



## Soil salinity mapping using spatio-temporal kriging and Bayesian maximum entropy with interval soft data

Ahmed Douaik<sup>a,b,\*</sup>, Marc Van Meirvenne<sup>a</sup>, Tibor Tóth<sup>c</sup>

<sup>a</sup>Department of Soil Management and Soil Care, Ghent University, 653 Coupure Links, Ghent, Belgium

<sup>b</sup>Research Unit on Environment and Conservation of Natural Resources (UR-ECRN), National Institute of Agricultural Research (INRA), CRRRA Rabat, BP 415, Rabat, Morocco

<sup>c</sup>Research Institute for Soil Science and Agricultural Chemistry, Hungarian Academy of Sciences, 15 Herman O. Street, PO Box 35, 1525 II Budapest, Hungary

Available online 4 May 2005

### Abstract

We used geostatistical tools to identify the spatio-temporal variability of soil salinity with both field and laboratory measurements. This analysis used kriging and Bayesian maximum entropy (BME) to predict soil salinity at unobserved spatial locations and time instants. We compared the accuracy of the mapped predictions from BME using soft interval data, kriging with either hard and soft data (HSK), or hard data only (HK). A large spatio-temporal database on soil salinity data was available. It consists of 413 sites where the apparent or bulk soil electrical conductivity ( $EC_a$ ) was measured with electrical probes over an area of 25 ha. These measurements are our 'data set to be calibrated'. On a limited subset of these sampling sites (13–20), electrical conductivity was determined by laboratory analysis from 1:2.5 soil–water suspensions ( $EC_{2.5}$ ), which is a simple representation of the electrical conductivity of the water-saturated soil-paste extract ( $EC_e$ ). These are our 'calibration data set'. The whole procedure was repeated 19 times between November 1994 and June 2001.

The methods of prediction were compared quantitatively by mean error (ME) and mean squared error (MSE). The errors are the differences between the measured electrical conductivity in the laboratory on samples from March and June 2001 (which were not used in previous computations) and their cross-validation estimates. The MSE was divided further into three components revealing different aspects of the discrepancy between the observed and the estimated values of electrical conductivity. The BME predictions were less biased and more accurate than those from kriging. The MSE decomposition showed that the kriging with soft data (HSK) provided more biased estimates and failed to reproduce the magnitude of fluctuation in the observed soil salinity. The difference in incorporating soft data into the analysis was confirmed and was more acute when only the largest intervals from the soft data were used. In this situation BME produced very reliable estimates whereas HSK failed to predict soil salinity accurately. © 2005 Elsevier B.V. All rights reserved.

**Keywords:** Bayesian maximum entropy; Kriging; Soft data; Soil salinity; Space–time

\* Corresponding author. Department of Soil Management and Soil Care, Ghent University, 653 Coupure Links, Ghent, Belgium. Tel.: +32 92646042; fax: +32 92646247.

E-mail address: [ahmdk@hotmail.com](mailto:ahmdk@hotmail.com) (A. Douaik).

## 1. Introduction

Soil salinity limits food production in many countries of the world. There are mainly two kinds of soil salinity: naturally occurring dryland salinity and human-induced salinity caused by the low quality of water. In both cases the development of plants and soil organisms are limited leading to low yields. In Hungary, where more than 10% of the land is affected by salt, groundwater is the major cause of salinization.

Saline and sodic soils have particular physical and chemical properties that require specific management. As a first step for the better management of salt-affected soils, soil salinity needs to be monitored in space as well as in time to determine where it is, where it is spreading to, and the rate at which it is spreading. Therefore, we need to sample the soil for laboratory analysis to determine the electrical conductivity of the saturated soil paste extract ( $EC_e$ ). The latter is a measure of soil salinity. This conventional procedure (Soil and Plant Analysis Council, 1992) is expensive, time-consuming, and provides an incomplete view of the extent of soil salinity.

An alternative to laboratory analysis is to assess soil salinity in the field by determining the apparent electrical conductivity ( $EC_a$ ). This can be done using sensors such as the four-electrode probes (Rhoades and van Schilfgaarde, 1976) or by electromagnetic induction instruments (McNeil, 1980). This procedure is cheaper and less time-consuming than the conventional one and the sensors can be mounted on a small vehicle enabling a more intensive survey of the study area.

Lesch et al. (1998) used a classical statistical method to monitor the temporal change of soil salinity between two time periods. The approach can be applied easily to a few measurement times. However it is of limited practical use for many time periods as the procedure must be repeated for each pair of times. In addition, the technique takes no account of any possible temporal correlation between two or more successive measurements.

Douaik et al. (2004) proposed an alternative approach. They rescaled the  $EC_a$  measurements into  $EC_{2.5}$  (the electrical conductivity determined by laboratory analysis from 1:2.5 soil–water suspensions, which is a simple representation of the electrical conductivity of the water-saturated soil-paste extract,

$EC_e$ ) using calibration equations based on regression models. This was followed by spatio-temporal kriging to predict soil salinity at unknown places and times. The approach takes into account the spatial and the temporal correlations between the soil salinity measurements. However, the resulting  $EC_{2.5}$  values from the calibration equations are estimates of the actual soil salinity. This means that they have some degree of uncertainty, which needs to be considered in the analysis.

The method of Bayesian maximum entropy (BME) (Christakos, 1990, 2000) enabled a rigorous analysis of our data by distinguishing formally between the accuracy of the laboratory and the field electrical conductivity measurements. The former are direct and accurate measurements of the soil salinity; they are considered as hard data. The latter are indirect measurements that represent uncertain estimates of soil salinity. They provide less accurate values of soil salinity and can be considered as soft data.

Bayesian maximum entropy provides a general framework for space–time interpolation. It can incorporate different physical knowledge bases such as statistical moments (not limited to the second-order), multipoint statistics, physical laws, hard and soft data, etc. Kriging, the classical geostatistical method of interpolation, is a special case of BME. When physical knowledge is restricted to the second-order statistical moments (mean and covariance or variogram functions) and to the hard data, kriged and BME predictions are equivalent (Christakos and Li, 1998; Lee and Ellis, 1997).

This study has two main objectives: (i) to apply the BME method to soil data using interval soft data and (ii) to compare the prediction performance of BME with two types of kriging: ordinary kriging with hard data only (HK) and ordinary kriging with hard data and the mid-interval of the soft data (HSK).

## 2. Materials

### 2.1. Study site description

Salt-affected soils cover more than 1 million hectares in Hungary and more than 95% of these soils are in the Great Hungarian Plain (GHP) (Szabolcs,

1989). The Hortobagy National Park (HNP) forms a subregion of this plain. The characteristic soil-forming loess material of the GHP was deposited during the Quaternary period in the glacial eras. Water also played a decisive role in the formation of the parent material (Tóth et al., 1991). The elevation of the HNP ranges from 88 to 92 m. However, Tóth and Rajkai (1994) noted that even a small difference in elevation (of the order of  $10^{-1}$  m) results in large difference in salt accumulation. The drought index, the ratio of potential evaporation to precipitation, exceeds unity for eight months of the year (from March to October). This has an important impact on the salinization of the area.

Hortobagy is a discharge area of saline groundwater originating from the northern mountains. This groundwater is the main source of salt accumulation in the area. Waterlogging induces a rise in the groundwater level during the wet season. This results in a flow towards salic solonetz elevated zones (Tóth and Jozefaciuk, 2002). During the dry season, however, the groundwater level is lowered and salts remain in the upper horizons. Water table depth has played a major role in the formation of salt-affected soils in the GHP.

## 2.2. Data description

The study area covers about 25 ha and is located in the HNP ( $47^{\circ}30''$  N,  $21^{\circ}30''$  E), in the east of Hungary. The locations of the sampling sites are shown in Fig. 1.

Initially the sampling design was chosen so that any temporal change in soil salinity between two sampling campaigns could be determined. The first data set, the 'data set to be calibrated', comprises measurements of apparent electrical conductivity ( $EC_a$ ) in the field at 413 locations. These sites are more or less on a grid of  $25 \times 25$  m, with some locations further apart, mainly in the middle of the study area. The  $EC_a$  was measured using a conductivity meter with four electrodes (Rhoades and van Schilfgaarde, 1976). The electrodes were inserted in the soil at two depths: 8 and 13 cm which correspond to  $EC_a$  values for the 0–20 cm and 0–40 cm soil depths. At the calibration sites there were always three measurements of  $EC_a$ , but only one for the other sites.

An algorithm was used to select a minimal number of calibration sites, based on the response surface design approach (Lesch et al., 1995). The selection was based on the spatial configuration of locations at

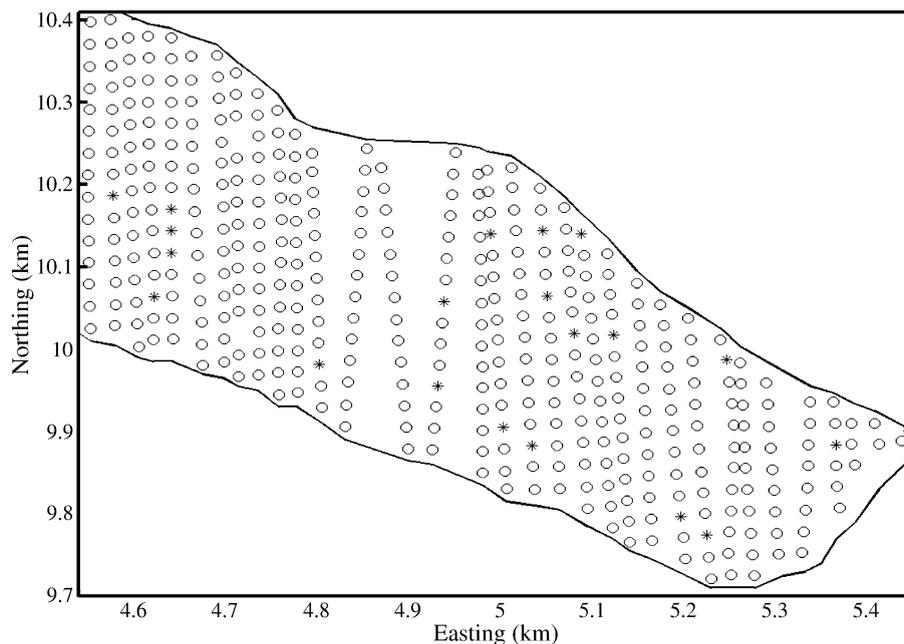


Fig. 1. Sample locations. Circles are the locations of the soft data and asterisks are the locations of the hard data.

which  $EC_a$  was measured and also on the values of the measurements. They are representative spatially of the study area and enable accurate estimation of the calibration parameters. The soil samples at the calibration points were taken at 10-cm increments to a depth of 40 cm. The samples were bulked from the cores of two holes located between the pairs of electrodes and separated by 50 cm. The selected sites form our second data set, ‘calibration data set’. Soil samples from these sites were obtained and analyzed in the laboratory. Samples were air dried and crushed to pass through a 2-mm mesh. The 1:2.5 soil–water suspensions were prepared and pH and electrical conductivity ( $EC_{2.5}$ ,  $dS\ m^{-1}$ ) were measured after 16 h. The  $EC_{2.5}$  is given after conversion to the standard temperature of 25 °C. Gravimetric moisture was determined by drying the soil samples in air-tight containers at 105 °C.

Sampling the ‘data set to be calibrated’ and ‘calibration data set’ was repeated 19 times from November 1994 to June 2001. The most frequent temporal lag was 3 months, but it ranged from 2 to 9 months. The  $EC_{2.5}$ , which represents the hard data, was measured at 13–20 locations, depending on the sampling period.

### 3. Methods

#### 3.1. Space–time random field model

The distribution of the electrical conductivity values in space and time is represented adequately by a space–time random field (STRF),  $X(\mathbf{p})$ , which takes values at points  $\mathbf{p}=(s,t)$  in a space–time domain, where  $s=(s_1,s_2)$  represents spatial location and  $t$  is the time (Christakos, 1992). The values taken by the STRF,  $X(\mathbf{p})$ , at a given space–time point  $\mathbf{p}_i$  is a random variable  $x_i=X(\mathbf{p}_i)$  and its corresponding realization is  $\chi_i$ . The random field  $X(\mathbf{p})$  at  $m+1$  points  $\mathbf{p}_i$  ( $i=1,\dots,m,k$ ) ( $m$  observation points and one estimation point) is denoted by the random variables  $x_{\text{map}}=[x_1,\dots,x_m,x_k]'$  and their realization is  $\chi_{\text{map}}=[\chi_1,\dots,\chi_m,\chi_k]'$ ; the prime represents the transpose of the vector.

A random variable  $x_i$  may acquire any one value from a distribution of values. The distribution of values  $\chi_i$  that the random variable  $x_i$  may take is

described by the cumulative distribution function (cdf):  $F_x(\chi_i)=Prob[x_i\leq\chi_i]$ . The derivative of the cdf,  $F_x(\chi_i)$ , with respect to  $\chi_i$  is the probability density function (pdf):

$$f_x(\chi_i) = \delta F_x(\chi_i) / \delta \chi_i. \tag{1}$$

The random variable  $x_i$  is fully described by its cdf or pdf.

The random field  $X(\mathbf{p})$  is a collection of random variables  $x_{\text{map}}=[x_1,\dots,x_m,x_k]'$  at points  $\mathbf{p}_i$  ( $i=1,\dots,m,k$ ). It is completely described by the multivariate cdf:

$$F_x(\chi_1, \chi_2, \dots, \chi_m, \chi_k) = Prob[x_1\leq\chi_1, x_2\leq\chi_2, \dots, x_m\leq\chi_m, x_k\leq\chi_k]. \tag{2}$$

Its derivative with respect to  $\chi_{\text{map}}=[\chi_1,\dots,\chi_m,\chi_k]$  is the multivariate pdf:

$$f_x(\chi_1, \chi_2, \dots, \chi_m, \chi_k) = \delta^{m+1} F_x(\chi_1, \chi_2, \dots, \chi_m, \chi_k) / \delta \chi_1 \delta \chi_2 \dots \delta \chi_m \delta \chi_k. \tag{3}$$

As the multivariate pdf is, in general, not known a priori, the random field can be characterized by its moments of order one, which is the mean function,  $m_x(\mathbf{p})$ , and/or order two, which is the covariance function,  $c_x(\mathbf{p},\mathbf{p}')$ .

If the two moments are invariant for all translations in space and time, the STRF is spatially homogeneous and temporally stationary. In this case the mean function is constant and the covariance function depends only on the spatial lag  $r=s-s'$  and temporal lag  $\tau=t-t'$  between any two points  $\mathbf{p}=(s,t)$  and  $\mathbf{p}'=(s',t')$ :

$$c_x(\mathbf{p},\mathbf{p}') = c_x(s-s',t-t') = c_x(r,\tau). \tag{4}$$

The random field is said to be isotropic when the covariance function depends only on the spatial separating distance and not on its direction:

$$c_x(r,\tau) = c_x(|r|,\tau). \tag{5}$$

The STRF is spatially homogeneous and temporally stationary only if any spatial or temporal trend is absent. In the presence of trends the mean and covariance functions depend on the location and time of measurements and the covariance function does not tend towards zero as the spatial or

temporal lag increases. The trend can be estimated (see Section 3.3) and subtracted from the original spatially non-homogeneous and temporally non-stationary random field to obtain a residual random field,  $R(\mathbf{p})$ , that is now spatially homogeneous and temporally stationary:

$$R(\mathbf{p}) = R(s, t) = X(\mathbf{p}) - m_x(\mathbf{p}). \tag{6}$$

Kriging is a well-known geostatistical prediction technique (Goovaerts, 1997) and so is not discussed. Bayesian maximum entropy is a more recent approach and we describe it briefly in the following section.

### 3.2. Bayesian maximum entropy

The BME approach, promoted by Christakos (1990, 2000), provides a systematic and rigorous way to incorporate soft data and other sources of information, in addition to hard data, into the analysis.

In space–time mapping we want to predict the values of the random field,  $X(\mathbf{p})$ , at a point,  $\mathbf{p}_k (k \neq i = 1, \dots, m)$ , given data at space–time points,  $\mathbf{p}_i (i = 1, \dots, m)$ . The joint cdf for the  $m$  data points and the prediction point,  $\mathbf{p}_k$ , is defined by Eq. (2) and its corresponding pdf by Eq. (3). This pdf forms the prior pdf. It should be derived by an estimation process that takes into account physical constraints given by prior information or knowledge. In the BME context, Shannon’s information criterion (Shannon, 1948) is used as a measure of information:

$$\text{Info}(x_{\text{map}}) = -\log[f_G(x_{\text{map}})], \tag{7}$$

where  $f_G(x_{\text{map}})$  has the same definition as Eq. (3), except that  $x$  is replaced by  $G$ . This means that the pdf is defined based only on general knowledge, before any use is made of the site-specific data. The expected information is defined as:

$$E[\text{Info}(x_{\text{map}})] = -\int \log[f_G(x_{\text{map}})]f_G(x_{\text{map}})dx_{\text{map}}. \tag{8}$$

Eq. (8) is the Shannon entropy function. The expected information or, conversely, the entropy function needs to be maximized.

As the estimator,  $\hat{X}(\mathbf{p}_k)$ , of the random field,  $X(\mathbf{p})$ , at the point,  $\mathbf{p}_k$ , is in general expressed in terms of expectations of some function of  $X(\mathbf{p})$  (not limited to

linear combinations), the physical constraints imposed by  $G$  are given by:

$$E[g_\alpha] = \int g_\alpha(\chi_{\text{map}})f_G(\chi_{\text{map}})d\chi_{\text{map}}, \quad \alpha = 0, \dots, N_c \tag{9}$$

where  $g_\alpha$  are functions chosen such that the general knowledge base,  $G$ , is taken account of in full in the prediction process, and their expectations,  $E[g_\alpha]$ , provide the space–time statistical moments of interest.

The BME analysis is done in three main stages (Christakos, 1998, 2000):

(1) Structural or prior stage: The goal is to maximize the information content using the general knowledge only. The total physical knowledge in the BME context includes site-specific and general knowledge. The latter, denoted  $G$ , includes general information that can characterize more than one STRF, such as physical laws, statistical moments, multipoint statistics, etc. In our case  $G$  represents the moments of order two (space–time mean and covariance functions). Eq. (10) gives the prior (or  $G$ -based) multivariate pdf:

$$f_G(\chi_{\text{map}}) = Z^{-1} \exp \left[ \sum_{\alpha=1}^{N_c} \mu_\alpha g_\alpha(\chi_{\text{map}}) \right] \tag{10}$$

where  $Z = \exp(-\mu_0)$  is a normalization constant and  $\mu_\alpha$  are Lagrange multipliers.

(2) Meta-prior stage: In the BME framework  $\chi_{\text{map}}$  includes the data values and the value to be predicted:  $\chi_{\text{map}} = [\chi_{\text{data}}, \chi_k]'$ . The data vector,  $\chi_{\text{data}} = [\chi_1, \dots, \chi_m]'$ , forms the available data, or site-specific knowledge denoted by  $S$ . The total knowledge is  $K = G \cup S$ . The available data can be divided into two main types:

- Hard data:  $\chi_{\text{hard}} = [\chi_1, \dots, \chi_h]'$  and
- Soft data:  $\chi_{\text{soft}} = [\chi_{h+1}, \dots, \chi_m]'$ , such that  $\chi_{\text{data}} = [\chi_{\text{hard}}, \chi_{\text{soft}}]'$ .

In our case study,  $EC_{2.5}$  are the hard data available at up to  $h = 20$  locations in space. The lower and upper confidence interval limits, calculated with the calibration equations, form the soft data. As we have soft interval data,  $\chi_{\text{soft}}$  can be written as:

$$\chi_{\text{soft}} = \left\{ [\chi_{h+1}, \dots, \chi_m]' : \chi_i \in I_i = [l_i, u_i], \right. \\ \left. i = h + 1, \dots, m \right\}. \tag{11}$$

This means that the unobserved exact values  $\chi_i$  have probabilities of one within known intervals  $I_i$  with  $l_i$  and  $u_i$  the lower and upper limits, respectively.

(3) Integration or posterior stage: In this last step the two knowledge bases ( $G$  and  $S$ ) are integrated. The aim is to maximize the posterior pdf given the total knowledge  $K$ . The prior pdf is updated by taking into account the available site-specific knowledge (the data). The posterior and the prior pdfs are related through the conditional probability law (Christakos, 1990), based on Bayes theorem, hence B in BME:

$$f_K(\chi_k|\chi_{\text{data}}) = f_G(\chi_{\text{map}})/f(\chi_{\text{data}}), \tag{12}$$

where  $f_K(\chi_k|\chi_{\text{data}})$  and  $f_G(\chi_{\text{map}})$  are the posterior and the prior pdfs, respectively.

The posterior pdf should be maximized with respect to  $\chi_k$ . This stage gives the  $K$ -based pdf:

$$f_K(\chi_k) = A^{-1} \int_I f_G(\chi_{\text{map}}) d\chi_{\text{soft}}, \tag{13}$$

where  $A = \int_I f_G(\chi_{\text{data}}) d\chi_{\text{soft}}$  is a normalization coefficient and both of the integrations are in the defined domain for the soft interval data, Eq. (11), such that  $I = I_{h+1} \cup I_{h+2} \cup \dots \cup I_m$ . When we substitute the prior pdf (Eq. (10)) into the  $K$ -based pdf (Eq. (13)), we get the posterior or BME pdf:

$$f_K(\chi_k) = (AZ)^{-1} \int_I \exp \left[ \sum_{z=1}^{N_c} \mu_z g_z(\chi_{\text{map}}) \right] d\chi_{\text{soft}}. \tag{14}$$

This posterior pdf, which is not necessarily Gaussian, describes fully the STRF at the target point. It provides a complete picture of the prediction situation, instead of one statistic (the expected value), as the full statistical distribution is defined as well as the different estimators of the STRF and the estimation uncertainty. Among the possible estimators, the mode represents the most probable realization. It is the value that maximizes the posterior pdf. The mean estimate is defined as:

$$\bar{x}_{k|K} = \int f_K(\chi_k) \chi_k d\chi_k. \tag{15}$$

It is, in general, a nonlinear function of the available data and is suitable for situations where one is interested in minimizing the mean squared estimation error.

A guide to the uncertainty associated with the estimated values is given by the variance of the estimation error. This is provided by BME, which is data-dependent, whereas in kriging it is data-free. Moreover, BME enables a more accurate assessment of the estimation error from the posterior pdf, Eq. (14). From this, one can calculate the confidence intervals (Serre and Christakos, 1999), which provide a more realistic assessment of the estimation error than the error variance. These intervals need not be symmetric about the estimated value.

When the general knowledge  $G$  is limited to the first two statistical moments (mean and covariance functions) and the site-specific knowledge is restricted to the hard data, the kriged estimates correspond to the BME mean estimates, Eq. (15). The difference between the two types of kriging is in the number of data considered during the analysis. Ordinary kriging (HK) is limited to the hard data only, whereas HSK treats both sets of data as hard essentially.

### 3.3. Data analysis

All analyses were done on logarithmically transformed  $EC_{2.5}$  and  $EC_a$  because of the pronounced skewness of their distributions.

The steps of the structural analysis have been discussed elsewhere (Douaik et al., 2004) and we summarize them here:

- (1) We estimated the space and time trends of the log-transformed data. The smoothed spatial components (one for each location) were calculated using an exponential spatial filter applied to averaged measurements over all the time periods. Also temporal components (one for each time instant) were computed using an exponential temporal filter applied to the averaged measurements over all the spatial locations. More detail on this step is given in the function ‘*stmean*’ in Christakos et al. (2002), page 160 and in the accompanying CD.
- (2) The above components were interpolated to the data grid giving the space–time trend function.
- (3) The trend,  $m_x(\mathbf{p})$ , was subtracted from the original  $EC_{2.5}$  data  $X(\mathbf{p})$ , which results in the residual random field  $R(\mathbf{p})$ , Eq. (6).

- (4) The residuals were used to compute the space–time covariance function,  $c_R(r; \tau)$ , Eq. (4), by replacing the original random field  $X(\mathbf{p})$  with the residual,  $R(\mathbf{p})$ . The available data were sparse and so we were restricted to an isotropic model,  $c_R(r; \tau) = c_R(|r|, \tau)$ , Eq. (5).
- (5) Finally, we fitted a theoretical model to the experimental residual covariance function.

Our two data sets were separated into hard and soft electrical conductivity data. The measurements of  $EC_{2.5}$ , which provide accurate and direct values of soil salinity, form the hard data. The calibration data set was used to calculate the soft interval data. The pairs of data values of  $EC_a$  and  $EC_{2.5}$  were used to determine the calibration equations, one for each time instant, by calculating simple ordinary least squares regression models:

$$\log_e(EC_{2.5}) = a + b \log_e(EC_a), \quad (16)$$

where  $a$  is the intercept and  $b$  is the slope of the regression model.

These calibration equations were applied to the ‘data set to be calibrated’ to give the expected values and their standard deviations for all locations (413) and 17 time periods; the two remaining periods (sampling campaigns 18 and 19, which correspond to March and June 2001, respectively) were kept for validation. These parameters were used to determine the 95% confidence intervals; their lower and upper limits form our soft interval data.

We compared 3 methods of space–time prediction, which differ in the way the soft data are processed:

- (1) Ordinary kriging using only the hard data (HK), which provides no direct way of integrating soft data and ignores them.
- (2) Ordinary kriging using hard data and the midpoint of the soft interval data regards the latter as if it were a hard datum (HSK) and disregards their uncertainty.
- (3) Bayesian maximum entropy using the hard and soft interval data (BME), which integrates the soft interval data in the prediction as they are, and maintaining the difference in the degree of uncertainty between hard and soft data.

The methods were compared by cross-validation with the validation data of sampling campaigns 18 and 19.

### 3.4. Validation and comparison criteria

Soil salinity was predicted at each of the sites (19 for March 2001 and 20 for June 2001) for which we had measurements by deleting in turn the value of each location where the prediction was being made. This gave pairs of estimated–observed soil salinity values for the two time periods. Three quantitative criteria were computed from these pairs of values: the Pearson correlation coefficient ( $r$ ), the mean error or bias (ME), and the mean squared error (MSE). The first,  $r$ , measures the strength of the linear relation between the estimated and the observed soil salinity values and should be close to one for an accurate prediction. The ME should be close to zero and the MSE should be as small as possible. We also represented graphically the distribution of the estimation errors for a visual comparison of the three methods.

The MSE can be divided further into components that identify and quantify the deviation of estimated values from the observations. They represent different aspects of the discrepancy between the estimates and the measurements (Kobayachi and Us Salam, 2000). Let  $x_i$  and  $y_i$  ( $i=1, \dots, n$ ) represent the estimated and the observed soil salinity values, respectively, and  $d_i = x_i - y_i$  the deviation of the estimated values from the observations. The mean error (ME) or bias is defined by:

$$ME = \frac{1}{n} \sum_{i=1}^n (x_i - y_i) = \bar{x} - \bar{y}, \quad (17)$$

where  $\bar{x}$  and  $\bar{y}$  represent the means of the estimated and the observed values, respectively, and  $n$  is the number of locations for which observations are available. The mean squared error (MSE) is:

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)^2 = (\bar{x} - \bar{y})^2 + \frac{1}{n} \sum_{i=1}^n [(x_i - \bar{x}) - (y_i - \bar{y})]^2. \quad (18)$$

The first term on the right is the square of the bias (SB):

$$SB = (\bar{x} - \bar{y})^2 = ME^2. \quad (19)$$

The second term is the mean squared difference between the estimates and the measured values with respect to the deviation from means. It is known as the mean squared variation (MSV) and represents the proportion of the MSE that is not due to the bias. A larger MSV indicates that the model did not estimate the variability of the observed values around their mean adequately, i.e. the precision of the predicted values is poor.

Eq. (18) can be rewritten as:

$$MSE = SB + MSV. \quad (20)$$

The MSV can be divided, in turn, into two components:

$$\begin{aligned} MSV &= (SD_e - SD_o)^2 + 2SD_eSD_o(1 - r) \\ &= SDSD + LCS, \end{aligned} \quad (21)$$

where

$$SD_e = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}, \quad (22)$$

which is the standard deviation of the estimated values, and

$$SD_o = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}, \quad (23)$$

is the standard deviation of the observed values. The SDSD is the difference in the magnitude of fluctuation between the estimated and measured values. A larger value implies that the model failed to estimate the magnitude of fluctuation among the measurements. The LCS is the lack of positive correlation ( $1 - r$ ) weighted by the standard deviations. A large value means that the model did not estimate the degree of fluctuation in the observations.

Eq. (20) can be rewritten as:

$$MSE = SB + SDSD + LCS. \quad (24)$$

All the analyses were done using the BMelib toolbox (Christakos et al., 2002) written for Matlab (MathWorks, 1999).

## 4. Results and discussion

### 4.1. Descriptive statistics

Table 1 gives the summary statistics of soil salinity ( $EC_{2.5}$ ) for the different time periods. The mean values, which vary between 1.42 dS  $m^{-1}$  for November 1994 and 3.30 dS  $m^{-1}$  for September 1998, suggest the presence of strong temporal variability. There is also considerable spatial variation shown by the large ranges between the minimum and maximum values for the different time periods. For example, December 2000 has a range of 7.38 dS  $m^{-1}$ , which is the largest for the data examined.

The Pearson correlation coefficients between  $EC_a$  and  $EC_{2.5}$  are strong. They vary from 0.83 to 0.97.

### 4.2. Covariography

The spatial and temporal components of the mean trend were computed and interpolated to the grid data. They were then subtracted from the  $EC_{2.5}$  values to

Table 1  
Summary statistics for the hard data  $EC_{2.5}$  (dS  $m^{-1}$ )

$EC_{2.5}$	$N$	Mean	SD	Range	$r$	Skewness	Kurtosis
<i>Calibration data</i>							
Nov 1994	13	1.42	0.34	0.96	0.85	-0.71	-1.00
Mar 1995	20	2.29	0.70	2.23	0.91	0.66	-0.78
Jun 1995	20	2.02	1.10	3.66	0.88	0.80	-0.50
Sep 1995	20	2.01	0.99	3.73	0.94	0.89	0.67
Dec 1995	20	1.84	0.90	2.59	0.92	0.45	-1.31
Mar 1996	16	2.07	0.81	2.57	0.87	0.08	-1.26
Jun 1996	20	1.83	0.90	2.86	0.87	0.28	-1.12
Mar 1997	20	1.61	0.63	2.20	0.89	-0.08	-0.87
Jun 1997	15	1.77	0.97	3.12	0.83	0.78	-0.23
Sep 1997	20	1.63	1.24	4.25	0.94	1.06	0.20
Dec 1997	20	1.60	1.04	3.61	0.90	0.51	-0.62
Sep 1998	20	3.30	2.17	6.95	0.85	0.50	-0.69
Apr 1999	20	1.84	1.74	6.41	0.93	1.70	2.48
Jul 1999	13	2.27	1.57	4.86	0.91	0.73	-0.52
Sep 1999	20	2.29	1.89	6.52	0.91	1.07	1.10
Apr 2000	18	2.11	1.63	6.32	0.94	1.19	1.51
Dec 2000	20	2.32	1.91	7.38	0.93	1.31	1.71
<i>Validation data</i>							
Mar 2001	19	1.80	1.21	4.42	0.97	1.05	1.18
Jun 2001	20	1.99	1.75	5.83	0.86	1.22	0.42

$N$  is the number of observations, SD is the standard deviation, and  $r$  is the Pearson correlation coefficient.

give the residuals on which we computed and modelled the space–time covariance function. The latter is shown in Fig. 2a as function of the spatial lag,

$r$ , and in Fig. 2b as function of the temporal lag,  $\tau$ . Its full representation as function of both spatial and temporal lags is shown in Fig. 3.

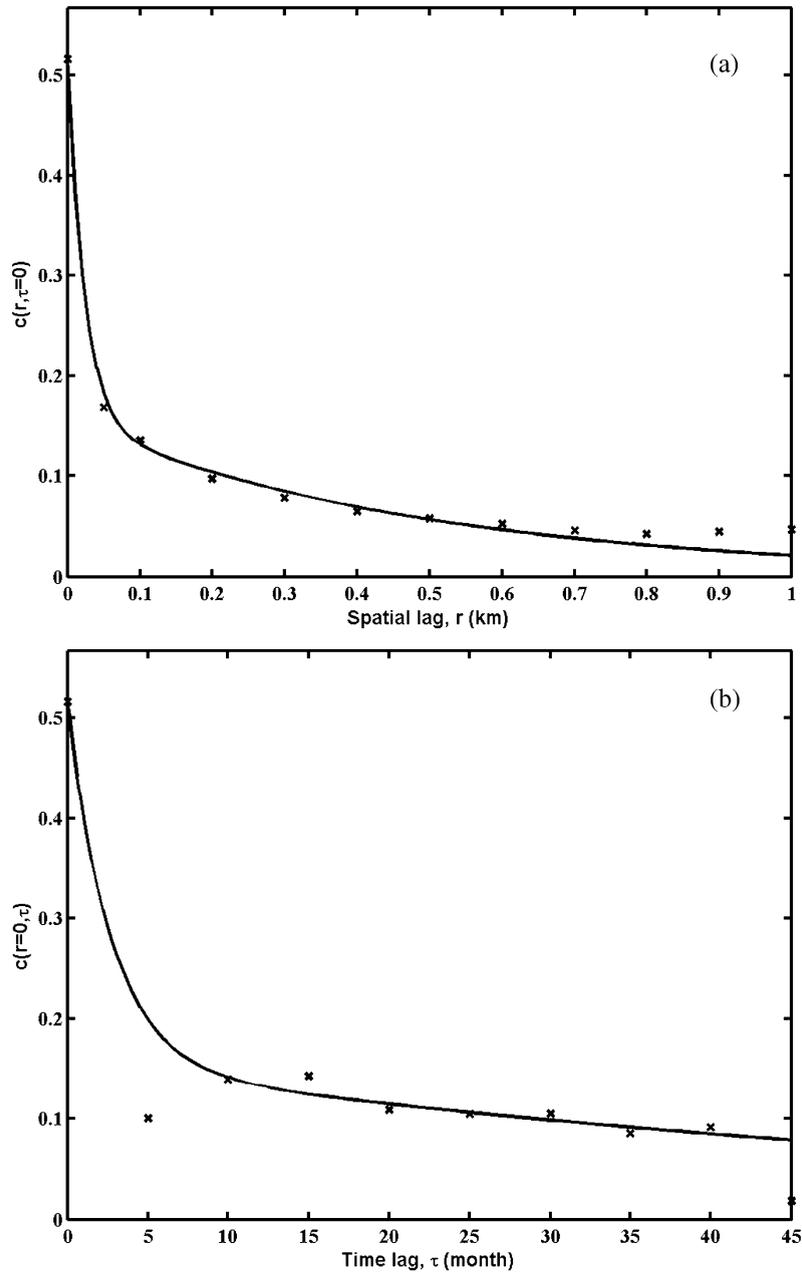


Fig. 2. Covariograms: (a) spatial covariance and (b) temporal covariance. Crosses are the experimental values and the solid line is the fitted model.

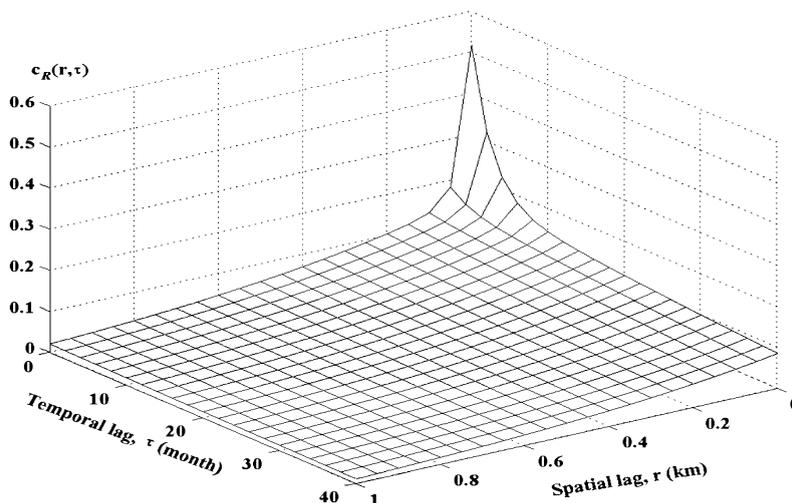


Fig. 3. Space–time covariance of the residual soil salinity data,  $R(p)$ .

The covariance model fitted to the experimental values was a non-separable space–time model that is the sum of two nested exponential models:

$$c_R(r, \tau) = \sum_{k=1}^2 C_{0k} \exp \left[ -\frac{3r}{as_k} - \frac{3\tau}{at_k} \right].$$

The first nested model with a sill,  $C_{01}$ , of  $0.27 \text{ (dS m}^{-1}\text{)}^2$ , a spatial range,  $as_1$ , of 250 m, and a temporal range,  $at_1$ , of 8 months corresponds to short-scale fluctuations. The long-range fluctuations are represented by the second component of the model with a sill,  $C_{02}$ , of  $0.07 \text{ (dS m}^{-1}\text{)}^2$ , a spatial range,  $as_2$ , of 1500 m, and a temporal range,  $at_2$ , of 200 months. The contribution of the second structure to the total variance of  $0.34 \text{ (dS m}^{-1}\text{)}^2$  is small; it is only 20.6%. The ranges of both the spatial and temporal covariograms of this second structure are beyond the sampling distances.

### 4.3. Comparison of results

Soil salinity was predicted for two time periods, March and June 2001, using the three approaches discussed above. The cross-validation criteria ME, MSE, and  $r$  are given in Table 2 for both times. Fig. 4 gives the distributions of the errors for March 2001.

The HSK results are the poorest; they have the largest bias (ME) compared to the two other approaches, although it is still not significantly

different from zero, and the largest MSE for both time periods. Fig. 4 shows that HSK has the broadest error distribution, mostly on the negative side of the curve. This implies that this method is likely to produce larger errors than the other two. The errors for BME have a higher mode and a narrower distribution compared with both kriging techniques (Fig. 4): this is confirmed by its having the smallest MSE values (Table 2). The MSE for HK is between those of BME and HSK; therefore this method provides more accurate estimates than HSK but less accurate ones than BME. However, HK gives estimates that are less biased than BME (mainly for March 2001). The estimates are strongly correlated with the observations for the three techniques and both time periods.

When BME is used without any soft data, the results are strictly equivalent to HK. This accords with the theory (Christakos and Li, 1998). Bayesian

Table 2  
Quantitative criteria for the comparison of the three approaches

Criterion	Time	HK	HSK	BME
$r$	March 2001	0.87	0.92	0.93
	June 2001	0.93	0.93	0.95
ME (dS m <sup>-1</sup> )	March 2001	-0.176	-0.323	-0.226
	June 2001	0.062	-0.057	-0.017
MSE (dS m <sup>-1</sup> ) <sup>2</sup>	March 2001	0.489	0.650	0.387
	June 2001	0.378	0.513	0.337

$r$  is the Pearson correlation coefficient, ME is the mean error, and MSE is the mean squared error.

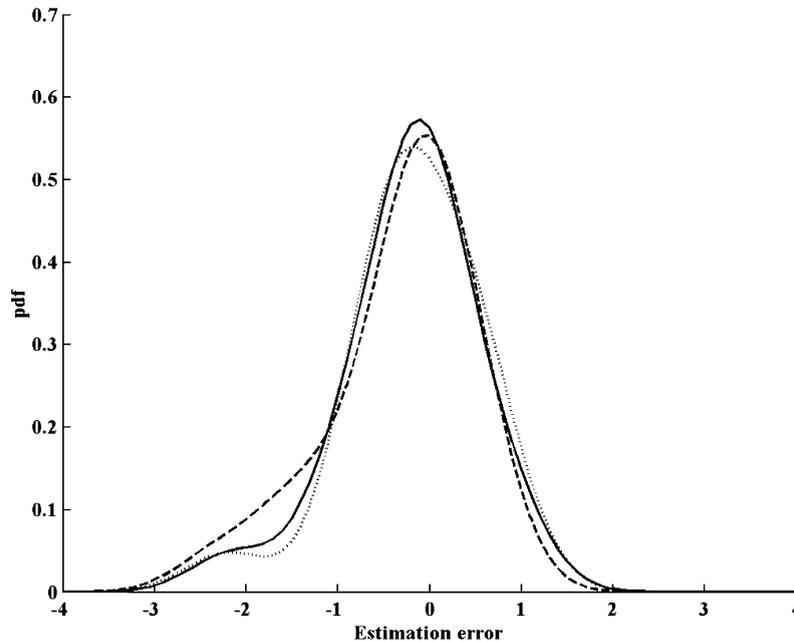


Fig. 4. Distributions of estimation error for March 2001. The solid line is for BME, the dotted line is for HK, and the dashed line is for HSK.

maximum entropy can also give estimates of soil salinity in the absence of hard data; the results of this cross-validation are given in Table 3. The estimates show more bias (but still negligible) and are slightly less accurate (particularly for March 2001), but the differences in the results when hard (the neighbourhood search was limited to a maximum of the 10 nearest data) and soft data were used are not significantly greater (last column of Tables 2 and 3). This is a useful feature of BME and D'Or and Bogaert (2003) used this property to map soil texture using only the intervals defined from a textural triangle.

To investigate further the incorporation of soft data by HSK and BME, we analyzed, in addition to the hard data, only data with the largest intervals (from 30 spatial locations for each time period) rather than the full set of 393 spatial locations for each time period

Table 3  
Quantitative criteria for BME when hard data were excluded

Criterion	March 2001	June 2001
$r$	0.91	0.94
ME	0.294	-0.075
MSE	0.701	0.348

$r$  is the Pearson correlation coefficient, ME is the mean error, and MSE is the mean squared error.

(see Table 2 and Fig. 4). Table 4 gives the results for March 2001 and the distributions of the estimation errors are shown in Fig. 5.

The results for HK (Table 4) are the same as before as this technique takes no account of the soft data. The ME for BME, -0.192, is not markedly different from that in Table 2 when all the interval data were used (-0.226) but the MSE is larger, 0.612 instead of 0.387 for the full interval data. The results for HSK show that the estimates are biased and inaccurate, giving erroneous estimates for some spatial locations. The quantitative criteria confirm the graphical representation in Fig. 5.

The components of MSE can provide more information about the difference between the estimated and the observed values. They are given in

Table 4  
Quantitative criteria to compare the three methods of prediction using only the largest interval data

Criterion	HK	HSK	BME
$r$	0.87	0.23	0.86
ME	-0.176	-88.4	-0.192
MSE	0.489	106079.4	0.612

$r$  is the Pearson correlation coefficient, ME is the mean error, and MSE is the mean squared error.

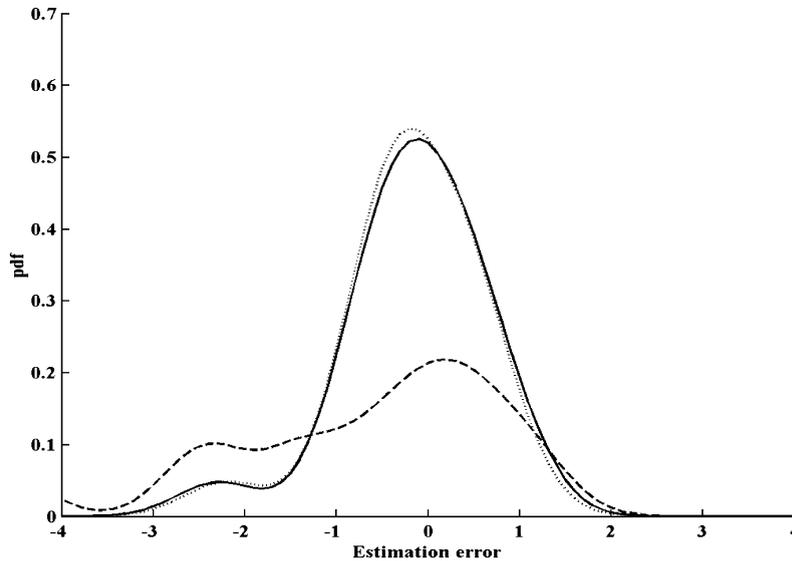


Fig. 5. Distributions of the estimation errors for March 2001, using only the largest interval data. The solid line is for BME, the dotted line is for HK, and the dashed line is for HSK.

absolute values in Table 5 and shown as proportions in Fig. 6. The HSK has the largest MSE for both time periods, whereas BME has the smallest values. This suggests that BME provides more accurate predictions than the two methods of kriging used. In addition, the contribution of the bias to the MSE is almost zero for the three approaches for June 2001, but for March 2001 it is largest for HSK (16%), followed by BME (13.2%), and finally HK (6.3%). This confirms that HK estimates are the least unbiased and those of HSK are the most biased.

The lack of positive correlation (LCS) is the component that contributed the most to the MSE of

the three interpolation methods for both time periods (Table 5 and Fig. 6); the larger contribution is for June 2001. In particular, it contributed the most to the MSE of the HK estimates with 87% and 95%, whereas it contributed least to that of HSK (47% and 62%) and its contribution was intermediate for BME (63% and 76%); the first value refers to March and the second to June 2001. This suggests that HK failed to estimate the degree of fluctuation in the observed soil salinity even if its MSE is smaller than that of HSK. However, the SDSD contributed more to the MSE of HSK (36.8% and 36.5%) than to that of HK (6.7% and 4%) and BME (23.8% and 24.3%), which indicates that HSK failed to estimate the magnitude of fluctuation in the measured electrical conductivity. The three components of MSE (SB, SDSD, and LCS) for BME are intermediate to those of HK and HSK. The MSE of BME is the smallest, indicating that it performs better than the two kriging techniques, and the SDSD shows that it represents the degree of fluctuation in the observations reasonably.

The small MSE for BME in March 2001 can be explained as follows. The standard deviation  $SD_o$  for this time period is  $1.18 \text{ dS m}^{-1}$ . Since SB (0.051) and SDSD (0.092) are negligible, LCS (0.244) is the component that contributes most to the MSE (0.387). As the Pearson correlation coefficients (0.92 for HSK

Table 5  
The components of MSE as absolute values

Criterion	Time	HK	HSK	BME
SB	March 2001	0.031	0.104	0.051
	June 2001	0.003	0.004	0.000
SDSD	March 2001	0.033	0.239	0.092
	June 2001	0.015	0.187	0.082
LCS	March 2001	0.425	0.307	0.244
	June 2001	0.360	0.322	0.255
MSE	March 2001	0.489	0.650	0.387
	June 2001	0.378	0.513	0.337

SB is the squared bias, SDSD is the squared difference between standard deviations, and LCS is the weighted lack of positive correlation.

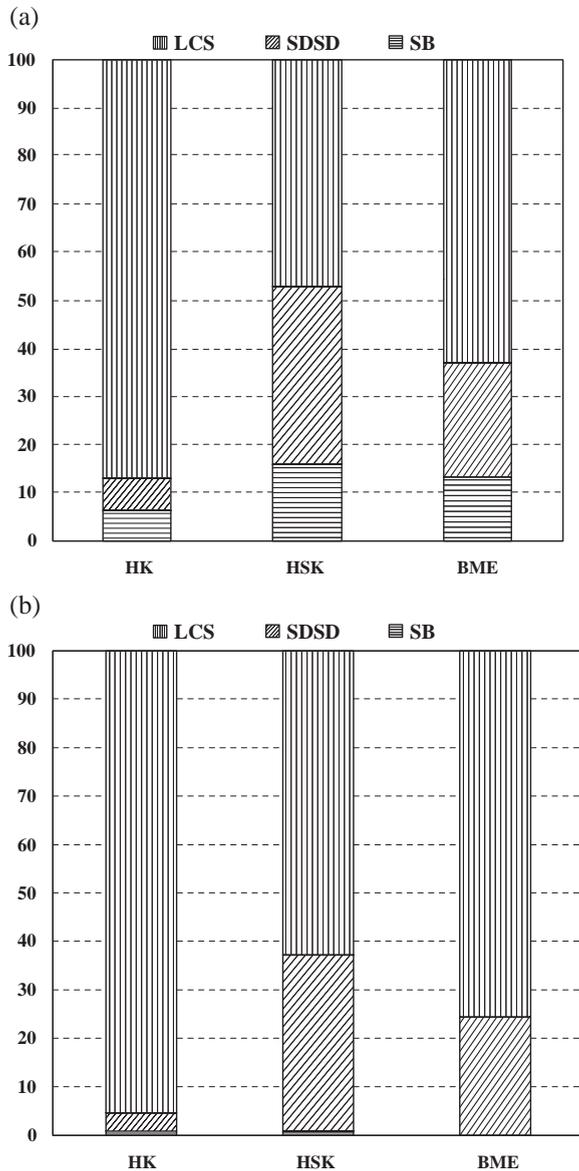


Fig. 6. Contribution of the components of MSE to its total: (a) March 2001 and (b) June 2001. The SB is the squared bias, SDSD is the squared difference between standard deviations, and LCS is the weighted lack of positive correlation.

and 0.93 for BME) are the same and  $SD_o$  is the same, the larger  $SD_e$  for HSK (1.67) led to a larger LCS and hence larger MSE. The smaller  $SD_e$  for BME (1.48) resulted in a small LCS and hence small MSE. The overall deviation (MSE) between the measurements and the BME predictions for March 2001 is small

Table 6

The components of MSE as absolute values, using only the largest interval data for March 2001

Component	HK	HSK	BME
SB	0.031	7813.7	0.037
SDSD	0.033	97695.5	0.086
LCS	0.425	570.2	0.489
MSE	0.489	106079.4	0.612

SB is the squared bias, SDSD is the squared difference between standard deviations, and LCS is the weighted lack of positive correlation.

because the predicted soil salinity shows limited variation (compared to HSK estimates) for the 19 sites and BME predicted the observations with a reduced bias ( $-0.226$  against  $-0.323$  for HSK).

Using only the largest interval data instead of all of them, Table 6 and Fig. 7, the SDSD becomes the major component of MSE for HSK (92.1%). This means that HSK failed to estimate the magnitude of fluctuation accurately among the observed electrical conductivity values. The minimum and maximum values of the latter for March 2001 are  $0.16$  and  $4.58$   $dS\ m^{-1}$ , respectively, whereas for the HSK estimates they are  $0$  and  $1408$   $dS\ m^{-1}$ . This large range for the latter resulted in the large  $SD_e$  of  $313.74$   $dS\ m^{-1}$  compared to the small  $SD_o$  of  $1.18$   $dS\ m^{-1}$ ; these values explain the large SDSD. In comparison BME performed well. Its minimum and maximum predictions are  $0.10$  and  $5.24$   $dS\ m^{-1}$ , respectively. This is a consequence of how the soft data are integrated into

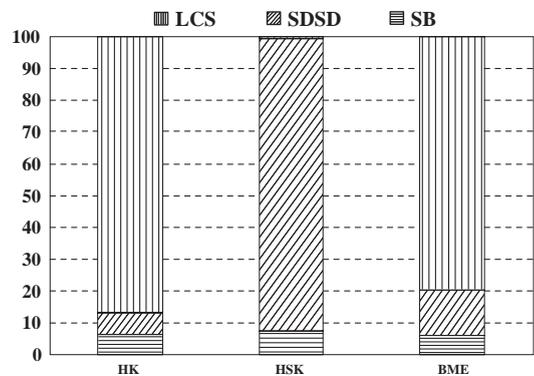


Fig. 7. Contribution of the components of MSE to its total, using only the largest interval data for March 2001. The SB is the squared bias, SDSD is the squared difference between standard deviations, and LCS is the weighted lack of positive correlation.

the prediction process. The HSK used only the midpoint of intervals, disregarding their range and the uncertainty associated with them, whereas BME considers the full information provided by the soft data. It takes into account the upper and lower limits of the interval data and the uncertainty associated with them. More importantly, BME distinguishes clearly between accurate (hard) and uncertain (soft) data and processes them differently.

## 5. Conclusions

The main aim of this work was to compare the performance of three prediction techniques: BME which incorporates soft interval data and two variants of kriging (one using only hard data and the other using hard data as well as the midpoint value of soft interval data, treating them as if they were hard data). The three approaches were evaluated by cross-validation for two different time periods, which had not been used in the previous analyses.

The BME provided reliable estimates even in the absence of any hard data. When no soft data were used, the BME estimates were strictly equivalent to those from kriging (HK). Based on the ME and MSE, we can conclude that the predictions from BME are less biased and more accurate than those from the two kriging techniques. Of these two techniques, the one using the soft data (HSK) resulted in more bias and less accuracy in the predictions. The results showed that BME improved substantially the accuracy of the predictions compared to kriging by taking into account soft interval data.

The Pearson correlation coefficients were of the same magnitude for HSK and BME. However, by dividing the MSE into three components, we found that HSK gave more biased estimates (large SB) and failed to reproduce the true magnitude of fluctuation among the observations.

The failure of HSK to incorporate the soft information was more pronounced when we used only the largest interval data, in addition to the hard data, instead of the full interval data. In this case HSK produced some unrealistic predictions of electrical conductivity (very large and unreliable predictions). In contrast, the way that BME integrates soft data into the prediction process resulted in more accurate

predictions, whether we used the full interval data or only the largest ones.

Ancillary data are cheap and readily available, sometimes for the whole study area (exhaustive secondary information). This secondary information can be used in an efficient way to complement the scarcity of the direct measurements of a soil property. This work showed that BME could incorporate and process soft data rigorously, leading to more accurate predictions.

## Acknowledgements

The authors acknowledge the two reviewers and Dr. Oliver for their comments, which improve the paper. They are also indebted to Dr. Oliver for the improvement of the English writing. Tibor Tóth acknowledges the material assistance of grants HNSF T37731 and OM-00124/2001.-NKPF/4.

## References

- Christakos, G., 1990. A Bayesian/maximum entropy view to the spatial estimation problem. *Mathematical Geology* 22, 763–776.
- Christakos, G., 1992. *Random Field Models in Earth Sciences*. Academic Press, San Diego.
- Christakos, G., 1998. Spatiotemporal information systems in soil and environmental sciences. *Geoderma* 85, 141–179.
- Christakos, G., 2000. *Modern Spatiotemporal Geostatistics*. Oxford University Press, New York.
- Christakos, G., Li, X., 1998. Bayesian maximum entropy analysis and mapping: a farewell to kriging estimators. *Mathematical Geology* 30, 435–462.
- Christakos, G., Bogaert, P., Serre, M.L., 2002. *Temporal GIS*. Springer-Verlag, New York.
- D'Or, D., Bogaert, P., 2003. Continuous-valued map reconstruction with the Bayesian maximum entropy. *Geoderma* 112, 169–178.
- Douaik, A., Van Meirvenne, M., Tóth, T., 2004. Spatiotemporal kriging of soil salinity rescaled from bulk soil electrical conductivity. In: Sanchez-Vila, X., Carrera, J., Gomez-Hernandez, J. (Eds.), *GeoEnv IV: Geostatistics for Environmental Applications*. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 413–424.
- Goovaerts, P., 1997. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York.
- Kobayachi, K., Us Salam, M., 2000. Comparing simulated and measured values using mean squared deviation and its components. *Agronomy Journal* 92, 345–352.
- Lee, Y.-M., Ellis, J.H., 1997. On the equivalence of kriging and maximum entropy estimators. *Mathematical Geology* 29, 131–151.

- Lesch, S.M., Strauss, D.J., Rhoades, J.D., 1995. Spatial prediction of soil salinity using electromagnetic induction techniques: 2. An efficient spatial sampling algorithm suitable for multiple linear regression model identification and estimation. *Water Resources Research* 31, 387–398.
- Lesch, S.M., Herrero, J., Rhoades, J.D., 1998. Monitoring for temporal changes in soil salinity using electromagnetic induction techniques. *Soil Science Society of America Journal* 62, 232–242.
- MathWorks, 1999. Using Matlab, Version 5. The MathWorks Inc., Natick, MA.
- McNeil, J.D., 1980. Electromagnetic Terrain Conductivity Measurement at Low Induction Numbers: Technical Note TN-6. GEONICS Limited, Ontario, Canada (15 pp.).
- Rhoades, J.D., van Schilfgaarde, J., 1976. An electrical conductivity probe for determining soil salinity. *Soil Science Society of America Journal* 40, 647–651.
- Serre, M.L., Christakos, G., 1999. Modern geostatistics: computational BME in the light of uncertain physical knowledge—the Equus beds study. *Stochastic Environmental Research and Risk Assessment* 13 (1/2), 1–26.
- Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423.
- Soil and Plant Analysis Council, 1992. Handbook on Reference Methods for Soil Analysis. Georgia University Station, Athens, Georgia.
- Szabolcs, I., 1989. Salt-affected Soils. CRC Press, Florida.
- Tóth, T., Rajkai, K., 1994. Soil and plant correlations in a solonchic grassland. *Soil Science* 157, 253–262.
- Tóth, T., Jozefaciuk, G., 2002. Physicochemical properties of a solonchic toposequence. *Geoderma* 106, 137–159.
- Tóth, T., Csillag, F., Biehl, L.L., Micheli, E., 1991. Characterization of semi-vegetated salt-affected soils by means of field remote sensing. *Remote Sensing of Environment* 37, 167–180.