ELSEVIER

Geoderma 103 (2001) 3–26

GEODERMA

www.elsevier.com/locate/geoderma

Geostatistical modelling of uncertainty in soil science

P. Goovaerts *

Department of Civil and Environmental Engineering, The University of Michigan, EWRE Building, Room 117, Ann Arbor, MI 48109-2125, USA

Received 12 December 1999; received in revised form 22 June 2000; accepted 9 February 2001

Abstract

This paper addresses the issue of modelling the uncertainty about the value of continuous soil attributes, at any particular unsampled location (local uncertainty) as well as jointly over several locations (multiple-point or spatial uncertainty). Two approaches are presented: kriging-based and simulation-based techniques that can be implemented within a parametric (e.g. multi-Gaussian) or non-parametric (indicator) frameworks. As expected in theory and illustrated by case studies, the two approaches yield similar models of local uncertainty, yet the simulation-based approach has several advantages over kriging: (1) it provides a model of spatial uncertainty, e.g. the probability that a given threshold is exceeded jointly at several locations can be readily computed, (2) conditional cumulative distribution function (ccdf) for supports larger than the measurement support (e.g. remediation units or flow simulator cells) can be numerically approximated by the cumulative distribution of block simulated values that are obtained by averaging values simulated within the block, and (3) the set of realizations allows one to study the propagation of uncertainty through global GIS operations or complex transfer functions, such as flow simulators that consider many locations simultaneously rather than one at a time.

The other issue is the evaluation of the quality or "goodness" of uncertainty models. Two new criteria (exceedence probability plot and narrowness of probability intervals that include the true values) are presented to assess the accuracy and precision of local uncertainty models using cross-validation. According to the second criterion, multi-Gaussian kriging performs better than indicator kriging for the hydraulic conductivity (HC) data set. However, looking at the distribution of flow simulator responses, sequential indicator simulation (sis) yields better results than sequential Gaussian simulation (sGs) that does not allow for significant correlation of extreme values (destructuration effect).

Keywords: Geostatistics; Uncertainty; Indicator kriging; Stochastic simulation; Accuracy

* Tel.: +1-734-936-0141; fax: +1-734-763-2275.
E-mail address: goovaert@engin.umich.edu (P. Goovaerts).

0016-7061/01/$ - see front matter
PII: S0016-7061(01)00067-2

1. Introduction

The last decade has witnessed an increasing awareness of the importance of assessing the uncertainty about the value of soil properties at unsampled locations, and of the need to incorporate this assessment in subsequent decision-making processes, such as delineation of contaminated areas or identification of zones that are suitable for crop growth. Until the mid-1990s, uncertainty assessment has been essentially performed using non-linear kriging approaches (e.g. disjunctive or indicator kriging) that aim at evaluating the probability for the target attribute to be no greater than any threshold value at a specific unmonitored location (Webster and Oliver, 1989; Smith et al., 1993; Goovaerts and Journel, 1995). Nowadays, stochastic simulation algorithms are becoming more common—or maybe just more "fashionable"—for uncertainty modelling in soil science (Vanderborght et al., 1997; Pachepsky and Acock, 1998; McKenna, 1998; Bloom and Kentwell, 1999; Broos et al., 1999). The basic idea is to generate a set of equiprobable representations (realizations) of the spatial distribution of soil attribute values and to use differences among simulated maps as a measure of uncertainty. Again, there is a vast toolbox of algorithms available: sequential indicator or Gaussian simulation, turning band method, p-field simulation, simulated annealing, to name a few.

Several authors have attempted to compare the performance of the many flavors of non-linear kriging and stochastic simulation algorithms. To my knowledge, the most complete comparison of non-linear kriging algorithms has been conducted by Papritz and Dubois (1999). They found that all the non-linear methods modelled the local conditional distributions of probability equally well and no method was consistently superior to others. Comparison of stochastic simulation algorithms has received more attention in the literature (Deutsch, 1994; Gotway and Rutherford, 1994; Srivastava, 1996; Goovaerts, 1999a), but the conclusion is essentially the same: there is no simulation algorithm that is best for all cases, but rather a toolbox of alternative algorithms from which to choose or to build the algorithm best suited for the problem at hand (Gómez-Hernández, 1997a). Although most of the aforementioned papers have been published in congress proceedings that are not readily accessible to soil scientists, the aim of this paper is not to compare or even review the different geostatistical algorithms available for uncertainty modelling. Readers should refer to recent books on the subject, such as Goovaerts (1997), Deutsch and Journel (1998) or Chilès and Delfiner (1999). The focus is here on the first-order decision of whether a kriging-based or a simulation-based approach should be adopted.

The choice of an approach for uncertainty modelling should be guided by the answer to the questions discussed below.

(1) *What type of uncertainty model is being sought*? In some applications, the objective is to estimate the uncertainty prevailing at a particular location (local

uncertainty), while other problems like flow modelling requires the uncertainty about soil attribute values to be assessed at many locations simultaneously (multiple-point or spatial uncertainty).

(2) *What is the support for the uncertainty assessment*? Decision-making is frequently performed over supports (remediation units or agricultural fields) that are much larger than the measurement supports, hence some aggregation or up-scaling must be conducted.

(3) *Why do we model uncertainty*? Assessing uncertainty about soil attribute values is rarely a goal per se, rather it is often a preliminary step to evaluate the risk involved in any decision-making process or to investigate how prediction errors propagate through complex functions, such as crop models, cost curves for wrong remediation decision or flow simulators.

Each issue will be reviewed in this paper and illustrated using two case studies, see Fig. 1. The first data set, kindly provided by M.J.-P. Dubois of the Swiss Federal Institute of Technology, consists of 259 topsoil Cd concentrations

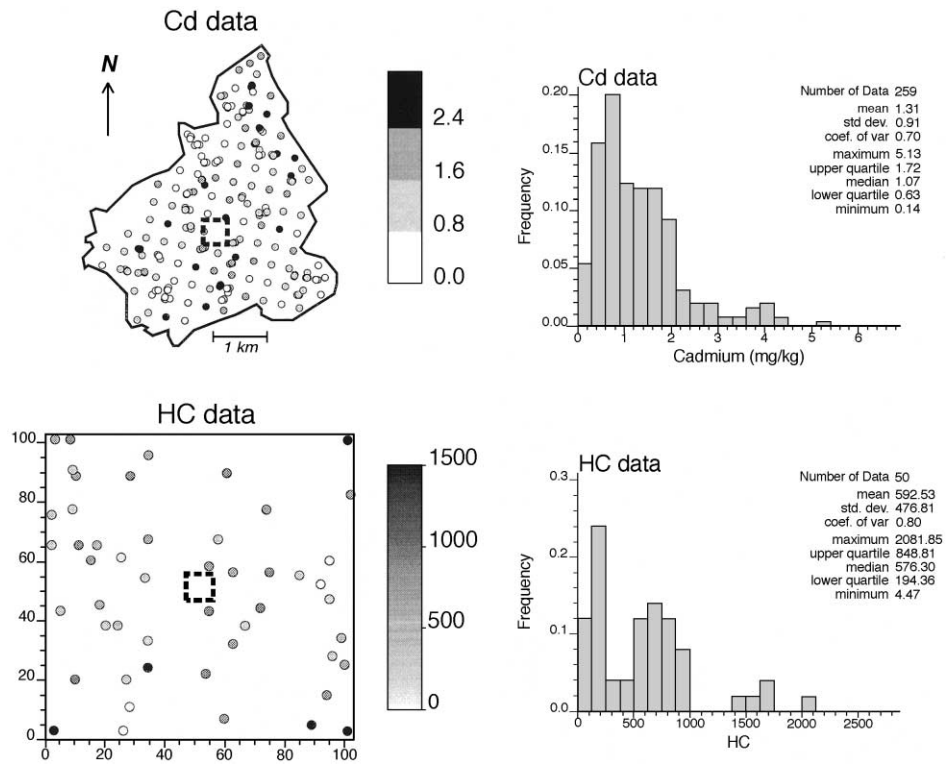


Fig. 1. Two data sets: 259 topsoil Cd concentrations measured in a 14.5-km^2 area (units = mg/kg), and 50 hydraulic conductivity (HC) data originating from the scanning of a block of sandstone. The dashed box depicts a 0.25-km^2 remediation unit for cadmium and flow simulator cell for HC.

measured in a 14.5-km^2 area in the Swiss Jura (Atteia et al., 1994; Webster et al., 1994). The objective is to assess the uncertainty about Cd concentration at unsampled locations and to assess its impact on remediation decisions. The second data are interpreted as 50 hydraulic conductivity (HC) data measured in a cross-section of a soil profile. This data set is not synthetic, but results from the scanning of a block of sandstone (Srinivasan and Journel, 1998) and the 50 values have been randomly drawn from the 102×102 original image; reference true values will be used to assess the quality of predictions. Because the initial measurements (porosity values) have been rescaled, the units are meaningless and are not included in this paper. The objective here is to assess the uncertainty about HC values at unsampled locations and its impact on the prediction of flow or solute transport.

Once the choice of a kriging-based or a simulation-based approach has been matured, there remains the problem of picking the algorithm that would yield the "best" model of uncertainty. Since there is no "best" algorithm for all situations, different algorithms should be applied to the data and their respective performances should be compared. Such a sensitivity analysis is frequently implemented for spatial interpolation: kriging estimates are compared with observations that have been either temporarily removed one at a time (leave-one-out or cross-validation) or set aside for the whole analysis (jacKnife), and statistics such as mean square or mean absolute errors of prediction are computed (e.g. see Davis, 1987; Wackernagel, 1998, p. 91). Assessing the goodness of uncertainty models has received far less attention. For example, Goovaerts and Journel (1995) compared the performances of various indicator kriging algorithms for predicting probabilities of deficiencies at test locations (jacKnife approach) by computing the ratio of probabilities at locations that are actually sufficient or deficient: the larger the ratio the better the discrimination between the two types of location. More sophisticated measures of the goodness of local uncertainty models have been recently proposed by Deutsch (1997) or Papritz and Dubois (1999). These measures are briefly reviewed in this paper, and two new criteria (exceedence probability plot and narrowness of probability intervals that include the true values) are proposed to assess the accuracy and precision of uncertainty models.

2. Modelling the local or the spatial uncertainty?

2.1. Local uncertainty

Consider the simplest situation where the uncertainty prevailing at a single location $\mathbf{u}$ is to be modelled. The probabilistic way to model the uncertainty about a continuous attribute z at $\mathbf{u}$ consists of: (1) viewing the unknown value

$z(\mathbf{u})$ as the realization of a random variable $Z(\mathbf{u})$, and (2) deriving the conditional cumulative distribution function (ccdf) of $Z(\mathbf{u})$:

$$F(\mathbf{u}; z \mid (n)) = \text{Prob}\{Z(\mathbf{u}) \leq z \mid (n)\}, \quad (1)$$

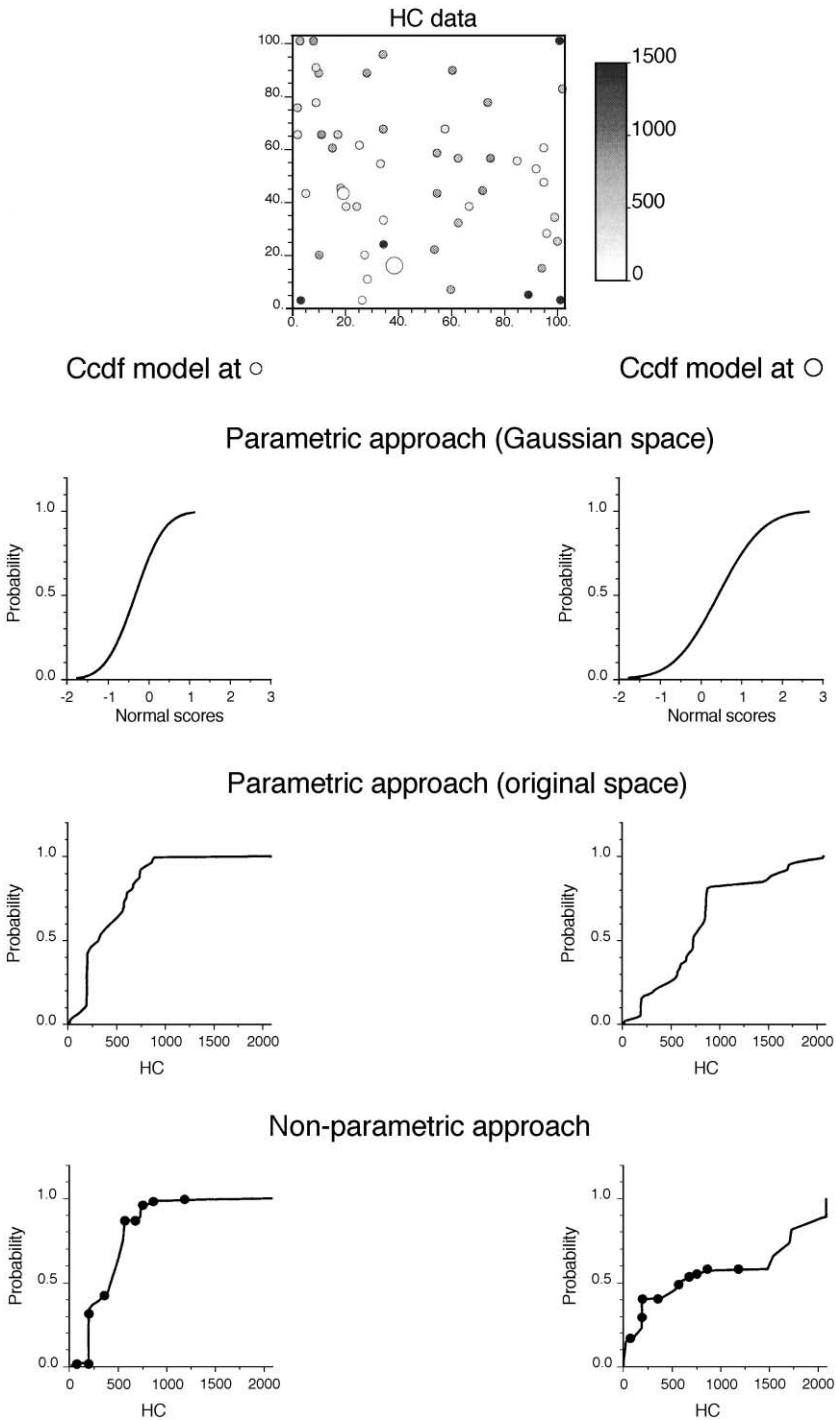
where the notation " $\mid (n)$" expresses conditioning to the local information, say, n neighboring data $z(\mathbf{u}_\alpha)$ The ccdf fully models the uncertainty at $\mathbf{u}$ since it gives the probability that the unknown is no greater than any given threshold z.

Ccdfs can be modelled using either parametric (a model is assumed for the multi-variate distribution of the random function $Z(\mathbf{u})$, commonly the multi-Gaussian distribution is adopted) or non-parametric (indicator) approaches (e.g. see Goovaerts, 1997, Chap. 7). Their most salient features are discussed below.

- Under the multiGaussian model, the ccdf at any location $\mathbf{u}$ is Gaussian and fully characterized by its mean and variance, which correspond to the simple kriging estimate and variance at $\mathbf{u}$. The approach typically requires a prior normal score transform of data to ensure that at least the univariate distribution (histogram) is normal. The normal score ccdf then undergoes a back-transform to yield the ccdf of the original variable.
- The indicator approach also begins with a non-linear transform of the data. In this case, each observation is transformed into a set of K indicator values corresponding to K threshold values (e.g. 9 deciles of the sample histogram). K ccdf values at $\mathbf{u}$ are estimated by a kriging of these indicator data, and the complete function is obtained by interpolation/extrapolation of the estimated probabilities, see Goovaerts (1999b) for more details on implementation of the indicator approach.

Both approaches are used to derive the ccdf of HC values at two locations $\mathbf{u}_1$ (small open circle) and $\mathbf{u}_2$ (large open circle) depicted at the top of Fig. 2. For multi-Gaussian kriging, the mean and the variance of the ccdf in the normal space (top graph) was identified with the simple kriging estimate and variance, then the horizontal axis was rescaled (middle graph) through a normal score back-transform. Bottom graphs show the model fitted to the nine ccdf values estimated at $\mathbf{u}_1$ and $\mathbf{u}_2$ using ordinary indicator kriging.

Comparison of ccdfs in Fig. 2 shows that both parametric and non-parametric approaches yield similar models at $\mathbf{u}_1$ (small circle). The steep ccdf indicates that the unknown HC value is more likely smaller than 1000, and the proximity to observations explains the smaller uncertainty prevailing at $\mathbf{u}_1$ relatively to $\mathbf{u}_2$. Differences between uncertainty models are more pronounced at $\mathbf{u}_2$. The spread of the parametric ccdf is mainly controlled by the kriging variance, which is larger than at $\mathbf{u}_1$ because of the greater distance to neighboring observations. The large spread of the non-parametric ccdf is caused by the wide range of surrounding HC values, which leads to a bimodal probability distribution with a high probability to be either below 250 or above 1500. Note that the relative goodness of the parametric and non-parametric ccdfs can be assessed and


HC data
1500
1000
500
0
Ccdf model at ○
Ccdf model at ○
Parametric approach (Gaussian space)
Probability
Normal scores
Parametric approach (original space)
HC
Non-parametric approach

compared only if the actual HC value at $\mathbf{u}_2$ is known, and this issue will be discussed later in Section 5.

2.2. Spatial uncertainty

The two ccdfs of Fig. 2 provide information on the probability of any given threshold to be exceeded at $\mathbf{u}_1$ and $\mathbf{u}_2$ taken *separately*. For applications like the prediction of the time for fertilizers or contaminants to percolate across a soil profile, it is critical to assess the probability that a series of locations are all above a high threshold or are all below a small HC value: such strings of high or low values may represent flow paths or barriers that control most of the flow and transport. For example, modelling the uncertainty prevailing *jointly* at $\mathbf{u}_1$ and $\mathbf{u}_2$ would require the inference of the two-point ccdf:

$$F(\mathbf{u}_1, \mathbf{u}_2; z_1, z_2 \mid (n)) = \text{Prob}\{Z(\mathbf{u}_1) \leq z_1, Z(\mathbf{u}_2) \leq z_2 \mid (n)\}. \quad (2)$$

Inference of the joint probability (2) is straightforward if the two locations are uncorrelated: $F(\mathbf{u}_1, \mathbf{u}_2; z_1, z_2 \mid (n))$ is simply the product of the two one-point probabilities $F(\mathbf{u}_1; z_1 \mid (n))$ and $F(\mathbf{u}_2; z_2 \mid (n))$. This situation is however of little interest, and in presence of spatial correlation joint probabilities are estimated using either an analytical (in most situations, a Gaussian model is adopted for the multivariate distribution of the random function) or a numerical (stochastic simulation) approach.

Under the multi-Gaussian model, the two-point probability distribution is bivariate Gaussian and fully characterized by the mean and the variance of each one-point ccdf, plus the covariance between $\mathbf{u}_1$ and $\mathbf{u}_2$. Two shortcomings of the analytical approach are: (1) the spatial uncertainty assessment becomes very complex as the number of grid nodes increases, and (2) it is cumbersome to check in practice the validity of the bi-Gaussianity assumption, and data sparsity prevents us from performing such checks for more than two locations at a time.

The numerical approach consists of generating a set of L pairs of correlated values $\{z^{(l)}(\mathbf{u}_1), z^{(l)}(\mathbf{u}_2)\}$, $l = 1, \ldots, L$, then estimating the joint probability from the set of simulated values as:

$$F(\mathbf{u}_1, \mathbf{u}_2; z_1, z_2 \mid (n)) \approx \frac{1}{L} \sum_{l=1}^{L} i^{(l)}(\mathbf{u}_1; z_1) \cdot i^{(l)}(\mathbf{u}_2; z_2), \quad (3)$$

where the indicator value $i^{(l)}(\mathbf{u}_j; z_j)$ is 1 if the simulated z-value at $\mathbf{u}_j$ does not exceed the threshold z_j, and 0 otherwise for $j = 1,2$. At least 100 simulations should be performed ($L \geq 100$), which renders the approach computationally

Fig. 2. Location map of 50 HC data, and conditional cdf models provided by: (1) multi-Gaussian kriging in the normal space (top graph) and after normal score back-transform (middle graph), and (2) indicator kriging (bottom graph) at two grid nodes depicted by open circles.

more demanding than the analytical alternative. The procedure can however be readily generalized to any number J of locations:

$$\text{Prob}\left\{Z(\mathbf{u}_j) \le z_j,\ j=1,\ldots,J \mid (n)\right\} \approx \frac{1}{L}\sum_{l=1}^{L}\prod_{j=1}^{J} i^{(l)}(\mathbf{u}_j; z_j). \quad (4)$$

Typically, the spatial distribution of soil attribute values is simulated over the entire domain of interest, then the joint probabilities can be computed for the target set of grid nodes. For example, Fig. 3 shows HC maps generated using sequential simulation that proceeds as follows:

- Define a random path visiting only once each node to be simulated.
- Model the ccdf at the first visited location $\mathbf{u}'_1$ conditional to the n original data $z(\mathbf{u}_\alpha)$.
- Draw from that ccdf a realization $z^{(l)}(\mathbf{u}'_1)$, which becomes a conditioning datum for all subsequent drawings.

$$\vdots$$

- At the ith node $\mathbf{u}'_i$ visited, model the conditional cdf of $Z(\mathbf{u}'_i)$ given the n original data and all $(i-1)$ values $z^{(l)}(\mathbf{u}'_j)$ simulated at the previously visited locations $\mathbf{u}'_j$, $j=1,\ldots,i-1$.
- Draw from the ccdf a realization $z^{(l)}(\mathbf{u}'_j)$, which becomes a conditioning datum for all subsequent drawings.
- Repeat the two previous steps until all N nodes are visited and each has been given a simulated value.

The resulting set of simulated values $\{z^{(l)}(\mathbf{u}'_j),\ j=1,\ldots,N\}$ represents one realization of the random function $\{Z(\mathbf{u}),\ \mathbf{u}\in\mathscr{A}\}$ over the N nodes $\mathbf{u}'_j$. Any number L of such realizations, $\{z^{(l)}(\mathbf{u}'_j),\ j=1,\ldots,N\}$, $l=1,\ldots,L$, can be obtained by repeating L times the entire sequential process with possibly different paths to visit the N nodes.

Two major classes of sequential simulation algorithms can be distinguished, depending on whether the series of conditional cdfs are determined using the multi-Gaussian (sGs = sequential Gaussian simulation) or the indicator (sis = sequential indicator simulation) formalism. Fig. 3 shows that the two algorithms yield quite different results although the use of the same random path and sequence of random drawings for the paired sGs and sis realizations creates similarities in the location of small and large simulated HC values.

The simulation procedure generates, at each location, a distribution of L values which can be used to approximate numerically the ccdf. For example, Fig. 4 (top graph) shows the cumulative distributions of 100 HC values (black dots) simulated at $\mathbf{u}_1$ and $\mathbf{u}_2$ using sequential Gaussian simulation. The joint uncertainty prevailing at these two locations can be described by the scattergram of the 100 pairs of simulated values, see Fig. 4 (left bottom graph). Because the separation distance between $\mathbf{u}_1$ and $\mathbf{u}_2$ is larger than the range (24 units) of the

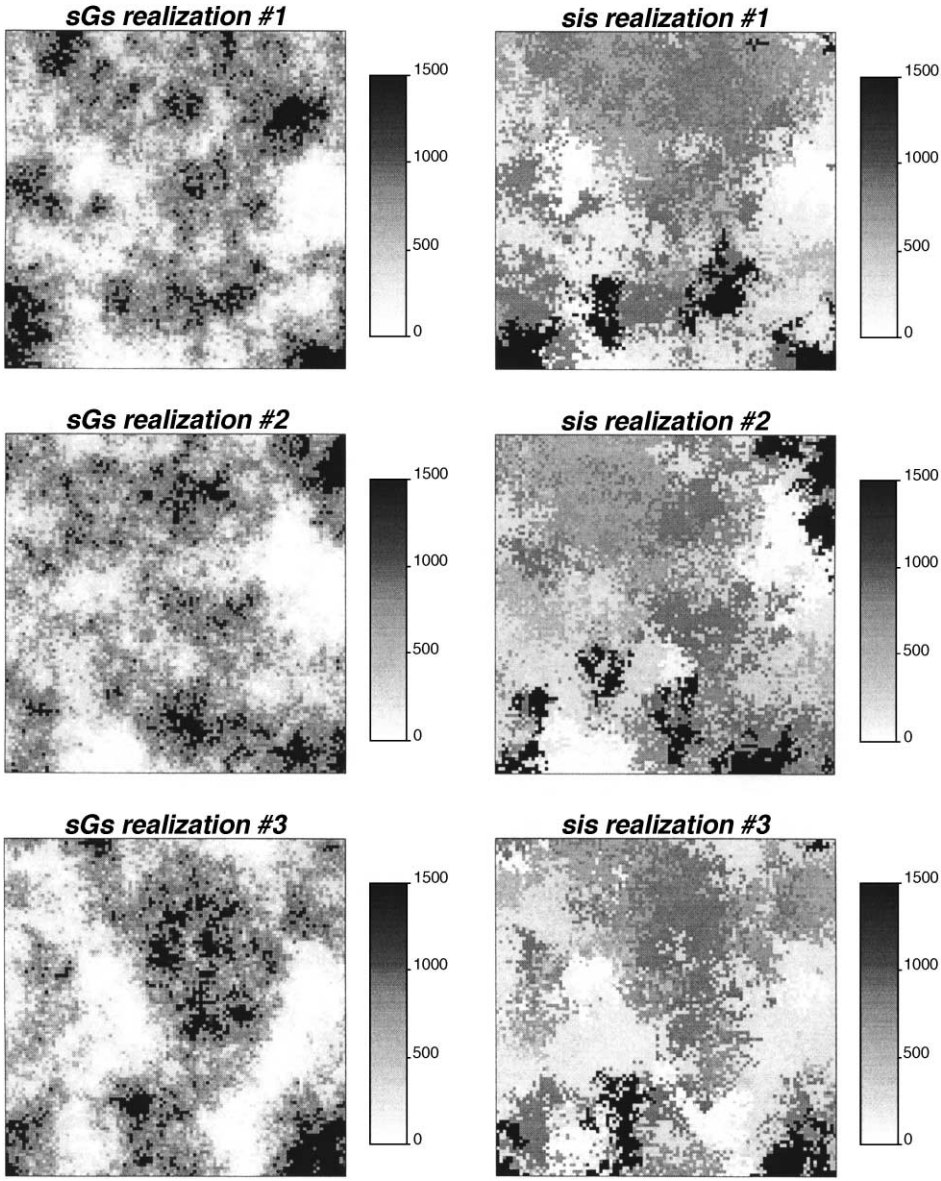


Fig. 3. Sets of three realizations generated using sequential Gaussian simulation (sGs) and sequential indicator simulation (sis) conditionally to the 50-HC data of Fig. 1.

normal score semivariogram model, the scattergram of simulated values displays no correlation; hence, in this case, the joint probability distribution could be inferred directly from the two one-point distributions. This is not so if $\mathbf{u}_1$ is paired with the adjacent location $\mathbf{u}_3 = \mathbf{u}_1 + (1,0)$, as illustrated by the correlation of the corresponding simulated values (Fig. 4, right bottom graph).

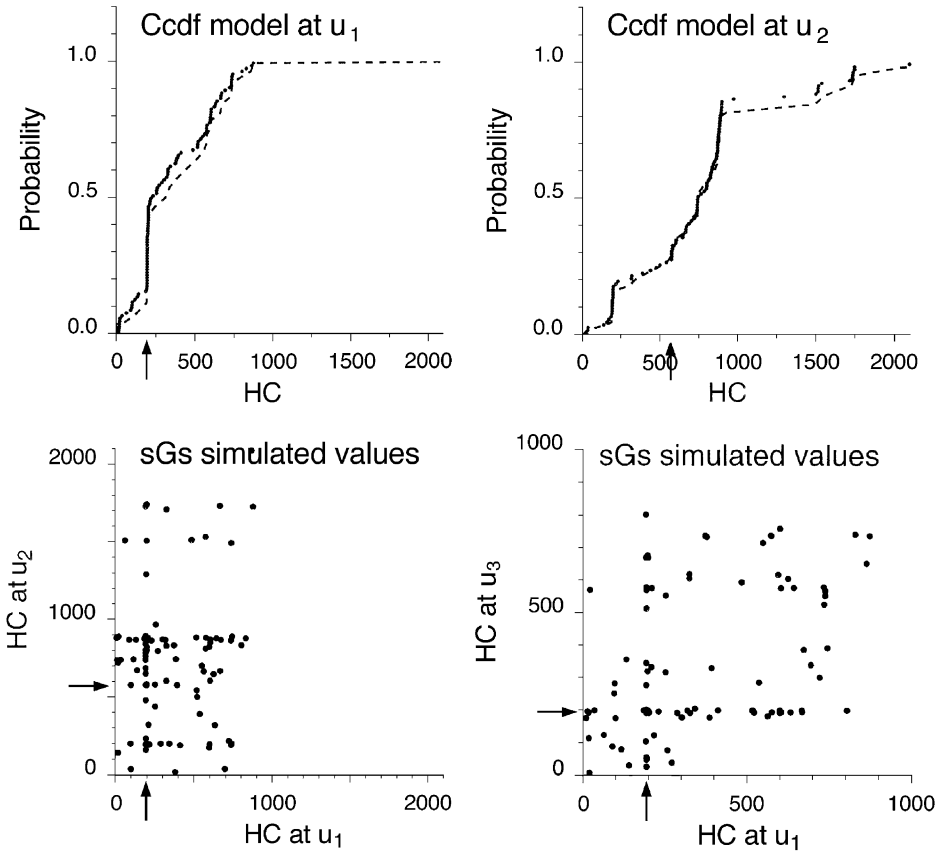


Fig. 4. Conditional cdf models inferred by multi-Gaussian kriging (dashed line) and numerically through the generation of 100-HC values (black dots) using sequential Gaussian simulation at the two grid nodes displayed in Fig. 2. The bottom graphs show the scattergrams of 100 pairs of HC values jointly simulated at two remote locations ($\mathbf{u}_1$, $\mathbf{u}_2$) or at two adjacent locations ($\mathbf{u}_1$, $\mathbf{u}_3$).

2.3. Stochastic simulation versus kriging

In several applications published so far in soil science, stochastic simulation has been used merely to assess local uncertainty from the local distribution of simulated values; that is the ccdf at $\mathbf{u}$ is approximated by:

$$F(\mathbf{u}; z \mid (n)) \approx \frac{1}{L} \sum_{l=1}^{L} i^{(l)}(\mathbf{u}; z), \tag{5}$$

where $i^{(l)}(\mathbf{u}; z) = 1$ if $z^{(l)}(\mathbf{u}) \leq z$, and 0 otherwise. In theory, as the number of realizations tends to infinity ($L \to \infty$), the local distribution of simulated values should match that provided by kriging within a similar framework. For example, the ccdfs inferred using multi-Gaussian kriging at $\mathbf{u}_1$ and $\mathbf{u}_2$ are overlaid over the cumulative distributions of 100 simulated HC values (black dots) in Fig. 4

(top graph). Clearly, the kriging-based and simulation-based approaches yield similar uncertainty models. Thus, if one seeks only a location-specific assessment of uncertainty (e.g. derivation of probability maps) the use of stochastic simulation entails an unnecessary waste of CPU time and disk storage.

3. Change of support

All the techniques presented in Section 2 allow one to model uncertainty over supports (area or block) similar to the observations' one, which is usually small if soil cores have not been bulked. In many situations, the uncertainty need to be assessed for much larger supports (Heuvelink and Pebesma, 1999), such as remediation units for cadmium or flow simulator cells for hydraulic conductivity, see Fig. 1. The objective is then to assess the probability that the average z-value over a block of size V centered on $\mathbf{u}$ is no greater than any threshold z, that is to model the block ccdf:

$$F_V(\mathbf{u}; z \mid (n)) = \text{Prob}\{Z_V(\mathbf{u}) \leq z \mid (n)\}. \quad (6)$$

Under the multiGaussian model, ccdfs can be inferred using a procedure similar to the one described above, except that simple block kriging is now used to derive the mean and variance of block ccdfs in the Gaussian space before the back-transform of the horizontal axis (e.g. see Chilès and Delfiner, 1999, p. 445). A limitation of the technique is that the attribute values must average linearly in space which is, for instance, not the case for hydraulic conductivity. Also, the multi-Gaussian assumption may not be appropriate for the data under study.

Limitations inherent in multiGaussian block kriging are not shared by the volume–variance correction (VVC) approach (Isaaks and Srivastava, 1989, Chap. 9), which proceeds in two steps: (1) the point ccdf $F(\mathbf{u}; z \mid (n))$ is first inferred at the central location $\mathbf{u}$ of the block of size V, and (2) the point ccdf is corrected for the support effect to yield the block ccdf $F_V(\mathbf{u}; z \mid (n))$. Such correction typically involves a decrease of the ccdf variance going from point to the block support, while the shape of the ccdf remains the same (permanence of distribution) or becomes less skewed (more symmetric with shorter tails). The ratio of the block and point ccdf variances, known as the variance reduction factor f, is estimated from the z-semivariogram model as (Journel, 1987):

$$f = \frac{\text{Var}\{Z_V(\mathbf{u}) \mid (n)\}}{\text{Var}\{Z(\mathbf{u}) \mid (n)\}} = 1 - \frac{\bar{\gamma}(V, V)}{\bar{\gamma}(\mathscr{A}, \mathscr{A})} \leq 1, \quad (7)$$

where $\bar{\gamma}(V, V)$ and $\bar{\gamma}(\mathscr{A}, \mathscr{A})$ are the average semivariogram values within the block V and the study area $\mathscr{A}$, respectively. If the area $\mathscr{A}$ is large with regard to

the range of the semivariogram, the $\bar{\gamma}(\mathcal{A}, \mathcal{A})$ can be set equal to the semivariogram sill, and f is estimated as:

$$f = 1 - \bar{\gamma}_S(V, V), \tag{8}$$

where $\bar{\gamma}_S(V, V)$ is the average value of the standardized z-semivariogram (unit sill) within the block V and is approximated by the arithmetic average of the standardized semivariogram values defined between any two discretizing points $\mathbf{u}'_i$ and $\mathbf{u}'_j$ within that block:

$$\bar{\gamma}_S(V, V) \approx \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \gamma_S(\mathbf{u}'_i - \mathbf{u}'_j).$$

For a given reduction factor f, the block ccdf can be derived from the point ccdf using various algorithms that differ in the degree of symmetrization imparted to that block ccdf. The affine correction amounts at keeping unchanged both the mean and the shape of the point ccdf, regardless of the magnitude of the change of support. This permanence hypothesis is realistic only for small blocks with little variability of z-values within them. Intuitively, one would expect a gradual symmetrization of the block ccdf as the block size increases, which is achieved by the indirect lognormal correction implemented in Gslib software, e.g. see Isaaks and Srivastava (1989, p. 473–475) for details.

A more versatile—yet computationally intensive—alternative to the block kriging and VVC approaches is provided by stochastic simulation. The block ccdf is numerically approximated by generating many simulated block values as (non)linear averages of simulated point values (Journel and Huijbregts, 1978, p. 511):

$$F_V(\mathbf{u}; z \mid (n)) \approx \frac{1}{L} \sum_{l=1}^{L} i_V^{(l)}(\mathbf{u}; z), \tag{9}$$

where $i_V^{(l)}(\mathbf{u}; z) = 1$ if $z_V^{(l)}(\mathbf{u}) \leq z$, and 0 otherwise. For example, the simulated average Cd concentration within the 0.25-km^2 block of Fig. 1 is computed as the arithmetic average of the $J = 25$ simulated point values falling into it:

$$z_V^{(l)}(\mathbf{u}) = \frac{1}{J} \sum_{j=1}^{J} z^{(l)}(\mathbf{u}'_j). \tag{10}$$

The attractive features of the simulation approach are that: (1) non-linear averaging functions (e.g. geometric or harmonic mean) can be considered, (2) no assumption is made regarding the impact of the change of support on the shape and variance of the block ccdf, and (3) once the grid of point values has been simulated, ccdfs for various block sizes and shapes can be derived at little computational cost.

Fig. 5 (top left graph, solid line) shows the cumulative distribution of 100 simulated block values computed from the set of 100 sis Cd maps; the first three

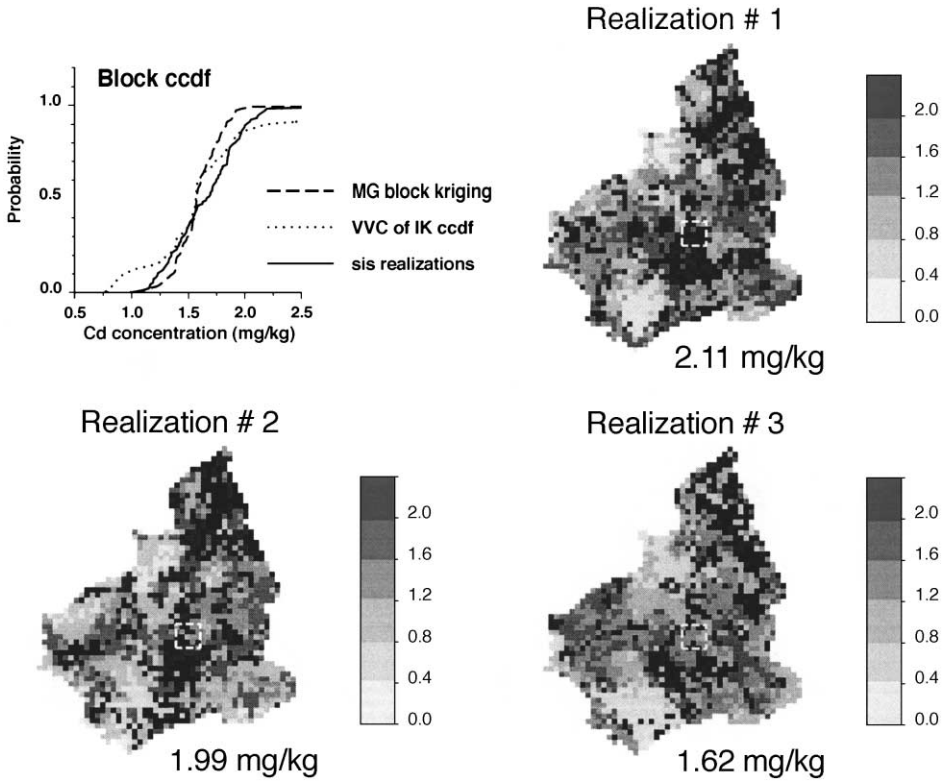


Fig. 5. Block ccdfs modelling the uncertainty about the average Cd concentration over the 0.25-km^2 block delineated by the dashed line in Fig. 1. Three different approaches are used: multi-Gaussian block kriging, correction of an indicator-based point support ccdf model, and distribution of 100 block Cd concentrations simulated using sequential indicator simulation (first three realizations are displayed).

realizations with the corresponding block simulated values are also displayed in Fig. 5. Such an uncertainty assessment is non-parametric: no prior assumption is made about the shape of the distribution of possible values. For the same block, the ccdf is also derived by multi-Gaussian block kriging (dashed line) and indirect lognormal correction of the ccdf inferred using indicator kriging at the center of the block (dotted line, $f = 0.24$). Although the three block ccdfs have similar median values, the upper and lower tails strongly differ. In particular, the VVC approach yields a much larger block variance with a higher probability of occurrence for both small and large average Cd concentrations.

4. How does the uncertainty propagate?

Our representations of the spatial distribution of soil attribute values are frequently imported into functions of various complexity, such as economic

remediation models, crop models or flow simulators. The uncertainty about the input soil attribute values will propagate through these "transfer" functions, leading to uncertain response values such as (non)remediation costs, crop yields or flow travel times. Depending on the nature of the transfer function, various approaches can be used to assess the response uncertainty.

4.1. Local transfer functions

A transfer function is said to be local if it is applied to a single location $\mathbf{u}$ or block $V(\mathbf{u})$ at a time. For example, the cost of classifying a location as safe with respect to Cd could be modelled as a function of the concentration $z(\mathbf{u})$ at that particular location:

$$y(\mathbf{u}) = \begin{cases} 0 & \text{if } z(\mathbf{u}) \leq z_c \\ z(\mathbf{u}) - z_c & \text{otherwise} \end{cases}, \tag{11}$$

where $z_c = 0.8$ mg/kg is the regulatory threshold for Cd. Similarly, the impact of soil acidity on the growth of a particular crop can be modelled as a function of the local pH value $z(\mathbf{u})$:

$$y(\mathbf{u}) = \begin{cases} 1 & \text{if } z(\mathbf{u}) \leq 5 \\ 0 & \text{if } z(\mathbf{u}) > 7 \\ (7 - z(\mathbf{u}))/2 & \text{if } 5 < z(\mathbf{u}) \leq 7 \end{cases}. \tag{12}$$

The propagation of uncertainty from z- to y-values can be approached either analytically or numerically. Heuvelink (1998) provides a comprehensive review of analytical propagation of local uncertainty. Because these methods rely on stringent assumptions on the distribution of input values, one may prefer a numerical approach (Monte-Carlo simulation), whereby the ccdf of z at $\mathbf{u}$ is randomly sampled many times, yielding a set of simulated z-values $\{z^{(l)}(\mathbf{u}), l = 1, \ldots, L\}$, which are then fed into functions of type (11) or (12) to generate a set of simulated y-values $\{y^{(l)}(\mathbf{u}), l, \ldots, L\}$. For example, the distribution of possible health costs in Fig. 6 (right bottom graph) is obtained through a random drawing of 1000 simulated Cd values and their post-processing by function (11). Random sampling of ccdfs can be very time-consuming, and a more efficient strategy is Latin hypercube sampling (LHS) that consists of dividing each distribution into N equiprobable classes (100, for example), and sampling these classes without replacement to generate a set of N outcomes (stratified random sampling). In this way, the input distributions are represented in their entirety, and it requires a much smaller sample than the Monte Carlo approach for a given degree of precision (Mckay et al., 1979; Helton, 1997). Recent application of LHS in soil science can be found in Hansen et al. (1999), Pebesma and Heuvelink (1999), Van Merveinne and Goovaerts (2001).

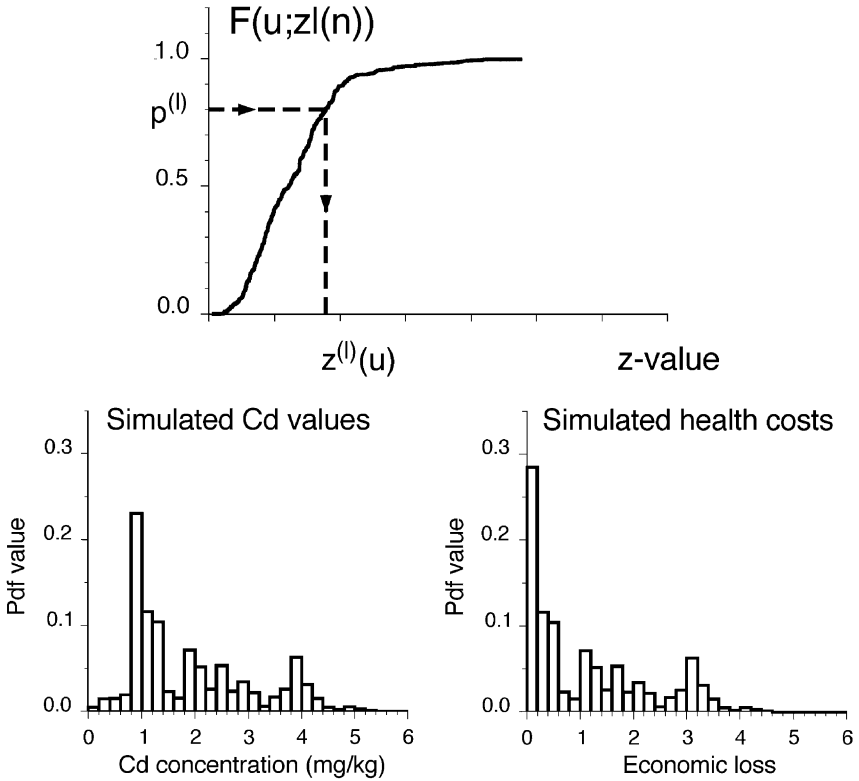


Fig. 6. Use of Monte-Carlo simulation (random sampling of ccdf) to propagate the uncertainty through local transfer functions. Application to the assessment of the uncertainty about health costs arising from the uncertainty about the Cd concentration at a particular location.

Uncertainty about the parameters of the transfer function itself can be accounted for by sampling probability distributions of these parameters and combining both simulated parameter and input *z*-values. For many parameters (say more than five), the number of possible combinations becomes quickly prohibitive, although an LHS strategy may greatly alleviate the computational requirements. When these parameters are correlated, their distributions cannot be sampled independently and particular sampling techniques must be used, see Viscarra Rossel et al. (2001).

4.2. Global transfer functions

Complex transfer functions, such as flow simulators, truly propagate the uncertainty across the study area, which precludes the use of analytical procedures in most situations—see Dettinger and Wilson (1981) for the use of the first order analysis as an alternative. The common approach is to generate many realizations of the spatial distribution of attribute values which are then fed into

the flow simulator to yield a distribution of response values, such as travel times for flow and solute transport. For example, Fig. 7 shows the histograms of water travel times across the HC data set (SW corner to NE corner) obtained by post-processing two sets of 100 sGs and sis realizations. These graphs confirm the visual interpretation of Fig. 3 that the two sequential simulation algorithms generate different sets of realizations, although they are all based on the same 50 HC data. In particular, the response distribution is much narrower for sis realizations.

The characterization of the space of uncertainty is rendered difficult by the fact that only a limited number of realizations is usually generated. A frequent and still open question relates to the number of realizations needed to sample fairly this space. For a given simulation algorithm and response variable, the impact of the number of realizations on the accuracy (inclusion of the true value) and precision (narrowness) of the response distribution can be investigated using the procedure described in Goovaerts (1999a).

4.3. Stochastic simulation versus kriging

Stochastic spatial simulation is usually not recommended for the local transfer functions, since an LHS of ccdf provides similar results with much less CPU costs. There are, however, two situations where a simulation-based approach is warranted: transfer functions are available only for supports much larger than the measurement support (e.g. remediation costs are known only for remediation units of a given size), which requires a prior simulation-based derivation of the block ccdf, and (2) the histogram of response values averaged over the study area is sought. Consider, for example, the problem of assessing the economic impact of declaring the entire study area safe with respect to Cd. The distribu-

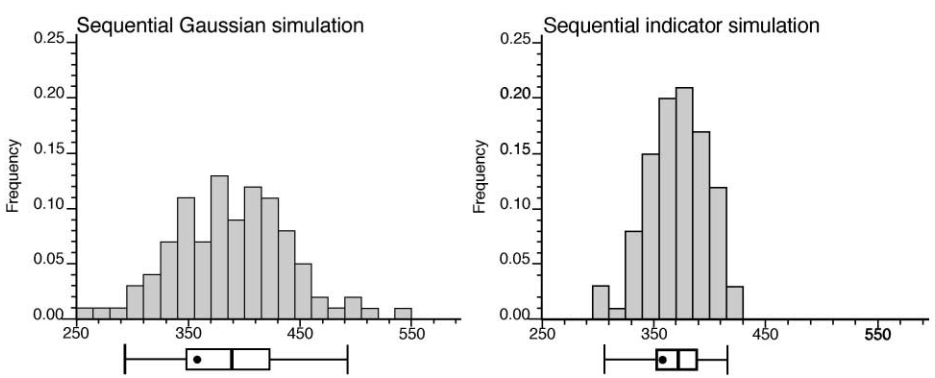


Fig. 7. Histograms of water travel times obtained by post-processing 100-sGs and 100-sis HC maps through a flow simulator. The black dot in the box plot below each histogram is the true travel time. Five vertical lines are the 0.025 quantile, lower quartile, median, upper quartile, and 0.975 quantile of the distribution.

tion of global cost would be built by applying the cost function (11) to the set of simulated point values and averaging the local costs for each realization, see Goovaerts (1999b, Fig. 7).

5. How good is the model of uncertainty?

The previous sections have shown that different methods yield models of uncertainty that can greatly differ, and a legitimate question is whether the choice of a technique (e.g. multi-Gaussian versus indicator kriging) can be supported by the data. Like spatial interpolation, alternative techniques can be used to build uncertainty models which are then compared with observations that have been either temporarily removed one at a time (cross-validation) or set aside for the whole analysis (jacKnife). The remaining issue is the selection of performance criteria for uncertainty assessment.

5.1. Local uncertainty

At any test location $\mathbf{u}$, knowledge of the ccdf $F(\mathbf{u};\ z\,|\,(n))$ allows the computation of a series of symmetric p-probability intervals (PI) bounded by the $(1-p)/2$ and $(1+p)/2$ quantiles of that ccdf. For example, the 0.5-PI is bounded by the lower and upper quartiles $[F^{-1}(\mathbf{u};\,0.25\,|\,(n)),\ F^{-1}(\mathbf{u};\,0.75\,|\,(n))]$. A correct modelling of local uncertainty would entail that there is a 0.5 probability that the actual z-value at $\mathbf{u}$ falls into that interval or, equivalently, that over the study area, 50% of the 0.5-PI include the true value. If a set of z-measurements and independently derived ccdfs are available at N locations, $\mathbf{u}_j$, $\{[z(\mathbf{u}_j),\ F(\mathbf{u}_j,\ z\,|\,(n))],\ j=1,\ldots,\ N\}$, the fraction of true values falling into the symmetric p-PI can be computed as:

$$\bar{\xi}(p) = \frac{1}{N}\sum_{j=1}^{N} \xi(\mathbf{u}_j;\, p) \forall p \in [0,1] \tag{13}$$

with:

$$\xi(\mathbf{u}_j;\, p) = \begin{cases} 1 & \text{if } F^{-1}\big(\mathbf{u}_j;(1-p)/2\big) < z(\mathbf{u}_j) \le F^{-1}\big(\mathbf{u}_j;(1+p)/2\big) \\ 0 & \text{otherwise} \end{cases}.$$

The scattergram of the estimated versus expected fractions is called "accuracy plots". Fig. 8 (left top graph) shows, for the HC data set, accuracy plot computed for both parametric and non-parametric ccdfs using a cross validation of the 50 observations. Most of the points fall below the 45° line, i.e. $\bar{\xi}(p) < p$ for most p, which reflects the inaccuracy of the probabilistic models. For example, the 0.5 probability intervals derived from the parametric approach only

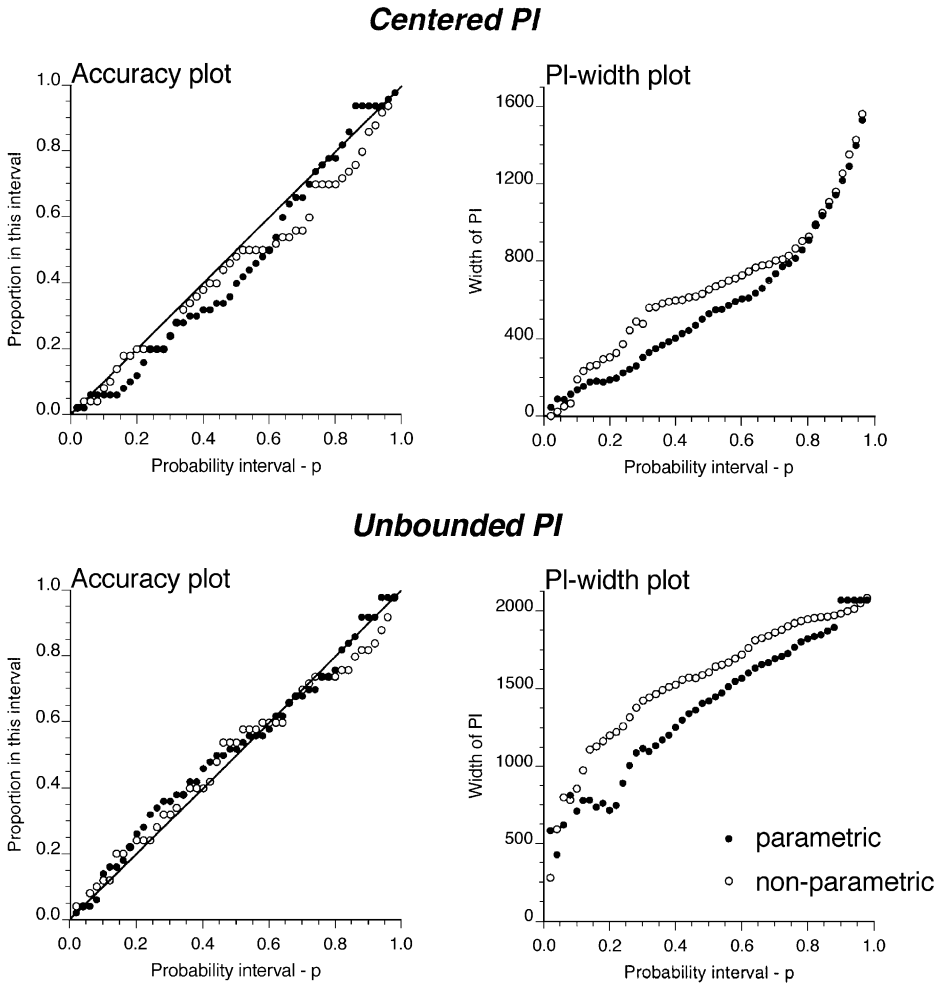


Fig. 8. Plots of the proportion of true HC values falling within probability intervals (accuracy plot), and the width of these intervals versus the probability p. Top graphs relate to symmetric intervals centered on the ccdf median, while bottom graphs depict results for unbounded intervals. Parametric (multi-Gaussian) and non-parametric (indicator) algorithms are used to derive ccdf models in a cross-validation approach.

contain 40% of the true values. While the parametric models yield the largest deviations between the estimated and theoretical fractions for small p-values, they perform better than the non-parametric models for large p-values. Deutsch (1997) proposed to assess the closeness of the estimated and theoretical fractions using the following "goodness" statistics:

$$G = 1 - \int_0^1 [3a(p) - 2]\left[\bar{\xi}(p) - p\right] \mathrm{d}p, \tag{14}$$

where the indicator function $a(p)$ is defined as:

$$a(p) = \begin{cases} 1 & \text{if } \bar{\xi}(p) \geq p \\ 0 & \text{otherwise} \end{cases}.$$

Twice more importance is given to deviations when $\bar{\xi}(p) < p$ (inaccurate case). The weight $|3a(p) - 2| = 2$ instead of 1 for the accurate case, that is the case where the fraction of true values falling into the p-PI is larger than expected. For the cross validation of Fig. 8, the differences between the two approaches, as assessed by the goodness statistics G, are negligible: 0.902 (non-parametric) and 0.899 (parametric), which agrees with Papritz and Dubois' (1999) conclusions.

Not only should the true value fall into the PI according to the expected probability p, but this interval should be as narrow as possible to reduce the uncertainty about that value. In other words, among two probabilistic models with similar goodness statistics, one would privilege the one with the smallest spread (less uncertain). Different measures of ccdf spread can be used: variance, interquartile range, entropy. In this paper, it is proposed to plot, for a series of probabilities p, the average width of the PIs that include the true values; see Fig. 8 (right top graph). For a probability p, the average width $\overline{W}(p)$ is computed as:

$$\overline{W}(p) = \frac{1}{N\bar{\xi}(p)} \sum_{j=1}^{N} \xi(\mathbf{u}_j; p) \cdot \left[F^{-1}\left(\mathbf{u}_j; (1+p)/2\right) - F^{-1}\left(\mathbf{u}_j; (1-p)/2\right) \right]. \quad (15)$$

Fig. 8 indicates that the parametric intervals are narrower until $p = 0.8$, which may cause the proportion of true values falling into these intervals to be smaller than for the non-parametric models.

For applications where the exceedence of specific threshold values is the main concern, median-centered PIs could be replaced by unbounded PIs; for example, the 0.5-PI is now defined as $[F^{-1}(\mathbf{u}; 0.5 \mid (n)), +\infty]$ or, in practice, $[F^{-1}(\mathbf{u}; 0.5 \mid (n)), z_{\max}]$. A correct modelling of local uncertainty would entail that over the study area, 50% of the true values locally exceed the ccdf median. Fractions (Eq. (13)) are thus computed as:

$$\bar{\xi}(p) = \frac{1}{N} \sum_{j=1}^{N} \xi(\mathbf{u}_j; p) \forall p \in [0,1] \quad (16)$$

with:

$$\xi(\mathbf{u}_j; p) = \begin{cases} 1 & \text{if } F^{-1}(\mathbf{u}_j; p') < z(\mathbf{u}_j) \\ 0 & \text{otherwise} \end{cases}$$

where $p' = 1 - p$. Particular types of accuracy plots, called "exceedence probability plots", are created by plotting the estimated fractions versus the expected ones; see Fig. 8 (bottom left graph). Both parametric and non-parametric models yield similar fractions that are slightly larger than the expected ones for small probabilities p corresponding to high HC values, p'-quantiles, while the proportion of true values that exceed low HC thresholds (high p values) is smaller than expected. The last graph shows the width of probability intervals which corresponds here to the difference between the maximum z-value (2100) and the p'-quantile of the ccdf; for example, for $p = 0.1$, the PI is bounded by the 9th decile and the maximum value. As for symmetric PIs, multi-Gaussian kriging performs better than indicator kriging in that the probability intervals are narrower (larger precision), while including similar proportions of true values (similar goodness).

5.2. Spatial uncertainty

Unlike models of local uncertainty that can be compared to actual observations, there is no reference spatial distribution of values to use as yardstick for models of spatial uncertainty. Classical checks, which amount at verifying the quality of the reproduction of histogram and semivariogram by the realizations, provide no indication on the goodness of the set of realizations as a whole. In some instances, information might be available on the actual response value for a particular transfer function. For example, field tests may allow an accurate prediction of the travel time of contaminants or flow between two locations. This field observation can then be compared with the distribution of times derived by post-processing a set of HC realizations through a flow stimulator. In this particular case, one may also adopt geostatistical inverse methods that allow the use of the output variable to calibrate the geostatistical model in a cokriging environment (Zhang and Yeh, 1997) or actually condition simulated realizations (Gómez-Hernández et al., 1997b).

In this paper, the true travel time has been computed from the reference 102×102 image and is depicted by a black dot on the histograms of Fig. 7. For the HC data set, sequential indicator simulation clearly yields the best results since the response distribution is narrower (greater precision), while including the true value (accuracy). Thus, it appears that the greater precision (narrower PI) of the parametric ccdfs does not entail a greater precision of the model of response uncertainty because flow simulation is influenced primarily by the spatial distribution of HC values.

6. Conclusions

Uncertainty assessment, which should become a key step of any geostatistical analysis of soil data, can take various forms, such as mapping the probability of

exceeding threshold values or generating sets of realizations of the spatial distribution of attribute values. As for spatial prediction, the users are faced with the difficult choice of an algorithm within the geostatistical toolbox, and the aim of this paper was to review briefly the tools available to address various issues, such as the modelling of uncertainty over supports of different sizes or the propagation of uncertainty through local or global transfer functions.

The primary question is whether stochastic simulation is required for uncertainty assessment, or whether similar information can be acquired by the computationally cheaper kriging approach. There are two situations where there is no good alternative to stochastic simulation: (1) propagation of uncertainty through multiple-point (dynamic) transfer functions, such as flow simulators, and (2) the modelling of uncertainty over supports much larger than the measurement supports for attributes (like hydraulic conductivity) that do not average linearly in space. In all other situations—which may represent the majority of current applications of geostatistics to soil science—a kriging-based modelling of ccdf and its subsequent Latin-hypercube sampling for numerical propagation of uncertainty should be the first choice.

For both stochastic simulation and kriging-based approaches, the common issue is whether a parametric (multi-Gaussian) or non-parametric (indicator) approach should be adopted. There is no straight answer to that question, and for ccdf modelling it is recommended to perform a cross-validation to assess the goodness and precision of models of local uncertainty. For example, accuracy plots can be computed using a Gslib-type program available on http://www.ualberta.ca/~cdeutsch/. For spatial uncertainty and its propagation through complex transfer functions, there is usually no way to assess the quality of models. Users should then keep in mind that the multi-Gaussian model does not allow for any significant spatial correlation of very large or very small values (destructuration effect), which may not be desirable (conservative) for the application of interest. In the case of the HC data set, a parametric approach was better for modelling local uncertainty, while for propagation of spatial uncertainty, best results were obtained using a non-parametric algorithm.

In this paper, the focus has been on the modelling of uncertainty for single continuous soil attributes. Many soil attributes are categorical and uncertainty assessment through indicator-based kriging or stochastic simulation has received an increasing attention the last few years, e.g. see Bierkens and Burrough (1993a,b), Goovaerts (1996), Oberthür et al. (1999). Again, the issue of assessing the goodness of such models has rarely been addressed and criteria similar to the ones introduced here could be developed. In several situations, there is also a need for modelling the joint uncertainty about the values of multiple soil attributes at a single location or across the entire domain. Similarly, transfer functions, such as crop models, frequently include many input variables which requires a multivariate propagation of uncertainty. In the future, more research should be devoted to issues such as the multivariate modelling of ccdfs or the

spatial cosimulation of variables, and the joint sampling of multiple ccdfs using LHS.

References

Atteia, O., Dubois, J.-P., Webster, R., 1994. Geostatistical analysis of soil contamination in the Swiss Jura. Environ. Pollut. 86, 315–327.

Bierkens, M.F.P., Burrough, P.A., 1993a. The indicator approach to categorical soil data: I. Theory. J. Soil Sci. 44, 361–368.

Bierkens, M.F.P., Burrough, P.A., 1993b. The indicator approach to categorical soil data: II. Application to mapping and land use suitability analysis. J. Soil Sci. 44, 369–381.

Bloom, L., Kentwell, D.J., 1999. A geostatistical analysis of cropped and uncropped soil from the Jimperding Brook catchment of Western Australia. In: Gómez-Hernández, J., Soares, A., Froidevaux, R. (Eds.), geoENV II—Geostatistics for Environmental Applications. Kluwer Academic Publishing, Dordrecht, pp. 369–379.

Broos, M.J., Aarts, L., van Tooren, C.F., Stein, A., 1999. Quantification of the effects of spatially varying environmental contaminants into a cost model for soil remediation. J. Environ. Manage. 56, 133–145.

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

Davis, B.M., 1987. Uses and abuses of cross validation in geostatistics. Math. Geol. 19, 241–248.

Dettinger, M.D., Wilson, J.L., 1981. First order analysis of uncertainty in numerical models of groundwater flow: Part I. Mathematical development. Water Resour. Res. 17, 149–161.

Deutsch, C.V., 1994. Algorithmically-defined random function models. In: Dimitrakopoulos, R. (Ed.), Geostatistics for the Next Century. Kluwer Academic Publishing, Dordrecht, pp. 422–435.

Deutsch, C.V., 1997. Direct assessment of local accuracy and precision. In: Baafi, E.Y., Schofield, N.A. (Eds.), Geostatistics Wollongong ’96. Kluwer Academic Publishing, Dordrecht, pp. 115–125.

Deutsch, C.V., Journel, A.G., 1998. GSLIB: Geostatistical Software Library and User’s Guide. 2nd edn. Oxford Univ. Press, New York, 369 pp.

Gómez-Hernández, J.J., 1997a. Issues on environmental risk assessment. In: Baafi, E.Y., Schofield, N.A. (Eds.), Geostatistics Wollongong ’96. Kluwer Academic Publishing, Dordrecht, pp. 15–26.

Gómez-Hernández, J.J., Sahuquillo, A., Capilla, J.E., 1997b. Stochastic simulation of transmissivity fields conditional to both transmissivity and piezometric data: I. Theory. J. Hydrol. 203, 162–174.

Goovaerts, P., 1996. Stochastic simulation of categorical variables using a classification algorithm and simulated annealing. Math. Geol. 28, 909–921.

Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford Univ. Press, New York, 512 pp.

Goovaerts, P., 1999a. Impact of the simulation algorithm, magnitude of ergodic fluctuations and number of realizations on the spaces of uncertainty of flow properties. Stochastic Environ. Res. Risk Assess. 13, 161–182.

Goovaerts, P., 1999b. Geostatistics in soil science: state-of-the-art and perspectives. Geoderma 89, 1–45.

Goovaerts, P., Journel, A.G., 1995. Integrating soil map information in modelling the spatial variation of continuous soil properties. Eur. J. Soil Sci. 46, 397–414.

Gotway, C.A., Rutherford, B.M., 1994. Stochastic simulation for imaging spatial uncertainty:

comparison and evaluation of available algorithms. In: Armstrong, M., Dowd, P.A. (Eds.), Geostatistical Simulations. Kluwer Academic Publishing, Dordrecht, pp. 1–21.
Hansen, S., Thorsen, M., Pebesma, E.J., Kleeschulte, S., Svendsen, H., 1999. Uncertainty in simulated nitrate leaching due to uncertainty in input data: a case study. Soil Use Manage. 15, 167–175.
Helton, J.C., 1997. Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. J. Stat. Comput. Simul. 57, 3–76.
Heuvelink, G.B.M., 1998. Error Propagation in Environmental Modelling with GIS. Research Monographs in GIS Series, London, 127 pp.
Heuvelink, G.B.M., Pebesma, E.J., 1999. Spatial aggregation and soil processes modelling. Geoderma 89, 47–65.
Isaaks, E.H., Srivastava, R.M., 1989. An Introduction to Applied Geostatistics. Oxford Univ. Press, New York, 561 pp.
Journel, A.G., 1987. Geostatistics for the Environmental Sciences. EPA project no CR 811893. Technical report, US Environmental Protection Agency, EMS Laboratory, Las Vegas, NV.
Journel, A.G., Huijbregts, C.J., 1978. Mining Geostatistics. Academic Press, New York, 600 pp.
McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.
McKenna, S.A., 1998. Geostatistical approach for managing uncertainty in environmental remediation of contaminated soils: case study. Environ. Eng. Geosci. 4, 175–184.
Oberthür, T., Goovaerts, P., Dobermann, A., 1999. Mapping soil texture classes using field texturing, particle size distribution and local knowledge by both conventional and geostatistical methods. Eur. J. Soil Sci. 50, 457–479.
Pachepsky, Y., Acock, B., 1998. Stochastic imaging of soil parameters to assess variability and uncertainty of crop yield estimates. Geoderma 85, 213–229.
Papritz, A., Dubois, J.-P., 1999. Mapping heavy metals in soil by (non-)linear kriging: an empirical validation. In: Gómez-Hernández, J., Soares, A., Froidevaux, R. (Eds.), geoENV II—Geostatistics for Environmental Applications. Kluwer Academic Publishing, Dordrecht, pp. 429–440.
Pebesma, E.J., Heuvelink, G.B.M., 1999. Latin hypercube sampling of Gaussian random fields. Technometrics, in press.
Smith, J.L., Halvorson, J.J., Papendick, R.I., 1993. Using multiple-variable indicator kriging for evaluating soil quality. Soil Sci. Soc. Am. J. 57, 743–749.
Srinivasan, S., Journel, A.G., 1998. Direct simulation of permeability field conditioned to well test data. In: Report 11, Stanford Center for Reservoir Forecasting, Stanford, CA.
Srivastava, M.R., 1996. An overview of stochastic spatial simulation. In: Mowrer, H.T., Czapleweski, R.L., Hamre, R.H. (Eds.), Spatial Accuracy Assessment in Natural Resources and Environmental Sciences: Second International Symposium, General Technical Report RM-GTR-277. US Department of Agriculture, Forest Service, Fort Collins, pp. 13–22.
Vanderborght, J., Jacques, D., Mallants, D., Tseng, P.H., Feyen, J., 1997. Analysis of solute redistribution in heterogeneous soil: II. Numerical simulation of solute transport. In: Soares, A., Gómez-Hernández, J., Froidevaux, R. (Eds.), geoENV I—Geostatistics for Environmental Applications. Kluwer Academic Publishing, Dordrecht, pp. 283–295.
Van Meirvenne, M., Goovaerts, P., 2001. Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold. Geoderma 102, 63–88.
Viscarra Rossel, R.A., Goovaerts, P., McBratney, A.B., 2001. Assessment of the production and economic risks of site-specific liming using geostatistical uncertainty modelling. Environmetrics 12 (5), in press.
Wackernagel, H., 1998. Multivirate Geostatistics. 2nd edn. Springer, Berlin, 291 pp.

Webster, R., Oliver, M.A., 1989. Optimal interpolation and isarithmic mapping of soil properties: VI. Disjunctive kriging and mapping the conditional probability. J. Soil Sci. 40, 497–512.
Webster, R., Atteia, O., Dubois, J.-P., 1994. Coregionalization of trace metals in the soil in the Swiss Jura. Eur. J. Soil Sci. 45, 205–218.
Zhang, J., Yeh, T.-C.J., 1997. An iterative geostatistical inverse method for steady flow in the vadose zone. Water Resour. Res. 33, 63–71.