Geostatistical Tools for Deriving Block-Averaged Values of Environmental Attributes

Pierre Goovaerts

Department of Civil and Environmental Engineering,
The University of Michigan,
Ann Arbor, MI 48109-2125, USA.

Abstract

This paper reviews four different approaches for estimating the block-averaged value of an environmental attribute from point data: 1) the sample arithmetic average, 2) the declustered mean, 3) block kriging, and 4) stochastic simulation. The first approach is straightforward and well suited for estimation over large blocks that contain many randomly located observations. Declustering techniques can be used to correct for preferential sampling of specific subareas of such large blocks. The last two techniques, kriging and simulation, account for the pattern of spatial dependence of observations and allow one to compensate for the shortage of data inside small blocks by incorporating observations outside the block. The major advantage of stochastic simulation is that it provides a non-parametric measure of the uncertainty attached to the prediction of a single block or multiple spatially dependent blocks.Stochastic simulation can also be used to upscale properties like permeability that do not average linearly in space, whereas the first 3 techniques are only valid for linear averaging parameters. The different techniques are illustrated using a soil data set related to heavy metal contamination over a 14.5km^2 area in the Swiss Jura.

I. INTRODUCTION

Over the last fifteen years, kriging has been increasingly preferred to traditional interpolation methods, such as moving averages or inverse distance methods, for predicting environmental attribute values, and in particular soil properties, at unsampled locations [9,14,19]. The main advantage of geostatistical interpolation is that it accounts for the pattern of spatial variability of observations modeled through the semivariogram, while providing a measure of estimation variance. Soil properties are typically measured on small cores, whereas land managers or decision makers are interested in the average attribute value over larger surfaces such as 1 ha plots, hereafter called "blocks". Determination of average values over supports larger than the measurement or data support is generally referred to as "upscaling",particularly when the attribute considered does not average linearly in space. Other terms like coarse-graining or aggregation are also used in soil science [18].

A particular feature of kriging is that it allows estimation of the target attribute on a support that is different from the data support, and so upscaling of soil properties has been naturally performed using block kriging [4]. A recent paper by Brus and de Gruijter [3] has, however, recalled that the simple arithmetic (equal-weighted) average of observations is a valuable alternative to kriging when the block contains many observations collected according to a random sampling design. A preferential sampling of low- or high-valued parts of the block can be corrected using weighted averages provided by declustering algorithms [9,11].

Upscaling issues are not specific to soil science. Hydrologists, mining and petroleum engineers have faced the problem of change of support for a long time [15 p.511—515], [5,23].The characterization of petroleum reservoirs requires the building of numerical models with grid cells of size several orders of magnitude greater than the volume support of the available core data. An additional constraint is that the block values must reproduce the pattern of spatial variability of petrophysical properties (e.g. porosity, permeability) in order to achieve realistic predictions of flow responses. In petroleum engineering as in soil science, the upscaling issue has been addressed using geostatistics, but simulation algorithms are typically preferred to block kriging [10]. The basic idea consists of generating realizations of the spatial distribution of the target attribute which reproduce the pattern of spatial variability of point measurements. Block simulated values are then computed using linear or non-linear averages of the point simulated values inside each block [6,7,13,17,19].

The objective of this paper is to present a practical overview of the geostatistical techniques currently available for determining block-averaged values of environmental attributes, and to emphasize the merits of the lesser known stochastic simulation approach over block kriging.

1082-4006/99/0502-88$5.00

II. SETTING THE PROBLEM

Consider the problem of estimating the average value of a soil attribute z, say Cd concentration, over a block of 2km^2 located within a larger 14.5km^2 area, see Figure1. The information available consists of 259 point Cd concentrations $z(\mathbf{u}_\alpha)$, 38 of those being located inside the 2km^2 bloc.A detailed description of the sampling, field, and laboratory procedures is given in [1,22]. Each observation $z(\mathbf{u}_\alpha)$ refers to a single soil core which can be assimilated to a point of coordinate $\mathbf{u}_\alpha$ with respect to the size of the block.

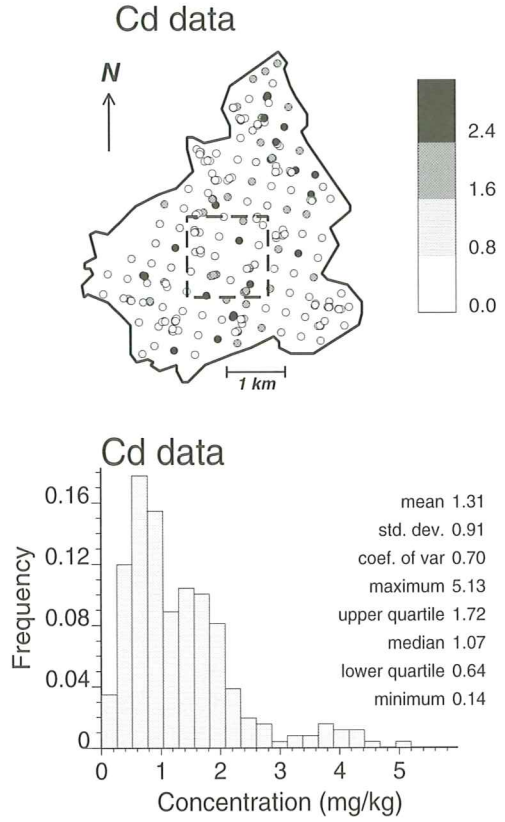

Figure 1. Location map of the 259 point Cd concentrations available for estimating the average concentration over the 2km^2 block delineated by the dashed line. The bottom graph shows the sample histogram.

Provided the averaging process is linear, the most direct estimate is the equal-weighted arithmetic average of the n=38 observations that fall inside the block B:

$$\hat{m} = \frac{1}{n}\sum_{\alpha=1}^{n} z(\mathbf{u}_\alpha) = 1.55\text{mg/kg} \qquad (1)$$

Because the block estimate is based on a limited number of observations,one would not expect it to identify exactly the true block value.Thus, it is often more informative to state an interval within which the unknown block value would be expected to lie with a specified level of certainty or confidence.For example, the classical 95% confidence interval (e.g., see[2p.18]) is defined as:

$$[\hat{m} - t_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}, \hat{m} + t_{1-\alpha/2}\sqrt{\frac{\hat{\sigma}^2}{n}}] = [1.30, 1.80] \qquad (2)$$

where the value of the Student's t statistics, $t_{1-\alpha/2}$, is 2.02 for n=37 and a probability α=0.05, and the variance is computed as:

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{\alpha=1}^{n}[z(\mathbf{u}_\alpha) - \hat{m}]^2 = 0.585(\text{mg/kg})^2 \qquad (3)$$

The derivation of the confidence interval (2) is based on the critical assumption that the n observations are independent and that the sample mean follows a normal distribution. In a recent discussion paper, Brus and de Gruijter [3] stated that independence can be created through randomization of sampling locations, regardless of the pattern of spatial correlation of soil attributes. According to the Central Limit Theorem, the second requirement is satisfied even if the histogram of point data is asymmetric, provided a large number of randomly selected observations are available.Thus, the suitability of this approach relies mainly on the number and layout of observations inside the block.

In many situations, only a limited number of laboratory measurements can be afforded and so each block contains only a few observations making the inference of the sample variance $\hat{\sigma}^2$ unreliable. Also, these observations are rarely randomly distributed in space. Typically, specific subareas of low or high values are preferentially sampled, e.g., areas around sources of pollution receive more attention than remote locations where exceedence of the regulatory threshold is less likely. This purposive sampling is indeed more cost-effective than random sampling for small sample sizes. Whenever data locations are not randomly spread inside the block, the representativity of the sample statistics should be questioned, and the data configuration must be accounted for in the analysis.

III. HISTOGRAM DECLUSTERING

A second approach, which allows one to account for data locations in the prediction of block values, is based on the concept of "data declustering"[8p.77-82], [12]. The idea consists of replacing the equal-weighted average of observations by a weighted average such that data in densely sampled areas receive less weight than isolated observations:

$$\hat{m}_D = \sum_{\alpha=1}^{n}\omega_\alpha z(\mathbf{u}_\alpha) \quad \text{with:} \quad \sum_{\alpha=1}^{n}\omega_\alpha = 1 \qquad (4)$$

Several algorithms can be used to compute the declustering weights ω_α:

1. The polygonal method (or Thiessen polygons) first

delineates the polygon of influence of each datum location $\mathbf{u}_\alpha$, that is, the area constituted by all locations $u \in B$ closer to $\mathbf{u}_\alpha$ than to any other datum location.The relative area of the polygon centered at location u_α is then used as a declustering weight for datum value $z(\mathbf{u}_\alpha)$.

2. The cell-declustering approach calls for dividing the block B into rectangular cells, and counting the number K of cells that contains at least one datum and the number n_k of data falling within each cell k. Each datum location $\mathbf{u}_\alpha$ then receives a weight $\omega_\alpha = 1/(K \cdot n_k)$, which gives more importance to isolated locations.

The weights ω_α can be used to "decluster" the sample histogram. This "declustered" marginal distribution has for mean the weighted mean $\hat{m}_D$ and its variance is computed as:

$$\hat{\sigma}_D^2 = \sum_{\alpha=1}^{n} \omega_\alpha \cdot [z(\mathbf{u}_\alpha) - \hat{m}_D]^2 \qquad (5)$$

Because preferential sampling is usually performed empirically, that is outside any statistical framework, its properties cannot be properly defined. The declustering approach should thus be viewed as an heuristic estimation method, and one cannot ensure that the negative effect of preferential sampling is fully corrected. Moreover, confidence intervals of type (2) cannot be computed since they require the assumption that the n observations are a simple random sample, which is not the case here.

The 38 Cd data of Figure 1 were declustered using square cells of 250 m which corresponds to the spacing of the sampling grid. The declustered mean is 1.63 mg/kg, which is slightly larger than the arithmetic mean because of the preferential location of data clusters in low-valued areas. The declustered variance is also larger, $\hat{\sigma}_D^2 = 0.933 > \hat{\sigma}^2 = 0.585$, because of the smaller weight given to clusters of similar concentrations. This approach still suffers from the shortcomings of the previous approach, that is, the pattern of spatial dependence of Cd concentrations as well as observations outside the block are not accounted for.

IV. BLOCK KRIGING

Kriging is a generic name adopted by geostatisticians for a family of generalized least-squares regression algorithms [21]. The basic idea is to estimate the average value of a continuous soil attribute z over the block B centered at $\mathbf{u}$ as a linear combination of neighboring point observations:

$$z_B^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) z(\mathbf{u}_\alpha) \qquad (6)$$

where the weights are chosen so as to minimize the estimation or error variance $\sigma_E^2(\mathbf{u}) = Var\{Z_B^*(\mathbf{u}) - Z_B(\mathbf{u})\}$ under the constraint of unbiasedness of the estimator. These weights are obtained by solving a system of linear equations, known as "block ordinary kriging system" [8, p.154]:

$$\begin{cases} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta(\mathbf{u}) \gamma(\mathbf{u}_\alpha - \mathbf{u}_\beta) - \mu(\mathbf{u}) = \bar{\gamma}(\mathbf{u}_\alpha, B(\mathbf{u})) & \alpha = 1, \ldots\ldots n(\mathbf{u}) \\ \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta(\mathbf{u}) = 1 \end{cases} \qquad (7)$$

where $\mu(\mathbf{u})$ is a Lagrange parameter. It is worth recalling that,like the two previous approaches, block kriging is valid only for linear averaging processes.

Using matrix notation, the vector of kriging weights is computed as:

$$\begin{bmatrix} [\lambda_\beta(\mathbf{u})]^T \\ \mu(\mathbf{u}) \end{bmatrix} = \begin{bmatrix} [\gamma(\mathbf{u}_\alpha - \mathbf{u}_\beta)] & [1]^T \\ [1] & 0 \end{bmatrix}^{-1} \begin{bmatrix} [\bar{\gamma}(\mathbf{u}_\alpha, B(\mathbf{u}))]^T \\ 1 \end{bmatrix}$$

Like the declustering weights in expression(4), the kriging weights account for data clustering through the data semivariogram matrix $[\gamma(\mathbf{u}_\alpha - \mathbf{u}_\beta)]$which informs the system on the redundancy of neighboring observations.Major differences are that:

1. The measure of redundancy depends on the pattern of spatial variability of observations instead of the mere Euclidian distance between observations.
2. The kriging weights also account for the proximity of observations to the center of the block through the point-to-block semivariogram values $\bar{\gamma} = (\mathbf{u}_\alpha, B(\mathbf{u}))$. Observations outside the block can thus be accounted for, although their influence will tend to be screened by those located inside the block.
3. The kriging weights are chosen so as to minimize the estimation variance; that minimum error variance is called the kriging variance and is computed as:

$$\sigma_B^2(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) \bar{\gamma}(\mathbf{u}_\alpha, B(\mathbf{u})) - \mu(\mathbf{u}) - \bar{\gamma}(B(\mathbf{u}), B(\mathbf{u})) \qquad (8)$$

Semivariogram Inference

The only information required by the kriging system and equation (8) is the point-to-point, point-to-block, and within-block semivariogram values: $\gamma(\mathbf{u}_\alpha - \mathbf{u}_\beta)$, $\bar{\gamma}(\mathbf{u}_\alpha, B(\mathbf{u}))$, and $\bar{\gamma}(B(\mathbf{u}), B(\mathbf{u}))$,respectively. Point-to-point semivariogram values are readily retrieved from the model fitted to the experimental semivariogram of point observations computed as:

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(h)} [z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})]^2 \qquad (9)$$

where $N(\mathbf{h})$ is the number of data pairs for a given separation vector $\mathbf{h}$. The semivariogram is but a measure of the average dissimilarity between data separated by a vector $\mathbf{h}$; as intuitively expected,the dis-

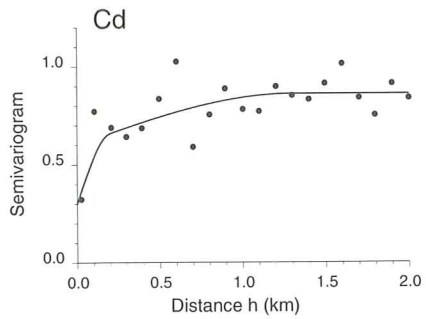

Figure 2. Experimental semivariogram of Cd concentrations (dots) and the model fitted (solid line).

similarity increases with the distance $|\mathbf{h}|$ between observations, see Figure 2.

The point-to-block semivariogram values are approximated by the arithmetic average of the point support semivariogram values $\gamma(\mathbf{u}_\alpha - \mathbf{u}'_i)$ defined between location $\mathbf{u}_\alpha$ and N points $\mathbf{u}'_i$ discretizing the block $B(\mathbf{u})$:

$$\bar{\gamma}(\mathbf{u}_\alpha, B(\mathbf{u})) \cong \frac{1}{N}\sum_{i=1}^{N}\gamma(\mathbf{u}_\alpha - \mathbf{u}'_i) \qquad (10)$$

Similarly, the within-block semivariogram value is approximated as the arithmetic average of the point semivariogram values $\gamma(\mathbf{u}'_i - \mathbf{u}'_j)$ defined between any two discretizing points $\mathbf{u}'_i$ and $\mathbf{u}'_j$ inside the block:

$$\bar{\gamma}(B(\mathbf{u}), B(\mathbf{u})) \cong \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\gamma(\mathbf{u}'_i - \mathbf{u}'_j) \qquad (11)$$

A rule of thumb is to select $N=(4)^K$ discretizing points, where K is the number of dimensions, 2 or 3, of the block [11,15].The level of discretization should also depend on the size of the block relative to the range of the semivariogram model. A good practice consists of computing quantities (10) and (11) for increasing numbers of N discretizing points: at some stage increasing the density of points will not significantly modify the results of the approximation.

Figure 3 illustrates the impact of the level of discretization ($\sqrt{N}$) on the block kriging results. Using only $N=9$ discretizing points leads to a severe overestimation of the kriging variance; the kriging estimate is much less sensitive to block discretization. In this example, a 8×8 grid yields good approximations at reasonable computational cost.

In theory, the block $B(\mathbf{u})$ can be any shape as long as the discretizing points are uniformly distributed within that block.However, for reasons of computational efficiency, most available programs of block kriging can handle only rectangular blocks.A straigthforward approach for estimating the average value of an irregularly shaped block consists of taking the linear average of point kriging estimates

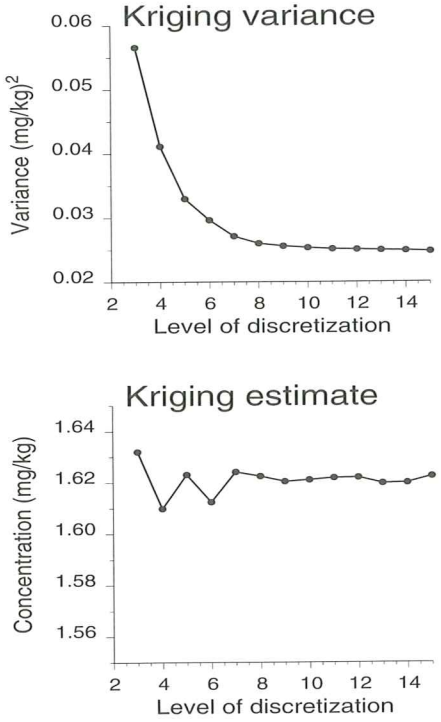

Figure 3. Impact of the level of discretization (square root of the number of gridded discretizing points) on the computation of the block kriging variance and estimate.

$z^*(\mathbf{u}_i)$ at N grid nodes $\mathbf{u}'_i$ discretizing the block. Provided the same data are used for the N point kriging systems, the so-called combination of kriged estimates [15 p.321) will provide the same estimate as block kriging using the same data and discretizing points. The main difficulty lies in the computation of the variance of the global estimator because it cannot be derived as a mere combination of the kriging variances at each discretizing point, see [15] p.410-412.

Confidence Interval

Assuming that the kriging prediction error is normally distributed, the block kriging estimate and variance can be combined to derive a confidence interval of type (2). For example,the 95% probability interval is typically computed as [16 p.68]:

$$[z^*_B(\mathbf{u}) - 2\sqrt{\sigma^2_{\mathrm{B}}(\mathbf{u})}, z^*_B(\mathbf{u}) + 2\sqrt{\sigma^2_{\mathrm{B}}(\mathbf{u})}] \qquad (12)$$

The main difference with the two previous approaches is that the estimation variance $\hat{\sigma}^2/n$ based on the sample variance (3) or (5) has been replaced by the model-based kriging variance (8) which is independent of the data values but account for the specific data configuration through the semivariogram model itself inferred from the data. So, besides the assumption of normality, the assumption of homoscedasticity must hold in order to compute the confidence inter-

val (12) [11], p.517-519.

If observations are spatially independent, the semivariogram value $\gamma(\mathbf{h})$ is constant and equal to the stationary variance $C(0)=\mathrm{Var}\{Z(\mathbf{u})\}$ whatever the separation vector $\mathbf{h}$ (pure nugget effect). All kriging weights are then equal to $1/n(\mathbf{u})$, and the block kriging estimate is but the arithmetic average of these $n(\mathbf{u})$ neighboring observations:

$$z_B^*(\mathbf{u}) = \frac{1}{n(\mathbf{u})} \sum_{\alpha=1}^{n(\mathbf{u})} Z(\mathbf{u}_\alpha) \qquad (13)$$

The kriging variance (8) becomes:

$$\sigma_B^2(\mathbf{u}) = \left[\sum_{\alpha=1}^{n(\mathbf{u})} \frac{C(0)}{n(\mathbf{u})} \right] - \mu(\mathbf{u}) - C(0)$$

$$= -\mu(\mathbf{u}) = \frac{C(0)}{n(\mathbf{u})} \qquad (14)$$

an expression similar to that used to compute the estimation variance of the arithmetic or declustered mean.Once again, the main difference is that the sample variance $\hat{\sigma}^2 = 0.585$, or better the declustered variance $\hat{\sigma}_D^2 = 0.933$, has been replaced by the model-based variance $C(0)=0.86$, which corresponds to the sill of the bounded semivariogram of Figure 2.

For the example of Figure 1, block kriging using only the observations inside the block ($n(\mathbf{u})=38$) and $N=64$ discretizing points yields a block estimate of 1.62 mg/kg, which is very close to the declustered mean $\hat{m}_D = 1.63$ mg/kg .The block kriging variance is 0.026 $(\mathrm{mg/kg})^2$, incidentally very close to the estimation variance for the declustered mean, 0.025=0.933/38.

An advantage of kriging over previous techniques is that observations outside the block can be accounted for in the prediction, which is particularly important for small blocks which usually contain too few observations for reliable estimation of the local (within-block) variance. In this case, it is more appropriate to rely on a global variance modeled from the entire data set, that is the sill of the bounded semivariogram model. The trade-off cost is the assumptionof stationarity of the variance across the study area.As the size of the block and the number of data increase, block kriging requires the solving of a large kriging system and knowledge of the semivariogram model for large distances. Moreover, the many observations within large blocks tend to screen the influence of outside data, in which case the additional complexity of the kriging approach might not be worth.

V. STOCHASTIC SIMULATION

In the last approach, the spatial distribution of the soil attribute across the region is first simulated, i.e. a simulated value is derived at each node of a grid that discretizes the region.The block simulated value is then computed as the arithmetic average of point simulated values within that block.

Many different simulation algorithms are available, and most of them are reviewed in [6,8].This paper considers only sequential indicator simulation because this algorithm does not require any multiGaussian assumption and allows one to account for class-specific patterns of spatial continuity (see hereafter). The algorithm proceeds as follows:

- Discretize the range of variation of the attribute z into $(K+1)$ classes using K threshold values z_k. Then, transform each datum $z(\mathbf{u}_\alpha)$into a vector of indicator data defined as:

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{u}_\alpha) \le z_k \\ 0 & \text{otherwise} \end{cases} \qquad k = 1,\ldots,K \qquad (15)$$

- For each threshold z_k, compute the semivariogram of the correspondingindicator data and model it.
- Define a random path visiting only once each node to be simulated.
- At each node $\mathbf{u}'$:
  1. Determine the K conditional probabilities $[F(\mathbf{u}'; z_k|(n))] = \mathrm{Prob}\{z(\mathbf{u}') \le z_k|(n)\}$ using ordinary indicator kriging. The conditioning information (n) consists of indicator transforms of neighboring original z-data and previously simulated z-values.
  2. Correct for any order relation deviations, then build a complete (for all z) conditional cumulative distribution function (ccdf) $F(\mathbf{u}'; z|(n))$ by interpolation/ extrapolation of the previously calculated K probability values.
  3. Draw a simulated value $z^{(l)}(\mathbf{u}')$ from that ccdf.
  4. Add the simulated value to the conditioning data set.
  5. Proceed to the next node along the random path, and repeat steps 1 to 4.

Other realizations $\{z^{(l')}(\mathbf{u}'_j),\ j=1,\ldots,N\}$, $l' \neq l$,are generated by repeating the entire sequential process with a different random path.

One hundred realizations of the spatial distribution of Cd values over the study area were generated using sequential indicator simulation and five threshold values corresponding to the 1st, 3rd, 5th, 7th and 9th deciles of the sample distribution of 259 cadmium data. The corresponding indicator semivariograms are displayed in Figure 4. The resolution of the discrete ccdf was increased by performing a linear interpolation between tabulated bounds provided by the sample histogram [6 p.134-137]. Figure 5 shows the first four realizations. Each realization is a plausible representation of the unique and unknown distribution of Cd

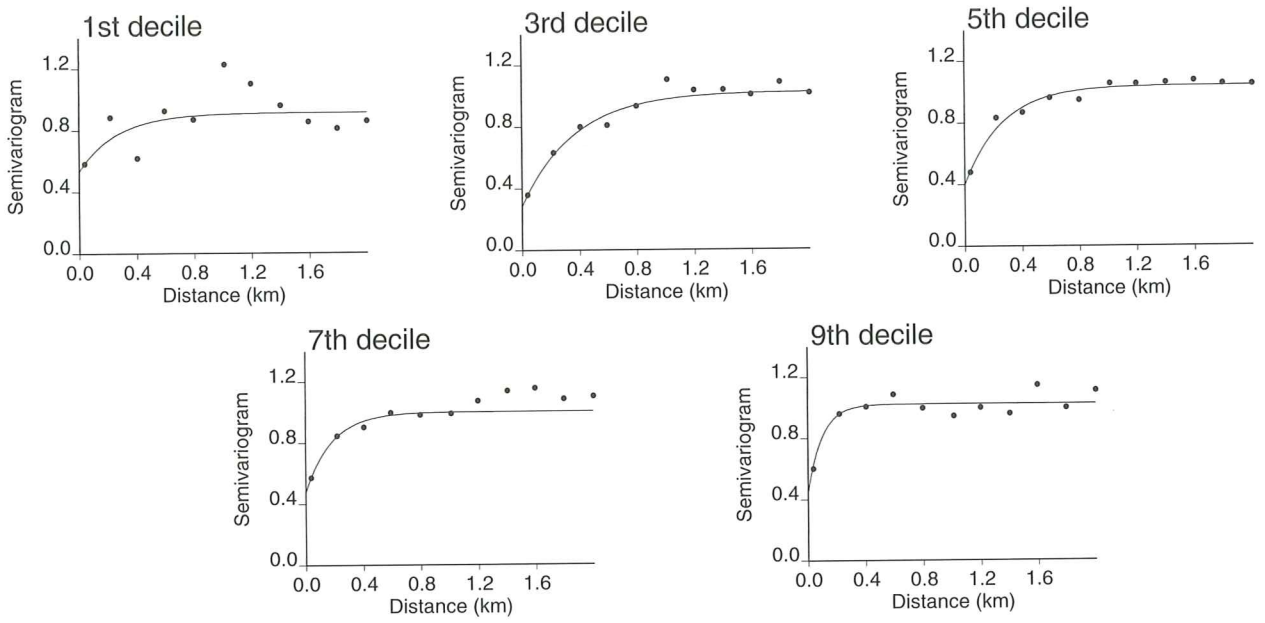

Figure 4. Standardized experimental indicator semivariograms computed for five deciles of the histogram of Figure 1.

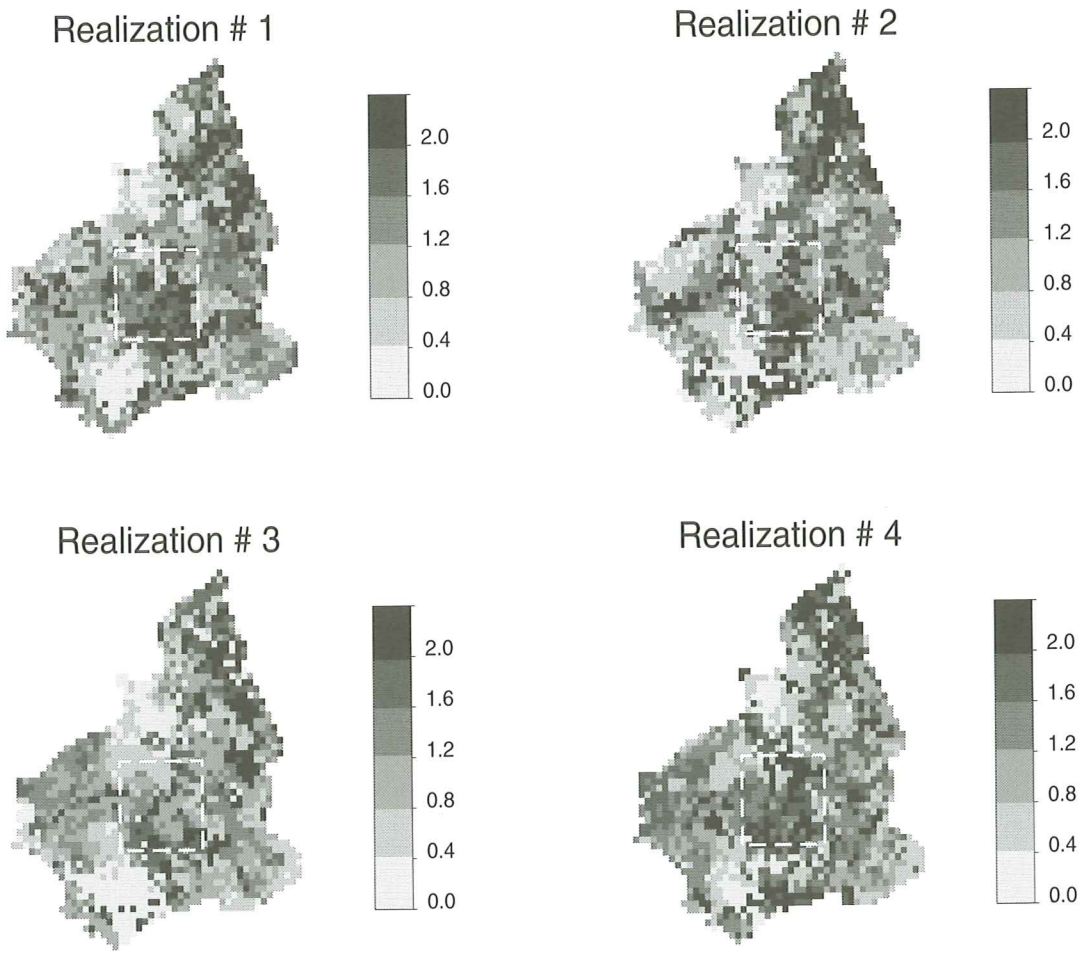

Figure 5. Four realizations of the spatial distribution of Cd values over the study area which were generated using sequential indicator simulation.

values across the region in that each simulated map honors the 259 data and reproduces approximately the sample histogram of Figure 1 and the indicator semivariogram models of Figure 4. Differences between the four realizations thus provide a measure of spatial uncertainty.

The simulated average Cd concentration within the 2km^2 block is computed as the arithmetic average of the J=225 simulated point values falling into it:

$$z_B^{(l)}(\mathbf{u}) = \frac{1}{J}\sum_{j=1}^{J} z^{(l)}(\mathbf{u}'_j) \qquad (16)$$

Figure 6 shows the histogram of the 100 simulated block values computed from the set of 100 simulated maps. This histogram depicts the uncertainty about the unknown average Cd concentration over the block. Such an uncertainty assessment is non-parametric: no prior assumption is made about the shape of the distribution of possible values. Note that, although each block value is computed as the linear average of many (J=225) simulated point values, the histogram is not symmetric.The uncertainty assessment is also independent of any particular"best" estimate retained for the unknown block value. Instead of a 95% confidence interval centered on the block Cd estimate, 95% probability intervals can be built by identifying the lower and upper bounds to the 2.5 and 97.5 percentiles of the distribution of simulated block values, respectively. For block Cd concentrations, that probability interval is [1.23, 1.76].

Simulation versus Block Kriging

Under the multiGaussian assumption, the distribution of simulated block values should be Gaussian with for mean and variance the block kriging estimate and variance. Thus, there would be no benefit in using stochastic simulation for deriving a block estimate and the attached uncertainty. However, whenever the multiGaussian assumption is inappropriate, as in the case of Cd concentrations, a non-parametric (i.e. indicator) approach must be adopted. Indicator block kriging cannot be used to model the uncertainty about the unknown block value since the indicator variable of type (15) is a non-linear transform of the original variable $z(\mathbf{u}_\alpha)$[8 p.305]. The only option is thus the use of non-parametric simulation algorithms, such as sequential indicator simulation.

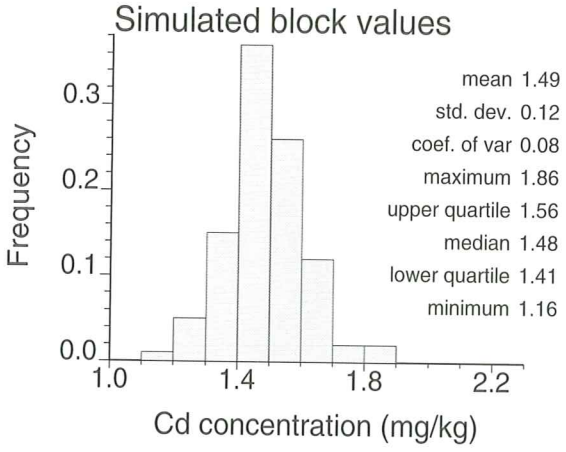

Figure 6. Histogram of block Cd concentrations computed from the set of 100 simulated maps.

Even under the multiGaussian assumption, the simulation approach is often a better alternative than block kriging for several reasons:

1. the shape of the block may be irregular, which requires either the modification of common block kriging programs or the discretization of the block and solving of punctual kriging systems at each discretizing point, recall previous discussion. In the latter case, the problem is the computation of the block kriging variance from the set of point kriging variances. Stochastic simulation allows one to compute this variance numerically from the distribution of block values.
2. some attributes such as permeability do not average linearly in space, making block kriging irrelevant. Non-linear upscaling is straightforward in a stochastic simulation approach since the averaging functions (e.g., geometric or harmonic mean) can be applied directly to the simulated point values inside the block.
3. certain applications require the determination of attribute values over many small blocks; the resulting map is then fed into transfer functions, such as flow simulator or runoff model, which heavily rely on the reproduction of the pattern of spatial dependence of block values. Since the simulated point values are spatially correlated, so are their local arithmetic averages.
4. risk assessment may involve the computation of the joint frequency of occurrence of events related to different blocks. For example, the probability that a regulatory threshold in soil pollution is simultaneously exceeded in two neighboring blocks B and B' is readily retrieved from L equiprobable realizations as the linear average of a product of indicator values:

$$\text{Prob}\{Z_B(\mathbf{u}) > z_c, Z_{B'}(\mathbf{u}') > z_c\} = \frac{1}{L}\sum_{l=1}^{L} i_B^{(l)}(\mathbf{u}; z_c)\cdot i_{B'}^{(l)}(\mathbf{u}'; z_c) \qquad (17)$$

where $i_B^{(l)}(\mathbf{u}; z_c) = 0$ if $i_B^{(l)}(\mathbf{u}) \le z_c$, and 1 otherwise.

The probability (17) cannot be easily inferred using non-simulation approaches, except if the two block values were deemed independent one from the other. In this particular case, the joint prob-

ability is but the product of the two block probability of exceeding the regulatory threshold.

VI. CONCLUSIONS

Prediction of environmental attributes over supports larger than the measurement support can be performed using a variety of approaches, ranging from the straightforward arithmetic average to the more demanding stochastic simulation. There is no such a thing as a "best" approach for all situations. Rather,the user should select the upscaling algorithm according to the objective of the study, the size of the block to be estimated, the sampling design, the type of averaging process involved, and the computational resources available.

Computation of the arithmetic mean offers a straightforward way to estimate a linear average attribute value over a large block provided the sampling scheme has been carefully designed to avoid any bias. In the common situation where specific subareas have been preferentially sampled, declustering techniques could be used to correct for the bias in the estimation of both mean and variance, keeping in mind that these techniques lack a firm statistical basis .

As the block gets smaller, observations inside the block usually do not suffice for a reliable estimation of the mean and variance,and it becomes critical to account for data outside the block.Provided the averaging process is linear, block kriging allows one to capitalize on the spatial correlation between attribute values to compensate for the lack of data inside the block. Block kriging does not require any specific sampling design since the kriging weights naturally correct for any redundancy of clustered observations.

For decision making, block estimates must be supplemented with a measure of the uncertainty attached to their prediction. Unlike the previous approaches, stochastic simulation provides a non-parametric assessment of that uncertainty, hence does not rely on any prior assumption about the shape of the distribution of block values. The trade-off cost is the computational demand for generating many realizations of the spatial distribution of point attribute values within the block(s). Also, stochastic simulation allows one to consider non-linear averaging processes.

An additional advantage of stochastic simulation is that it yields a set of maps of spatially correlated block values, thereby allowing statistical inference and decision making involving several blocks simultaneously. Information such as the joint probability of exceedence of a regulatory threshold in several blocks cannot be obtained from non-simulation approaches, except in the very particular case of independent block values or under stringent multiGaussian assumptions. Another application of stochastic simulation is the upscaling of soil infiltration properties in order to model solute transport in the vadoze zone. In such application, it is critical that the set of block values fed into the flow simulator reproduce the spatial variability actually prevailing in the field and modeled from field observations.

REFERENCES

[1] Atteia,O.,J.P. Dubois, and R. Webster. 1994. Geostatistical analysis of soil contamination in the Swiss Jura, *Environmental Pollution*, 86:315-327.

[2] Berthouex, P.M., and L.C. Brown. 1994. *Statistics for Environmental Engineers*, Lewis Publishers, Ann Arbor.

[3] Brus, D.J., and J.J. de Gruijter. 1997.Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion),*Geoderma*, 80:1-59.

[4] Burgess, T.M., and R. Webster. 1980. Optimal interpolation and isarithmic mapping of soil properties, II. Block kriging, *Journal of Soil Science*, 31:333-341.

[5] Christie, M.A., 1996. Upscaling for reservoir simulation, *Journal of Petroleum Technology*, 48:1004-1007.

[6] Deutsch, C.V., and A.G. Journel. 1998. *GSLIB: Geostatistical Software Library and User's Guide: second edition,* Oxford Univ. Press, New-York.

[7] Gómez-Hernández, J.J., 1991. *A Stochastic Approach to the Simulation of Block Conductivity Fields Conditioned upon Data Measured at a Smaller Scale*, PhD thesis, Stanford University, Stanford, CA.

[8] Goovaerts, P., 1997. *Geostatistics for Natural Resources Evaluation*, Oxford Univ. Press, New-York.

[9] Goovaerts, P., 1998. Geostatistics in soil science: state-of-the-art and perspectives, *Geoderma*, 89(1-2):1-45.

[10] Haldorsen, H., and E. Damsleth. 1990. Stochastic modeling, *Journal of Petroleum Technology*, 404-412.

[11] Isaaks, E.H. and R.M. Srivastava. 1989. *An Introduction to Applied Geostatistics*, Oxford Univ. Press, New York.

[12] Journel, A.G., 1983. Non-parametric estimation of spatial distributions, *Mathematical Geology*, 15:445-468.

[13] Journel, A.G., 1992. Geostatistics: roadblocks and challenges. In Geostatistics Tória '92, ed. A. Soares, Dordrecht: *Kluwer Academic Publisher*, 213-224.

[14] Journel, A.G., 1997. Geostatistics: tools for advanced spatial modeling in GIS, In Application of GIS to the Modeling of Non-point Source Pollutants in the Vadose Zone, ed. D.L. Corwin and K. Loague, *Special SSSA Publications*,39-55.

[15] Journel, A.G., and C.J. Huijbregts.1978. *Mining Geostatistics*. Academic Press, New York.

[16] Kitanidis, P.K., 1997. *Introduction to Geostatistics: Applications in Hydrogeology*, Cambridge University Press, New York.

[17] Kyriakidis, P.C., 1997. Selecting panels for remediation

in contaminated soils via stochastic imaging. In Geostatistics Wollongong '96, ed. E.Y. Baafi and N.A. Schofield, Dordrecht: *Kluwer Academic Publisher*, 973-983.

[18] McBratney, A.B., 1998. Some considerations on methods for spatially aggregating and disaggregating soil information,*Nutrient Cycling in Agroecosystems,* 50:51-62.

[19] Miller, S.M., 1997. Geostatistical simulation for upscaling field measurements of unsaturated hydraulic conductivity. In Geostatistics Wollongong '96, ed. E.Y. Baafi and N.A. Schofield, Dordrecht: *Kluwer Academic Publisher*, 1098-1111.

[20] Oliver, M.A., and R. Webster. 1991. How geostatistics can help you, *Soil Use and Management*, 7:206-217.

[21] Webster, R., 1996. What is kriging? *Aspects of Applied Biology*, 46:57-66.

[22] Webster, R.,O. Atteia, and J.P. Dubois, 1994. Coregionalization of trace metals in the soil in the Swiss Jura, *European Journal of Soil Science*, 45:205-218.

[23] Wen, X.H., and J.J. Gómez-Hernández.1996. Upscaling hydraulic conductivities in heterogeneous media: An overview,*Journal of Hydrology*, 183:9-32.

[24] Wood, G., Oliver, M.A. and Webster, R. 1990. Estimating soil salinity by disjunctive kriging, *Soil Use and Management*, 6:97-104.