05JEI00041
1726-2135/1684-8799

www.iseis.org/jei

Journal of Environmental Informatics 5 (1) 9-16 (2005)

A Multivariate Approach for the Analysis of Spatially Correlated Environmental Data

A. Lamberti[1*] and E. Nissi[2]

[1]ISTAT - Via C. Balbo, 16 - 00184 Roma, Italy
[2]Dipartimento di Metodi Quantitativi e Teoria Economica, Viale Pindaro, 42 - 65127 Pescara, Italy

ABSTRACT. The formulation and the evaluation of environmental policy depend upon a general class of latent variable models known as multivariate receptor models. Estimation of the number of major pollution sources, the source composition profiles and the source contributions are the main interests in multivariate receptor modelling. Many different approaches have been proposed both when the number of sources is unknown (explorative factorial analysis) and when the number and the type of sources are known (regression models). The objective of this work is to propose a flexible approach to the multivariate receptor models that incorporates the extra variability due to the spatial dependence. The method is applied to Lombardia air pollution data.

Keywords: Covariance modelling, environmental data, latent variable models, multivariate receptor models, spatio-temporal modelling

1. Introduction

In the past few years interest in air quality monitoring has increased, specifically pertaining to the identification of pollution sources and their information needed to implement air pollution control programs. Since observing the quantity of various pollutants emitted from all potential pollution sources is virtually impossible, receptor models are used to analyze concentrations of pollutants or particles measured over time in order to gain insight concerning the unobserved pollution sources. Multivariate receptor modeling aims to identify the pollution sources and assess the amounts of pollution by resolving the measured mixture of chemical species into the contributions from the individual source types. The basic physical model comes from the laws of chemistry. The number of sources is the first problem we encounter. When the number and the composition of pollution sources are unknown, factor analytic approaches have been employed in order to identify pollution sources. As in the factor analysis models, the choice of the number of pollution sources (factors) used in receptor models is crucial.

Generally, the number of sources is chosen using one of many methods (often ad- hoc methods) suggested in the literature. Park, Henry and Spiegelman (1999) provide a review with discussion of several of these methods. However, these methods often are not satisfying and in many papers the number of pollution sources is fixed on the basis of previous studies and/or specific assumptions made by the researcher. Once a model with k sources has been fitted, interest often lies in describing the composition of each pollution source and the amount of pollution emitted from each source. Such information is of great value when formulating and evaluating air quality policy.

To make sound decisions from the data, it is necessary to make inferences about the fitted model; however statistical tools for such data have not received much attention in the literature. Pollution data collected over time and/or space often exhibit dependence which needs to be accounted for in the procedures for inference on model parameters.

The objective of this paper is to present a flexible approach to multivariate receptor models for incorporating the spatial dependence exhibited by the data and then show the usefulness of the procedure using air pollution data from Lombardia area.

The paper is organized as follow: in section 2 we restate the model from a statistical point of view, section 3 contains the methodological issues related to spatial covariance estimation and in section 4 we present the application of the proposed model to the air pollution data.

2. A flexible multivariate receptor model

Let p be the number of pollutants and k be the number of sources. Based on the chemical mass balance equation and assuming that the relative amounts of the pollutants are approximately the same traveling from sources to receptor, a multivariate receptor model can be written as follow (Park et al., 2002):

$$y_t = \sum_{j=1}^{k} \Lambda_j f_{tj} + e_t \qquad t = 1, \ldots, M \tag{1}$$

* Corresponding author: aldo.lamberti@istat.it

where $y_t = (y_{t1}, ..., y_{tp})$ is the t-*th* observation at the receptor, $\Lambda_j = (\Lambda_{j1}, ..., \Lambda_{jp})$ is the *j*-th source composition profile composed of the portion of each pollutant in the emission from the j-*th* source at time *t* and $e_t = (e_{t1}, ..., e_{tp})$ is the measurement error in the *t*-th observation.

In vector form, model (1) can be written as:

$$y_t = \mathbf{\Lambda} \mathrm{f}_t + e_t \tag{2}$$

where $\mathbf{f}_t = (f_{t1}, ..., f_{tk})$ and $\mathbf{\Lambda}$ is a $p \times k$ non-negative source composition matrix and its columns are the source composition profiles.

From a statistical perspective, model (2) can be viewed as a latent variable model (Bartholomew and Knot, 1999), in particular as a factor analysis model where $\mathbf{y}$ is a set of p variables that can be directly observed, $\mathbf{f}$ is a set of k latent variables or factors (unobservable), $\mathbf{\Lambda}$ is the unknown $p \times k$ factor loading and *k* is the unknown number of factors.

Two different approaches have been used in the literature depending on the knowledge of the number and the nature of the pollution sources.

When the number and nature of the pollution sources are known (in this case this means that $\mathbf{\Lambda}$ is known), the pollution source contributions can be estimated using regression or measurement error models. Conversely when some of the elements of the source composition matrix $\mathbf{\Lambda}$ are not known, the estimates of the pollution sources contributions can be obtained using linear factor analysis models.

Much of the multivariate receptor modeling studies in the literature use exploratory factor analytic techniques to identify the number of pollution sources, the pollution source compositions and the source contributions. However, this goal cannot be achieved without additional assumptions on the model. The unknown number of sources (factors) *k*, is the first problem because $\mathbf{\Lambda}$ and $\mathbf{f}_t$ depend on k in the model (2). Secondly, the parameters in the model (2) are not uniquely defined, even under the assumption that *k* is known. This means that there may be other parameterizations that produce the same data (rotational indeterminacy of factors plays here a major role).This is called nonidentifiability in latent variable models and additional restrictions on the parameters are required to remove it. Park et al. (2001) discussed a wide range of identifiability conditions for multivariate receptor models when the number of sources k is assumed to be known. In a receptor modelling feasibility study Javitz, Watson, Guertin and Muller (1988) made several recommendations for future developments in receptor modeling. They noted the non-unique nature of exploratory factor analysis fit of the multivariate receptor model (2) when matrix $\mathbf{\Lambda}$ is unknown, and they also noted that it is often impossible to obtain complete and accurate source composition information necessary to fit a chemical mass balance model using regression. These authors noted the need for future development of a physically meaningful hybrid model which could be used with only partial source composition information and pointed out the importance of estimates of uncertainties associated with the model which are necessary for inference. The use of a flexible latent variable model allows the researcher to incorporate physical constraints, past data or other subject matter knowledge in the model and guarantees valid model fits using only limited information about the relationship between the observed ambient species and the pollution sources.

The main difference between the use of multivariate receptor models in the literature and the use of linear factor analysis models is that the observations in a pollution data set are rarely if ever independent. Multivariate receptor models, in fact, are used to model data that exhibit temporal and/or spatial dependence. Several potential hazards arise when factor analysis ignores dependence structure, most of them related to invalidity of inferential techniques.

3. Spatial covariance estimation

The hazard of ignoring temporal and spatial dependence is implicitly assumed in almost every study involving receptor modelling. One exception was given by Park et al. (2001) who incorporated temporal dependence structure directly into hierarchical model and then estimated the model parameters using Markov Chain Monte Carlo methods.

Spatial dependence can be incorporated in the model finding an appropriate estimate of the spatial covariance matrix that take into account the spatial structure of the data.

In this paper we account for dependence structure using the method proposed by Nott and Dunsmuir (1998) for estimating non stationary spatial covariance structure from space time data. The methods are computationally attractive and can be extended to the assessment of covariance for multivariate processes.

Departing from the estimated spatial covariance structure, we can use the classical multivariate receptor/latent variable model, which can be fit using existing software package, that yield a unique model fit based on only partial source profile information.

An advantage of our approach is that it requires no assumption about distributional form and prior distributions for parameters.

We must introduce some additional notation in order to formulate the general problem of spatial covariance estimation for multivariate space-time data.

$$\text{Let } Y = \left\{ Y(s, t), s \in \Re^2, t \in \Re_+ \right\} \tag{3}$$

be a multivariate spatio-temporal process with

$$Y(s, t) = \left(Y^1(s, t), ..., Y^q(s, t) \right)^T \tag{4}$$

We assume that $\mathbf{Y}$ can be observed at a collection of sites $I = \{s_1, ..., s_n\}$ and for a collection of times $T = \{t_1, ..., t_M\}$.

Write

$$Y_i = \left(Y^1(s_1, t_i), \ldots, Y^1(s_n, t_i), \ldots, Y^q(s_1, t_i), \ldots, Y^q(s_n, t_i) \right)^T$$

for the vector of the spatial measurements at time $\mathbf{t}_i$, and write $\mathbf{y}_i$ for an observed realization of $\mathbf{Y}_i$. If $\mathbf{Y}$ is temporally ergodic then we can define the $q \times q$ matrix valued spatial covariance function of $\mathbf{Y}$ as:

$$R_Y(s, u) = \left[\mathrm{Cov}\left(Y^l(s, t), Y^j(u, t) \right) \right]_{l \le i, j \le q} \tag{5}$$

and we can estimate spatial covariance between pairs of monitored sites by averaging over time.

In this paper, following the approach suggested by Nott and Dunsmuir (1998), we estimate site means by averaging over time. In particular we write $\bar{y}$ for the spatial mean vector obtained in this way:

$$\bar{y} = \frac{1}{M} \sum_{i=1}^{M} y_i$$

with spatial trend estimated by site means, a spatial covariance matrix $\mathbf{\Gamma}$ can be estimated as:

$$\mathbf{\Gamma} = \frac{1}{M} \sum_{i=1}^{M} (y_i - \bar{y})(y_i - \bar{y})^{\mathrm{T}} \tag{6}$$

If $\mathbf{\Gamma}$ is partitioned into nxn blocks of size qxq, each block can be interpreted as an empirical spatial covariance or cross-covariance matrix for the components of $\mathbf{Y}$ at the monitored sites.

Given an estimate of $\mathbf{\Gamma}$ it is important in many spatial modeling problem to estimate valid (non-negative definite) covariance function of $\mathbf{Y}$ based on the information in $\mathbf{\Gamma}$.

Following the method suggested by Nott and Dunsmuir (1998), one way of estimating a valid non-negative definite spatial covariance function from $\mathbf{\Gamma}$ is by reproducing $\mathbf{\Gamma}$ at monitored sites and then describing conditional behaviour given monitoring sites values by a stationary process or collection of a stationary processes.

To describe the idea of Nott and Dunsmuir more precisely, we need some more notations. Let $\{W(s);\ s \in \Re^2\}$ be a multivariate zero mean, stationary Gaussian process with q components, $W(s) = \left(W^1(s), \ldots, W^q(s) \right)^T$, with covariance function R(h). Let $\mathbf{W}$ denote the vector:

$$\mathbf{W} = \left(W^1(s_1), \ldots, W^1(s_n), W^q(s_1), \ldots, W^q(s_n) \right)^T$$

and write $\mathbf{C}$ for the covariance matrix of $\mathbf{W}$.

Also write the $nq \times q$ cross-covariance matrix between $\mathbf{W}$ and $W(s)$ as

$$c(s) = \begin{pmatrix} \mathrm{Cov}\left(W^1(s_1), W^1(s), \ldots, W^1(s_n), W^1(s) \right), \ldots, \\ \mathrm{Cov}\left(W^q(s_1), W^q(s), \ldots, W^q(s_n), W^q(s) \right) \end{pmatrix}^T$$

If we observe values of the process at the monitored sites, $\mathbf{W} = \mathbf{w}$ say, then for an arbitrary collection of sites we can write down the joint distribution of $\mathbf{W}$ at these sites. These are the finite dimensional distributions of a random field which describes the conditional behaviour of $W(\cdot)$ given $\mathbf{W} = \mathbf{w}$, and such a random field has a representation:

$$c(s)^T C^{-1} w + \delta(s) \tag{7}$$

where $\delta(\mathbf{s})$ is a zero mean Gaussian process with covariance function:

$$R_\delta(s, u) = R(u - s) - c(s)^T C^{-1} c(u)$$

One simple way of constructing a valid non-negative definite covariance function which reproduces the empirical spatial covariance matrix $\mathbf{\Gamma}$ at the monitored sites is to replace $\mathbf{w}$ in the above representation by a random vector $\mathbf{W}^*$ which has zero mean and the covariance matrix $\mathbf{\Gamma}$ independent of $\delta(\mathrm{s})$.

Intuitively, we are constructing a process with the covariance matrix $\mathbf{\Gamma}$ at the monitored sites but with the conditional distribution given values at monitored sites the same as those of the stationary random field $W(\mathrm{s})$.

It must be emphasized that the covariance function of W(s) is not the model used for the unconditional covariance but is merely a part of a construction to obtain valid, non-negative definite non-stationary spatial covariance function model for $\mathbf{Y}$:

$$\begin{aligned} &\mathrm{Cov}\left(Y(s, t), Y(u, t) \right) \\ &\quad = R(u - s) - c(s)^T C^{-1} (\Gamma - C) C^{-1} c(u) \end{aligned} \tag{8}$$

This covariance function reproduces $\mathbf{\Gamma}$ (that is, evaluating (8) at $(s, u) = (s_i, s_j)$ gives $\mathbf{\Gamma}_{ij}$) since $c(\mathrm{s}_i)$ is simply the i-th column of $\mathbf{C}$. Hence one simple way of constructing a non-negative definite estimate of spatial covariance is to fit a stationary model to $\mathbf{\Gamma}$ and to then compute (8) with $R(u - s)$, $\mathbf{C}$, $c(s)$ and $c(u)$ evaluated according to the fitted stationary model.

4. Analysis of Milano-Bergamo air pollution data

We apply the model to air pollution data in the Milano-Bergamo districts. In particular we consider the daily average of CO, NO_X, NO_2 and SO_2 obtained from 23 monitoring sites during May - June 2000 (Figure 1).

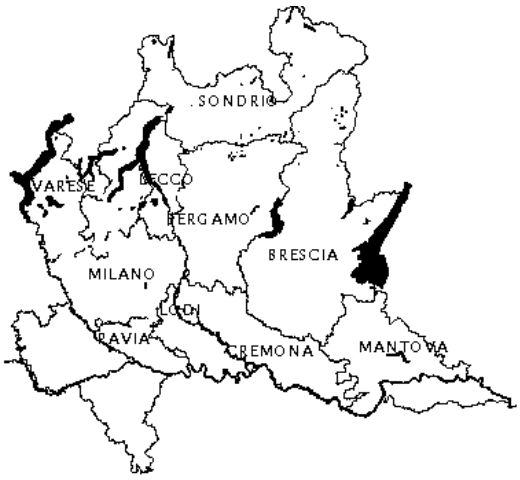

Figure 1. Geographic map of the Lombardia Region.

The monitoring sites are: Zavattari, Verziere, Limito, Melegnano, Corsico, Pero, Legnano S. Magno, Carate Brianza, Vimercate, Arese, Settimo, S.Giuliano, Cormano1, Magenta, Ponte S. Pietro Nembro, Seriate Treviglio, Ciserano, San Giorgio, Costa Volpino, Garibaldi, Goisis (Figure 2).

The goal of the analysis is to identify major sources of the pollutant variables. Here, the 23 monitoring sites play the role of the variables in our basic multivariate receptor model.

The source profile, consisting of the relative amount of pollutant that are conveyed to the 23 monitoring sites in this case represents the spatial pattern underlying CO, NO_X, NO_2 and SO_2 concentration from each source (Figure 3).

The underlying assumptions for this approach are:

1) there are few underlying spatial patterns and they do not vary over time;

2) the environmental factors such as wind do not interact with **Λ,** the overall spatial wind flow pattern (on which the spatial source pattern depend) are approximately constant.

After trend removal (Tables 1 and 2) and missing data reconstruction, the first step of the analysis was the estimation of the spatial covariance matrix.

Table 1. Estimated Variogram Parameters

	Nugget	Sill	Range
CO	0.051	0.129	0.12
NO	8.4	42.55	34.15
NO_x	9.05	42.90	39.60
SO_2	0	2670	104

Table 2. Estimated Trend Surface Parameters

	β_0	β_1	β_2	β_3	β_4	β_5
CO	-9901.00	193.9	333.00	14.00	47.00	26.00
NO	-89031.00	427.7	222.7	66.00	125.00	30.00
NO_x	-23813.00	342.00	841.00	4.00	9.00	7.00
SO_2	-66413.00	49.2.00	248.20	10.00	4.00	24.00

The use of the method described above requires calculat-

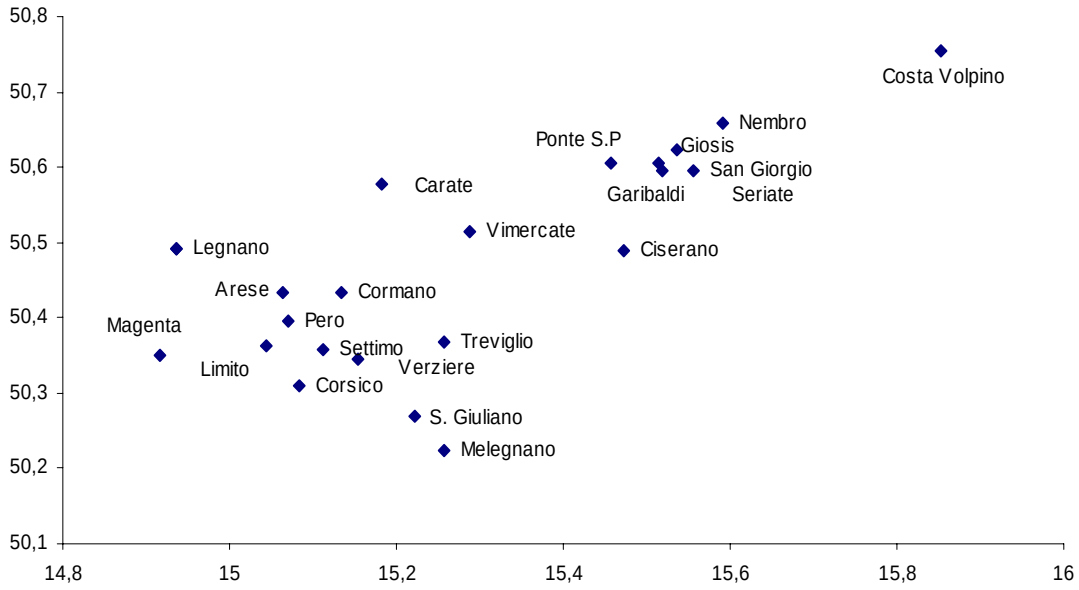

Figure 2. Monitoring sites in Milano Bergamo district.

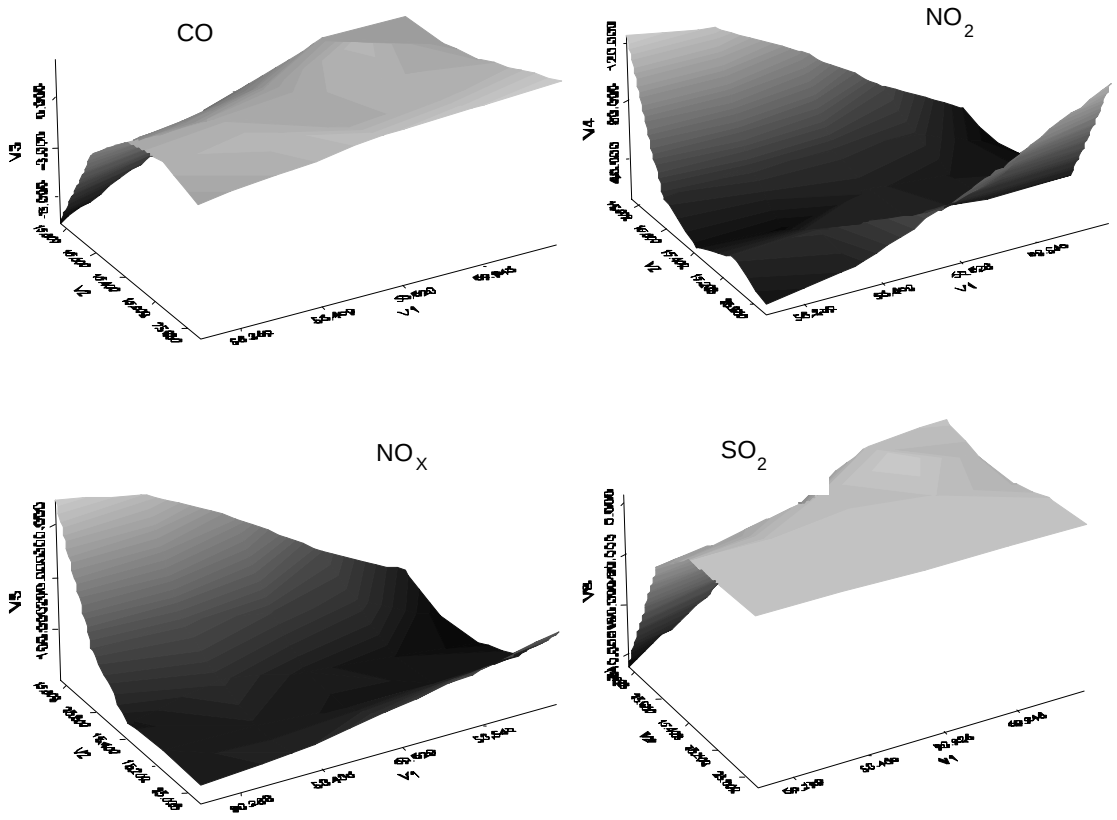

Figure 3. Plot of the original data (1st June, 2000).

ing the empirical spatial covariance structure. Figure 4 shows the empirical and the fitted variogram using the exponential model.

Departing from the estimated spatial covariance structure we have applied the multivariate receptor models with $k = 3$ sources (factors) for each variable assuming the 1st of June as reference day. As we can see in Table 3 the analysis carried out with $k = 3$ sources is quite satisfying, with cumulative variance explained ranging from 81.5% (NO_X) to 96.5% (SO_2).

For this type of data the three major pollution sources are: vehicle exhaust, industrial emissions and non-industrial emissions. In Figure 5, we can see the plot of the factors for each variable considering the first six loadings in order of importance.

For comparison purposes we applied the model with $k = 4$ sources (factors) to the same data but the gain in terms of explained variance is negligible. Then, we can say that the model with $k = 3$ sources is appropriate to describe the data we used for the analysis and this is in accordance with past information about this kind of data. Carrying out the analysis without taking into account the dependence exhibited by the data could be very misleading.

5. Conclusions

Identification of major pollution sources and their contributions can be assessed by a class of latent variable models known as multivariate receptor models. Using very limited information on the pollution sources, it is possible to fit a multivariate receptor model that is uniquely identified and the model parameter estimates have meaningful interpretations. Air quality data exhibit temporal and/or spatial dependence that is often ignored at the expense of valid inference. In this paper we incorporate dependence structure estimating a non-stationary spatial covariance matrix for multivariate space-time data where a given spatial covariance matrix is reproduced at a collection of monitored sites and conditional behaviour, given monitored site values, is described by a stationary process. Using this spatial covariance estimate in the model gives good results. A possible

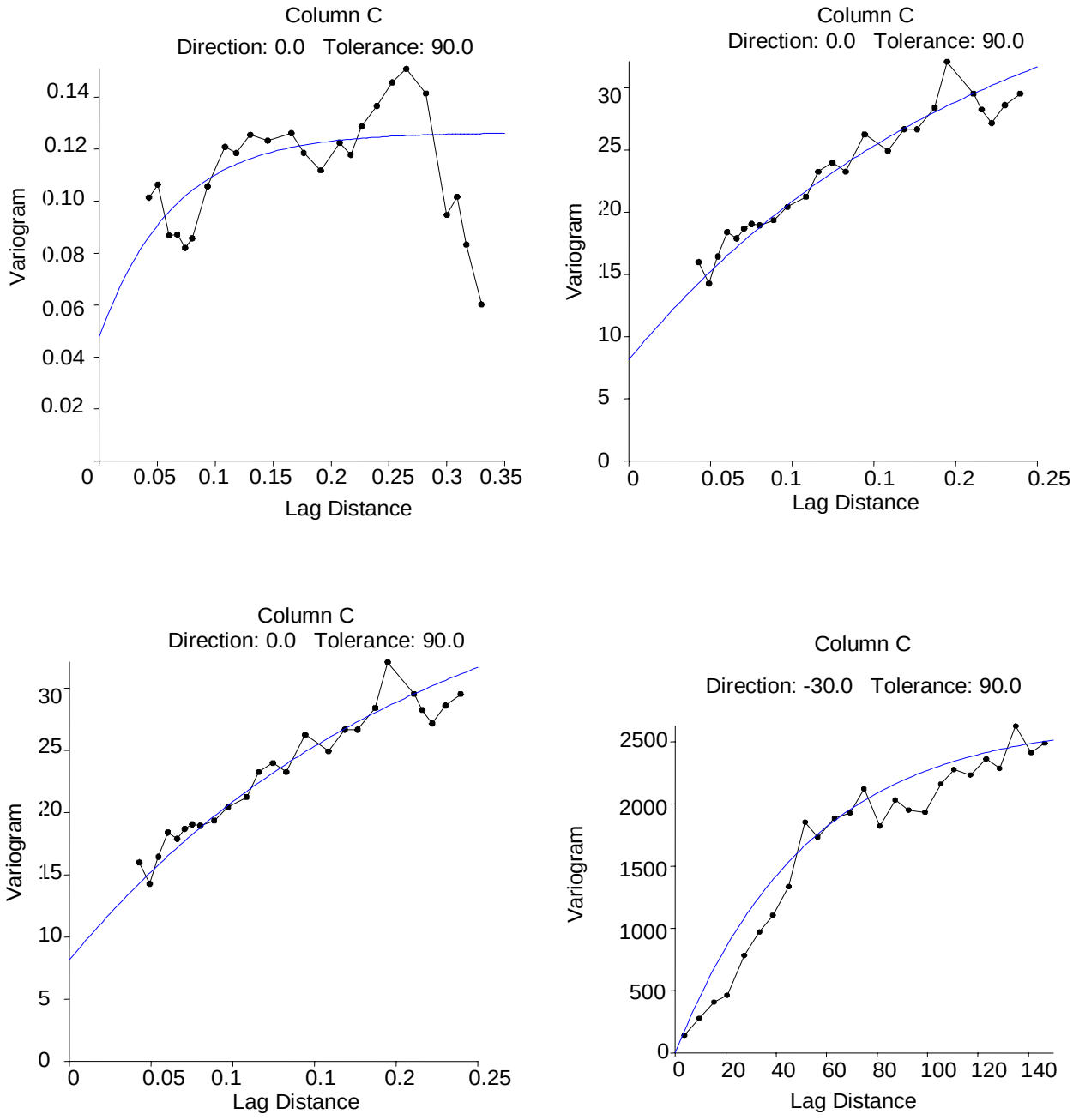

Figure 4. Empirical and fitted semivariogram.

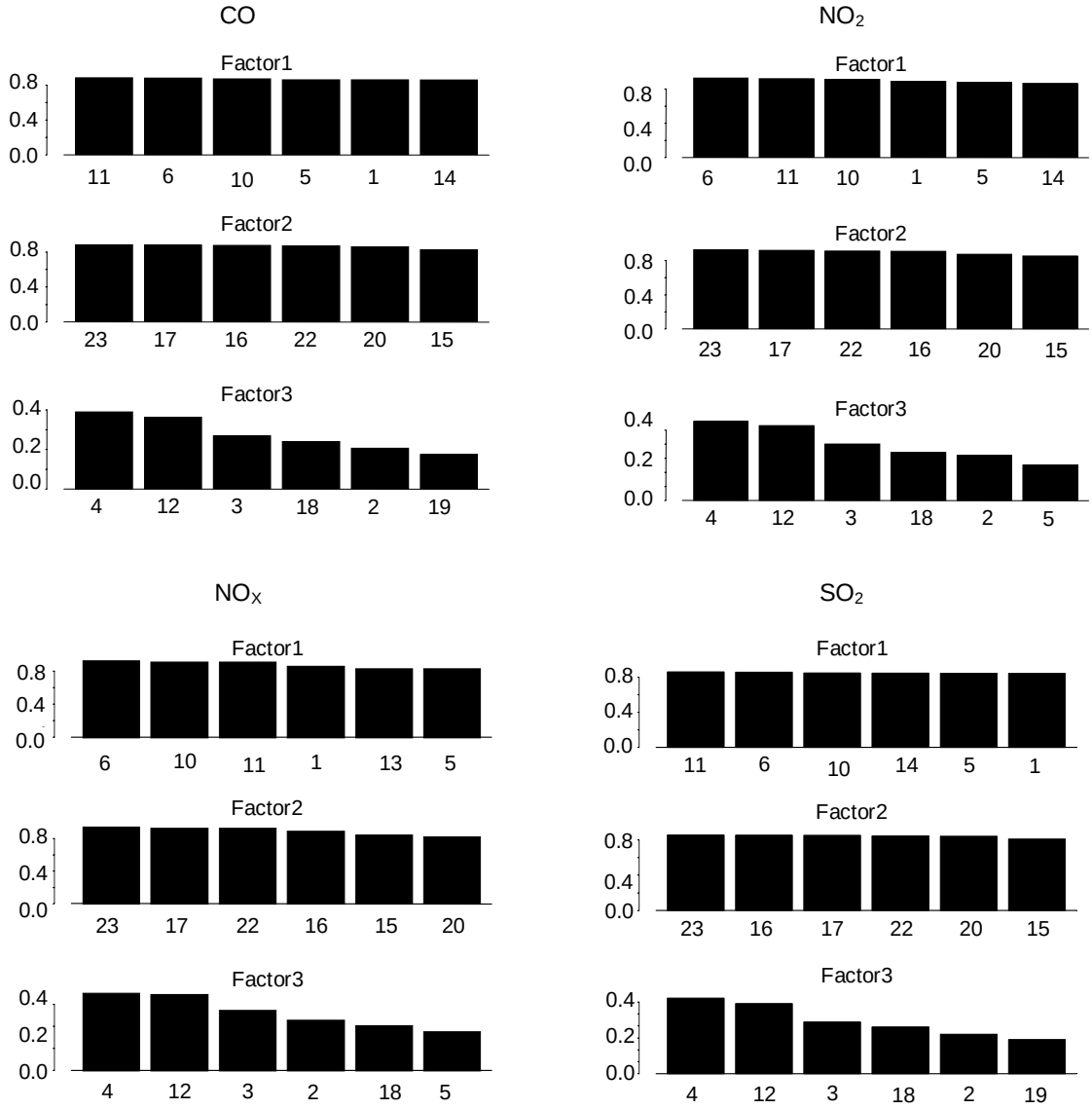

Figure 5. Plots of factors for each variable.

extension for future works is to take into account meteorological variables in the model and to compare the behaviour of the different pollutants considering more than one day.

References

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis*, 2nd Edition, John Wiley & Sons, New York, USA.

Bartholomew, D.J. and Knot, M. (1999). *Latent Variable Models and Factor Analysis*, 2nd Edition, Oxford University Press, New York.

Gleser, L.J. (1997). Some thoughts on chemical mass balance models. *Chemom. Intell. Lab. Syst.*, 37, 15-22.

Guttorp, P. and Sampson P.D. (1994). Methods for estimating heterogeneous spatial covariance functions with environmental applications, in G.P. Patil and C.R. Rao (Eds.), *Handbook of Statistics XII: Environmental Statistics*, Elsevier/North Holland, New York, pp. 663-690.

Henry, R.C. (1987). Current factor analysis models are ill-posed. *Atmos. Environ.,* 21, 1815-1820.

Henry, R.C. (1997). History and fundamentals of multivariate air quality receptor models, *Chemom. Intell. Lab. Syst.*, 37, 37-42.

Henry, R.C., Park, E.S. and Spiegelman, C.H. (1999). Comparing a new algorithm with the classic methods for estimating the number of factors. *Chemom. Intell. Lab. Syst.*, 48, 91-97.

Henry, R.C. (2002). Multivariate receptor models: current practice and future trends. *Chemom. Intell. Lab. Syst.*, 60, 43-48.

Henry, R.C. (2003). Multivariate receptor modelling by N-dimensional edge detection. *Chemom. Intell. Lab. Syst.*, 65, 179-189.

Hopke, P.K. (1991). An introduction to receptor modelling. *Chemom. Intell. Lab. Syst.*, 10, 21-43.

Hopke, P.K. (1997). Receptor modelling for air quality management, in R.E. Hester and R.M. Harrison (Eds.), *Issues in Environmental Science*, Issue 8, Royal Society of Chemistry, Cambridge UK, pp. 95-117.

Hopke, P.K. (2003). Recent developments in receptor modelling. *J. Chemom.*, 17, 255-265.

Javitz, H.S., Watson, J.G., Guertin, J.P. and Mueller, P.K. (1988). Results of a receptor modelling feasibility study. *J. Air Pollut. Control Assoc.*, 38, 661-667.

Kim, E., Hopke, P.K., Paatero, P. and Edgerton, E.S. (2003). Incorporation of parametric factors into multilinear receptor model studies of Atlanta aerosol. *Atmos. Environ.*, 37, 5009-5021.

Loader, P.S. (1992). Spatial covariance estimation for monitoring data, in A. Walden and P. Guttorp (Eds.), *Statistics in Environmental and Earth Sciences*, Edward Arnold, London, pp. 52-70.

Meiring, W., Sampson, P.D. and Guttorp, P. (1998). Space-time estimation of grid-cell hourly ozone levels for assessment of a deterministic model. *Environ. Ecol. Stat.,* 5, 197-222.

Nott, D.J., Dunsmuir, W.T.M., Speer, M.S. and Glowacki, T.J. (1998). *Non-stationary Multivariate Covariance Estimation for Monitoring Data*, Technical Report S98-14.

Paatero, P. and Hopke, P.K. (2002). Utilizing wind direction and wind speed as independent variables in multilinear receptor modelling studies. *Chemom. Intell. Lab. Syst.*, 60, 25-41.

Paatero, P., Hopke, P.K., Hoppenstock, J. and Eberly, S.I. (2003). Advanced factor analysis of spatial distributions of PM2.5 in the eastern United States. *Environ. Sci. Technol.*, 37, 2460-2476.

Park, E.S., Henry, R.C. and Spiegelman, C.H. (1999). *Determining the Number of Major Pollution Sources in Multivariate air Quality Receptor Models*, NRCSE, TSR No.34.

Park, E.S., Henry, R.C. and Spiegelman, C.H. (2000). Estimating the number of factors to include in a high-dimensional multivariate bilinear model. *Commun. Stat.*, 29(B), 723-746.

Park, E.S., Guttorp, P. and Henry, R.C. (2001). Multivariate receptor modelling for temporal correlated data by using MCMC. *J. Am. Stat. Assoc.,* 96, 1171-1183.

Park, E.S., Oh, M.S. and Guttorp, P. (2002). Multivariate receptor models and model uncertainty. *Chemom. Intell. Lab. Syst.*, 60, 49-67.

Sampson, P.D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *J. Am. Stat. Assoc.*, 87, 108-119.

Spiegelman, C.H. and Dattner, S. (1993). Multivariate chemometrics, a case study: applying and developing receptor models for the 1990 El Paso winter PM10 receptor modelling scoping study, in G.P. Patil and C.R. Rao (Eds.), *Multivariate Environmental Statistic*s, Elsevier Science publishers, New York, pp. 509-524.