Stoch Envir Res and Risk Ass (2004) 18: 219–227
DOI 10.1007/s00477-004-0177-5


ORIGINAL PAPER

A. Douaik · M. van Meirvenne · T. Tóth · M. Serre

# Space-time mapping of soil salinity using probabilistic bayesian maximum entropy

**Abstract** The mapping of saline soils is the first task before any reclamation effort. Reclamation is based on the knowledge of soil salinity in space and how it evolves with time. Soil salinity is traditionally determined by soil sampling and laboratory analysis. Recently, it became possible to complement these hard data with soft secondary data made available using field sensors like electrode probes. In this study, we had two data sets. The first includes measurements of field salinity ($EC_a$) at 413 locations and 19 time instants. The second, which is a subset of the first (13 to 20 locations), contains, in addition to $EC_a$, salinity determined in the laboratory ($EC_{2.5}$). Based on a procedure of cross-validation, we compared the prediction performance in the space-time domain of 3 methods: kriging using either only hard data (HK) or hard and mid interval soft data (HMIK), and Bayesian maximum entropy (BME) using probabilistic soft data. We found that BME was less biased, more accurate and giving estimates, which were better correlated with the observed values than the two kriging techniques. In addition, BME allowed one to delineate with better detail saline from non-saline areas.

**Keywords** Bayesian maximum entropy · Electrical conductivity · Geostatistics · Kriging · Soil salinity · Space–time variability

A. Douaik (✉) · M. van Meirvenne
Department of Soil Management and Soil Care,
Ghent University, Coupure Links 653, Gent, Belgium

A. Douaik
Department of Computer Science and Biometry,
Institut National de la Recherche Agronomique,
Avenue de la victoire, BP 415, Rabat, Morocco

T. Tóth
Research Institute for Soil Science and
Agricultural Chemistry of the Hungarian Academy of Sciences,
Herman O. út 15, PO Box 35, 1525 II Budapest, Hungary

M. Serre
Department of Environmental Sciences and Engineering,
University of North Carolina at Chapel Hill,
120B Rosenau Hall, Chapel Hill, 27599–7400, NC, USA

**Abbreviations** *BME* Bayesian maximum entropy · *EC* Electrical conductivity · *HK* Kriging with hard data · *HMIK* Kriging with hard and mid interval soft data · *ME* Mean error · *MSE* Mean square error · *R* Pearson correlation coefficient

## 1 Introduction

Worldwide, soil salinity (especially sodium salinization) is a major hazard to agriculture. It is a main limiting factor for agricultural productivity as salts cause limited uptake of water by plants and soluble or adsorbed sodium disperses soil aggregates, creating a weak structural stability, low hydraulic conductivity and reduced infiltration rate. In addition, leached salts may reach the ground and surface waters and contribute to pollution.

Consequently saline and sodic soils need specific management approach. This approach is based largely on the exact determination of the magnitude and the extent of soil salinity in space as well as in time.

Conventionally (Soil and Plant Analysis Council 1992) soil salinity is determined by laboratory analysis (electrical conductivity of saturated soil paste extract, $EC_e$). This procedure is expensive and time consuming. Alternatively soil salinity can be evaluated in the field (bulk soil or apparent electrical conductivity, $EC_a$) using, for example, electrode probes or electromagnetic induction techniques. The field approach is cheaper and easier than the laboratory analysis and, as a consequence, it is now widely used for monitoring soil salinity in space and time.

The appraisal of space-time variability of soil salinity has been approached in different ways. Lesch et al. (1998) used a classical statistical method by which they tried to detect if there is a temporal change in soil salinity between two time instants. The drawbacks of this approach are that they needed to repeat the analysis many times as they compared only two time instants. In addition the temporal autocorrelation, which may exist

between two or more successive measurements, was not accounted for.

An alternative to this approach has been proposed by Douaik et al. (2003). They first rescaled $EC_a$ into $EC_e$ using a simple regression model, then they applied space-time kriging to the new data matrix of 'estimated $EC_e$'. They were able to include the spatial as well as the temporal dependencies in their analysis, which represents an advantage over the approach of Lesch et al. (1998). However as $EC_e$ data are only estimates, they are not precise and entailed with uncertainty that needs to be taken into account.

In this work we use the Bayesian maximum entropy (BME) (Christakos 1990; Serre et al. 1998) framework to rigorously analyze our data by formally incorporating the difference in accuracy between laboratory and field soil salinity. The former is an accurate measure and can be considered as hard data and the latter is a cheaper and less accurate measurement of soil salinity that can be considered as soft data.

The BME framework is a spatio-temporal estimation method of modern geostatistics that can incorporate a wide variety of knowledge bases (statistical moments of any order, multipoint statistics, physical laws, hard and soft data, etc.) and kriging is considered to be a special and limiting case of it under restrictive assumptions about the knowledge base considered.

BME has been successfully applied in different areas. D'Or et al. (2001), Bogaert and D'Or (2002), and D'Or and Bogaert (2003) mapped soil texture using BME in the space domain. Serre and Christakos (1999) studied the water-table elevations of an aquifer in Kansas while Christakos and Serre (2000) analyzed the distribution of particulate matter in North Carolina. These two studies have been done in the space-time context. Bogaert (2002) extended the approach to include categorical variables.

The main objectives of this study are: (i) to apply BME to a space–time soil salinity data set, (ii) to determine the probability that soil can be considered saline in a study site, and (iii) to compare BME with two kriging techniques.

## 2 Data description

The study area, which covers around 25 ha, is located in the Hortobagy National Park (47°30′ N and 21°30′ E), a small zone of the larger Great Hungarian Plain, in the east of Hungary.

Soil salinity/sodicity and its correlation with the vegetation have been studied in the area by many researchers, among them Tóth et al. (1991, 2002), and Van Meirvenne et al. (1995).

Two data sets were available. The first, called 'data set to be calibrated', consisted of the values of the apparent electrical conductivity ($EC_a$, $dSm^{-1}$) which were measured using electrical probes equipment (with 4 probes) at 413 locations (Fig. 1). The probes were inserted in the soil to two depths (8 and 13 cm), which correspond to $EC_a$ measured over 0–20 and 0–40 cm soil depths respectively. These data and statistical parameters derived from them are called soft data since the field measured $EC_a$ is determined, not only by the $EC_e$ of the soil, but also by the soil moisture content, temperature, particle size distribution, etc.

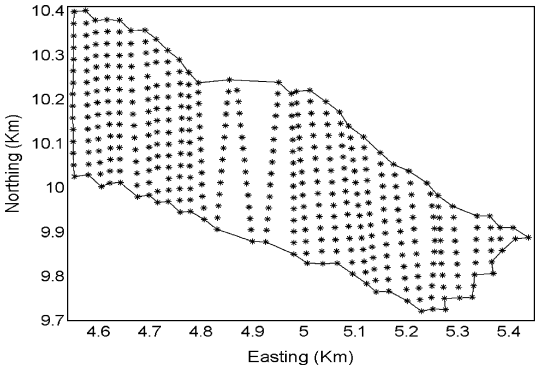


**Fig. 1** Spatial location of the samples where $EC_a$ was measured. The calibration data set is a sub-sample of these locations

The second data set, termed 'calibration data set', which is a subset of the first (13 to 20 locations) involves (in addition to $EC_a$) the measurements of soil salinity in 1:2.5 soil suspensions in the laboratory ($EC_{2.5}$, a simple proxy for $EC_e$, in $dS\ m^{-1}$), the soil moisture content (%), and the soil pH. The soil samples were taken at 4 depths from 0 to 40 cm with an increment of 10 cm. At each location and for each depth, two augerings were bulked to get a soil sample. The locations for this data set were selected using the response surface design algorithm (Lesch et al. 1995) in such a way that the selected samples were representative of all other locations from two points of view: the spatial configuration and possible salinity. The $EC_{2.5}$ data were considered to be hard data, since they were measured with high precision in standard laboratory conditions.

The whole process was repeated in time, from November 1994 to June 2001, at 19 time instants, with an average temporal lag of 3 months but intervals ranging from 2 to 9 months.

## 3 Bayesian maximum entropy approach

The concept of BME appeared more than one decade ago (Christakos, 1990) and was continuously developed and expanded (Christakos 1998, 2000; Christakos and Li 1998). It involves three basic steps.

The space-time random field (STRF) representing soil salinity is denoted by $X(\mathbf{p})$, with $\mathbf{p} = (\mathbf{s},t)$ where $\mathbf{s}$ denotes the spatial coordinates and $t$ the temporal coordinate. In our case, the general knowledge $G$ consists of the mean

$$m_{\underline{X}}(\mathbf{p}) = E[X(\mathbf{p})]$$

and covariance

$$c_{\underline{X}}(\mathbf{p},\mathbf{p}') = E\{[X(\mathbf{p}) - m_{\underline{X}}(\mathbf{p})][X(\mathbf{p}') - m_{\underline{X}}(\mathbf{p}')]\}$$

where $E$ denotes the stochastic expectation. The mean function represents the trends and systematic structures, and the covariance function expresses spatio-temporal interaction and dependencies. The vector of random variables $\mathbf{x}_{\text{map}} = (x_1, x_2, \ldots, x_{mh}, x_{mh+1}, \ldots, x_m, x_k)$ represents the random field $X(\mathbf{p})$ at the mapping points $\mathbf{p}_{\text{map}} = (\mathbf{p}_{\text{hard}}, \mathbf{p}_{\text{soft}}, \mathbf{p}_k)$, where $\mathbf{p}_{\text{hard}} = (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_{mh})$ are the hard data points, $\mathbf{p}_{\text{soft}} = (\mathbf{p}_{mh+1}, \mathbf{p}_{mh+2}, \ldots, \mathbf{p}_m)$ are the soft data points, and $\mathbf{p}_k$ is the estimation point; such that $\mathbf{x}_{\text{map}} = (\mathbf{x}_{\text{hard}}, \mathbf{x}_{\text{soft}}, \mathbf{x}_k) = (\mathbf{x}_{\text{data}}, \mathbf{x}_k)$. The vector of values corresponding to a specific realization of $\mathbf{x}_{\text{map}}$ is denoted $\boldsymbol{\chi}_{\text{map}} = (\boldsymbol{\chi}_{\text{hard}}, \boldsymbol{\chi}_{\text{soft}}, \boldsymbol{\chi}_k)$.

At the prior step, the objective is to maximize the information content using only $G$. The expected information is:

$$E[Info(\mathbf{x}_{\text{map}})] = -\int \ln[f_G(\boldsymbol{\chi}_{\text{map}})] f_G(\boldsymbol{\chi}_{\text{map}})\, \mathrm{d}\boldsymbol{\chi}_{\text{map}} \qquad (1)$$

where $f_G(\boldsymbol{\chi}_{\text{map}})$ represents the prior probability density function (pdf) of $\mathbf{x}_{\text{map}}$.

The maximization is done under the set of constraints

$$E[g_\alpha] = \int g_\alpha(\boldsymbol{\chi}_{\text{map}}) f_G(\boldsymbol{\chi}_{\text{map}})\, \mathrm{d}\boldsymbol{\chi}_{\text{map}}, \quad \alpha = 0, \ldots, N_c$$

with $g_\alpha$ the chosen functions such that their expectations $E[g_\alpha]$ are space–time statistical moments known from the general knowledge (Serre and Christakos 1999). In our case:

$g_0(\boldsymbol{\chi}_{\text{map}}) = 1 \Rightarrow E[g_0] = 1$,

which is a normalization constraint,

$g_\alpha(\chi_i) = x_i \Rightarrow E[g_\alpha] = E[x_i]$,

mean at point $p_i$, $\alpha = 1, \ldots, m+1$,

$g_\alpha(\chi_i, \chi_j) = [\chi_i - E(x_i)]\,[\chi_j - E(x_j)] \Rightarrow$

$E[g_\alpha] = E\{[x_i - E(x_i)]\,[x_j - E(x_j)]\}$,

covariances $c(x_i, x_j)$ with $\alpha = m+2, \ldots, (m+1)(m+4)/2$.

When Eq. (1) is solved, the $G$-based multivariate pdf is as follows

$$f_G(\boldsymbol{\chi}_{\text{map}}) = Z \exp[\Sigma \mu_\alpha g_\alpha(\boldsymbol{\chi}_{\text{map}})] \qquad (2)$$

where $Z = \exp(\mu_0)$ is a normalization constraint, and $\mu_\alpha$ the Lagrange multipliers.

At the meta-prior stage, the site-specific knowledge $S$ is collected and organized. In our case it involves the following hard and probabilistic soft data

$$S: \boldsymbol{\chi}_{\text{data}} = (\boldsymbol{\chi}_{\text{hard}}, \boldsymbol{\chi}_{\text{soft}}) = (\chi_1, \chi_2, \ldots, \chi_{mh}, \chi_{mh+1}, \ldots, \chi_m),$$

where the hard data $\mathbf{x}_{\text{hard}} = (x_1, x_2, \ldots, x_{mh})$ are the soil electrical conductivity measured in the laboratory with $m_h = 13$ to 20 depending on the time instant; and the soft data have the specific pdf $f_S(\xi)$ expressed as

$$\boldsymbol{\chi}_{\text{soft}} : P_S(x_{\text{soft}} <= \xi) = \int_{-\infty}^{\xi} f_S(\boldsymbol{\chi}_{\text{soft}}) \mathrm{d}\boldsymbol{\chi}_{\text{soft}}$$

During the last step, called integration or posterior stage, the prior pdf is updated considering the site-specific knowledge $S$ available. The objective is cogency (maximization of the posterior pdf given the total knowledge $K = G \cup S$). This step leads to the $K$-based pdf

$$f_K(\chi k) = A^{-1} \int f_G(\boldsymbol{\chi}_{\text{map}})\, \mathrm{d}\boldsymbol{\chi}_{\text{soft}} \qquad (3)$$

where $A = \int f_G(\boldsymbol{\chi}_{\text{data}}) \mathrm{d}\boldsymbol{\chi}_{\text{soft}}$ is a normalization coefficient and both the integrations are in the definition domain $I$ for the probabilistic soft data $\boldsymbol{\chi}_{\text{soft}}$.

The substitution of the $G$-based pdf, Eq. (1), into the $K$-based pdf, Eq. (3), leads to the posterior or BME pdf

$$f_K(\chi k) = (AZ)^{-1} \int \exp[\Sigma \mu_\alpha g_\alpha(\boldsymbol{\chi}_{\text{map}})] \mathrm{d}\boldsymbol{\chi}_{\text{soft}}$$

This pdf rigorously incorporates the space-time dependency of the data and the difference in accuracy between the hard and soft data. It is in general non-Gaussian, and provides a full stochastic description of the estimated value at the estimation point, which constitutes a crucial advantage over kriging, as for the latter the difference between hard and soft data is ignored, and we can only obtain the estimate and its variance. From the posterior pdf (3) we can compute different statistics, such as the conditional mean, the mode, the median or any other quantile. The uncertainty can be assessed using the variance of the posterior pdf, or better the BME confidence interval (Serre and Christakos 1999).

## 4 Data analysis

The distributions of both $EC_a$ and $EC_{2.5}$ were skewed. Therefore we applied a natural logarithmic transformation to both variables to obtain less asymmetric distributions.

The structural analysis (data detrending, experimental and fitted covariograms) has been reported elsewhere (Douaik et al. 2003). By way of summary, the steps of the structural analysis are as follow:

- The space-time mean trend of the log-transformed salinity is estimated. The smoothed spatial components were computed using an exponential spatial filter applied to the averaged measurements (for each location, over all the time instants). We computed also the smoothed temporal components using an exponential temporal filter applied to the averaged measurements (for each time instant, over all the locations);
- These components were interpolated to the data grid giving the space–time mean trend $m(\mathbf{p})$;

- The residuals were calculated as the space–time mean trend subtracted from the original log-transformed data: $R(\mathbf{s},t) = R(\mathbf{p}) = X(\mathbf{p}) - m(\mathbf{p})$
- The residuals were used to compute the spatio-temporal covariogram:

  $c_R(r,\tau) = cov[R(\mathbf{s}+r, t+\tau), R(\mathbf{s},t)]$

  $r$ and $\tau$ are the spatial and temporal lags, respectively, and cov is the covariance function.
- Finally we fitted a theoretical model to the computed experimental covariogram The mean trend and the covariance function obtained in the structural analysis were used as general knowledge for all three of the space-time interpolation methods presented in this study.

Next we organized the site-specific knowledge into hard and soft data of soil salinity. The hard data were the measurements of $EC_{2.5}$, which provide accurate values for the soil salinity. The soft data (interval midpoints as well as probabilistic pdf) were obtained from regression equations. As we had a 'calibration data set', we used the pairs of data $EC_a$–$EC_{2.5}$, to determine the calibration equations, one for each time, by computing simple ordinary least squares regression models:

$$\log(EC_2.5) = a + b\log(EC_a)$$

with $a$: the intercept, and $b$: the slope.

Using the calibration equations and applying them to the 'data set to be calibrated', we obtained the lower and upper 95% confidence intervals at the locations and time instants where measurements of $EC_a$ were available. The interval midpoint data used in method (2) were calculated simply as the average of the lower and upper limits. Using the same calibration equations, the expected values and their corresponding standard errors for the 413 locations and 17 time instants used for the computations were determined. These two statistics were used to determine the soft pdfs assuming a Gaussian distribution for each of the $413 \times 17$ points. These soft pdfs were used as probabilistic soft data in the method (3). For illustration, the probabilistic soft data for some points are reported in Fig. 2.

The soft data from location 1 at November 1994 is smaller in magnitude than the 3 others, however it is less uncertain as its pdf is less dispersed around the mean value.

Because BME processes the full pdf of the soft data rather than just the mid point of its confidence interval, it uses more information, which will lead to more accurate estimations.

Using the hard and soft data, we compared 3 methods of space-time data interpolation: (1) classical ordinary kriging using only the hard data (HK); (2) kriging using hard data and the interval soft data midpoints (HMIK); and (3) BME using hard data and the full probabilistic soft data.

BME was presented in Sect. 3. Kriging is, in what follows, described briefly (Christakos, 2000). When the

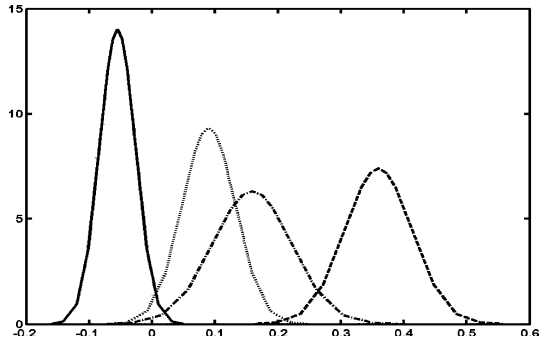


**Fig. 2** Examples of probabilistic soft data (based on the residuals). Full curve: location 1 in November 1994; dashed curve: location 2 in March 1995; dash-dotted curve: location 3 in June 1995; dotted curve: location 4 in September 1995

site-specific knowledge is limited to a set of hard data $\boldsymbol{\chi}_{\text{hard}}$ of the variable $X(\mathbf{p})$ at the space-time points $\mathbf{p}_i$, $(i = 1, \ldots, m)$, the best minimum mean squared error (MMSE) estimator of $X(\mathbf{p}_k)$ at point $\mathbf{p}_k (k \neq i)$, is the conditional mean

$$\chi_{k,MMSE} = E[X(\mathbf{p}_k)|\boldsymbol{\chi}_{\text{hard}}] \tag{4}$$

When the space–time random field is Gaussian, Eq. (4) becomes linear and optimal among all MMSE estimators, and expressed as

$$\chi_{k,\text{MMSE}} = \lambda' \boldsymbol{\chi}_{\text{hard}}$$

where $\lambda$ is a vector of weights associated with the data points and involving the space–time mean and covariance functions. This estimator corresponds to the BME mode estimate when the general knowledge corresponds to space-time mean and covariance functions (in this case the pdf is Gaussian and its mean is equivalent to its mode), and site-specific knowledge is reduced to the hard data. The only difference between HK and HMIK is the availability of hard data; for HMIK, there is more data available since the mid interval of soft data is considered as hard data.

To compare the different approaches, we used the cross-validation procedure for one time instant: the sampling campaign no. 18 which has not been used in all the previous computations (structural analysis and neighborhood). Therefore we estimated soil salinity at each of the 19 locations, for which we have the observed measurements, by deleting in turn the value of each location where the estimate is being calculated.

At the end of this step, 19 pairs of estimated-observed soil salinity values existed. From these we determined three quantitative criteria: the mean error (ME), the mean square error (MSE), and the Pearson correlation coefficient (R). ME should be near zero for a non biased estimator, MSE must be as small as possible for an accurate estimate, and R, which measures the linearity between the estimated and the observed values, should be close to one with a small ME.

Additional to the above statistics, the distribution of the estimation errors was presented graphically for a visual comparison of the 3 methods.

The validation study was followed by predicting soil salinity for time instant no. 12 (September 1998) along with the corresponding estimation variances. The data for this time instant were included in the structural analysis. We used a fine estimation grid ($10 \times 10$ m). The probability that the soil salinity could exceed 4 dS m$^{-1}$ was mapped, 4 dS m$^{-1}$ being the threshold between saline and non-saline soils (Spaargaren 1994; USDA 1996).

All the computations were done using the BMElib toolbox (Christakos et al. 2002) working under Matlab (MathWorks 1999).

## 5 Results and discussion

### 5.1 Exploratory data analysis

Some statistics about the hard data ($EC_{2.5}$) are reported in Table 1. There is a significant temporal variability in the soil salinity as it is indicated by the differences between mean values (ranging from 1.42 dS m$^{-1}$ for November 1994 to 3.30 dS m$^{-1}$ for September 1998). It was also noted the presence of a strong spatial variability. For example in December 2000, there was a big difference between the extreme values (minimum: 0.12, maximum: 7.50 dS m$^{-1}$).

Pearson correlation coefficients between $EC_a$ and $EC_{2.5}$ indicated a strong correlation between these two variables as it ranged between 0.83 and 0.97.

**Table 1** Statistical parameters of salinity data ($EC_{2.5}$ in dS m$^{-1}$)

| $EC_{2.5}$ | $N$ | Mean | Std | Min | Max | $R$ |
|---|---|---|---|---|---|---|
| Calibration data | | | | | | |
| Nov-94 | 13 | 1.42 | 0.34 | 0.86 | 1.82 | 0.85 |
| Mar-95 | 20 | 2.29 | 0.7 | 1.37 | 3.6 | 0.91 |
| Jun-95 | 20 | 2.02 | 1.1 | 0.58 | 4.24 | 0.88 |
| Sep-95 | 20 | 2.01 | 0.99 | 0.83 | 4.56 | 0.94 |
| Dec-95 | 20 | 1.84 | 0.9 | 0.74 | 3.33 | 0.92 |
| Mar-96 | 16 | 2.07 | 0.81 | 0.86 | 3.43 | 0.87 |
| Jun-96 | 20 | 1.83 | 0.9 | 0.63 | 3.49 | 0.87 |
| Mar-97 | 20 | 1.61 | 0.63 | 0.43 | 2.63 | 0.89 |
| Jun-97 | 15 | 1.77 | 0.97 | 0.47 | 3.59 | 0.83 |
| Sep-97 | 20 | 1.63 | 1.24 | 0.15 | 4.4 | 0.94 |
| Dec-97 | 20 | 1.6 | 1.04 | 0.14 | 3.75 | 0.9 |
| Sep-98 | 20 | 3.3 | 2.17 | 0.31 | 7.26 | 0.85 |
| Apr-99 | 20 | 1.84 | 1.74 | 0.17 | 6.58 | 0.93 |
| Jul-99 | 13 | 2.27 | 1.57 | 0.39 | 5.25 | 0.91 |
| Sep-99 | 20 | 2.29 | 1.89 | 0.12 | 6.64 | 0.91 |
| Apr-00 | 18 | 2.11 | 1.63 | 0.1 | 6.42 | 0.94 |
| Dec-00 | 20 | 2.32 | 1.91 | 0.12 | 7.5 | 0.93 |
| Validation data | | | | | | |
| Mar-01 | 19 | 1.80 | 1.21 | 0.16 | 4.58 | 0.97 |

*N* number of observations, *std* standard deviation, *min* minimum, *max* maximum, *R* correlation coefficient between $EC_a$ and $EC_{2.5}$

### 5.2 Structural analysis

The space–time covariance $c_R(r,\tau)$ of the log-transformed mean-trend removed salinity STRF $R(\mathbf{p})$, is shown in Fig. 3a as a function of the spatial lag $r$ (for $\tau = 0$), and in Fig. 3b as a function of the temporal lag $\tau$ (for $r = 0$). The experimental covariance values that were obtained using the BMElib library (Christakos et al. 2002) are shown with circles. The covariance model selected to fit these experimental values is the following non-separable space–time model that is the sum of two nested exponential models

$$c_R(r,\tau) = c_{01}\exp(-3r/as_1)\exp(-3\tau/at_1) + c_{02}\exp(-3r/as_2)\exp(-3\tau/at_2)$$

where $c_{01}$ and $c_{02}$ are the sills of the nested models, $as_1$ and $as_2$ their spatial ranges, and $at_1$ and $at_2$ their temporal ranges. This model is shown in Fig. 3a and b, and the function of both $r$ and $\tau$ in Fig. 4.

The non-separable space–time covariance of Fig. 4 provides a more accurate representation of the correlation structure of salinity in both space and time than that described by a purely spatial covariance model, or a covariance model where time is taken as an additional spatial coordinate. Its parameters are reported in Table 2. The first nested model corresponds to short-scale fluctuations with spatial range of 250 m and a temporal range of 8 months, and it has a sill of 0.27 (dS m$^{-1}$)$^2$ that explains a large part of the total salinity variance (0.34 dS m$^{-1}$)$^2$. The second nested model corresponds to large-scale fluctuations that are greater than

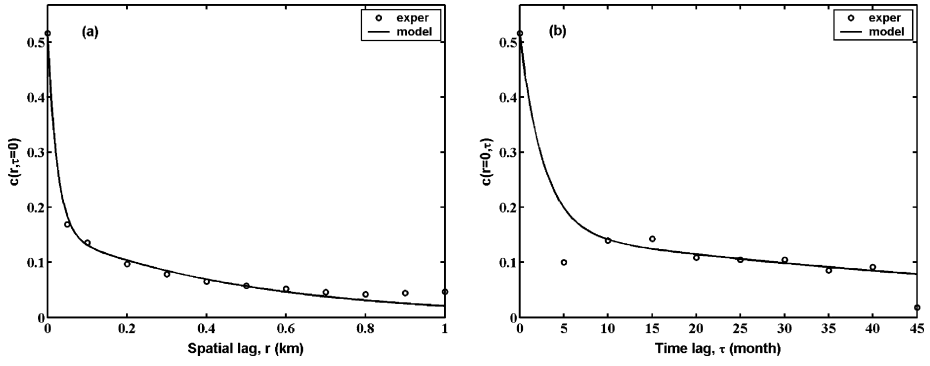


**Fig. 3** **a** Spatial covariance; **b** temporal covariance. *Circles*: experimental data; *curve*: fitted model

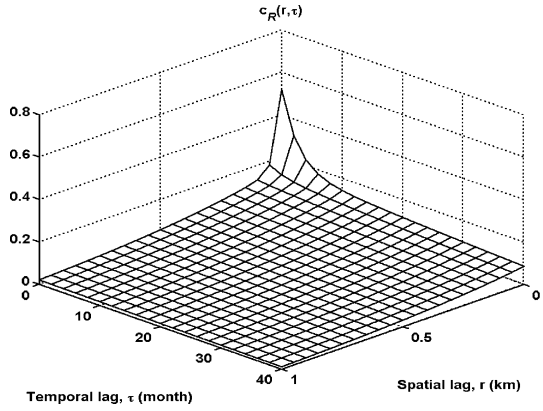


**Fig. 4** Spatio-temporal covariance of the residual data $R(\mathbf{s},t)$

**Table 2** Parameters of the fitted spatio-temporal covariance model

| Component | Spatial range (m) | Temporal range (month) | Sill $(\text{dS m}^{-1})^2$ |
|---|---|---|---|
| First nested model | 250 | 8 | 0.27 |
| Second nested model | 1500 | 200 | 0.07 |

**Table 3** Values for ME, MSE, and $R$ for soil salinity computed from the three approaches

| Criterion/method | HK | HMIK | BME |
|---|---|---|---|
| ME | −0.176 | −0.323 | −0.173 |
| MSE | 0.489 | 0.650 | 0.254 |
| $R$ | 0.87 | 0.92 | 0.94 |

the spatial and temporal size of the study area, with spatial range of 1500 m and a temporal range of 200 months. This nested model has a sill of only 0.07 $(\text{dS m}^{-1})^2$, which explains a smaller part of the total salinity variance.

## 5.3 Cross-validation results

Using the theoretical space-time covariance, we predicted soil salinity at the 20 locations where observed values were available, for the sampled time instant no. 18 (March 2001), using the three different approaches discussed above. Table 3 summarizes the ME, MSE and R for these approaches.

All the ME values are not significantly different from zero. This conclusion was reached using Student $t$ test for correlated samples (McClave and Sincich 2000; p391). However those for HK and BME are smaller than the one corresponding to HMIK. Thus we got indications of less biased estimates with BME and hard kriging compared to kriging using hard and soft data.

The smallest MSE corresponds clearly to BME and the largest to HMIK while HK is intermediate. This means that BME is more accurate than HK which, in turn, is more accurate than HMIK. The results reflect the additional contribution of probabilistic soft data, in the BME approach, to the estimation of the scarce laboratory soil salinity (hard data). HMIK is the worst even if it uses the additional data. This is due to the fact that it used only the mid intervals ignoring the uncertainty related to, and the width of, these intervals. These results showing that BME performs better than kriging are in agreement with the results obtained by D'Or and Bogaert (2001) in the context of soil texture mapping, and by Serre and Christakos (1999) in the context of water surface elevation mapping in aquifers.

The Pearson correlation coefficients between measured and estimated data increased with the incorporation of the soft data. It changed from 0.87 for HK, to 0.92 for HMIK, and finally to 0.94 for BME. This again confirms the usefulness of the soft data and more precisely, the importance in the way that they were processed.

The distributions of the estimation errors are reported in Fig. 5. The curve for BME has a higher peak around zero, meaning that the estimates are more accurate using BME than for the two other methods. This finding corroborates the conclusions reached with MSE. D'Or et al. (2001) arrived at the same results when they compared kriging with hard data and BME with interval soft data applied to soil texture in the spatial domain.

## 5.4 Space–time mapping of soil salinity

We predicted the soil salinity for September 1998 to compare the three methods presented in this study graphically. The soil salinity maps are given in Fig. 6 and their corresponding estimation variances in Fig. 7.

The soil salinity is strongly smoothed in the map obtained by HK. This is the consequence of the limited number of hard data (19, in fact as we interpolated in the space-time domain, we have the 19 locations of some previous time instants as well; however, geographically, we used the same 19 locations). In contrast, the maps of HMIK and BME show much more detail due to the additional data ('hardened' for HMIK or used as soft

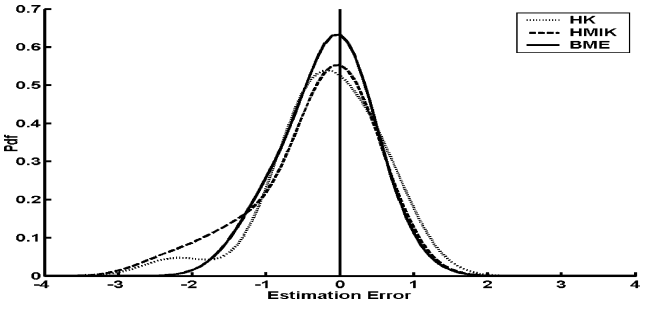


**Fig. 5** Estimation error distribution for HK, HMIK, and BME

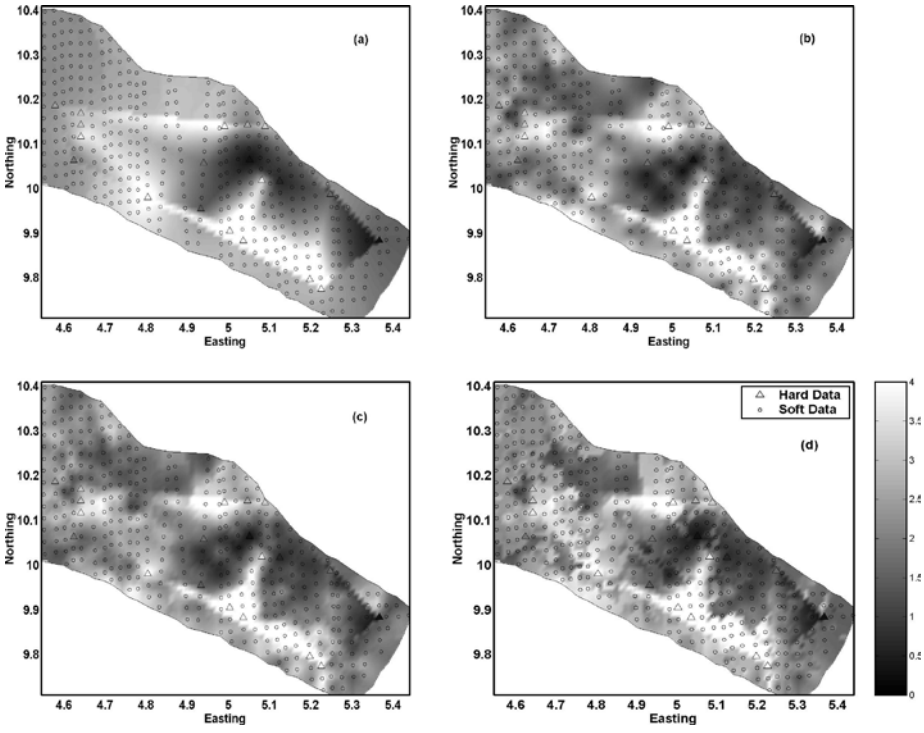


**Fig. 6** Soil salinity ($EC_{2.5}$ in dS $m^{-1}$) estimates for September 1998; **a** *HK*, **b** *HMIK*, **c** BME mean, **d** BME mode. On the maps, triangles and circles indicate points where hard and soft data were available, respectively

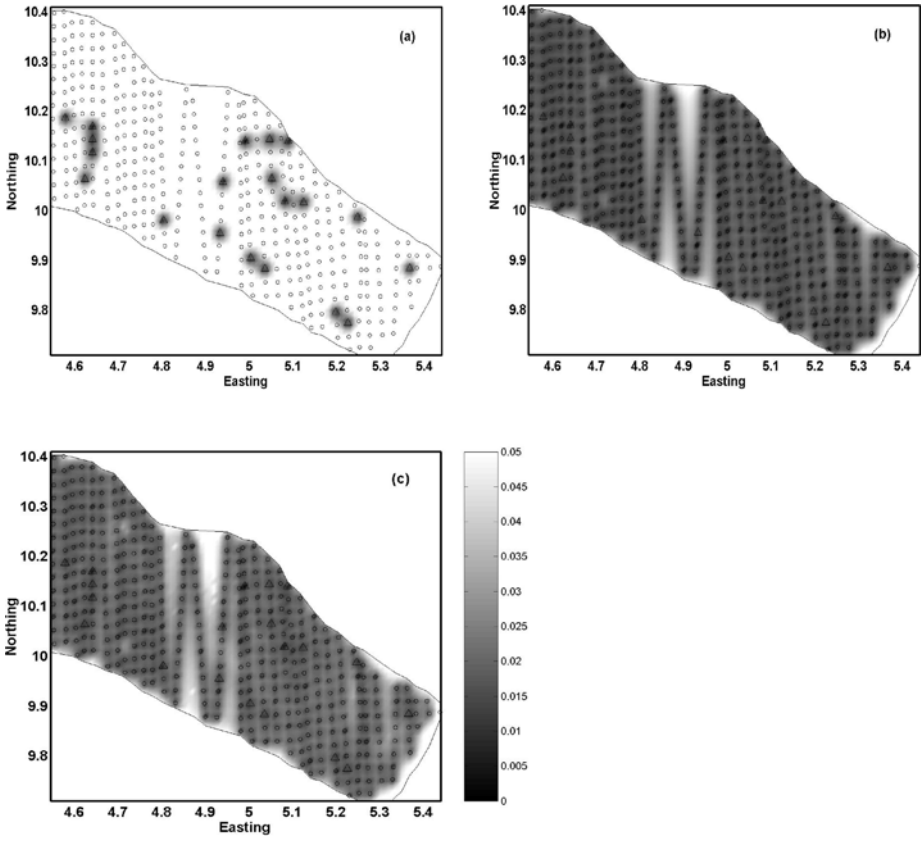


**Fig. 7** Soil salinity estimation variances for September 1998; **a** HK, **b** HMIK, **c** BME

for BME). Furthermore, for the BME, we mapped the conditional mean (minimizing a least square criterion) (Fig. 6c) as well as the mode (the most probable value) of the posterior pdf (Fig. 6d). The map of the BME mode estimate is less smooth than that of the BME mean estimate.

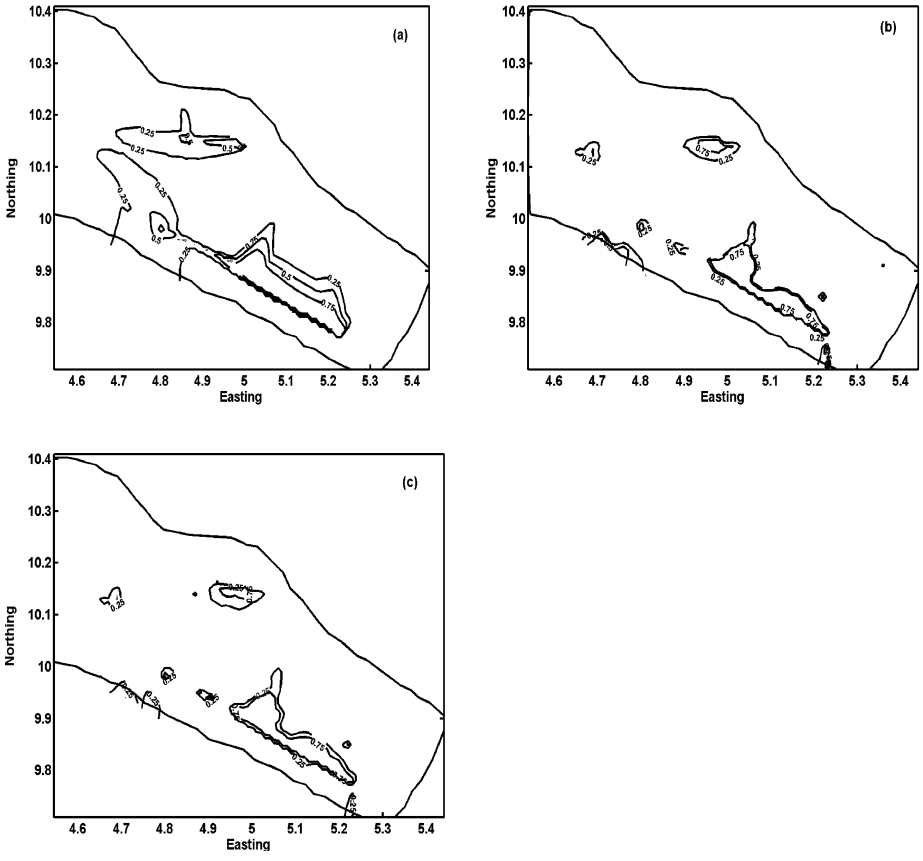


**Fig. 8** Probability that estimated soil salinity exceeds 4 dS $m^{-1}$ for September 1998; **a** HK, **b** HMIK, **c** BME mean

The estimation variance maps (Fig. 7) reflect the difference in data availability. The HK map shows a zero estimation variance at the locations where hard data were present, and a gradual increase in the estimation variances as one goes further away from these hard data locations. In the case of HMIK, the estimation variance is zero for the hard as well as soft data points, resulting in a small estimation variance for the entire study area (except in the centre where the locations are more spaced), which results in an under-prediction of the true uncertainty. On the other hand, BME map has a zero estimation variance at the hard data points only, and a small but non-zero estimation variance at the soft data points, which is a better representation of the true uncertainty in the estimated map.

We also mapped the probability that the estimated soil salinity exceeds 4 dS $m^{-1}$, a critical threshold separating non-saline from saline conditions (Fig. 8). As the estimation standard errors for HK are higher than those for HMIK and BME, and as the estimated soil salinity is smoothed, a larger area of saline soil was delineated. In contrast we obtained, for HMIK and BME, clearly delimited and smaller areas of saline soil.

## 6 Conclusions

The main objective of this work was to compare the prediction performance of BME using uncertain soft data with two kriging techniques. This evaluation was done by means of cross-validation procedure for one time instant which had not been used in the previous steps of the analysis.

This study shows that BME can readily be applied to a soil space–time data set. We conclude that BME estimates were less biased compared to HMIK, and more accurate and better correlated with the observed values than the two kriging approaches. Also the BME estimation error distribution showed a higher peak around zero than the two other techniques, indicating that the probability of obtaining an estimate equal to the observed soil salinity is higher for BME than for the two versions of kriging. In addition, BME allows one to delineate more rigorously saline areas from non-saline zones.

This study demonstrated the possibility of using cheap, dense and easily obtained data (like $EC_a$), to estimate with less bias and more accuracy, a scarce, time consuming and expensive soil property of interest (such as $EC_{2.5}$).

**Acknowledgements** The first author appreciates the financial support made available in the frame of the scientific and technical bilateral cooperation project between the Ministry of Flemish Community (Ghent University, Belgium) and the Republic of Hungary (Research Institute for Soil Science and Agricultural Chemistry of the Hungarian Academy of Sciences: RISSAC-HAS) (project no. 011S1203). He also acknowledges his 2 months stay at RISSAC, Budapest. Collection of data and evaluation were

assisted by Hungarian Scientific Fund Project No. T 037731 and NKFP-4/030/2001. This work was also supported in part by grants from the US National Institute of Environmental Health Sciences (Grant no. P42-ES05948 and P30-ES10126).

## References

Bogaert P (2002) Spatial prediction of categorical variables: the Bayesian maximum entropy. Stochastic Environ Res Risk Assessment 16: 425–448

Bogaert P, D'Or D (2002) Estimating soil properties from thematic soil maps: the Bayesian maximum entropy. Soil Sci Soc Am J 66: 1492–1500

Christakos G (1990) A Bayesian/maximum entropy view to the spatial estimation problem. Math Geol 22: 763–776

Christakos G (1998) Spatiotemporal information systems in soil and environmental sciences. Geoderma 85(2–3): 141–179

Christakos G (2000) Modern spatiotemporal geostatistics. Oxford University Press, New York

Christakos G, Bogaert P, Serre ML (2002) Temporal GIS. Springer-Verlag, New York

Christakos G, Li X (1998) Bayesian maximum entropy analysis and mapping: a farewell to kriging estimators? Math Geol 30: 435–462

Christakos G, Serre ML (2000) BME analysis of spatiotemporal particulate matter distributions in North Carolina. Atmospheric Environ 34: 3393–3406

D'Or D, Bogaert P (2001) Fine scale soil texture estimation using soil maps and profile descriptions. In: Monestiez P et al. (eds) GeoEnv III – Geostatistics for Environmental applications. Kluwer Academic Publishers: Dordrecht, The Netherlands pp 453–462

D'Or D, Bogaert P (2003) Continuous-valued map reconstruction with the Bayesian Maximum Entropy. Geoderma 112: 169–178

D'Or D, Bogaert P, Christakos G (2001) Application of the BME approach to soil texture mapping. Stochastic Environ Res Risk Assessment 15: 87–100

Douaik A, Van Meirvenne M, Tóth T (2003) Spatio-temporal kriging of soil salinity rescaled from bulk soil electrical conductivity. In: Sanchez-Vila X, Carrera J, Gomez-Hernandez J (eds) Quantitative geology and geostatistics, GeoEnv IV: 4th European conference on geostatistics for enironmental applications, Kluwer Academic Publishers: Dordrecht, The Netherlands

Lesch SM, Herrero J, Rhoades JD (1998) Monitoring for temporal changes in soil salinity using electromagnetic induction techniques. Soil Sci Soc Am J 62(1): 232–242

Lesch SM, Strauss DJ, Rhoades JD (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 Resour Res 31: 387–398

MathWorks (1999) Using Matlab, version 5. The MathWorks Inc, Natick : MA

McClave JT, Sincich T (2000) Statistics. 8th edn, Prentice Hall, Upper Saddle River: NJ

Serre ML, Bogaert P, Christakos G (1998) Latest Computational Results in Spatiotemporal Prediction Using the Bayesian Maximum Entropy Method. In: Buccianti A, Nardi G, Potenza R (eds), Proceedings of IAMG '99 - Fifth Annual Conference of the International Association for Mathematical Geology, vol 1, De Frede Editore, Napoli, pp 117–122

Serre ML, Christakos G (1999) Modern geostatistics: computational BME in the light of uncertain physical knowledge – The Equus Beds study. Stochastic Environ Res Risk Assessment 13(1/2): 1–26

Soil and Plant Analysis Council (1992) Handbook on reference methods for soil analysis. Georgia University Station, Athens GA p 202

Spaargaren OC (1994) World Reference Base for soil Resources. ISSS/ISRIC/FAO, Wageningen/Rome

Tóth T, Csillag F, Biehl LL and Micheli E (1991) Characterization of semi-vegetated salt-affected soils by means of field remote sensing. Remote Sens Environ 37: 167–180

Tóth T, Kuti L, Fórizs I, Kabos S, Douaik A (2002) Spatial and temporal aspects of soil salinization in a sodic grassland. In: Cano ÁF, Silla RO, Mermut AR (eds) Sustainable use and management of soils in arid and semiarid regions, (vol I), 277–288

USDA (1996) Keys to soil taxonomy. 7th edn. USDA: NRCS

Van Meirvenne M, De Groote P, Kertesz M, Tóth T, Hofman G (1995) Multivariate geostatistical inventory of sodicity hazard in the Hungarian Puszta. In: Escadafal R, Mulders M, Thiombiano L (eds) Monitoring soils in the environment with remote sensing and GIS, ORSTOM éditions: Paris pp. 293–305