Revista de Matemática: Teoría y Aplicaciones 2000 **7**(1-2) : 153–164
cimpa – ucr – ccss issn: 1409-2433

SPATIAL INTERPOLATION OF DRY DEPOSITION USING EOF MODELS

Breda Muñoz-Hernández[*]

Received: 24 July 2000

Abstract

Random processes are monitored over space and time by a network of stations distributed across a spatial region. Auxiliary information is often gathered not only at the stations but at other points across the region. The incorporation of auxiliary information in some interpolation techniques has shown improvement on the interpolation results. The Empirical Orthogonal Functions (EOF) model is a well-known eigenvector based prediction technique, widely used in meteorology and oceanography for modeling the variability of the observed spatio-temporal random process. Similarity matrices are constructed using available auxiliary information and included in the EOF model to develop a spatial interpolation method. The resulting interpolation technique will be applied to a real data set and the results compared to ordinary kriging.

Keywords: EOF models, environmental monitoring, interpolation techniques, spatio-temporal data analysis, similarity matrix.

Resumen

Los procesos aleatorios son controlados (monitoreados) sobre el espacio y el tiempo por una red de estaciones distribuidas a lo largo de una región espacial. Información auxiliar es recogida a menudo no solamente en las estaciones sino también en otros puntos de la región. La incorporación de información auxiliar en algunas técnicas de interpolación ha mostrado que los resultados de la interpolación se mejoran. El modelo de Funciones Empíricas Ortogonales (EOF) es una conocida técnica de predicción basada en vectores propios, usada ampliamente en meteorología y oceanografía para modelar la variabilidad de procesos estocásticos espacio-temporales observados. Se construyen matrices de similitudes usando información auxiliar disponible e incluida

[*] CIMPA, Escuela de Matemática, Universidad de Costa Rica, 2060 San José, Costa Rica. E-Mail: bmunoz@cariari.ucr.ac.cr

en el modelo EOF para desarrollar un método de interpolación espacial. La técnica de interpolación resultante será aplicada a un conjunto de datos reales y los resultados serán comparados al 'kriging' ordinario.

Palabras clave: modelos EOF, control o monitoreo del ambiente, técnicas de interpolación, análisis de datos espacio temporales, matriz de similitudes.

Mathematics Subject Classification: 65D05, 92H25

1 Introduction

To model continuous spatio-temporal random processes, defined as $X(t, \mathbf{s})$, researchers developed analysis techniques that intend to capture the variability over space and time. These analysis techniques are used in order to model and understand the process, provide data reduction, prediction models and interpolation methods.

Data used to fit these models are collected using a finite network of sites, $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_n\}$, which are established in a region of interest $\mathcal{R}$ with the purpose of monitoring a process. The distribution of the stations on $\mathcal{R}$ captures information about the spatial variability of the process. If at each station, $\mathbf{s}_i \in \mathcal{S}$, data are collected over time, then information about the temporal variability may be incorporated in the model as well.

Time and budget constraints limit the observation of the process to the selected sites on $\mathcal{S}$ where the monitoring stations are placed. As the monitoring programs develop over time, many programs have incorporated new sites in the network. The increase in spatial coverage provides more information to be used for modeling techniques to describe the spatio-temporal variability of the process. However, at the same time, a problem of missing data results due to the different initiation dates for the stations that belong to the network. A malfunctioning station causes another type of missing data associated with a station network. In order to account for missing data in the model and the desire to obtain estimates at non coverage sites in $\mathcal{R}$, researchers were motivated to develop spatial interpolation techniques. Spatial interpolation refers to the ability of a model to predict the value of the random process $X(t, \mathbf{s}^*)$ at a fixed time t,for a given unsampled site $\mathbf{s}^*$ in $\mathcal{R}$, using the observed sample sites or stations from the network $\mathcal{S}$ (Okabe et. al., 1992; Cressie, 1991; Willmott and Robeson, 1995).

Spatial interpolation techniques are classified as global or local methods. The global method makes use of all stations providing available data in the network $\mathcal{S}$, while the local method uses only selected nearby stations that have available information. The local interpolation methods often identify a unique subregion of $\mathcal{R}$ in the vicinity of the unobserved site, $\mathbf{s}^*$. The information provided by the network stations that delimitate this specific subregion is used to calculate the interpolated value of the process at $\mathbf{s}^*$. Other spatial interpolation methods extend the concept above by incorporating stations that are located in nearby subregions (Okabe et. al., 1992; Watson and Philip, 1987).

Many of the spatial interpolation techniques calculate the interpolated value at the unobserved site $\mathbf{s}^*$ as a weighted sum of the data collected at the network stations,

$$\widehat{X}(\mathbf{s}^*, t) = \sum_{\mathbf{s}} w(\mathbf{s}, \mathbf{s}^*) X(\mathbf{s}, t) \tag{1}$$

where $w(\mathbf{s},\mathbf{s}^*)$ denotes the weight for station $\mathbf{s}$ related to the unobserved site $\mathbf{s}^*$, and $X(\mathbf{s},t)$ denotes the observed value at site $\mathbf{s}$ at fixed time t (Okabe et. al., 1992). Determination of the weights varies among different methods. Most spatial interpolation methods rely on the assumption that nearby sites tend to be more similar (high spatial correlation in terms of the random process of interest) than sites more distant from each other. Therefore, local interpolation estimates are based on observations of the monitored process at the nearby network stations and results depend on the between-station variability. This between-station variability is affected by the density of the network and may cause a reduction in the precision of the estimates.

The availability of auxiliary variables highly correlated with the random process of interest, and observed at a regional level, motivated researchers to attempt their incorporation into known interpolation methods (Willmott and Matsuura, 1995; Daly et. al., 1994; Robeson, 1997; Willmott and Robeson, 1995). Their objective was to reduce the between-station variability and improve the results of the interpolation techniques.

In this paper, a technique is introduced that incorporates auxiliary information in the Empirical Orthogonal Functions (EOF) model. The EOF model is a technique used to model the spatio-temporal random process. This analysis technique does not require any assumption on the distribution or the variance-covariance matrix of the random process (Obled and Creutin, 1986). The EOF model incorporates the spatio-temporal variability of the process captured by the network of stations. It is a popular analysis technique used in Meteorology and Oceanography for modeling as well as for data reduction. The design-based EOF model, proposed by Muñoz-Hernández et. al. (1999), is a modification to the usual EOF model analysis formulated under the assumption that a probabilistic sampling design was used to select the monitoring network sites. The spatial variability of the process is captured by the inclusion and joint inclusion density functions that characterized the probability sampling design involved in the location of the sites. The incorporation of the density functions into the EOF model reduces the instability in the calculations of the eigenvectors caused by the use of the finite network of stations.

We propose to include auxiliary information in the design-based EOF model and illustrate how this prediction model can also be considered for spatial interpolation purposes. Based on the auxiliary information, network sites that exhibit more similar attributes with the unobserved site $\mathbf{s}^*$ will be selected for the interpolation analysis. Under the assumption that the auxiliary variables are highly correlated with the random process, we will expect that the outcome of the random process at the unobserved site $\mathbf{s}^*$ will also be similar. A measure for this resemblance is called the coefficient of similarity. This coefficient takes values between 0 and 1. A coefficient with a higher value reflects a higher similarity between the two sites.

Approaches to calculate similarity coefficients depend whether the variables under consideration are categorical or continuous. (Gower, 1971; Everitt, 1993). Depending on the objective of the study and/or the numerical type of the variables, one method may be more appropriate than another. One alternative similarity measure that captures this imprecision of fuzziness is based in fuzzy set theory. The fuzzy set theory approach derives similarity matrices from fuzzy and/or precise data. This approach takes advantage of the assumption that auxiliary variables observed over the whole region of interest show a smooth variation from site to site, leading to more realistic similarity coefficients (Leung, 1988).

2 The empirical orthogonal functions model

Assume that a continuous, non-negative random process, $X(t, \mathbf{s})$ is observed at n different sites or locations in a region of interest $\mathcal{D}$. Assume that a probability sampling design was adopted in the selection of the sites and that the data is collected at each station at periodic intervals in time.

One of the methods available to analyze spatio-temporal data that does not require any assumption on the covariance matrix or on the variogram is the EOF model (Obled and Creutin, 1986). Under regular conditions, the random field $X(t, \mathbf{s})$ can be expanded as the double convergent series:

$$X(t, \mathbf{s}) = \sum_{k=0}^{\infty} Z_k(t)\varphi_k(\mathbf{s}) \tag{2}$$

where $\varphi_k(\mathbf{s})$ constitutes an orthogonal set of functions that captures the spatial variability of the process and are the solutions to the following second order Fredholm integral equation

$$\int_{\mathcal{D}} \mathcal{C}(\mathbf{s}, \mathbf{s}')\varphi_k(\mathbf{s}')d\mathbf{s}' = \lambda_k \varphi_k(\mathbf{s}) \tag{3}$$

where $\mathcal{C}(\mathbf{s}, \mathbf{s}') = \int_T X(t, \mathbf{s})X(t, \mathbf{s}')dt$ is the covariance function of the process. Equation (2) is known as the Karhunen-Loéve expansion. Observe that $\varphi_k(\mathbf{s})$ and λ_k constitute the eigenvectors and eigenvalues of the integral Equation (3), respectively.

The random variable $Z_k(t)$, $k = 1, \ldots,$ that captures the time variability of the process, is defined as:

$$Z_k(t) = \int_{\mathcal{D}} X(t, \mathbf{s})\varphi_k(\mathbf{s})d\mathbf{s}Z_k(t) = \int_{\mathcal{D}} X(t, \mathbf{s})\varphi_k(\mathbf{s})d\mathbf{s} \tag{4}$$

and satisfies

$$\int_T Z_j(t)Z_k(t)dt = \lambda_k \delta_{jk} \tag{5}$$

where $\delta_{jk} = 1$ if $j = k$ and 0 otherwise.

Researchers have proposed different methods to numerically solve Equations (3) and (4). Some numerical solutions proposed are based on some type of quadrature method (Buell, 1978) or the consideration of spline basis functions (Obled and Creutin, 1986; Wikle, 1995). Another feature to address in spatio- temporal analysis methods is the effect of the spatial distribution of the network stations in the estimation process. If a probabilistic sample of sites was selected, then the spatial distribution of the process is captured by the sampling design and incorporated in the analysis by the inclusion probabilities functions. This idea is presented in the design-based EOF models we proposed in an earlier paper (Muñoz-Hernández et.al., 1999).

3 The design-based EOF model

Let $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_n\}$ be a set of sites selected using a probability sampling design. Assume that the random process $X(t, \mathbf{s})$ is observed at these stations for a period of time T. Let $\pi(\mathbf{s})$ and $\pi(\mathbf{s},\mathbf{s}')$ denote the inclusion density function at site $\mathbf{s}$ and the joint inclusion density function for sites $\mathbf{s}$ and $\mathbf{s}'$, respectively. The design-based EOF model is determined by first solving the following eigenvector problem, which is the result of the application of a method of moments approach to solve Equation (5),

$$\mathbb{C}\boldsymbol{\Phi} = \boldsymbol{\Lambda}\boldsymbol{\Phi} \tag{6}$$

where the (i, j) element of the matrix $\mathbb{C}$ is defined as $\frac{\widehat{\mathcal{C}}(\mathbf{s}_i,\mathbf{s}_j)}{\pi(\mathbf{s}_i,\mathbf{s}_j)}, \widehat{\mathcal{C}}(\mathbf{s}_i, \mathbf{s}_j)$ is the (i, j)-entry of the sample covariance matrix, $\pi(\mathbf{s}_i, \mathbf{s}_j)$ is the joint inclusion density function for sites $\mathbf{s}_i$ and $\mathbf{s}_j$, $\boldsymbol{\Phi}$ is a matrix with its ith column equal to the ith eigenvector of $\mathbb{C}$, and $\boldsymbol{\Lambda}$ is a diagonal matrix where the ith diagonal entry is the eigenvalue corresponding to the ith eigenvector.

The solution of the eigenvector problem provides a total of n (the number of network stations) eigenvectors. After selecting K $(K < n)$ eigenvectors that retained more than 90% of the total variability of the process, $\widehat{\varphi}_k$, $k = 1, \ldots, K$, the unbiased estimator of the random variables $Z_k(t)$ is computed as follows:

$$\widehat{Z}_k(t) = \sum_{\mathbf{s}\in\mathcal{S}} \frac{X(t, \mathbf{s})}{\pi(\mathbf{s})} \widehat{\varphi}_k(\mathbf{s}). \tag{7}$$

The design-based EOF model is then obtained as:

$$\widehat{X}(\mathbf{s}, t) = \sum_{k=1}^{K} \widehat{Z}_k(t)\widehat{\varphi}_k(\mathbf{s}). \tag{8}$$

4 Similarity matrices. Standard classification and fuzzy set approaches.

Assume that $\mathcal{S} = \{\mathbf{s}_1, ..., \mathbf{s}_n\}$ is a collection of sites for which values of p auxiliary variables, $Y_1, \ldots, Y_p$, were recorded. We can calculate a degree of resemblance or similarity coefficient, denoted as r_{ijk}, between two sites $\mathbf{s}_i$ and $\mathbf{s}_j$ with respect to the kth auxiliary variable. This coefficient ranges between 0 and 1. Values close to 1 denote high similarity or resemblance between the two sites, $\mathbf{s}_i$ and $\mathbf{s}_j$, with respect to the kth auxiliary variable. Values close to 0 denote low similarity between the two sites $\mathbf{s}_i$ and $\mathbf{s}_j$ with respect to the kth auxiliary variable.

Calculation of the similarity coefficient varies depending on the nature of the auxiliary variable considered. A review of similarity measures can be found in Legendre and Legendre (1983). If the auxiliary variable is dichotomous (say presence and absence), Gower (1971) defined a similarity measure, r_{ijk}, as

$$r_{ijk} = \begin{cases} 1 & \text{if the characteristic is present in both sites} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

if the kth auxiliary variable is a multinomial, define

$$r_{ijk} = \begin{cases} 1 & \text{if both sites have the same value} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

and if the kth auxiliary variable is quantitative, the calculations of the similarity coefficient is done by measuring the distance or dissimilarity, d_{ijk}, between two sites. Distance is defined as a measure of the difference in the auxiliary variables between two sites. These quantities are converted into similarity coefficients, denoted $r_{ijk} = 1 - d_{ijk}$. For quantitative variables, the dissimilarity between sites $\mathbf{s}_i$ and $\mathbf{s}_j$ with respect to the kth variable is calculated as (Gower, 1971):

$$d_{ijk} = \frac{|x_{ki} - x_{kj}|}{\text{range}(x_k)} \tag{11}$$

were range(x_k) is the range of the kth auxiliary variable considering all network sites.

When more than one auxiliary variable is available, it is of interest to summarize for each pair of sites all the information provided by the different similarity coefficients calculated for each auxiliary variable. Sometimes the calculations of the similarity measure between two sites with respect to one auxiliary variable, r_{ijk}, is not possible as the result of missing information. To address this problem, define an indicator variable δ_{ijk}, as 1, if the comparison was possible between sites $\mathbf{s}_i$ and $\mathbf{s}_j$ with respect to the kth auxiliary variable, and 0 otherwise. Next, define the similarity or resemblance coefficient of $\mathbf{s}_i$ and $\mathbf{s}_j$ with respect to the p auxiliary variables as the sum of similarities divided by the number of auxiliary variables for which there is information available in both sites (Gower, 1971; Legendre and Legendre, 1983)

$$r_{ij} = \frac{\sum_{k=1}^{p} \delta_{ijk} r_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}} \qquad \text{for } i, j = 1, \ldots, n. \tag{12}$$

Another method to calculate similarity coefficients is based on fuzzy set theory which was introduced by Zadeh in 1965. This theory addresses the uncertainty that arises when objects are classified in groups according to observed outcomes of random variables. In fuzzy set theory, sets have no rigid boundaries, and therefore, elements can be classified as elements "in some degree" of many sets (Dubois and Prade, 1980).

The calculation of a fuzzy similarity measure requires the researcher to define a priori fuzzy sets based on the p auxiliary variables. Observe that one auxiliary variable, for example elevation, may generate two or more fuzzy sets. For example, low elevation, medium elevation and high elevation are three fuzzy sets that can be defined for elevation. The number of fuzzy sets depends on the nature of the auxiliary variable.

Each fuzzy set Y is characterized by a continuous function called a membership function, μ_Y, that takes values on the interval [0,1]. A membership function value of 0 denotes the element is not a member of the fuzzy set Y, values between 0 and 1 represent some degree of membership, while a value of 1 implies the element is a member of Y in the same sense of classic set theory.

Assume that m fuzzy sets, $Y_1, \ldots, Y_m$, are generated from the p auxiliary variables. After the m fuzzy sets and the membership functions are defined, the similarity coefficient of sites $\mathbf{s}_i$ and $\mathbf{s}_j$ with respect to the m fuzzy sets is calculated as:

$$r_{ij} = 1 - \frac{1}{m} \sum_{k=1}^{m} d_{ijk} \tag{13}$$

where $d_{ijk} = |\mu_{Y_k}(\mathbf{s}_i) - \mu_{Y_k}(\mathbf{s}_j)|$, and $Y_1, \ldots, Y_m$ are the m fuzzy sets defined using the p auxiliary variables.

5 Interpolation

Assume that $\mathbf{s}^*$ is an unsampled site located in the region of interest $\mathcal{R}$ and $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ are the network sites. After constructing a similarity matrix for the sites $\{\mathbf{s}^*, \mathbf{s}_1, \ldots, \mathbf{s}_n\}$ by using either of the two methods described above, the interpolation weights are calculated for each site $\mathbf{s}_i$ in the sample and the unobserved site $\mathbf{s}^*$ as the following ratio

$$w(\mathbf{s}_i, \mathbf{s}^*) = \frac{r_{i,*}}{\sum_{i=1}^{n} r_{i,*}} \tag{14}$$

where $r_{i,*}$ denotes the similarity between the network site $\mathbf{s}_i$ and the unobserved site $\mathbf{s}^*$. Note that the interpolation weights take values between 0 and 1 and gives more contribution in the interpolation process to those sites in the sample, more similar to $\mathbf{s}^*$. Define the interpolated process at the site $\mathbf{s}^*$ and time t, as:

$$\widetilde{X}(\mathbf{s}^*, t) = \sum_{i=1}^{n} \widehat{X}\,(\mathbf{s}_i, t)\, w(\mathbf{s}_i, \mathbf{s}^*). \tag{15}$$

where $\widehat{X}\,(\mathbf{s}_i, t)$ is the predicted design-based EOF model at site $\mathbf{s}_i$ and time t defined in Equation (8) and $w(\mathbf{s}_i, \mathbf{s}^*)$ is the interpolation weight for the network site $\mathbf{s}_i$ and the unobserved site $\mathbf{s}^*$.

5.1 Description of the data

Accurate long term measurements of dry deposition are difficult and expensive to make since they require frequent on-site supervision, special instruments and quality technical resources (Clarke et. al., 1997). For these reasons, inferential models are preferred and commonly used alternatives to direct measurement. Using data collected from the Clean Air Status and Trends Network (CASTnet), dry deposition data have been predicted by the United States (US) Environmental Protection Agency (EPA) using the Big Leaf model in the past, and currently the Multilayer model. CASTnet, originally called the National Dry Deposition Network (NDDN), is a monitoring program established in 1986 by EPA to characterize dry deposition patterns and trends across the United States.

Data used in this illustration consist of weekly atmospheric sulfur dioxide (SO_2) dry deposition calculated using the Multilayer model for CASTnet from 1987 to 1998. Most of the sites are located in eastern US and only 9 are located in the western US (Clarke et. al., 1997). Five auxiliary variables were available: latitude, longitude, elevation (meters), land use described as agricultural, range, forested and suburban and type of terrain classified as complex (local ground slopes greater than 15 °), rolling (local ground slopes between 5 ° and 15 °), flat and mountain-top (Clarke et.al., 1997). The data were made available by Dr. Henry Lee of the Environmental Protection Agency (EPA), Corvallis, Oregon.

5.2 Analysis of the data

CASTnet site locations were not selected using a probabilistic sampling design. Therefore in order to apply the design-based EOF model, the sites were stratified and a simple random sample was taken from the fixed number of sites available at each strata. This approach was used by Muñoz-Hernández et. al. (1999) to illustrate the applicability of the design-based EOF model to monitoring programs that do not use probability sampling designs to select the sites. Five strata were constructed using a hierarchical cluster analysis based on the five variables available. The sizes of the five strata were 11, 9, 4, 11 and 14.

Analysis will be performed using the whole network of sites and separately using only the sites located in eastern US. A random sample of size about 61% of the total number of sites considered was drawn using proportional allocation with the objective of obtaining a representative sample of CASTnet sites. To eliminate any possible spatial trend the time mean was removed from each network site. These zero mean data were then normalized as follows: $X^*(t,\mathbf{s}) = \frac{X(t,\mathbf{s}) - \overline{X}(\mathbf{s})}{\sigma_{\mathbf{s}}}$, where the site mean, $\overline{X}(\mathbf{s})$, and the site standard deviation, $\sigma_{\mathbf{s}}$, are calculated as: $\overline{X}(\mathbf{s}) = \frac{\sum_{t \in T} X(t,\mathbf{s})}{N(\mathbf{s})}$ and $\sigma_{\mathbf{s}} = \sqrt{\frac{\sum_{t \in T}(X(t,\mathbf{s}) - \bar{X}(\mathbf{s}))^2}{N(\mathbf{s})}}$. $N(\mathbf{s})$ is the total number of time points, for which the observation, $X(t,\mathbf{s})$, is not missing at the site $\mathbf{s}$. The sample variance-covariance matrix, $\widehat{\mathcal{C}}(\mathbf{s},\mathbf{s}')$ is calculated as: $\widehat{\mathcal{C}}(\mathbf{s},\mathbf{s}') = \frac{\sum_{t=1}^{N(\mathbf{s},\mathbf{s}')} X^*(t,\mathbf{s}) X^*(t,\mathbf{s}')}{N(\mathbf{s},\mathbf{s}')}$, where $N(\mathbf{s},\mathbf{s}')$ is the total number of observations for which the data points, $X^*(t,\mathbf{s})$ and $X^*(t,\mathbf{s}')$, are not missing at sites $\mathbf{s}$ and $\mathbf{s}'$, respectively. The matrix $\mathbb{C}$ is calculated as specified before and the eigenvector problem in Equation (6) is solved. After determining K $(K \leq n)$ non-zero significant eigenvectors, expressions for $Z_k(t)$ and the design-based EOF model predictors are calculated using Equations (7) and (8), respectively.

To implement the interpolation model using auxiliary variables, we calculated the similarity matrices using Gower and the fuzzy set approaches described in Equations (12) and (13), respectively. Finally, similarity coefficients were calculated following Equation (12).

Calculation of a fuzzy similarity matrix requires the definition of fuzzy sets and their corresponding membership functions. From exploratory analysis, three fuzzy sets were defined for longitude, latitude and elevation. The variable type of terrain was not considered in the calculations of the fuzzy similarity matrix based on insufficient information regarding each of the code definitions. We considered the following continuous membership function for each fuzzy set Y defined for any of the variables latitude, longitude and elevation using Bobrowicz et.al. (1990):

$$\begin{array}{rl} \mu_Y : & \mathcal{R} \rightarrow [0,1] \\ & \mathbf{s} \mapsto \frac{1}{1+a^b(\mathbf{s}-\mathbf{c})^b} \end{array} \tag{16}$$

where a, b and c characterize μ_Y by satisfying the following:

$$\begin{array}{c} \mu_Y(\mathbf{c}) = 1 \\ \mu_Y(\mathbf{c} - \frac{1}{a}) = \mu_Y(\mathbf{c} + \frac{1}{a}) = 0.5 \end{array} \tag{17}$$

and b determines the shape of the function. This membership function was selected based on the assumption that auxiliary variables range smoothly over the whole region $\mathcal{R}$. Three fuzzy sets were defined from the information on the auxiliary variable land use. The membership function for each fuzzy set was defined as a step function that takes values 0.25, 0.5 and 1, depending on the relation among the codes available.

Weights were calculated for both the Gower and fuzzy set approaches using Equation (14). A sample of sites of about 61% of the two sets of sites considered was selected at random to fit the design-based EOF model. Interpolation values for the non-selected sites were obtained at selected time points.

Interpolation results from the design-based EOF model are compared to those obtained using ordinary kriging. All models were fit using the same samples drawn from the whole network (49 sites) and from a subset of sites consisting of those sites located in eastern US (40 sites). Twenty-five time points with no missing data were selected among the 381 available for interpolation purposes. Selection of time points without missing data is a restriction imposed by the spatial statistics module of Splus.

Kriging is an optimal linear spatial interpolation method very popular in geostatistics. Kriging is optimal among the linear predictors in the sense that it is unbiased and has minimum variance for the prediction error. Exploratory data analysis was performed separately for each of the 25 time points. It was concluded, based on the small sample, that a spherical variogram captured the broad structure of the observed spatial dependence. The shape of the variograms appeared to be similar for all considered time points.

For a quantitative measure of the ability of a model to predict in space and time we considered the following statistic (Wikle and Cressie, 1997).

$$CR_3(\mathbf{s}) = \sqrt{\frac{1}{N(\mathbf{s})} \sum_{i=1}^{N(\mathbf{s})} \{X(t_i, \mathbf{s}) - \widetilde{X}(t_i, \mathbf{s})\}^2} \tag{18}$$

where $\widetilde{X}(t_i, \mathbf{s})$ is the model predicted value of the process at time t_i and site $\mathbf{s}$ and $N(\mathbf{s})$ is the number of time points for which $X(t_i, \mathbf{s})$ is not missing. The smaller the values of the spatial averages of $CR_3(\mathbf{s})$, the better the predicted values. $CR_3(\mathbf{s})$ is basically an indicator of the average precision of the model at each site.

Table 1 shows the validation statistic, $\overline{CR}_3$, for the design-based EOF model obtained with the three similarity matrices, Gower (EOF G5 and EOF G4) and fuzzy set (EOF F4), and the ordinary kriging results. EOF G4 denotes the design-based EOF model that incorporates only four auxiliary variables: latitude, longitude, elevation and land use by means of a similarity matrix calculated using Gower's approach. The fuzzy model interpolation results are denoted by EOF F4 and consider the same four auxiliary variables. By EOF G5 we denote the interpolation results based in Gower's approach that incorporates all five auxiliary variables available. Reduced information regarding the variable terrain type made impossible the definition of a proper fuzzy similarity measure, and the comparison between an EOF F5 and a EOF G5 model results. Observe that design-based EOF (EOF G4, EOF G5 and EOF F4) model approaches performed better than ordinary kriging (OK). A small difference is observed among the design-based prediction results for the different choices of similarity matrix. Observe that there was no significant difference in the performance of the model using the Gower method for the two sets of auxiliary variables. This suggests that type of terrain has little effect explaining the spatial variability of the random process of interest.

	EOF G5[1]	EOF G4[2]	EOF F4[2]	OK[3]
Whole data	0.3043	0.3042	0.405	1.272
Eastern data	0.1594	0.1596	0.208	0.942

Table 1: $\overline{CR}_3$ for the Design-based EOF (EOF G4, EOF G5 and EOF F4) Models and Ordinary Kriging (OK) Analysis on CASTnet Dry Deposition Data.

A better performance for all four models is observed when considering only the 40 sites in the eastern US. Since the variability among the auxiliary variables is smaller for this region of the US as compared to the entire data, better results in all these models are expected if additional information from auxiliary variables considered known to have influence in dry deposition of SO_2 is available (Clarke et. al., 1997; Holland et.al., 1999).

6 Summary

The incorporation of covariates into the design-based EOF model led to good space-time interpolation results. Small differences were observed in the results obtained by the two approaches used to develop the similarity matrices. The choice of the similarity matrix approach to use in the EOF model incorporating covariates would be based on the nature of the covariates available. If some of the covariates are of a non-precise nature, then the fuzzy set approach should be considered. The fuzzy set approach requires the specification of fuzzy sets and membership functions, which depend on the type of available auxiliary variables. Unfortunately, there is no specific rules for the determination of fuzzy sets and membership function.

The design-based EOF model can be considered as an alternative for spatial prediction not only for monitoring programs in environmental science but also for any lattice data where spatial data analysis is performed.

References

Bobrowiz, O.; Choulet, C.; Haurat, A.; Sandoz, F.; Tebaa, M. (1990) *A method to build membership functions. Application to numerical/symbolic interface building*, Lecture Notes in Computer Science, 521. Springer-Verlag, New York: 136–142.

Buell, C. E. (1971) "Integral equation representation for Factor Analysis", *Journal of Atmospheric Sciences* **28**: 1502–1505.

Clarke, J. F.; Edgerton, E. S.; Martin, B. E. (1997) "Dry deposition calculations for the clean air status and trends network", *Atmospheric Environment* **21**: 3667–3678.

Cressie, N.A.C. (1991) *Statistics for Spatial Data.* John Wiley & Sons, New York.

Daly, C.; Neilson, R.P.; Phillips, D.L. (1994) "A statistical-topographic model for mapping climatological precipitation over mountainous terrain", *Journal of Applied Meteorology* **33**: 140–158.

Dubois, D.; Prade, H. (1980) *Fuzzy Sets and Systems, Theory and Applications.* Mathematics in Science and Engineering, 144, Academic Press, New York.

Duckstein, L.; Blinowska, A.; Verroust, G. (1995) "Fuzzy classification of patient state with application to electrodiagnosis of peripheral polyneuropathy", *IEEE Transactions on Biomedical Engineering* **42**(8): 786–792.

Erickson, R.; Di Lorenzo, P.M.; Woodbury, M.A. (1994) "Classification of the taste responses in brain stem: membership in fuzzy sets", *The American Physiological Society*: 2139–2150

Everitt, B.S. (1993) *Cluster Analysis*, 3rd Edition. Edward Arnold, London.

Fang, J.H. (1997) "Fuzzy logic and geology mathematical modeling in geological systems", *Geotimes* **42**(10): 23–26.

Gower, J.C. (1971) "A general coefficient of similarity and some of its properties", *Biometrics* **27**: 857–874.

Guerra, T.M.; Loslever, P. (1993) "Two ways for getting connections between objective and subjective data sets in man-machine systems: multiple correspondence factor analysis and probabilistic fuzzy set theory", *Cybernetics and Systems: An International Journal* **24**: 217–242.

Hendricks, F.; van Eijnsbergen, A.C.; Stein, A. (1997) "Use of spatial prediction techniques and fuzzy classification mapping soil pollutants", *Geoderma* **77**: 243–262.

Hersh, H.M.; Caramazza, A. (1976) A fuzzy set approach to modifiers and vagueness in natural language", *Journal of Experimental Psychology: General* **105**(3): 254–276.

Hewitt, C.N. (1992) *Methods of Environmental Data Analysis.* Elsevier, Dordrecht.

Holland, D.M.; Principe, P.P.; Vorgurger, L. (1999) "Rural ozone: trends and exceedances at CASTnet sites", *Environmental Science and Technology* **33**: 43–48.

Juang, C.H.; Huang, X. H.; Holtz, R.D.; Chen, J. W. (1996) "Determining relative density of sands from CPT using fuzzy sets", *Journal of Geotechnical Engineering*, January: 1-6.

Legendre L.; Legendre P. (1983) *Numerical Ecology, Developments in Environmental Modelling 3.* Elsevier, Dordrecht.

Legge, H. A.; Krupa, S.V. (1990) *Acidic Deposition: Sulphur and Nitrogen Oxides.* Lewis Publishers

Leung, Y. (1988) "Spatial analysis and planning under imprecision", *Studies in Regional Science and Urban Economics* **14**: 66–139.

Muñoz-Hernández, Breda; Lesser, V.M.; Ramsey, F.L. (1999) "Design based empirical orthogonal functions", *Proceedings of the Section of Statistics and the Environment*, American Statistical Association.

Obled, Ch.; Creutin, J.D. (1986) "Some development in the use of empirical orthogonal functions for mapping meteorological fields", *Journal of Climate and Applied Meteorology* **23**(9): 1189–1204.

Okabe, A.; Boots, B.; Sugihara, K. (1992) *Spatial Tessellations. Concepts and Applications of Voronoi Diagrams.* John Wiley & Sons., New York.

Pleim, J.E.; Finkelstein, P.L.; Clarke, J.F.; Ellestad, T.G. (s.f.) "A technique for estimating dry deposition velocities based on similarities with latent heat flux", *Atmospheric Environment* **33**: 2257–2268.

Preisendorfer, R.W. (1988) "Principal component analysis in meteorology and oceanography", *Developments in Atmospheric Science* **17**: 192–199.

Robeson, S.M. (1997) "Spherical methods for spatial interpolation", *Review and Evaluation, Cartography and Geographic Information Systems* **24**(1): 3–20.

Watson, D.F.; Philip, G.M. (1987) "Neighborhood-Based Interpolation", *Geobyte* **2**(2): 12–16.

Wikle, C.K. (1996) *Spatio-Temporal Statistical Models with Applications to Atmospheric Processes.* Ph.D. dissertation, Iowa State University, Ames, Iowa.

Wikle, C.K.; Cressie, N. (1997) "A dimension-reduction approach to space-time Kalman filtering", Preprint No.97-2, Statistical Laboratory, ISU, Iowa State University, Iowa .

Willmott, C. J.; Robeson, S. (1995) "Climatological aided interpolation (CAI) of terrestrial air temperatures", *International Journal of Climatology* **15** 221–229.

Willmott, C. J.; Matsuura, K. (1995) "Sart Interpolation of annually averaged air temperature in the United States", *Journal of Applied Meteorology* **34**: 2577–2586.

Wirsam, B.; Uthus, E. O. (1996) "The use of fuzzy logic in nutrition", *American Institute of Nutrition*: 2337S–2341S.