiro/Seixal industrial area. In *Geostatistics*, Vol. 2, Soares A (ed.). Kluwer Academic Publishers: Dordrecht; 625–636.

Studi per la valutazione della qualità dell'aria nella provincia di Milano. 1993. Province of Milan and U.SS.L. 75/III (Local Health Department) of Milan.

