Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Geochemistry 20 (2005) 157–168
Applied Geochemistry
www.elsevier.com/locate/apgeochem

Interpolation and mapping of probabilities for geochemical variables exhibiting spatial intermittency

Eulogio Pardo-Igúzquiza, Mario Chica-Olmo*

Department of Geodynamics/CEAMA, University of Granada, Campus Fuentenueva s/n, 18071 Granada, Spain

Received 13 October 2003; accepted 20 May 2004
Editorial handling by A. Danielsson

Abstract

In monitoring a minor geochemical element in groundwater or soils, a background population of values below the instrumental detection limit is frequently present. When those values are found in the monitoring process, they are assigned to the detection limit which, in some cases, generates a probability mass in the probability density function of the variable at that value (the minimum value that can be detected). Such background values could distort both the estimation of the variable at nonsampled locations and the inference of the spatial structure of variability of the variable. Two important problems are the delineation of areas where the variable is above the detection limit and the estimation of the magnitude of the variables inside those areas. The importance of these issues in geochemical prospecting or in environmental sciences, in general related with contamination and environmental monitoring, is obvious. In this paper the authors describe the two-step procedure of indicator kriging and ordinary kriging and compare it with empirical maximum likelihood kriging. The first approach consists of using a binary indicator variable for estimating the probability of a location being above the detection limit, plus ordinary kriging conditional to the location being above the detection limit. An estimation variance, however, is not available for that estimator. Empirical maximum likelihood kriging, which was designed to deal with skew distributions, can also deal with an atom at the origin of the distribution. The method uses a Bayesian approach to kriging and gives intermittency in the form of a probability map, its estimates providing a realistic assessment of their estimation variance. The pros and cons of each method are discussed and illustrated using a large dataset of As concentration in groundwater. The results of the two methods are compared by cross-validation.

1. Introduction

In the study of the spatial distribution of some geochemical elements (especially the minor ones), the proportion of values below a limit of detection may be large. This often happens if a trace element is naturally very low in the composition of soils or groundwater, but its concentration is high at some places because of contamination or a geochemical anomaly. The anomaly may have a natural origin or be the result of human activities. Thus, monitoring or sampling of an area for that element may show a natural background well below the detection limit and a small number of values higher than the detection limit. This situation generates a particular structure of variability known as intermittency (by analogy with rainfall, where one may distinguish rainy and nonrainy areas). In geochemical monitoring, the values below the detection limit are assigned to the detection limit by the measurement device and a discrete probability mass appears in the probability density function at that particular value. This proportion of constant values, if partially considerable, may have an

* Corresponding author.
E-mail address: mchica@ugr.es (M. Chica-Olmo).

0883-2927/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.apgeochem.2004.05.007

influence masking the spatial structure of variability of the variable of interest and can affect the process of estimation (spatial interpolation).

Kriging is a distribution-free procedure. Its application is always possible using whatever variogram is estimated from the experimental data. Nonetheless, Barancourt et al. (1992) show that using indicator kriging for estimating intermittency, then estimating the variable inside the areas where the variable is above the detection limit, is preferable to ordinary kriging for the delineation of intermittency and for estimating the value of the variable inside the areas where it is estimated to be above the detection limit. First, one estimates intermittency, i.e. which areas have values above the detection limit. Secondly the magnitude of the variable inside those areas above the detection limit is estimated. The first step is performed by indicator kriging, whereas the second step is done by ordinary kriging conditional on the values being above the detection limit. The procedure is reviewed below and its drawbacks are highlighted. The aim of this paper is to present a novel procedure that can deal with intermittency, overcoming the weaknesses of the procedure of Barancourt et al. (1992). The two procedures are compared in a case study.

2. Methodology

In Geostatistics, a spatial variable (e.g., the concentration of a geochemical element) is modelled as a random function (RF) $Z(x)$ which, under the intrinsic hypothesis (Matheron, 1965), has the following mean and variance for the first-order increments:

$$E\{Z(x+h)-Z(x)\}=0 \quad \text{(the mean of the first-order increments is zero)} \tag{1}$$

and the variogram

$$\gamma(h)=\tfrac{1}{2}E\{[Z(x+h)-Z(x)]^2\} \tag{2}$$

(half the variance of the first-order increments is a function of the distance vector h between two spatial locations, and not of the particular locations themselves). This is not a full characterization of the RF, because its multivariate distribution law is undefined, but it suffices for performing spatial linear minimum mean square error interpolation by kriging.

From Eq. (1) the mathematical expectation of the RF is constant over the area of study, i.e. $\forall x \in D \subset R^2$, a region of the real plane, where x represents two cartesian coordinates $x=\{x_1,x_2\}$, easting and northing or longitude and latitude. The variance of $Z(x)$ does not need to be defined, so processes of infinite dispersion may be modelled by the intrinsic hypothesis. Nevertheless, a more restringent case, which suffices in many applications, is to consider second-order stationarity where a covariance function $C(h)$ exits. This requires a finite variance $C(0)$, leading to the well-known relation between covariance and variogram $C(h)=C(0)-\gamma(h)$. For interpolation, one is interested in a linear estimator of $Z(x)$ at unsampled locations x_0:

$$\hat{Z}(x_0)=\sum_{i=1}^{M} w_i^0 Z(x_i), \tag{3}$$

where M is the number of random variable neighbours of $Z(x_0)$ used in the interpolation.

The weights in Eq. (3) are obtained as the solution of the ordinary kriging system which is set up by imposing the condition of the estimator being unbiased and by minimization of the estimation variance (Journel and Huijbregts, 1978; Isaaks and Srisvastava, 1992; Goovaerts, 1997; Chilès and Delfiner, 1999). The estimation variance of the estimator in Eq. (3) can also be obtained through Ordinary kriging. This method is distribution-free and may be applied to any dataset if the variogram function is known or, as is most often the case in practice, can be estimated from the experimental data. In the presence of intermittency, i.e. when areas $\chi \subset D$ exit for which $Z(x)$ is zero or negligible (for example in geochemistry values below a detection limit), a binary RF can be defined to characterize the spatial variability of such intermittency. The binary RF is known as an indicator RF (Barancourt et al., 1992) $I(x)$:

$$I(x)=\begin{cases}1 & \text{if } Z(x)>z_c,\\ 0 & \text{otherwise,}\end{cases} \tag{4}$$

where z_c is, for example, the detection limit.

If $I(x)$ is second-order stationary, then it may be characterized by its variogram (or covariance) $\gamma_I(h)$:

$$\gamma_I(h)=\tfrac{1}{2}E\{[I(x+h)-I(x)]^2\}. \tag{5}$$

The variogram for distance h is related to the probability of transition from a value where the RF is below the detection limit to a value in which the RF is above the detection limit, for that distance h.

Moreover the expectation of the RF $I(x)$ is equal to the probability of $Z(x)$ being over the threshold limit

$$E\{I(x)\}=P\{Z(x)>z_c\}=1-\mathrm{F}(z_c), \tag{6}$$

where $F(x)$ is the cumulative distribution function of $Z(x)$. The probability density function has a finite probability mass at z_c. Thus, a new RF $S(x)$ can be defined which is equal to $Z(x)$ but without the probability mass at z_c. This new RF, defined over the whole domain D, is only known observable where $I(x)=1$.

The procedure described by Barancourt et al. (1992) involves the following steps.

Using the detection limit z_c, the indicator function $I(x)$ is calculated from the original RF $Z(x)$ and the variogram of the indicator $\gamma_I(h)$ is estimated. Next, the RF $S(x)$ is calculated (the subset of $Z(x)$ above the detection limit) and its variogram $\gamma_S(h)$ is estimated. This is followed by ordinary indicator kriging from the indicator data in yielding the map of indicator estimates $I^K(x)$. The next step is to perform the hard thresholding:

$$I^*(x) = \begin{cases} 1 & \text{if } I^K(x) > I_C, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where I_C is a threshold for the indicators such that

$$E\{I^*(x)\} = E\{I^K(x)\} \tag{8}$$

in order to have unbiasedness in the hard thresholding given in (7).

Next, $S(x)$ is estimated by ordinary kriging using the variogram $\gamma_S(h)$ and giving the estimated map $S^K(x)$, and finally the variable $Z(x)$ is estimated as

$$\hat{Z}(x) = I^*(x)S^K(x). \tag{9}$$

Although this procedure is preferable over ordinary kriging of the variable $Z(x)$ directly, it has several drawbacks:

(i) The hard thresholding implies that the procedure is unbiased with respect to the areas that are overestimated and underestimated. Yet this in turn implies bias in the total value of the variable inside those areas (known in mining geostatistics as the amount of metal in a given support). The total amount of the variable in overestimation is always small, because the overestimated cells are cells that, despite having a value equal to the detection limit, are estimated as being over the detection limit. Meanwhile, the cells that are estimated as being under the detection limit, when they are in fact over the detection limit (underestimation), can contain very high values of the variable. Thus, a probabilistic approach to intermittency would be preferable. Indeed, the estimated indicator map may be considered a probability map; as mentioned earlier, however, the indicator map does not take into account how far the variable is from the detection limit.

(ii) It is difficult to estimate the variogram of inner variability on areas over the detection limit, precisely because the limits of such areas are uncertain. This is particularly true when the indicator variogram has a nugget variance and with data scattered in the study area.

(iii) The most important drawback is that Barancourt et al. (1992) do not give a measure of uncertainty of their estimator, such as for example the estimation variance. Modern statistical practices call for not only an estimate but also a measure of the uncertainty of the estimate.

Although some modifications could be made to the previous procedure in order to overcome these difficulties, alternative approaches exist that can deal with intermittency, such as empirical maximum likelihood kriging (EMLK) (Pardo-Igúzquiza and Dowd, 2004). EMLK was designed primarily for dealing with skew distributions of the variable of interest but it can also account for a probability mass at the origin (the origin being the detection limit, not necessarily zero). Geochemical variables with intermittency are often highly skewed in their one-dimensional distribution.

The methodology of EMLK involves the following steps:

(1) *Normal score transformation.* A normal score transformation is used; that is, the experimental data are transformed to the univariate Gaussian domain

$$Y(x) = \varphi(Z(x)), \tag{10}$$

where $Y(x)$ is the normal score variable which is assumed to have a multivariate Gaussian distribution.

The general transformation to normality is defined as

$$y_{(i)} = \varphi(z_{(i)}) = \Phi^{-1}\left(\frac{i - 0.5}{n}\right), \tag{11}$$

where i is the position, or rank, of datum $z_{(i)}$ after sorting the data in increasing order; $(z_1, z_2, \ldots, z_n)$ are the original data; $(z_{(1)}, z_{(2)}, \ldots, z_{(n)})$ are the original data sorted in increasing order and $\Phi^{-1}(\cdot)$ is the inverse of the cumulative standard Gaussian distribution function.

As shown by Verly (1984), when the variable presents an atom at the origin, the normal score transformation of the values equal to the detection limit is done in such a way that their rank is related to the mean value of the data inside a window centred on their location.

(2) *Variogram inference from the normal score data.* The statistical inference of the variogram of the normal scores is generally easier than the original variables because the effect of the high values is damped by the Gaussian transformation.

(3) *Maximum likelihood estimation using a Bayesian approach, i.e. calculation of the posterior distribution function.* In the Gaussian domain, the simple kriging estimator (maximum likelihood estimator) is equal to the conditional mathematical expectation of the random variable at the unknown location, which is the optimal estimator (in mean square error terms). Nevertheless, a Bayesian approach is adopted because it introduces several advantages as seen below.

We are interested in the estimation of the variable of interest at an unsampled location. In the most general case it may be of interest to estimate an average value of

this variable on a given support V (e.g. mean value on a given area or volume):

$$Z_0 = \int_V Z_x \, \mathrm{d}x. \tag{12}$$

The likelihood function of the Gaussian variable $Y_0 = \varphi(Z_0)$ is

$$h\Big(Y_0\Big|(Z_1,\ldots,Z_p),\underline{\mu},\Sigma\Big) = (2\pi)^{-(p+1)/2}|\Sigma|^{-1/2} \times \exp\Big\{-\tfrac{1}{2}\Big[\phi(\underline{Z})-\underline{\mu}\Big]^t \times \Sigma^{-1}\Big[\phi(\underline{Z})-\underline{\mu}\Big]\Big\}, \tag{13}$$

with $\underline{Z} = (Z_0, Z_1, \ldots, Z_p)^t$ the $p+1$ original random variables;
$\varphi(\underline{Z}) = (\varphi(Z_0), \varphi(Z_1), \ldots, \varphi(Z_p))^t = (Y_0, Y_1, \ldots, Y_p)^t = \underline{Y}$ the $p+1$ normal scores and $\underline{\mu} = (\mu_0, \mu_1, \mu_2, \ldots, \mu_p)^t$ the mean of normal scores.

For second-order stationary data, this mean is zero, $\underline{\mu} = \underline{0}$, or, in the universal kriging case, a low degree polynomial (Pardo-Igúzqiza and Dowd, 1998).

Also we have that

$$\Sigma = \begin{bmatrix} \bar{C}_{00} & C_{01} & \cdots & C_{0p} \\ C_{10} & \sigma^2 & \cdots & C_{1p} \\ \vdots & \vdots & \cdots & \vdots \\ C_{p0} & C_{p1} & \cdots & \sigma^2 \end{bmatrix} \tag{14}$$

is the symmetric covariance matrix of the normal score variable with $C_{ij} = \mathrm{cov}(\varphi(Z_i), \varphi(Z_j)) = \mathrm{cov}(Y_i, Y_j) = E\{Y_i Y_j\} - \mu_i \mu_j$, and Z_i and Y_i are short notations for $Z(x_i)$ and $Y(x_i)$, respectively.

Moreover,

$$C_{ij} = C_{ji}$$

$$\bar{C}_{00} = \begin{cases} \sigma^2 & \text{if } Y_0 \text{ has point support,} \\ \bar{C}_V & \text{if } Y_0 \text{ has block support,} \end{cases}$$

where $\bar{C}_V$ is the mean value of the covariance inside support V.

The negative log-likelihood function (NLLF) of Y_0 is then

$$\ell(Y_0) = -\ln h\Big(Y_0\Big|(Z_1,\ldots,Z_p),\underline{\mu},\Sigma\Big) = \frac{p+1}{2}\ln(2\pi) + \frac{1}{2}\ln|\Sigma| + \frac{1}{2}\Big[\varphi(\underline{Z})-\underline{\mu}\Big]^t \times \Sigma^{-1}\Big[\varphi(\underline{Z})-\underline{\mu}\Big]. \tag{15}$$

The value Y^* for which the NLLF reaches its minimum is the maximum likelihood estimator of $\varphi(Z_0)$. The maximum likelihood estimator of Z_0 is $Z_0^* = \varphi^{-1}(Y^*)$. If, however, the original distribution of the random function is skewed, this estimator is conditionally biased. A Bayesian procedure can be used to obtain a conditionally unbiased estimate. From Bayes' theorem we have:

$$\tilde{f}(Y_0) \propto \hat{f}(Y_0) h(Y_0), \tag{16}$$

so the posterior distribution of $\varphi(Z_0) = Y_0$, $\tilde{f}(Y_0)$, is proportional to the product of prior distribution, $\hat{f}(Y_0)$, and the likelihood $h(Y_0)$. In the absence of prior knowledge of the values of the variable $\varphi(Z_0)$, an uniform distribution can be assumed for $\hat{f}(Y_0)$ and the posterior distribution will then have essentially the same form as the likelihood. Thus, the likelihood can be standardized, so that it integrates to 1 and can be interpreted as the *posterior distribution*.

(4) *Point estimates, interval estimates and measures of uncertainty derived from the posterior distribution.* The posterior distribution contains all the information required for estimation: point estimates, interval estimates and measures of uncertainty such as variances and confidence intervals. For example, the mean of the posterior distribution is used as an estimator instead of the mode of the posterior distribution (maximum likelihood or simple kriging estimator) because it has the property of being conditionally unbiased (unbiased in the real domain of the random function $Z(x)$).

In particular, conditional unbiasedness is achieved by using the mean of the posterior distribution, which minimizes the mean squared error $E\{[Z_0 - Z_0^*]^2\}$ with

$$\mathrm{Mean(PD)} = Z_0^* = E\{\phi^{-1}(Y)\} = \int_{-6}^{6} \phi^{-1}(y) f_Y(y)\, \mathrm{d}y, \tag{17}$$

where Mean(PD) is the mean of the posterior distribution estimator and $f_Y(y)$ is the posterior probability density function (i.e. Eq. (15) after normalization in order to integrate the area under the function to 1), and the integration limits of −6 and +6 suffice for practical purposes in lieu of the minus and plus infinite values.

Interval estimates are easily derived from the posterior probability density function as well as measures of uncertainty that can be attached to the estimates. The posterior probability function may be very asymmetrical and not necessarily Gaussian.

The complete formulation of the method is given in Pardo-Igúzquiza and Dowd (2004); and a computer program for normal score transformation with an atom at the origin, as well as for EMLK with the user guide, is in the public domain, available from the authors upon request. Another approach to Bayesian kriging where the uncertainty of the variogram parameters is included in the estimation may be seen in Pardo-Igúzqiza and Dowd (2003).

The procedure is no different from kriging in that one uses a moving neighbour around each location to be

estimated and then selects the experimental data inside that window. With the structure of spatial variability (variogram or covariance), the conditional Gaussian probability density function is fully specified. That posterior distribution contains all the information that is needed and point estimates, interval estimates, measures of uncertainty and the probability of the variable being larger (or smaller) than a given value may be easily calculated. A comparative study of the application of both techniques to a case study is given next.

3. Mapping areas of high arsenic content in Bangladesh

The dataset comprises 3453 values of As content in groundwater samples from Bangladesh (Reports of British Geological Survey on Arsenic Groundwater Contamination in Bangladesh). This dataset is chosen for several reasons:

- It is convenient because it is available on Internet to other researchers, and the proposed methodology and results are reproducible.
- It has an unusually high number of observations, allowing one to split the dataset into two subsets, one for estimation and the other for validation, though each subgroup maintains a sufficient number of data.
- From an environmental point of view, the problem of As contamination in Bangladesh is a serious problem and the authors wish to contribute to its study.

The data, while scattered, cover the country more or less uniformly (Fig. 1). A total of 1837 (53%) of the values of this dataset are below the detection limit of 6 $\mu g\,L^{-1}$.

The variogram of the normal scores of the experimental data – that is, of all values – may be seen in Figs. 2(a)–(d) together with the fitted model, for the 4 main geographical directions. It features nugget variance and two nested spherical structures:

$$\gamma_z(h) = 0.35 + 0.15\,\mathrm{f}(h; 15, 0°, 1) + 0.5\,\mathrm{Sph}(h; 240, 0°, 1.8). \quad (18)$$

The nugget variance is 35% of the total variance and each spherical structure is expressed in the standard form $a\,\mathrm{Sph}(h; b, c, d)$, with a as the sill (i.e. variance of that structure), b the long range, c the anisotropy angle in degrees and d the anisotropy ratio. Range b is the longest range of the anisotropy ellipse, and the

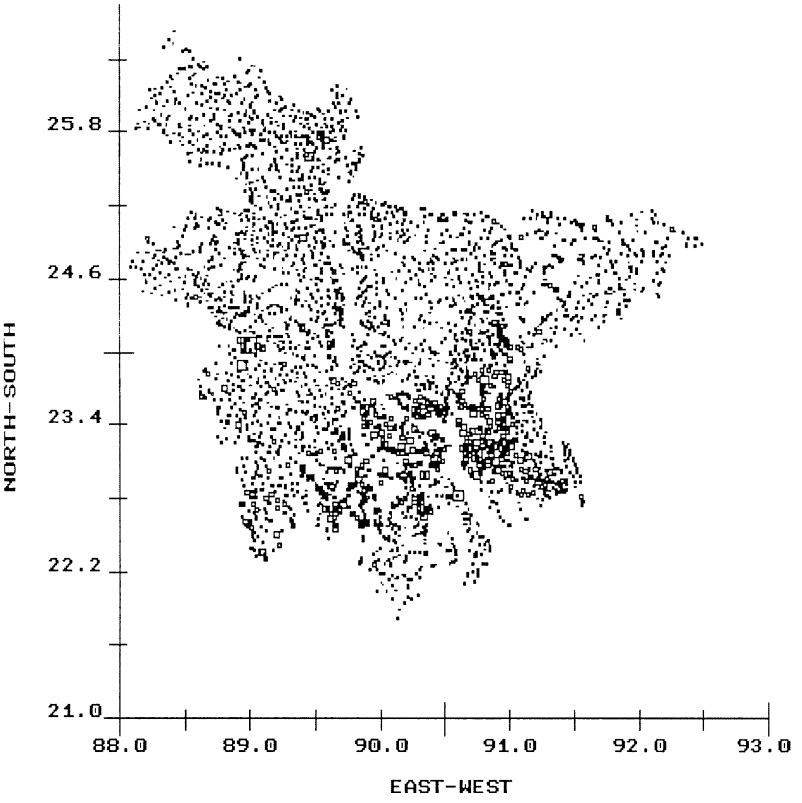

Fig. 1. Spatial location of experimental data.

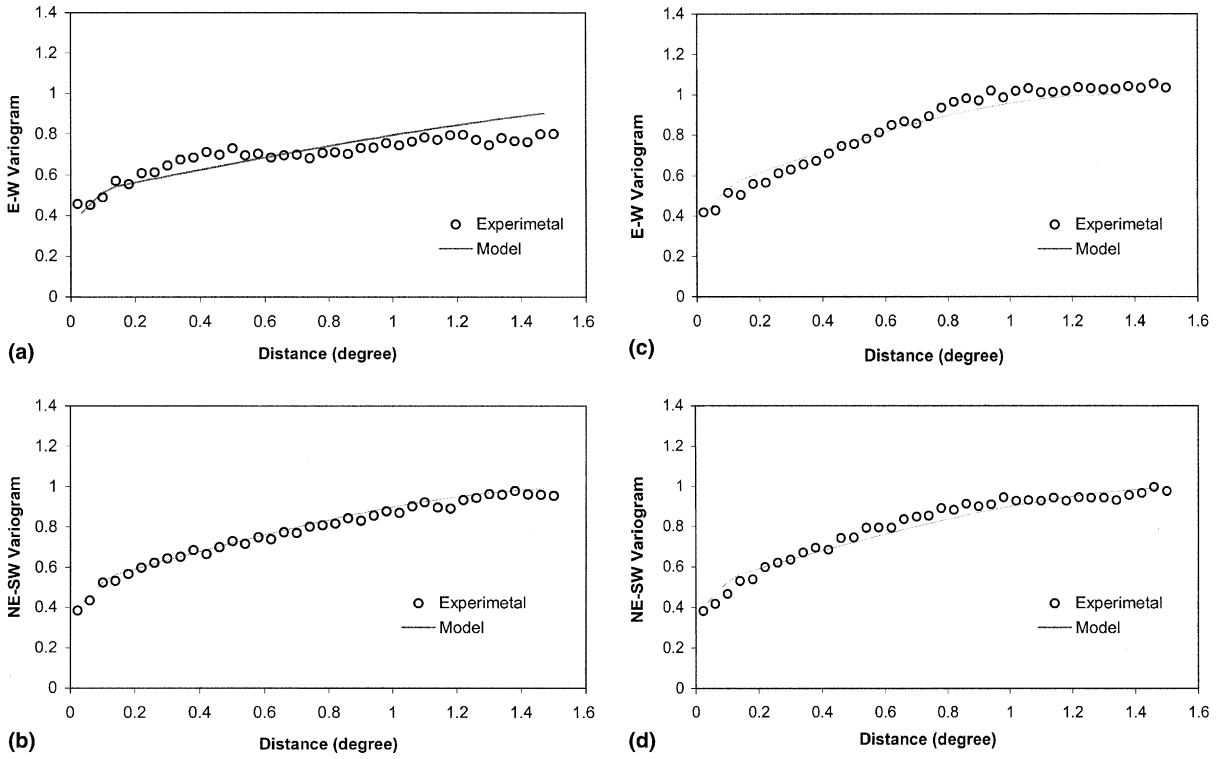

Fig. 2. Experimental variogram and fitted model for variable normal scores of $Z(x)$ for the 4 main geographical directions: (a) E–W, (b) NE–SW, (c) N–S and (d) NW–SE.

anisotropy angle is the angle between the X (or easting) axis and the longest range measured counterclockwise (Isaaks and Srisvastava, 1992).

The variance is given in $\mu g\,L^{-1}$ and the range in km (where each geographical degree has been taken as 100 km). The variogram for the indicator variable is given by (Figs. 3(a)–(d)):

$$\gamma_I(h) = 0.12 + 0.03\,\mathrm{Sph}(h; 16, 110°, 1.5) + 0.0976\,\mathrm{Sph}(h; 140, 110°, 2) \quad (19)$$

And the variogram for the variable $S(x)$ (Fig. 4) is

$$\gamma_S(h) = 13,100 + 4000\,\mathrm{Sph}(h; 15, 0°, 1) + 6000\,\mathrm{Sph}(h; 75, 0°, 1), \quad (20)$$

where the variogram structures are isotropic (the anisotropy ratio equals 1).

The original dataset was divided into two: a learning set and a validation set. The idea was to use the learning set as experimental data used for estimation and the validation set as locations where the variable would be estimated by both methods using the learning set. Because the variable at those locations is known, the error will likewise be known. The learning set consists of 500 data chosen at random from among the 3453 original data, and there are 2953 values for validation.

Fig. 5(a) shows the indicator kriging map. Since it is calculated that $I_C = 0.48$, the estimation of intermittency by thresholding the indicator map using the previous indicator threshold gives the estimation of intermittency of Fig. 5(b). The final estimate of As can be seen in Fig. 5(c). This estimate shows abrupt changes in the borders of the intermittency because of the hard thresholding of the indicator map. The main drawback is that there is no assessment of the reliability of this map; that is, there is no estimation variance map. In EMLK, all the information is obtained from the posterior distribution at each location of the estimation grid. In this sense, the probability map of the value of the variable is determined as being larger than the detection limit, seen in Fig. 5(d). The mean of the posterior distribution estimate is seen in Fig. 5(e), while the estimation variance calculated from the posterior distribution is given in Fig. 5(f). Intermittency is estimated in a probabilistic way using the probability map of Fig. 5(d). The estimate of the variable by EMLK has an associated estimation variance which makes this estimator more appealing.

Although one can certainly use the indicator estimated map as a map of probabilities of the variable being over the detection limit, this map entails two difficulties. First, it does not take into account the variance of the estimates and secondly it does not take into

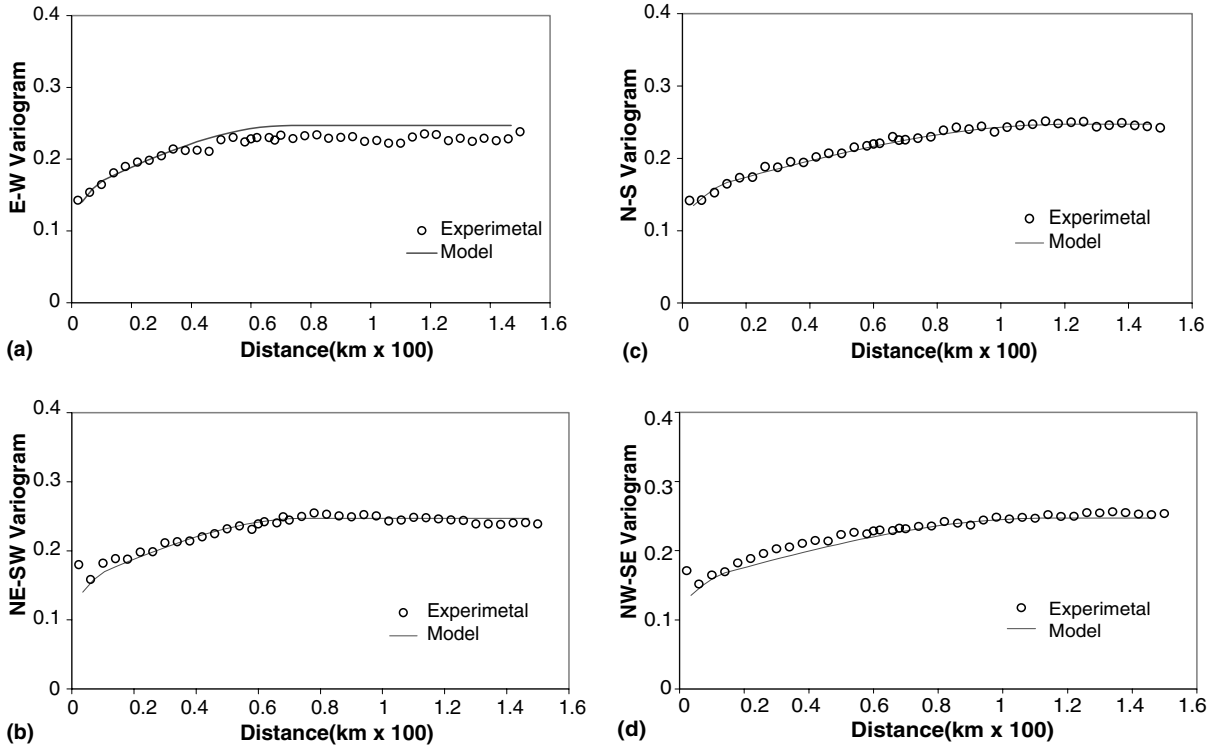

Fig. 3. Experimental variogram and fitted model for variable $I(x)$ for the 4 main geographical directions: (a) E–W, (b) NE–SW, (c) N–S and (d) NW–SE.

account the magnitude of the variable. For instance, a value above the threshold of 6 $\mu g\,L^{-1}$ of As will give a 1 in the indicator no mater if its value is 7 or 7000 $\mu g\,L^{-1}$, because both are larger than the detection limit. Clearly, however, it would be riskier to overlook the 7000 $\mu g\,L^{-1}$ datum if estimated as below the threshold than to make the same error with the 7 $\mu g\,L^{-1}$ datum.

The probability map given in Fig. 5(d) by EMLK, estimated from the posterior distribution, accounts for both the magnitude of the neighbourhoods and the

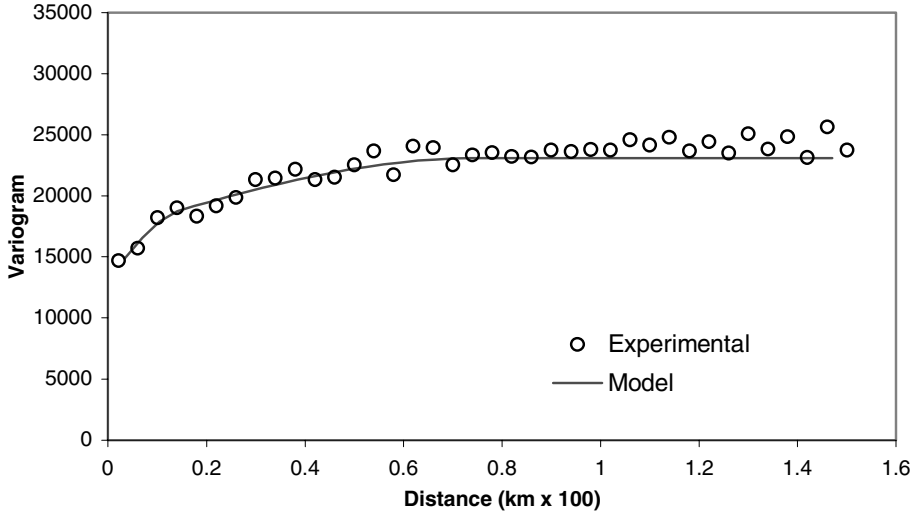

Fig. 4. Experimental variogram and fitted model for variable $S(x)$.

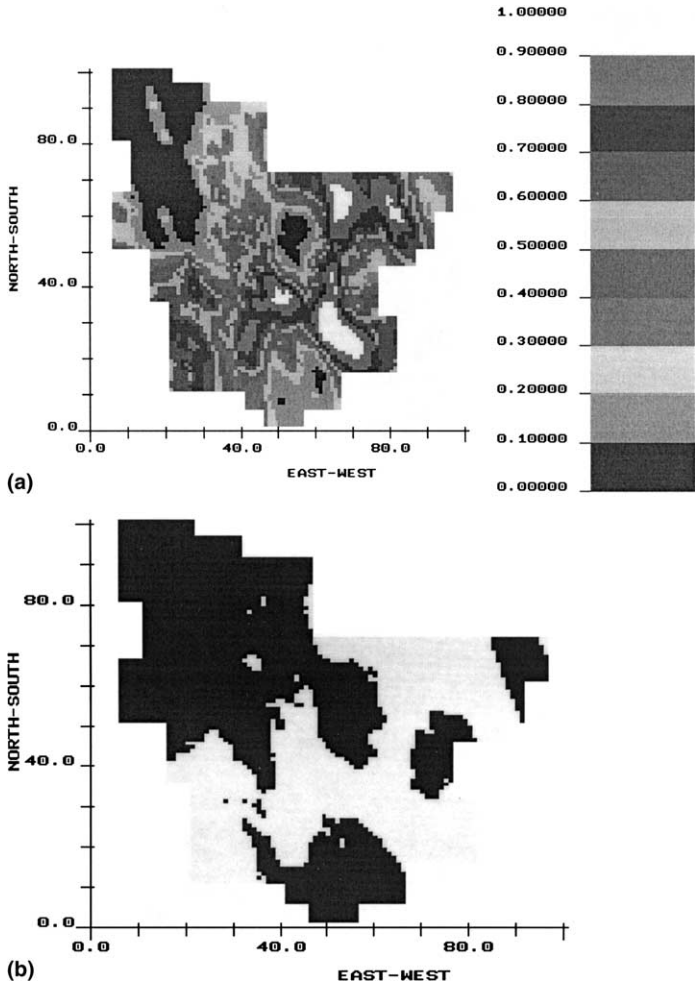

Fig. 5. Indicator map estimated by indicator kriging (a). Intermittency map calculated by applying the indicator threshold of 0.48 to the indicator map shown in (a). The light colour area is over the detection limit (b). Estimated map (logarithmic scale) by the method of Barancourt et al. (1992) (c). Probability map of the variable being larger than the detection limit obtained by EMLK (d). Estimated map by means of the posterior distribution in EMLK (e). Estimation variance from the posterior distribution in EMLK (f).

uncertainty of the estimation. From this map one can go on to intermittency maps taking into account the risk that one is willing to accept.

Because of the hard thresholding of the indicator map, there are areas above the threshold that are considered to be below the threshold and vice-versa. Although the procedure is unbiased with respect to the number of locations or surfaces in case of a nonpoint support (i.e. number of overestimated locations more or less equal to the number of underestimated locations), the method will be biased with respect to the total amount of the variable in those locations. The strong gradients seen in the estimated map of Fig. 5(c) are most likely due to misclassifications in the previous step of hard thresholding of the indicator map. The map given by EMLK (Fig. 5(e)) could be used with the probability map of Fig. 5(d) in order to produce more realistic images. Finally, the EMLK produces a map of estimation variances (Fig. 5(f)) that is lacking when the alternative methodology is used.

The validation of the previous maps was to be carried out in the 2953 validation locations after estimation by both methods using the 500 learning locations. The results are shown in Figs. 6(a) and (b). The coefficient of correlation is 0.53 for the method of Barancourt et al. (1992) and 0.55 for EMLK. The mean error for the first method is 1.734 and the mean squared error 10312, while the same statistics for EMLK are −8.995 and 9547, respectively. Furthermore, taking into account intermittency by Barancourt et al. (1992) in the validation

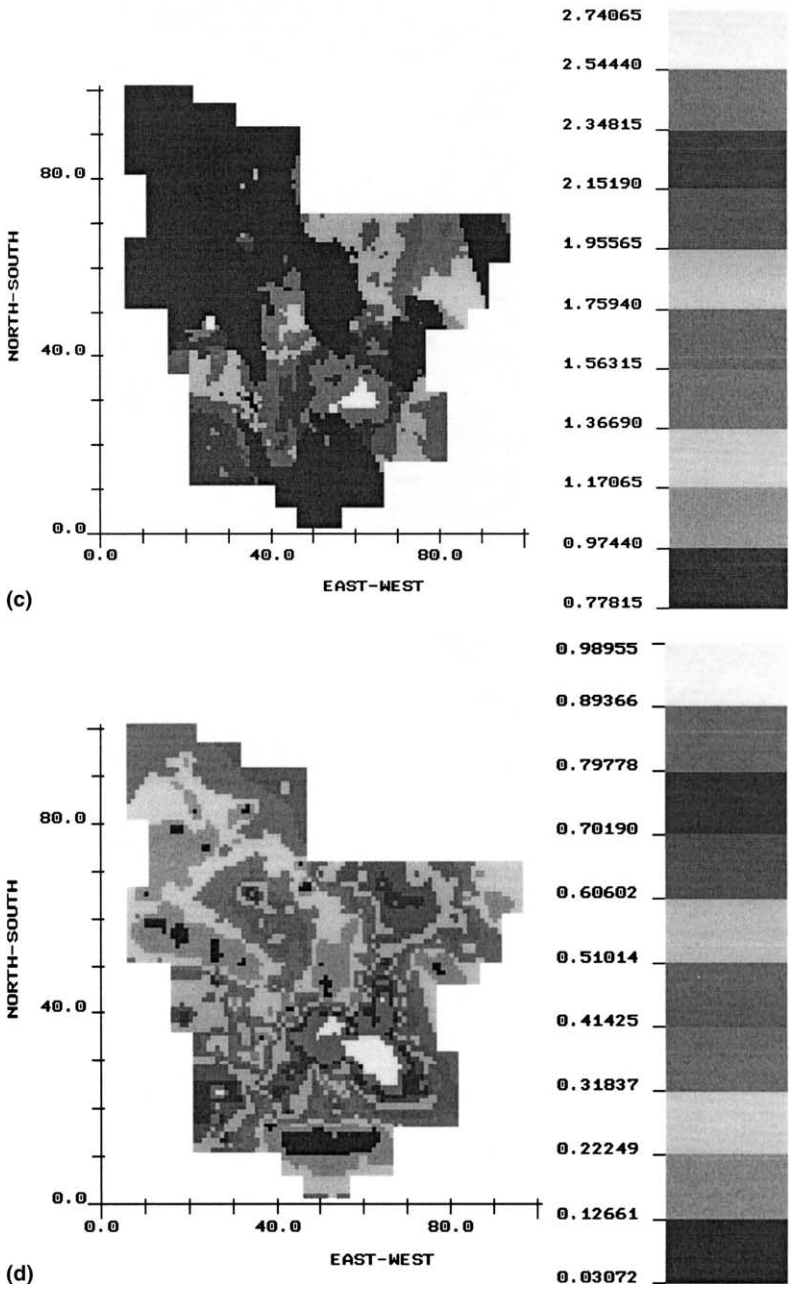

Fig. 5 (*continued*)

set, the number of underestimated locations is 421, while the number of overestimated locations is 349. The method itself is unbiased; but when taking into account the amount of As in the underestimated locations we arrive at a sum of 27,844, while the overestimated figure is 2094, clearly biased. For the underestimated locations, the Barancourt et al. (1992) method gives 2526 while the EMLK assigns a total of 11,277, a much more realistic figure. Finally, as may be seen in Fig. 6(a) the method of Barancourt et al. (1992) gives rise to two distinctive clusters in the validation results. The cluster labelled as (a) Fig. 6(a) is made up of the data which are below the detection limit but have been estimated with values over the detection limit (i.e. overestimation). The cluster labelled as (b) in Fig. 6(a) comprises the data that have been estimated as being below the detection limit while

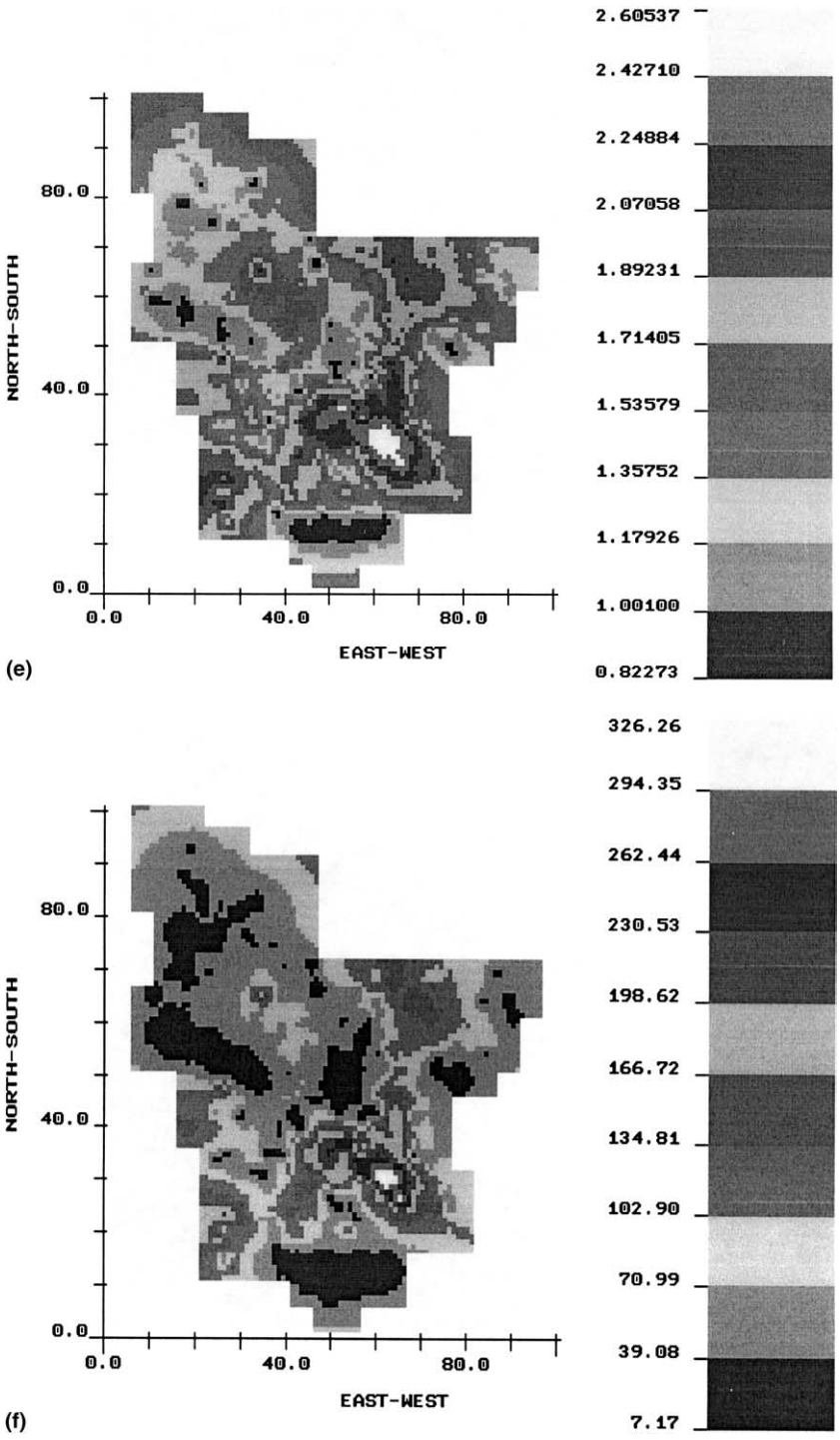

Fig. 5 (*continued*)

in reality they have values above the detection limit (i.e. underestimation). The results from EMLK seen in Fig. 6(b) show no type (b) cluster at all, while in the type (a) cluster the interval of values is closer to the true values than in Fig. 6(a). All these results of the case study underline the advantage of using EMLK.

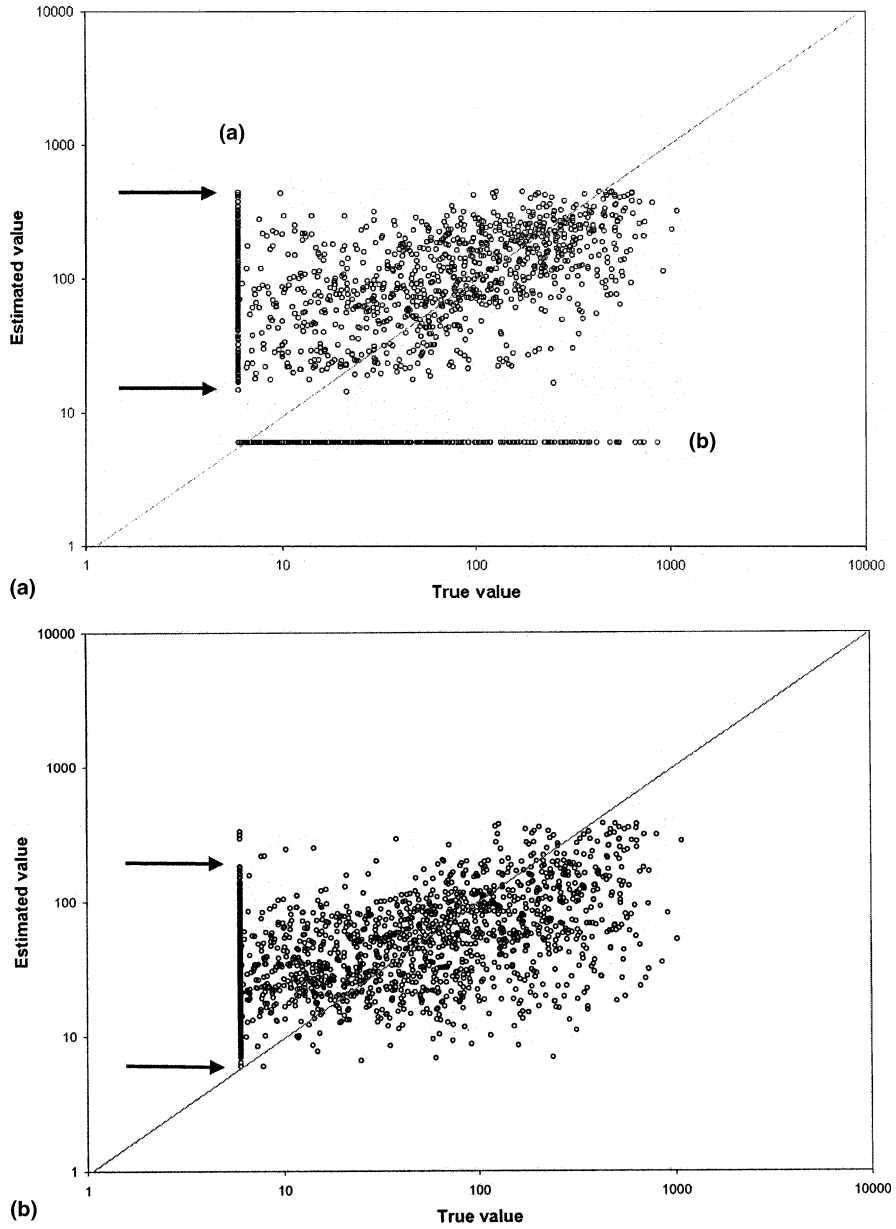

Fig. 6. Scatter plot of true values versus estimated values for the 2953 validation data using the methods of: (a) Barancourt et al. (1992) and (b) EMLK.

4. Discussion and conclusions

It is usual for geochemical variables of minor elements both to exhibit intermittency and to have a skewed distribution. Barancourt et al. (1992) showed that their method was superior than ordinary kriging for the purposes of identification of intermittency and estimating the variable inside the areas above the detection limit. Yet because the present authors have encountered several difficulties using the method of the above authors, results were checked against those of another method that can deal with intermittency and with skewed distributions: EMLK (Pardo-Igúzquiza and Dowd, 2004). The As dataset of Bangladesh was divided into two sets: 500 data for estimation and 2953 data for validation. First of all, EMLK gives slightly better validation statistics, which could be significant in view of the high number of data for validation. Also the cluster of values underestimated as being below the detection limit (labelled as (b) in Fig. 6(a)) is absent in the EMLK

method. The authors believe that EMLK is advantageous because of the rationale of the two methods: the hard thresholding of the indicator map implies that no uncertainty is attached to it. It was found preferable to use the probability map of EMLK, which gives the probability of the variable as larger than the detection limit, taking into account its location, the magnitude of the neighbouring data and the variability structure of the spatial variable. With respect to the map of estimates, the main advantage of EMLK is that it assigns an estimation variance to each estimate, whereas the method of Barancourt et al. (1992) does not provide for any assessment of the reliability of estimates. Finally, estimating the covariance or variogram inside areas over the detection limit will entail great difficulties in delineating the areas above the detection limit, especially when data are scattered and there is a nugget effect in the variogram. With EMLK, however, the normal score transformation facilitates the analysis of the structure of variability, as the effect of high values is minimized by the transformation.

Acknowledgements

The first author is a Ramon and Cajal Grant holder from the Ministry of Science and Technology of Spain (MCyT). We are grateful for the financial support given by the Spanish MCyT (Project BTE2002-00152) and Junta de Andalucía (Group RNM122). We also thank the British Geological Survey for making the data available at its home web page. We also thank the reviewers for their very constructive criticism.

References

Barancourt, C., Creutin, J.D., Rivoirard, J., 1992. A method for delineating and estimating rainfall fields. Water Resour. Res. 28, 1133–1144.

Chilès, J.-P., Delfiner, A., 1999. Geostatistics: Modelling Spatial Uncertainty. Wiley-Interscience, New York.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press, New York.

Isaaks, E.H., Srisvastava, R.M., 1992. Applied Geostatistics. Oxford University Press, Oxford.

Journel, A.G., Huijbregts, Ch., 1978. Mining Geostatistics. Academic Press, New York.

Matheron, G., 1965. Les variables regionnalisées et leut estimation Msasson, Paris.

Pardo-Igúzqiza, E., Dowd, P.A., 1998. The second-order stationary universal kriging model revisited. Math. Geol. 30, 347–378.

Pardo-Igúzqiza, E., Dowd, P.A., 2003. Assessment of the uncertainty of spatial covariance parameters of soil properties and its use in applications. Soil Sci. 168, 769–782.

Pardo-Igúzquiza, E., Dowd, P.A., 2004. Empirical maximum likelihood kriging: the general case. Math. Geol., in press.

Verly, G., 1984. The block distribution given a point multivariate normal distribution. In: Verly, G., David, M., Journel, A.G., Marechal, A. (Eds.), Geostatistics for Natural Resources Characterization, Part 1. Reidel, Dordrecht, Germany, pp. 495–515.