Mathematical Geology, Vol. 30, No. 3, 1998

An Alternative to Cokriging for Situations with Small Sample Sizes[1]

K. C. Abbaspour,[2] R. Schulin,[2] M. Th. van Genuchten,[3] and E. Schläppi[4]

Lack of large datasets in soil protection studies and environmental engineering applications may deprive these fields of achieving accurate spatial estimates as derived with geostatistical techniques. A new estimation procedure, with the acronym Co_Est, is presented for situations involving primary and secondary datasets of sizes generally considered too small for geostatistical applications. For these situations, we suggest the transformation of the secondary dataset into the primary one using pedotransfer functions. The transformation will generate a larger set of the primary data which subsequently can be used in geostatistical analyses. The Co_Est procedure has provisions for handling measurement errors in the primary data, estimation errors in the converted secondary data, and uncertainty in the geostatistical parameters. Two different examples were used to demonstrate the applicability of Co_Est. The first example involves estimation of hydraulic conductivity random fields using 42 measured data and 258 values estimated from borehole profile descriptions. The second example consists of estimating chromium concentrations from 50 measured chromium data and 150 values estimated from a relationship between chromium and copper concentrations. The examples indicate that in situations where the size of the primary data is small, Co_Est can produce estimates which are comparable to cokriging estimates.

KEY WORDS: cokriging, small sample size, pedotransfer functions, geostatistics, parameter uncertainty, measurement error.

INTRODUCTION

Practical applications of geostatistical techniques to environmental studies and engineering projects have at times met with limited success, often because of small sample sizes. The nature of information in environmental studies, which may include the fields of soil protection, water resources management, hydro-

[1] Received 24 January 1997; revised 5 June 1997.

[2] Swiss Federal Institute of Technology, Department of Soil Protection, Grabenstrasse 3, 8952 Schlieren, Switzerland. e-mail: abbaspou@baikal.ethz.ch

[3] U.S. Salinity Laboratory, USDA, ARS, 450 West Big Spring Road, Riverside, California, 92507.

[4] Colombi Schmutz Dorthe AG, Konsumstrasse 20, 3007 Bern, Switzerland.

geology, geotechnical engineering, and environmental engineering, has features which are often very distinct from that used in mining applications. Some of these features are: (1) most environmental data exhibit spatial or temporal autocorrelation; using this information in the autocorrelation function leads to more efficient use of available data (e.g., Abbaspour *et al.*, 1996); (2) environmental data exhibit natural heterogeneity, a feature which should be preserved in any realistic analysis; (3) geostatistical parameters (i.e., mean, variance, range, nugget, and shape of the autocorrelation structure) in environmental studies are almost always uncertain; ignoring parameter uncertainty may lead to severe underdesigns in many projects (e.g., Abbaspour and others, 1996); (4) measured data in environmental applications usually contain nonnegligible measurement errors which should be accounted for; (5) collection of a large amount of data is often not feasible because of cost, time constraints, and the destructive nature of data collection techniques; (6) collected data are usually not the end product, but are used often as input to sophisticated simulation programs where the different uncertainties and errors in them can further propagate; and (7) different types of data are usually available with each type often having only a limited quantity of data. Realistic environmental applications require estimation procedures which account for all of the above special features.

In response to the above situation, procedures such as cokriging (Wackernagel, 1988; Myers, 1982, 1984; Journel and Huijbregts, 1978), indicator kriging (Journel, 1983), indicator cokriging (Deutsch and Journel, 1992), soft kriging (Alabert, 1987; Journel, 1986), and other methods (see Deutsch and Journel, 1992) were developed to provide better estimates of a primary variable using one or more auxiliary variables. A common drawback of all of these procedures is that inference becomes extremely demanding as the number of variables increases. For example, a two-variable cokriging approach requires a semivariogram for each variable and a cross-semivariogram, while for three variables we need a semivariogram for each variable plus three cross-semivariograms, and so on. The above techniques improve estimations by including auxiliary or soft data, but the main problem of having a small sample size is not alleviated in any way. Furthermore, although some workers have focused on the issue of measurement errors (Royer, 1989; Cressie, 1986) and parameter uncertainty (Cui, Stein, and Myers, 1995; Kitanidis, 1986), most available geostatistical programs do not have provision for taking either measurement errors or parameter uncertainty into account, and hence are of limited use for practical environmental applications.

The objective of this paper is to demonstrate a new approach, referred to as Co_Est, where auxiliary data are transformed into estimated primary data. Combining estimated and measured primary data generates a larger dataset which can then be used in kriging-type estimation applications. Regardless of the number of auxiliary variables, the procedure only needs to determine one semivariogram on the basis of the larger primary dataset. Each data point in the larger set is assumed to have an error. For measured primary data, this is the measurement error, and for estimated primary data, this becomes the estimation error. Estimated values and associated estimation errors can be provided by pedotransfer functions. Pedotransfer functions (PTFs) are regression equations or models which relate hard-to-measure field properties to more basic, and generally more easily-measured properties. Literature abounds with such equations which have been derived for different properties (e.g., Batjes, 1996; Salchow, Fausey, and Ward, 1996, Wösten, Finke, and Jansen, 1995; Abbaspour and Moon, 1992). The following are some of the scenarios that can conceivably use Co_Est: (1) If there are only a limited number of primary data, but reasonably more secondary data, some of which are collocated with the primary data, then a local pedotransfer function can be modeled on the basis of collocated data. The model can then be used to estimate values of the primary data at locations for which secondary data, but not primary data are available. (2) The same as the above situation, but with too few collocated primary and secondary data to establish a meaningful correlation; in this case pedotransfer functions from the literature can be calibrated for local conditions on the basis of the available data, and used to generate values of the primary parameter. (3) If there are no data on the primary parameter, but only different types of secondary parameters, then pedotransfer functions from the literature must be used to transfer secondary parameters into the primary parameter, with subsequent use of Co_Est to obtain an improved estimation for the primary parameter. We note that Co_Est also provides capabilities to treat geostatistical parameters (i.e., mean, variance, nugget, range, and shape of the semivariogram) as uncertain random variables. This feature allows analysis of parameter uncertainty, which is inherently associated with semivariogram modeling. Finally, Co_Est should not be considered as an alternative to cokriging if enough data are available to infer a reliable coregionalization model. Rather, Co_Est is being proposed as an alternative to a nonspatial data analysis.

THEORETICAL BACKGROUND

If a local pedotransfer function is being developed, the first step is to establish a relationship between a primary variable, Z, and t auxiliary variables, Y_1, Y_2, Y_t, using n collocated measured values, in the form of

$$Z_j = \beta_0 + \sum_{i=1}^{t} \beta_i Y_{ij} + \varepsilon_j \qquad j = 1, 2, \ldots, n \tag{1}$$

where β_0 and β_is are regression constants, and ε_i is a random variable with mean zero and variance σ^2 (Draper and Smith, 1981). The value of σ^2 can be estimated by $S^2_{Z.Y_1,Y_2,\ldots,Y_t}$ the residual mean square which is the variance of Z after taking into account the dependence of Z on Y_1, Y_2, . . . , Y_t (Zar, 1984).

In the second step, Equation (1) is used to obtain predictions, $\hat{Z}$, of Z, at m locations for which measured values of Y_i, but not of Z, are available. For a given set of Y_{i0}, the predicted point, $\hat{Z}_0$, will have a variance given by the following expression (Zar, 1984):

$$S^2_{\hat{Z}_0} = S^2_{Z.Y_1,Y_2,\ldots,Y_t}\left[1 + \frac{1}{n} + \sum_{i=1}^{t}\sum_{k=1}^{t} c_{ik}y_iy_k\right] \tag{2}$$

where $\Sigma\,\Sigma\, y_iy_k$ are known as *corrected* sums of products, and c_{ik} is an inverted matrix of sums of square and sums of products (for more details see Zar, 1984, Chap. 20). For a situation with only one auxiliary (independent) variable, Y_1, expression (2) reduces to

$$S^2_{\hat{Z}_0} = S^2_{Z.Y_1}\left[1 + \frac{1}{n} + \frac{(Y_{10} - \overline{Y})^2}{\sum (Y_j - \overline{Y})^2}\right] \tag{3}$$

where

$$\overline{Y} = \frac{1}{n}\sum_{j=1}^{n} Y_{1j} \tag{4}$$

The above two steps result in a set of n values of Z with errors equal to measurement errors, augmented by a set of m values of Z with errors equal to the estimation errors for a total of $m + n = p$ values.

In the third step, the combined set of measured and estimated Z values are used to model the semivariogram or covariance function for the primary variable. Assuming second-order stationarity, the two spatial measures are related by

$$C(h) = \tilde{\omega}_z - \gamma(h) \tag{5}$$

where $\tilde{\omega}_z$ is the variance of the random process, γ is the semivariogram, C is the covariance function, and h is the lag.

In the fourth step, a set of linear measurement relations (Bryson and Ho, 1975) is invoked to estimate a k-component state vector X given a p-component array of known Z values containing random errors ε, where usually $k >> p$. The problem is defined as follows:

$$\{Z\} = [H]\{X\} + \{\varepsilon\} \tag{6}$$

where $[H] = (p \times k)$ matrix of regression coefficients, $\{\varepsilon\} = (p \times 1)$ vector of zero-mean measurement errors, and $[R] = (p \times p)$ matrix of covariances for random errors $\{\varepsilon\}$. Following the application of a Bayesian solution (for details see Bryson and Ho, 1975, Chap. 12), the above procedure leads to conditionally posterior (superscript *cp*) estimated values as follows:

$$\{\mu\}^{cp} = \{\mu'\} + [C][H]^T([H][C][H]^T + [R])^{-1}(\{Z\} - [H]\{\mu'\}) \tag{7}$$

$$[C]^{cp} = [C] - [C][H]^T([H][C][H]^T + [R])^{-1}[H][C] \tag{8}$$

where $\{\mu'\} = (k \times 1)$ array of prior estimates of the states and $[C] = (k \times k)$ covariance matrix of the state variable. Equations (7) and (8) were also used by Massmann and Freeze (1987) to obtain conditional hydraulic conductivity fields. They are mathematically equivalent but not quite the same as the simple kriging estimator commonly used in geostatistics. In the simple kriging estimator, the theoretical mean must be used, whereas the measurement relations above allow for the use of subjective prior estimates.

Uncertainty Analysis

Co_Est provides the option to invoke an uncertainty analysis. Such an analysis requires that all uncertain geostatistical parameters be depicted probabilistically. A worst-case scenario, in terms of data availability, would be to treat each uncertain parameter as having a uniform distribution within a given interval. All geostatistical parameters, i.e., mean (μ), variance ($\tilde{\omega}$), nugget (ν), range (ρ), and shape (s), can be treated in this manner. The variable shape can be thought of as being one of s possible semivariograms [spherical, circular, exponential, . . .]. Another possible probabilistic approach is to treat the mean and variance of a random process as having a joint normal-gamma distribution, but independent of the nugget and range, which in turn are depicted by simple uniform distributions. Based on our experience, the latter model works well for hydraulic conductivity. The normal-gamma joint distribution for mean and variance is represented by three parameters, i.e., mean, variance, and an equivalent prior sample size n' (Benjamin and Cornell, 1970). It is usually possible in a particular problem to find a normal-gamma distribution which approximates reasonably well an experimenter's actual prior distribution of the mean and variance (DeGroot, 1975). Having treated parameters in the geostatistical model as random, estimates of state variables have compound distributions (Benjamin and Cornell, 1970) with parameters which are in turn random variables. These type of distributions are referred to as Bayesian distributions of X, and are defined as

$$f_X^b(x) = \int f_X(x|\theta) f_\Theta(\theta)\, d\theta \tag{9}$$

where $f_X(x|\theta)$ is a model distribution of X, and $f_\Theta(\theta)$ contains information about the parameters. A Bayesian distribution can be interpreted as being a weighted average of all possible distributions $f_X(x|\theta)$ which are associated with different values of θ. We note that the unknown parameters do not appear in $f_X^b(x)$ because they have been integrated out of the equation, and also that $f_X^b(x)$ accounts for both spatial heterogeneity and statistical uncertainty.

To propagate uncertainty in the parameters through Equations (7) and (8), a procedure referred to here as exhaustive stratified sampling was employed. From the earlier discussion of the joint distribution of the mean and variance, it follows that the marginal distribution of variance is chi-squared, while the conditional distribution of μ, $f(\mu|\bar{\omega})$ is normal. A series of $N_{\bar{\omega}}$, $N_{\mu|\bar{\omega}}$, N_ν, N_ρ, and N_s equally likely realizations of $\bar{\omega}$, $\mu|\bar{\omega}$, ν, ρ, and s, respectively, were selected. State variables X were consequently simulated for each combination of the $N_{\bar{\omega}}$, $N_{\mu|\bar{\omega}}$, N_ν, N_ρ, and N_s realizations, for a total of N_T equally likely realizations, $N_T = N_{\bar{\omega}} \times N_{\mu|\bar{\omega}} \times N_\nu \times N_\rho \times N_s$. From these runs the following statistics were calculated: $E(\mu)$, $\text{Var}(\mu)$, $E(\bar{\omega})$, and $\text{Var}(\bar{\omega})$.

Realizations of $\bar{\omega}$ were made by dividing the cumulative chi-squared scale range (from 0 to 1) into $N_{\bar{\omega}}$ equally sized classes (Fig. 1). Realizations of $\mu|\bar{\omega}$ were obtained by similarly dividing the cumulative normal distribution scale range of $f(\mu|\bar{\omega})$ into $N_{\mu|\bar{\omega}}$ equally sized classes. Realizations of ρ and ν were generated by dividing the interval between $\rho_{\min}$ and $\rho_{\max}$ and $\nu_{\min}$ to $\nu_{\max}$ into N_ρ and N_ν equally sized classes, respectively. In each situation, the first moment of each interval (dashed lines in Fig. 1) was taken to represent that interval.

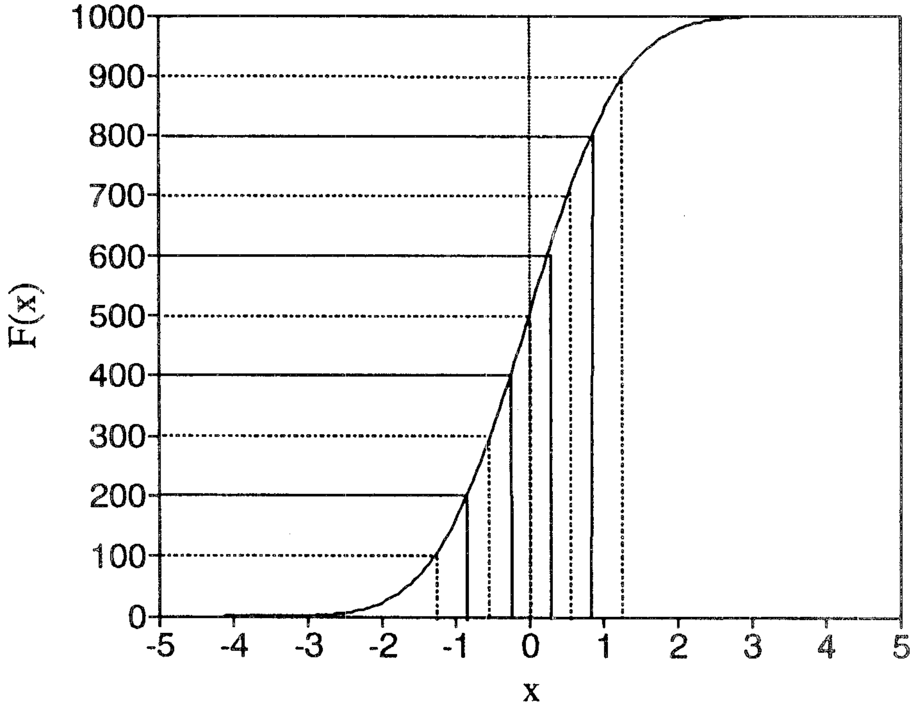


Figure 1. Division of a distributed parameter into equally-sized strata for exhaustive stratified sampling. Dashed lines locate the first moments of the strata.

EXAMPLES AND CO_EST RESULTS

Example 1: Saturated Hydraulic Conductivity of a Landfill Site

The first example involved a landfill site near Aarau, Switzerland. The landfill was used as a toxic waste repository and lies in a former clay pit excavated in aquitanian fresh water molasse. The molasse is composed primarily of marls and variegated clays interlayered with sandstone banks (Martinson, 1994). The site contained approximately 60 boreholes which had been drilled between 1970 and 1990 in an area of 800 m by 600 m (Fig. 2). Most boreholes were drilled to a depth of 30 m. Detailed lithologic descriptions were made for all boreholes (Fig. 3). Hydraulic conductivities were measured at several different depths in 17 boreholes, giving a total of 42 hard data points. Measurements were made over average vertical distances of 3 m using a double-packered technique. A frequency distribution of the hard conductivity data is shown in Figure 4a.

Boreholes with detailed profile descriptions were divided into 3 m intervals amounting to 298 sections. Of these, 42 sections were collocated with those for which hydraulic conductivities had been measured. The resulting set was used to establish the following relationship between hydraulic conductivity and

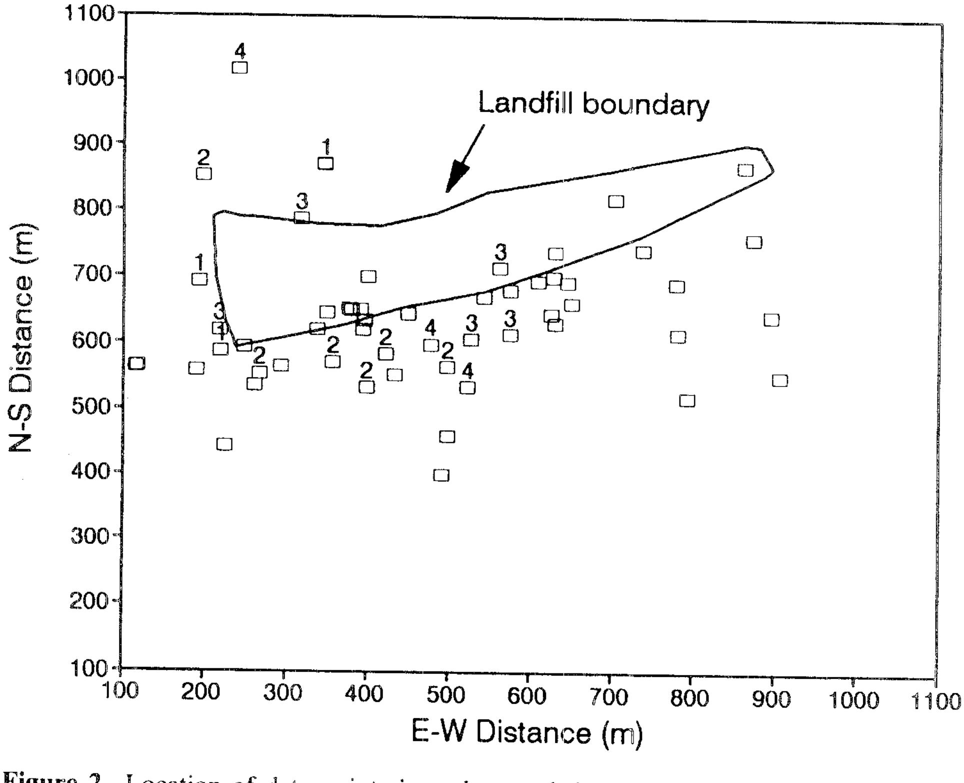


Figure 2. Location of data points in and around the landfill considered for example 1. The numbers indicate the number of measured data in the vertical direction.

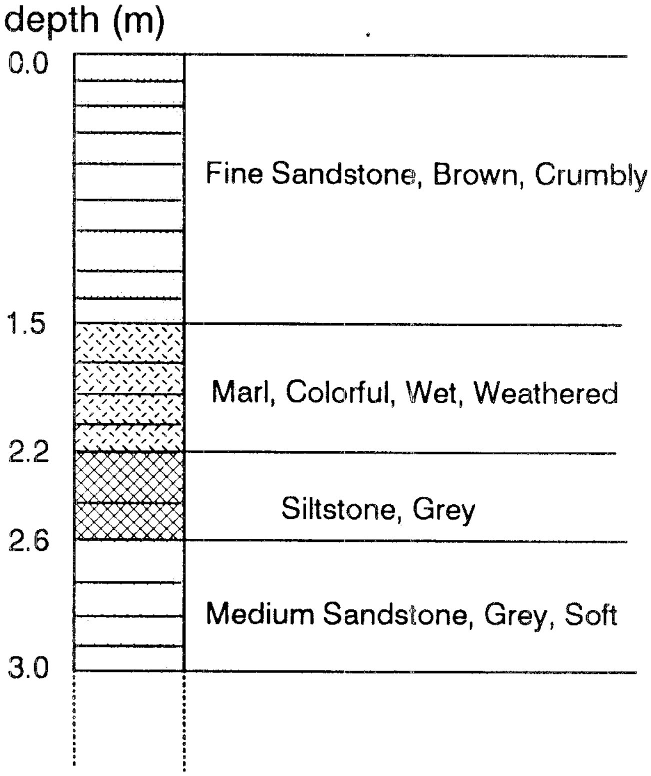
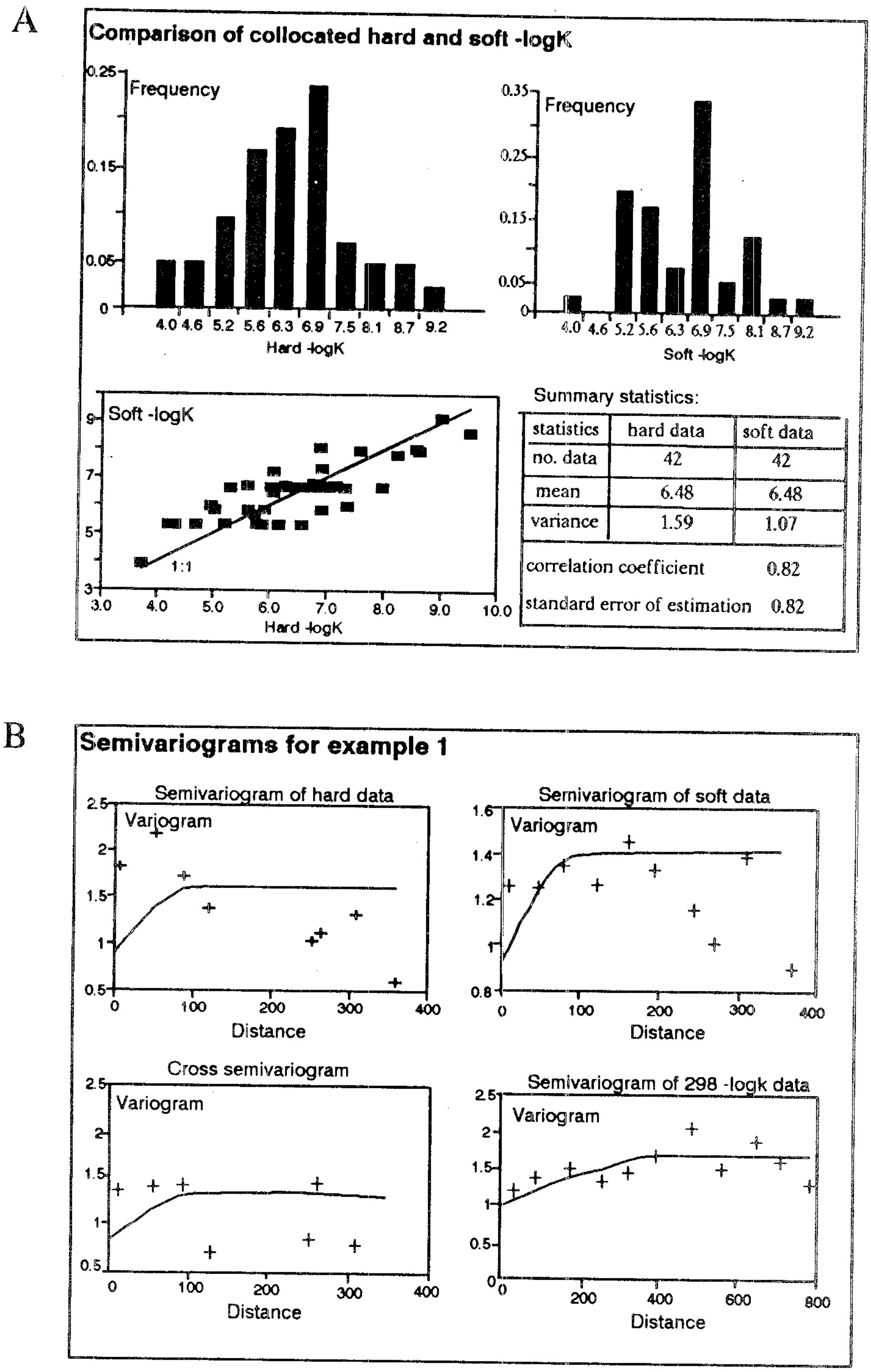

depth (m)
0.0
1.5
2.2
2.6
3.0
Fine Sandstone, Brown, Crumbly
Marl, Colorful, Wet, Weathered
Siltstone, Grey
Medium Sandstone, Grey, Soft

Figure 3. A typical segment of a chart used for profile description of a borehole.

borehole profile information (for more details of the regression procedure, see Abbaspour and Moon, 1992):

$$z = -\log K = 6.59 + 1.48\ \mathrm{FS} - 1.30\ \mathrm{MS} + 1.35\ \mathrm{SN} - 1.36\ \mathrm{CC} - 1.43\ \mathrm{CR1} + 1.16\ \mathrm{CY1} - 0.79\ \mathrm{CR2} \quad (10)$$

in which variables are defined as: FS = 1 if texture is fine sandstone, = 0 otherwise, MS = 1 if texture is medium sandstone, = 0 otherwise, SN = 1 if texture is siltstone, = 0 otherwise, CC = 1 if sandstone contains carbonate, = −1 if it does not, = 0 if texture is not sandstone, CR1 = 1 if fine sandstone is crumbly, = 0 otherwise, CY1 = 1 if fine sandstone is clayey, = 0 otherwise, and CR2 = 1 if medium sandstone is crumbly, = 0 otherwise. Equation (10) yielded a model correlation coefficient of 0.82, a standard error of estimation of 0.82, and a cross-validation correlation coefficient of 0.73. Figure 4a provides pertinent statistics for hard and soft datasets in this example. The correlation coefficient between hard and soft data generated by Equation (10) is significant, indicating a meaningful contribution from the borehole profile information to the estimation process. Using Equation (10), soft hydraulic conductivity data were calculated for the remaining 256 borehole sections.



statistics	hard data	soft data
no. data	42	42
mean	6.48	6.48
variance	1.59	1.07
correlation coefficient		0.82
standard error of estimation		0.82

Figure 4. A, pertinent statistics for the hard and soft data; and B, semivariograms used in the co-kriging and Co_Est estimations for example 1.

Next, we modeled four structure functions for use in Co_Est and cokriging as illustrated in Figure 4b. For each structure function, the two North–South and East–West directions with tolerances ±22.5° and the omnidirectional curve (0° ± 90°) were considered (only the omnidirectional direction is shown in Figure 4b along with the model fit). For cokriging we modeled two semivariograms for hard (γ_H) and soft (γ_S) data and one cross-semivariogram (γ_{HS}) by the linear model of coregionalization given in Equation 11. For Co_Est we only need one semivariogram and the model γ_Z in Equation 12 was inferred on the basis of experimental semivariograms calculated using the combined 42 hard and 256 soft data points.

$$\begin{cases} \gamma_H(h) = 0.9 + 0.7\text{Sph}(h/100) \\ \gamma_S(h) = 0.9 + 0.5\text{Sph}(h/100) \\ \gamma_{HS}(h) = 0.8 + 0.5\text{Sph}(h/100) \end{cases} \tag{11}$$

$$\gamma_Z = 1.0 + 0.7\text{Sph}(h/500) \tag{12}$$

where

$$\text{Sph}(h/\rho) = \begin{cases} 1.5\dfrac{h}{\rho} - 0.5\left(\dfrac{h}{\rho}\right)^3, & \text{if } 0 < h \leq \rho \\ 1, & \text{if } h > \rho \end{cases} \tag{13}$$

where h is the lag distance, and ρ is the range parameter. The determinants of the nugget coefficients and the spherical structure factors are positive, which indicates the model in Equation 11 is legitimate. Experimental semivariograms for hard and soft datasets based on 42 data points are very noisy (Figure 4b). This phenomenon is mostly due to the small number of pairs. On the other hand, the experimental semivariogram based on the complete set of 298 (42 measured plus 256 estimated from borehole descriptive variables) values for the logarithm of hydraulic conductivity is much better defined. The programs in GSLIB (Deutsch and Journel, 1992) were used to obtain semivariograms and to perform cokriging estimation. Measurement errors in this and the next example were set to zero as cokriging routines generally do not have any provisions for taking these errors into account.

Co_Est and cokriging were used to estimate a set of 42 data points by cross-validation, and compared these to the 42 measured data. Comparisons between the measured and estimated data were obtained by calculating the root mean square error (RMSE) to show the closeness of the two datasets, by determining the coefficient of correlation to show linear correlation between the two datasets, and by calculating the average estimation (kriging) variance to show the effect of data configuration.

Table 1. Comparison of Estimation Results Obtained with Cokriging and Co_Est (Example 1)

Statistics	Measured −logK	Cokriging results	Co_Est results	Co_Est with uncertainty analysis
Number of samples	42	42	42	5000 fields of 42 points
Mean, μ_Z (ms^{-1}), and its variance ( )	6.48	6.61	6.60	6.51 (0.047)
Variance, $\tilde{\omega}_Z$ and its variance ( )	1.59	0.51	0.39	1.57 (0.16)
Average estimation error, $\tilde{\omega}_E$, and its variance ( )	—	1.12 (0.002)	1.10 (0.0001)	— —
Root mean square error	—	1.25	1.10	—
Correlation coefficient	—	0.31	0.50	—

Results obtained with the Co_Est and cokriging runs are compared in Table 1. As expected, the mean of the random process, μ, is estimated relatively closely by both procedures. As is typical of linear estimation procedures, the variance of the random process, or the system heterogeneity, $\tilde{\omega}$, is not conserved. The estimation (kriging) variance, $\tilde{\omega}_E$, is a measure of the data configuration, and not of the local accuracy at a specific location, because $\tilde{\omega}_E$ does not depend on the actual data value (Journel, 1987). Because co-kriging and Co_Est both use the same hard and soft data configuration, estimation error is almost the same. Moreover, the RMSE and correlation coefficient in this situation are significantly better for Co_Est than for cokriging.

To alleviate the smoothing effects of linear estimation procedures, different stochastic simulation algorithms have been proposed (see Deutsch and Journel, 1992 for a comprehensive list). The Co_Est algorithm is particularly suited for simulation purposes and uncertainty analyses, since this method can consider all spatial parameters, i.e., mean, variance, nugget, range, and shape of the spatial structure, as being uncertain variables. Parameter uncertainties can then be propagated through a stochastic simulation model by drawing alternative, equally likely, and conditioned joint realizations of the random variables. In this example, the parameter uncertainty analysis option of Co_Est was invoked.

The mean and variance of the 42 measured −logK values were 6.48 and 1.59, respectively. The local geologists' prior belief was that the mean of the saturated hydraulic conductivity in a fresh water molasse zone typically found in Switzerland should be about twice as likely to lie inside the range 6.48 ± 1.0 than outside this range. This information suggested that the parameter n' (equivalent prior sample size) (Benjamin and Cornell, 1970) of the joint normal-gamma conjugate prior distribution should be taken as about 13. In consultation with local geologists, we decided to set the interval for the uncertainty in the nugget, ν, to [0.5, 1.5] and for the range, ρ, to [0.0, 1000]. By setting $N_{\mu|\omega}$ and N_{ω} to 10, N_{ρ} and N_{ν} to 5, and N_s to 2, (i.e., s = [spherical, exponential]) we hence obtained a sample space consisting of 5000 sets of parameters. Consequently, a total of 5000 realizations of the hydraulic conductivities were made at each of the 42 data locations. These realizations provided a numeric solution to the Bayesian distribution of Equation (9) for both the mean and the variance of the random process. After obtaining the cumulative distribution of both the mean and the variance of z, we again invoked the exhaustive stratified sampling procedure to obtain a set of 100 simulated Z values at each of the 42 points. Results of the uncertainty run are also summarized in Table 1, where statistics are based on 100 realizations. Here the variance of the random process is much larger than for the two estimation methods because of propagation of parameter uncertainties. Implications of this result are twofold. First, uncertainty in parameters in environmental projects does not need to be a hindrance for analysis. If properly quantified and propagated, parameter uncertainty can alleviate the often unrealistic and undesirable smoothing effect of traditional estimation procedures. Second, ignoring parameter uncertainty in light of its large effect on the variance may not be justifiable in practical environmental applications. It was shown elsewhere (Abbaspour and others, 1996) that ignoring parameter uncertainty can lead to severe underdesigns in geotechnical projects.

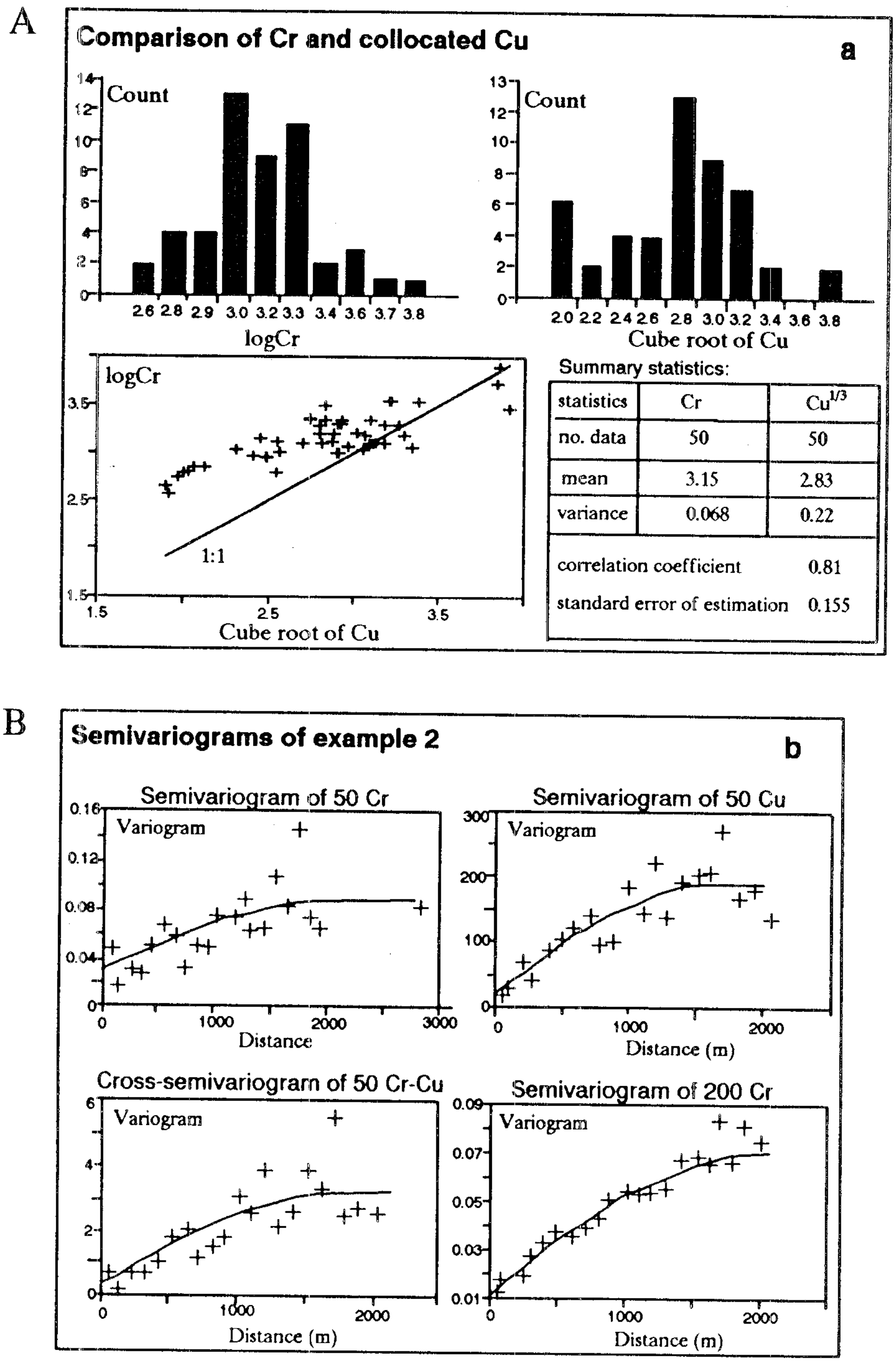
Example 2: Chromium Concentration in an Industrial Region

The second dataset consists of chromium and copper concentrations measured in the top (0–20) cm depth of an industrial region near Zurich, Switzerland. The primary variable in this example is chromium concentration, of which 120 measured points were available. The secondary data variable is concentration of copper, with 200 measurements being available, out of which 50 were collocated with chromium. The 50 collocated chromium and copper concentrations were used to establish the following pedotransfer function:

$$\log\mathrm{Cr} = 1.88 + 0.45\ \mathrm{Cu}^{1/3} \qquad (14)$$

Equation (14) yielded a model correlation coefficient of 0.81, and a standard error of estimation equal to 0.155. Figure 5a provides pertinent statistics for the two datasets of this example.

The remaining 150 copper data were used as soft data for estimation by cokriging, and also used to obtain estimated (soft) logCr data using the relationship in Equation 14 for estimation by Co_Est. The remaining 70 chromium data were retained for testing the estimates made by cokriging and Co_Est. As in the previous example, four structure functions were modeled as illustrated in Figure 5b, where only the omnidirectional experimental cases are shown along



statistics	Cr	$Cu^{1/3}$
no. data	50	50
mean	3.15	2.83
variance	0.068	0.22

correlation coefficient	0.81
standard error of estimation	0.155

Figure 5. A, pertinent statistics for the hard and soft data; and B, the semivariograms used in the co-kriging and Co_Est estimations for example 2.

with model fits. One semivariogram, γ_A, was modeled based on the combined 50 hard plus 150 soft chromium data for use in Co_Est (Eq. 15); two semivariograms, γ_{Cr} and γ_{Cu}, were modeled based on sets of 50 measured Cr and Cu data, respectively; and one cross-semivariogram, $\gamma_{Cr\text{-}Cu}$, was modeled based on the 50 measured Cr and Cu data (Eq. 16).

$$\gamma_A = 0.011 + 0.06\mathrm{Sph}(h/2000) \tag{15}$$

$$\begin{cases} \gamma_{Cr}(h) = 0.029 + 0.059\mathrm{Sph}(h/1800) \\ \gamma_{Cu}(h) = 16.7 + 174.3\mathrm{Sph}(h/1800) \\ \gamma_{Cr\text{-}Cu}(h) = 0.252 + 2.948\mathrm{Sph}(h/1800) \end{cases} \tag{16}$$

where the Sph() function is defined in Equation 13. Determinants of the nugget coefficients and the spherical structure factors are positive, indicating that the linear model of coregionalization in Equation 16 is also legitimate. As in the previous example, experimental semivariograms for Cr and Cu which are based on 50 data points are noisy, a phenomenon mostly due to the small number of pairs. By comparison, the experimental semivariogram based on the set of 200 chromium (50 measured data plus 150 data estimated from copper) data points is much better defined. Table 2 shows how well cokriging and Co_Est were able to estimate the 70 measured logCr concentrations. As in the previous example, the mean of chromium concentrations obtained with the two procedures were very close, whereas the variance was considerably underestimated. The average estimation (kriging) variance was much smaller with Co_Est, primarily because of the considerably smaller nugget and sill of the Co_Est semivariogram. Estimation based on Co_Est in this example also provided a smaller value of RMSE. Values of the correlation coefficient for cokriging and Co_Est were almost identical. Both examples suggest that the precision obtained by a better

Table 2. Comparison of Estimation Results Obtained with Cokriging and Co_Est (Example 2)

Statistics	Measured logCr	Cokriging estimation results	Co_Est estimation results
Number of samples	70	70	70
Mean, μ_Z (mg kg^{-1})	3.2	3.34	3.17
Variance, $\tilde{\omega}_Z$	0.083	0.050	0.038
Average estimation error, $\tilde{\omega}_E$,	—	0.042	0.022
and its variance ( )		(3.4E-5)	(2.4E-5)
Root mean square error	—	0.24	0.19
Correlation coefficient	—	0.74	0.75

semivariogram because of a larger number of data points in Co_Est, more than offsets the additional errors associated with the use of soft auxiliary data.

CONCLUSIONS

Use of a new procedure, Co_Est, for spatial estimation was considered for situations which are amenable to cokriging, but have an insufficient number of samples to make optimal use of geostatistics. Different types of data were correlated through regression analysis, and combined to form a larger set of data for the primary variable. Co_Est reduced to one the number of semivariograms needed for estimation, i.e., that of the primary variable only. Co_Est is also capable of performing stochastic simulation with uncertain parameters. Two very different examples were considered so as to illustrate use of Co_Est. For both examples, improvements could be made in the estimation process with Co_Est as compared to the traditional cokriging. Variance of a random process can be very much influenced by parameter uncertainty. This feature can have important implications in actual environmental applications. In conclusion, when the size of the primary data is too small to perform traditional geostatistical analyses, Co_Est can provide improved estimations. Co_Est is particularly useful for stochastic simulations if parameter uncertainty is to be considered.

ACKNOWLEDGMENTS

The authors wish to express their gratitude to R. Meuli for providing the data for example 2 and for fruitful conversations.

REFERENCES

Abbaspour, K. C., and Moon, D. E. 1992, Relationships between conventional field information and some soil properties measured in the laboratory: Geoderma, v. 55, no. 2, p. 119–140.

Abbaspour, K. C., Schulin, R., Schläppi, E., and Flühler, H., 1996, A Bayesian approach for incorporating uncertainty and data worth in environmental projects: Environmental Modeling and Assessment, v. 1, no. 3, p. 151–158.

Alabert, F., 1987, Stochastic Imaging of Spatial Distribution Using Hard and Soft Information: Unpubl. master's thesis, Stanford Univ., Stanford, CA, 185 p.

Batjes, N. H., 1996, Development of a world data set soil water retention properties using pedotransfer rules: Geoderma, v. 71, no. 1, p. 31–52.

Benjamin, J. R., and Cornell, C. A., 1970, Probability, statistics, and decision for civil engineers: McGraw-Hill Book Company, New York, 684 p.

Bryson, A. E., and Ho, Y. C., 1975, Applied optimal control: John Wiley & Sons, New York, 481 p.

Cressie, N., 1986, Kriging non-stationary data: Jour. American Statistical Assoc., v. 81, no. 2, p. 625–634.

Cui, H., Stein, S., and Myers, D. E., 1995, Extension of spatial information, Bayesian kriging and updating of prior variogram parameters: Environmetrics, v. 6, no. 4, p. 373–384.

DeGroot, M. H., 1975, Probability and statistics: Addison-Wesley Publishing Company, Reading, Massachusetts, 607 p.

Deutsch, C. V., and Journel, A. G., 1992, GSLIB Geostatistical software library and user's guide: Oxford University Press, New York, 340 p.

Draper, N. R., and Smith, H., 1981, Applied regression analysis: John Wiley & Sons, New York, 709 p.

Journel, A., 1983, Non-parametric estimation of spatial distributions: Math. Geology, v. 15, no. 3, p. 445–468.

Journel, A., 1986, Constrained interpolation and qualitative information: Math Geology, v. 18, no. 3, p. 269–286.

Journel, A. G., 1987, Geostatistics for the environmental sciences: EPA Project no. CR 811893, Technical Report, US EPA, EMS Lab, Las Vegas, 135 p.

Journel, A. G., and Huijbregts, C. J., 1978, Mining geostatistics: Academic Press, London, 600 p.

Kitanidis, P. K., 1986, Parameter uncertainty in estimation of spatial functions: Bayesian analysis: Water Resources Res., v. 22, no. 4, p. 499–507.

Martinson, C., 1994, Geochemical interactions of a saline leachate with Molasse at a landfill site: A case study: Eclogae Geological Helvetiae, v. 87, no. 2, p. 473–486.

Massmann, J., and Freeze, R. A., 1987, Groundwater contamination from waste management sites: The interaction between risk-based engineering design and regulatory policy, 1. Methodology: Water Resources Res., vol. 23, no. 2, p. 351–367.

Myers, D., 1982, Matrix formulation of co-kriging: Math Geology, v. 14, no. 3, p. 249–257.

Myers, D., 1984, Cokriging—new developments, *in* Verly G., *et al.*, eds., Geostatistics for natural resources characterization: Reidel Publishing, Dordrecht, p. 295–305.

Royer, J. J., 1989, Multivariate geostatistics and sampling problems, *in* Armstrong, M., ed., Geostatistics: Kluwer Academic Publishers, Dordrecht, p. 823–836.

Salchow, E., Lal, R., Fausey, N. R., and Ward, A., 1996, Pedotransfer functions for variable alluvial soils in southern Ohio: Geoderma, v. 73, no. 3, p. 165–181.

Wackernagel, H., 1988, Geostatistical techniques for interpreting multivariate spatial information, *in* Chung C., Fabbri, A. G., and Sinding-Larsen, R., eds., Quantitative analysis of mineral and energy resources: Reidel Publishing, Dordrecht, p. 393–409.

Wösten, J. H. M., Finke, P. A., and Jansen, M. J. W., 1995, Comparison of class and continuous pedotransfer functions to generate soil hydraulic properties. Geoderma, v. 66, no. 3, p. 227–237.

Zar, J. H., 1984, Biostatistical analysis: Prentice-Hall, Englewood Cliffs, NJ, 718 p.