TOTAL AIR POLLUTION AND SPACE-TIME MODELLING

S. DE IACO
Università "G. D'Annunzio"di Chieti
Facoltà di Economia, Via dei Vestini, 66013 Chieti, Italy
sdeiaco@tiscalinet.it

D. E. MYERS
University of Arizona
Dept of Mathematics, 85721 Tucson AZ, USA
myers@math.arizona.edu

AND

D. POSA
Università di Lecce
Facoltà di Economia, Via per Monteroni, 73100 Lecce, Italy
Istituto per Ricerche di Matematica Applicata (CNR), Via Amendola 122/I, 70126 Bari, Italy
posa@economia.unile.it

Abstract. Two general problems occur in the analysis of air pollution data; multiple contaminants and a dependence on both spatial location and time of observation. Principal Component Analysis (PCA) provides a tool for removing the interdependence of the contaminant concentrations, in addition an analysis of the principal components, eigenvectors and eigenvalues provides additional insight into the dispersion and occurrence of the pollution plume. New models for space-time variograms and techniques for modelling them have been introduced by De Iaco, Myers and Posa.

Hourly average concentrations for nitric oxide (NO), nitrogen dioxide (NO_2) and carbon monoxide (CO) measured at 30 stations in 1999 in the Milan district, Italy, were used for the analysis. These were converted to daily averages and PCA was applied to each of the 365 data sets (3 contaminants and 30 stations). The eigenvectors of the correlation matrices were used to generate principal components, which can be considered as measures of Total Air Pollution (TAP) in lieu of the separate contaminant concentrations. These components were treated as samples from unobserved variates defined over space and time. Space-time variograms were fitted to these new variates using the product sum model.

P. Monestiez et al. (eds.), geoENV III – Geostatistics for Environmental Applications, 45–56.

Although linked in these analyses, the principal components and their associated eigenvectors as well as the scores for each station vs the space-time variogram models provide two different pictures of the spatial and temporal dispersion of the contaminants as well as their interaction at different times of the year.

1. Introduction

Air pollution plumes will commonly consist of multiple contaminants with perhaps two main groupings. One group will consist of contaminants emitted by vehicles (or contaminants resulting from chemical reactions involving those emitted from vehicles) and a second will consist of contaminants emitted from industrial sources (and their by-products). For either group, the processes that result in the emission of one contaminant are the same or related to the emission or formation of others. Hence, rather than focusing on only one contaminant at a time, it is reasonable to consider some measure of TAP, e.g., a weighted linear combination.

This approach has been applied to an air pollution data set from Milan district, Italy, involving three contaminants of considerable interest, NO, NO_2 and CO. Ground level measurements are available at 30 locations, taken hourly for a year. For the purposes of this analysis, the hourly data have been converted to daily averages.

Let $\mathbf{R}(s,t)$ denote a vector valued random function, defined on space-time, with components $R_1(s,t)$, $R_2(s,t)$ and $R_3(s,t)$. These three components will represent the values of NO, NO_2 and CO at the point (s,t) in space-time. Using a Linear Coregionalization Model (LCM) for the matrix variogram $\mathbf{\Gamma}(h)$ of $\mathbf{R}(s,t)$ corresponds to assuming that $\mathbf{\Gamma}(h)$ can be diagonalized, Myers (1994). That is, there is a matrix $\mathbf{A}$ such that

$$\mathbf{A}^T\mathbf{\Gamma}(h)\mathbf{A} = \mathbf{D}(h).$$

The diagonal entries in $\mathbf{D}(h)$ are the variograms of uncorrelated random functions and such that each of the components of $\mathbf{R}(s,t)$ can be written as linear combinations of these uncorrelated random functions. To determine $\mathbf{A}$ or alternatively to construct the LCM would require modelling not only the variograms of the three components but also the cross-variograms for each pair. PCA provides an alternate tool for generating multiple linear combinations that are uncorrelated. Let $X_i, i, \ldots, 365$ denote the data array for day i. That is, the entries of X_i are the observed values of $\mathbf{R}(s,i)$. The columns correspond to the three contaminants and the rows correspond to the locations. Let Y_i denote the standardized data array (obtained by subtracting the column mean and dividing by the column standard deviation). Then $(1/N_i)Y_i^TY_i$ is the correlation matrix for the data array, where

N_i is the number of locations for day i. Let U_i, V_i, W_i denote the eigenvectors corresponding to the eigenvalues u_i, v_i, w_i of these correlation matrices. Then Y_iU_i, Y_iV_i and Y_iW_i are orthogonal weighted linear combinations of the standardized data for each day. As it will be shown later, the first two principal component explain a very large percentage of the variance. It is then reasonable for Y_iU_i, Y_iV_i to be considered as the observed values of two measures of TAP for day $i, i = 1, \ldots, 365$ at the N_i locations. These measures, selected in a suitable way, as it will be described herein, will be called TAP1 and TAP2. This data will then be used to model space-time variograms which in turn will allow interpolating TAP1 and TAP2 to non-data locations and also to predict their values at future times.

As an alternative to PCA one might model both variograms and cross-variograms to cokrige linear combinations but this will provide less insight into the choice of the linear combinations. PCA has been used previously to avoid the complexity of modelling cross-variograms, e.g., Davis and Greenes (1983), Myers and Carr (1984). The connections between the use of PCA, factorial kriging as well as the use of a LCM and diagonalization of the variogram matrix are discussed in Myers (1994).

These PCA results can be used in several ways to produce a better picture of the air pollution levels both in space and time. Since each Y_iU_i, Y_iV_i is a weighted linear combination of the columns of the (standardized) data, the row identifications are retained. That is, each row entry in Y_iU_i corresponds to a location. Hence, the selected combinations, Y_iU_i and Y_iV_i, might be considered as two samples from two different random functions defined in space-time. One possibility is to apply spatial analysis for each day separately, e.g., estimate and model (spatial) variograms for each day. One can then interpolate to construct a contour map of TAP for each day. A second possibility is to estimate and model a space-time variogram. The modelling results are given in section 4. These space-time models can be used for interpolation in space and prediction in time. An examination of the eigenvalues, the loadings/scores for the eigenvectors/variables and their behavior in time provides additional insight into the behavior of the air pollution patterns.

2. Air Pollution in Milan District

Air pollution in the Milan district may be attributed to different factors: emissions from motor vehicles, manufactoring work and heating systems during winter. The air pollution monitoring network for NO, NO_2 and CO during 1999 is shown in Figure 1. The same figure shows the following classification of the monitoring stations according to the Premier's Decree in 1991:

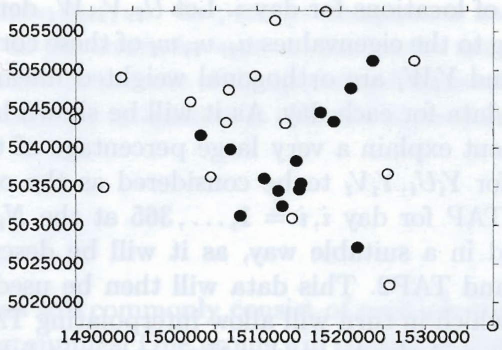

Figure 1. Posting map of 30 survey stations in Milan district and their classification.

– stations characterized by high density population (empty circles);
– stations characterized by heavy traffic (black circles).

NO is an index of air pollution generated by all types of heating systems and motor vehicles; NO_2 is a secondary pollutant resulting from the oxidation of NO in the air. CO is a direct index of contamination resulting from petrol-driven motor vehicles, in particular, CO emissions increase as the motor vehicle's speed decreases.

3. PCA Analysis

One of the purposes of the following analysis is to find simple underlying components and to attribute physical meaning to them. Figure 2 shows the eigenvalues, of the correlation matrices for the X_i's, viewed as a time series. Note that the first component explains approximately 70% of the total variance for each day although this dominance is reduced in summer. The second component is more important in summer and together the first two components explain more than 90% of the total variance for the whole year. In Figure 3 the loadings of NO and NO_2, for the first component, are compared with the loadings of CO over the whole year: note that the contribution of the 3 pollutants is approximately the same over the whole year. In particular, the loadings of CO are relatively smaller, especially in summer, than the loadings of the other 2 pollutants. There is an outlier at $i = 360$: on that day (the 26th of december) there was an Atlantic storm coming in from France and it was very foggy in the Milan district. The atmospheric conditions significantly affected the NO_2 daily behaviour and its relation to the other pollutants. For the first component, all the linear combinations for which the eigenvectors have the sign pattern (+,+,+) for

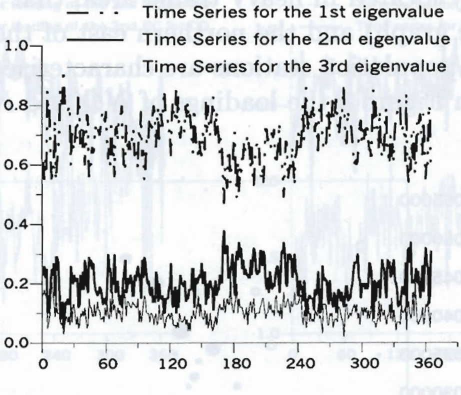

Figure 2. Time series for eigenvalues.

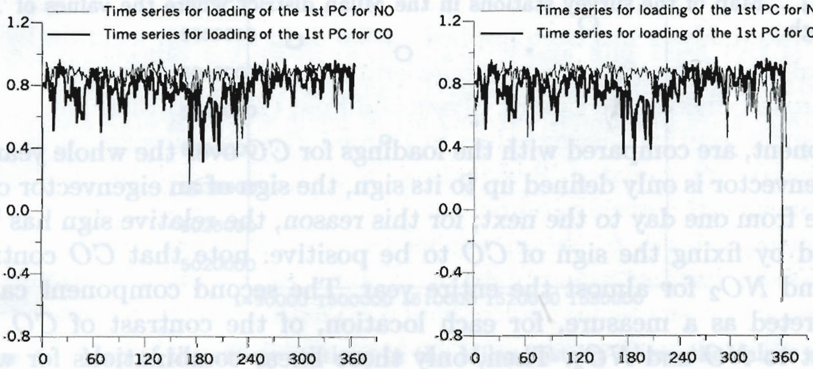

Figure 3. Time series for loadings of the 1st component.

NO, NO_2 and CO, will be considered. Note that only the 26th of December has been deleted. Since the loadings, selected in such a way, are all positive and almost equal, the above linear combinations can be interpreted as a measure of TAP generated by NO, NO_2 and CO and this measure will be called TAP1. In order to identify monitoring stations where TAP1 is critical, the quartiles of the its distribution have been considered. In Figure 4 the size of the posting symbol is proportional to the frequency of scores greater than the 3rd quartile for each monitoring station during the year. Moreover, the characterization of the stations (empty and black circles) has been preserved as in Figure 1. Note that the stations where TAP1 presents high values are primarily located in heavy traffic area, that is the city of Milan (central area in the graph) and the northern east of the district. Only two of the high density population stations are characterized by relatively high values of TAP1. In Figure 5 the loadings of NO and NO_2, for the second

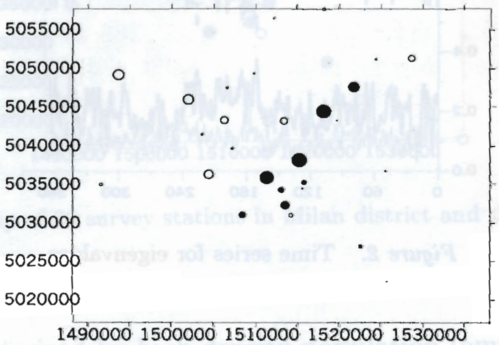

Figure 4. Map of the survey stations in the Milan district where the values of TAP1 were high.

component, are compared with the loadings for CO over the whole year. As an eigenvector is only defined up to its sign, the sign of an eigenvector could change from one day to the next; for this reason, the relative sign has been studied by fixing the sign of CO to be positive: note that CO contrasts NO and NO_2 for almost the entire year. The second component can be interpreted as a measure, for each location, of the contrast of CO with respect to NO and NO_2. Then, only those linear combinations for which the eigenvectors of the second component have the sign pattern (-,-,+) for NO, NO_2 and CO, respectively, will be retained. The corresponding linear combinations will be considered as data for TAP2. As for the first measure, the quartiles of the TAP2 distribution have been considered. In Figure 6 the size of the posting symbol is proportional to the frequency of scores greater than the 3rd quartile for each monitoring station during the year. Moreover, the characterization of the stations (empty and black circles) has been preserved as in Figure 1. Note that the stations where TAP2 presents high values, that is where the contrast of CO with respect to NO and NO_2 is more evident, are primarily located in the peripheral area of the district and they correspond to stations characterized by high density population. As this contrast explains a larger variance during summer, the following phisycal interpretation could be given: NO and NO_2 naturally record minimum values in peripheral urban centers, while high values of CO persist because of the usual urban traffic.

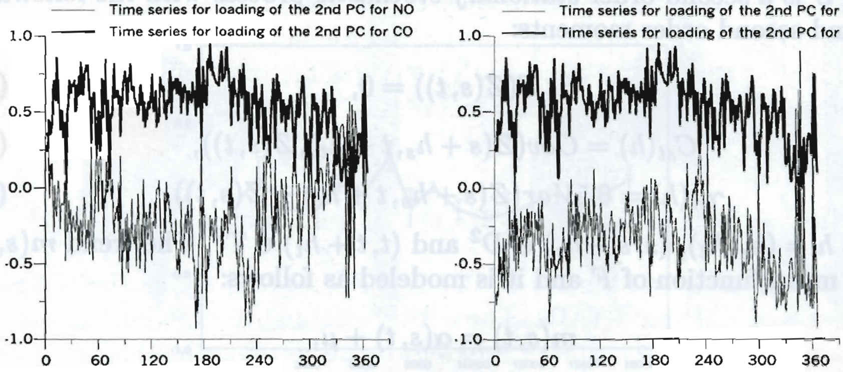

Figure 5. Time series for loadings of the 2nd component.

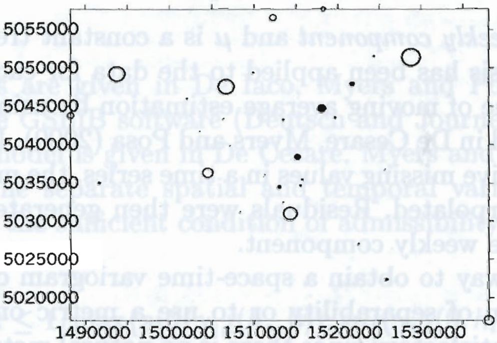

Figure 6. Map of the survey stations in the Milan district where the values of TAP2 were high.

4. Space-Time Correlation Models

The space-time analysis uses the data for TAP1 and TAP2 at 30 monitoring stations for 365 days of 1999. TAP1 and TAP2 are considered as space-time random fields, with general properties as follows:

$$F = \{F(s,t), (s,t) \in D \times T\}, \tag{1}$$

where $D \subset \Re^2$ and $T \subset \Re_+$. Assuming that the first and second moments of F exist, F can be decomposed as

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

where Z is a second order stationary stochastic process with the following first and second order moments:

$$E(Z(s,t)) = 0, \tag{3}$$

$$C_{st}(h) = Cov(Z(s+h_s, t+h_t), Z(s,t)), \tag{4}$$

$$\gamma_{st}(h) = 0.5Var(Z(s+h_s, t+h_t) - Z(s,t)), \tag{5}$$

where $h = (h_s, h_t)$, $(s, s+h_s) \in D^2$ and $(t, t+h_t) \in T^2$. The trend $m(s,t)$ is the mean function of F and it is modeled as follows:

$$m(s,t) = \alpha(s,t) + \mu,$$

where

1. $\alpha(s,t) = \alpha(s,t+7) \qquad \forall s \in D \quad \forall t, t+7 \in T,$
2. $\sum_{j=1}^{7} \alpha(s,j) = 0 \qquad \forall s \in D.$

$\alpha(s,t)$ is called *weekly component* and μ is a constant trend over the year. Time series analysis has been applied to the data for each location by the standard technique of moving average estimation by using the FORTRAN program described in De Cesare, Myers and Posa (2000). If there were fewer than five consecutive missing values in a time series, the missing values have been linearly interpolated. Residuals were then generated for all stations after removing the weekly component.

The simplest way to obtain a space-time variogram or covariance is to assume some form of separability or to use a metric on space-time. The latter is not too satisfactory since there is no natural metric for space-time. It is known that constructing a space-time model as the sum of a spatial covariance and a temporal covariance can result in a semi-definite function rather than a positive definite function, Rouhani and Myers (1990). While the product of a space covariance and a time covariance does produce a valid model, this method is somewhat restrictive. An example of this construction for air pollution data is found in De Cesare, Myers and Posa (1997). The product model was extended to the product-sum model in De Cesare, Myers and Posa (2001) and further generalized in De Iaco, Myers and Posa (2001). The generalized product-sum model is of the form:

$$\gamma_{s,t}(h_s, h_t) = \gamma_{s,t}(h_s, 0) + \gamma_{s,t}(0, h_t) - k\gamma_{s,t}(h_s, 0)\gamma_{s,t}(0, h_t), \tag{6}$$

where $\gamma_{s,t}(h_s, 0)$ and $\gamma_{s,t}(0, h_t)$ are valid spatial and temporal bounded variogram functions and:

$$k = \frac{(sill\gamma_{s,t}(h_s,0) + sill\gamma_{s,t}(0,h_t) - sill\gamma_{s,t}(h_s,h_t))}{(sill\gamma_{s,t}(h_s,0)sill\gamma_{s,t}(0,h_t))}. \tag{7}$$

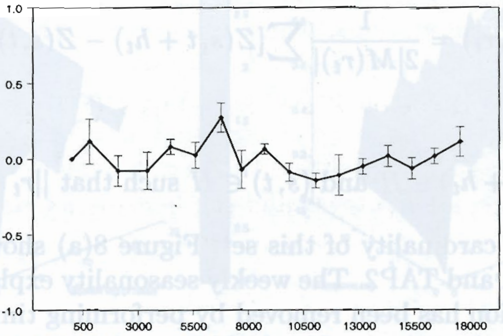

Figure 7. Error bar of monthly sample spatial cross-correlograms between TAP1 and TAP2.

Theoretical results are given in De Iaco, Myers and Posa (2001) and a modification of the GSLIB software (Deutsch and Journel, 1997) to apply the product-sum model is given in De Cesare, Myers and Posa (2000).

In modelling the separate spatial and temporal variograms, the sills are chosen so that the sufficient condition of admissibility for $\gamma_{s,t}(h_s, h_t)$ is satisfied, namely:

$$0 < k \leq 1/\max\{sill(\gamma_{s,t}(h_s, 0)); sill(\gamma_{s,t}(0, h_t))\}. \tag{8}$$

Note that k is selected in such a way to ensure that the global sill is fitted.

4.1. SPACE-TIME VARIOGRAM MODELLING

Before performing space-time variogram modelling for both deseasonalized measures of TAP, sample spatial cross-correlograms were computed. The error bar of monthly averaged cross-correlograms shows that TAP1 and TAP2 are orthogonal at any scale in space (Figure 7). Let H be the set of data locations, then the estimator for the spatial variogram at lag r_s, with spatial tolerance δ_s, is:

$$\widehat{\gamma}_{st}(r_s, 0) = \frac{1}{2|N(r_s)|} \sum [Z(s + h_s, t) - Z(s, t)]^2, \tag{9}$$

where the summation is over the set

$$N(r_s) = \{(s + h_s, t) \in H \text{ and } (s, t) \in H \text{ such that } \|r_s - h_s\| < \delta_s\},$$

and $|N(r_s)|$ is the cardinality of this set. Similarly

$$\widehat{\gamma}_{st}(0, r_t) = \frac{1}{2|M(r_t)|} \sum [Z(s, t + h_t) - Z(s, t)]^2, \tag{10}$$

where

$$M(r_t) = \{(s, t + h_t) \in H \text{ and } (s, t) \in H \text{ such that } ||r_t - h_t|| < \delta_t\},$$

and $|M(r_t)|$ is the cardinality of this set. Figure 8(a) shows the temporal variogram of TAP1 and TAP2. The weekly seasonality explained by the two measures of pollution has been removed by performing time series analysis for each location and residuals of both TAP1 and TAP2 have been used to compute sample spatial and temporal deseasonalized variograms (Figures 8(b,c)). Fitted models for these together with the resulting space-time models are given below

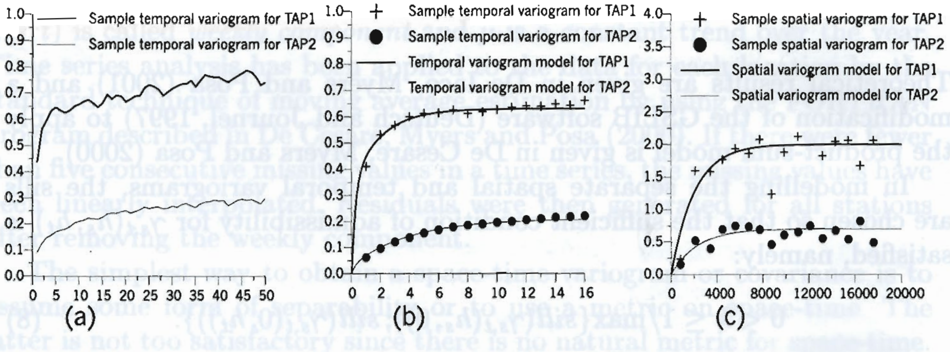

Figure 8. Sample temporal non deseasonalized (a), temporal (b) and spatial (c) deseasonalized variograms for TAP1 and TAP2.

$$\gamma_{s,t}^{(TAP1)}(h_s, 0) = 2(1 - exp(-h_s/2000)), \tag{11}$$

$$\gamma_{s,t}^{(TAP2)}(h_s, 0) = 0.7(1 - exp(-h_s/2300)), \tag{12}$$

$$\gamma_{s,t}^{(TAP1)}(0, h_t) = 0.34(1 - exp(-4h_t)) + 0.29(1 - exp(-h_t/2)), \tag{13}$$

$$\gamma_{s,t}^{(TAP2)}(0, h_t) = 0.03(1 - exp(-4h_t)) + 0.18(1 - exp(-h_t/4)), \tag{14}$$

and they are shown in Figure 8(b,c). For TAP1 and TAP2, the sill value $C_{st}(0, 0)$ of $\gamma_{st}(h_s, h_t)$ (called "global"sill in the literature) has been estimated graphically by plotting the spatial-temporal variogram surface of

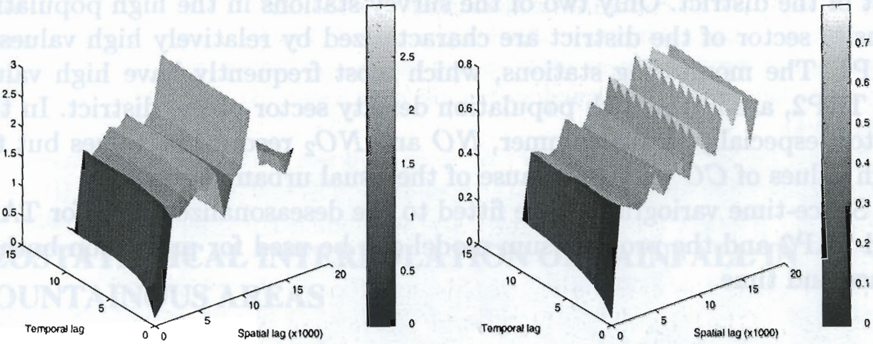

Figure 9. Sample space-time variogram surfaces of residuals for TAP1 and TAP2.

the residuals (see Figures 9), in order to compute (7) and generate the model (6) which can then be used for prediction in space and time. The "global"sill values for TAP1 and TAP2 are 2.3 and 0.7, respectively. The resulting space-time admissible models for TAP1 and TAP2 are:

$$\gamma_{s,t}^{(TAP1)}(h_s,h_t) = \gamma_{s,t}^{(TAP1)}(h_s,0) + \gamma_{s,t}^{(TAP1)}(0,h_t) + \\ -0.26[\gamma_{s,t}^{(TAP1)}(h_s,0)\gamma_{s,t}^{(TAP1)}(0,h_t)], \tag{15}$$

$$\gamma_{s,t}^{(TAP2)}(h_s,h_t) = \gamma_{s,t}^{(TAP2)}(h_s,0) + \gamma_{s,t}^{(TAP2)}(0,h_t) + \\ -1.43[\gamma_{s,t}^{(TAP2)}(h_s,0)\gamma_{s,t}^{(TAP2)}(0,h_t)]. \tag{16}$$

5. Summary

NO, NO_2 and CO air pollution patterns in the Milan district, Italy, during 1999 have been analyzed using two measures of TAP. These measures were constructed as linear combinations of NO, NO_2 and CO, the weights were determined by the use of PCA. TAP1 and TAP2 are considered as random functions defined in space-time. The first and second principal component, selected in a suitable way, are considered as daily data for TAP1 and TAP2. This data is used to model space-time variograms using a generalized product sum model. TAP1, which is modeled from the first principal component, is the most important because the daily first principal component explains about 70% of the total variance. The monitoring stations, which most frequently have high values for the first measure, are primarily located in heavy traffic area, that is the city of Milan and the northern east of the district. Only two of the survey stations in the high population density sector of the district are characterized by relatively high values of TAP1. The monitoring stations, which most frequently have high values for TAP2, are in the high population density sector of the district. In this sector, especially in the summer, NO and NO_2 record low values but the high values of CO persist because of the usual urban traffic.

Space-time variograms were fitted to the deseasonalized data for TAP1 and TAP2 and the product-sum model can be used for prediction both in space and time.

Acknowledgments

The authors would like to thank Dr. R. Gualdi, Dr. G. Tebaldi and Dr. M. Valentini of Environmental Protection Agency of Lombardy Region for providing the data set and all information about Milan district.

References

Davis, B. and Greenes, (1983), Estimation using distributed multivariate data: an example with coal quality, *Math. Geology* 15, 2, 287-300.

De Cesare, L., Myers, D. E. and Posa, D., (1997), Spatial Temporal Modeling of SO_2 in the Milan District, *Geostatistics Wollongong '96*, E.Y. Baafi and N.A. Schofield (eds), Kluwer Academic Publishers, 1031-1042.

De Cesare, L., Myers, D.E. and Posa, D., (2000), A FORTRAN program for Space-Time Modeling, submitted.

De Cesare, L., Myers and Posa, D., (2001), Estimating and modelling Space-Time Correlation Structures, *Statistics and Probability Letters*, 51, 1, 9-14.

De Iaco, S., Myers, D.E. and Posa, D., (2001), Space-Time analysis using a general product-sum model, *Statistics and Probability Letters*, 52,1, 21-28.

Deutsch, C. V. and Journel, A. G. (1997). *GSLIB: Geostatistical Software Library and User's Guide*, Oxford Univ. Press, New York.

Myers, D.E., (1994), The Linear coregionalization and simultaneous diagonalization of the variogram matrix function. *Sciences de la Terre, 32*, 125-139.

Myers, D.E. and Carr, J., (1984), Cokriging and Principal Component Analysis: Bentonite Data revisited. *Sciences de la Terre*, 21, 65-77.

Rouhani, S. and Myers, D.E., (1990), Problems in Space-Time Kriging of Hydrogeological data. *Math. Geology*, 22, 611-623.