

## REVIEW ARTICLE

P. Goovaerts

## Geostatistical tools for characterizing the spatial variability of microbiological and physico-chemical soil properties

Received: 6 January 1998

**Abstract** This paper reviews the main applications of geostatistics to the description and modeling of the spatial variability of microbiological and physico-chemical soil properties. First, basic geostatistical tools such as the correlogram and semivariogram are introduced to characterize the spatial variability of each attribute separately as well as their spatial interactions. Then, the key issue of fitting permissible models to experimental semivariograms is addressed for the univariate and multivariate situations. Capitalizing on this model of spatial dependence, the value of a soil property can be predicted at unsampled locations using only observations of this particular property (kriging) or incorporating additional information provided by other correlated properties (cokriging). Factorial kriging allows one to discriminate the different sources of spatial variation in soil on the basis of the scale at which they operate, and it often enhances relations between soil attributes which were blurred in a traditional correlation analysis where the different sources of variations are mixed. Geostatistics can also be used to assess the risk of exceeding critical values (regulatory thresholds, soil quality criterion) at unsampled locations, and to simulate the spatial distribution of attribute values. All the different tools are illustrated using two transects of 100 pH and electrical conductivity values measured in pasture and forest.

**Key words** Geostatistics · Spatial variability · Kriging · Scale-dependent correlation · Stochastic simulation

### Introduction

Soil scientists are aware that soil properties vary spatially, and they have already recorded strong fluctuations even over short distances (Trangmar et al. 1985; Warrick et al. 1986). Behind a locally erratic aspect, some spatial structure is often discerned and may be related to the combined action of several physical, chemical or biological processes that act at different spatial scales. The characterization of the spatial variability of soil attributes is essential to achieve a better understanding of complex relations between soil properties and environmental factors. Also, a model of spatial dependence between soil data can later be used to estimate attributes at unsampled locations, leading, for example, to better recommendations for the application of fertilizers or pesticides.

Geostatistics provides a set of statistical tools for incorporating the spatial coordinates of soil observations in data processing, allowing for description and modeling of spatial patterns, prediction at unsampled locations, and assessment of the uncertainty attached to these predictions. Since the publication of the first applications of geostatistics to soil data in the early 1980s (Burgess and Webster 1980a, b; Webster and Burgess 1980; Burgess et al. 1981), geostatistical methods have become popular in soil science, as illustrated by the increasing number of studies reported in the literature (Goovaerts 1998). The use of geostatistics is not confined to physico-chemical soil properties, but studies have also been conducted on the spatial patterns of soil micro-organisms (Wollum and Cassel 1984; Webster and Boag 1992), plants (Vieira et al. 1983; Sutherland et al. 1991), and other organisms living at the surface of the soil (Rossi et al. 1992).

The greater accessibility of geostatistical software (e.g., Englund and Sparks 1991; Pannatier 1996; Deutsch and Journel 1998) has increased the risk that geostatistical tools are used without a good command of the underlying theory, in particular in the field of

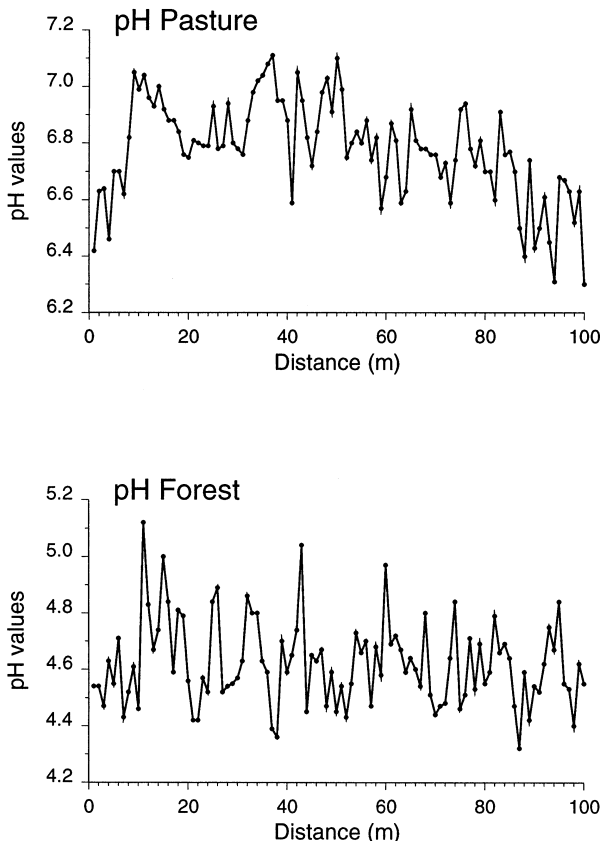
P. Goovaerts  
Department of Civil and Environmental Engineering,  
The University of Michigan, EWRE Bldg, Room 117,  
Ann Arbor, MI 48109-2125, USA  
e-mail: goovaert@engin.umich.edu, Tel.: +001-734-9360141,  
Fax: +001-734-7632275

semivariogram modeling. Moreover, in several situations, inappropriate or non-optimal approaches are adopted because scientists are unaware of recent developments, such as factorial kriging, indicator geostatistics or stochastic simulation.

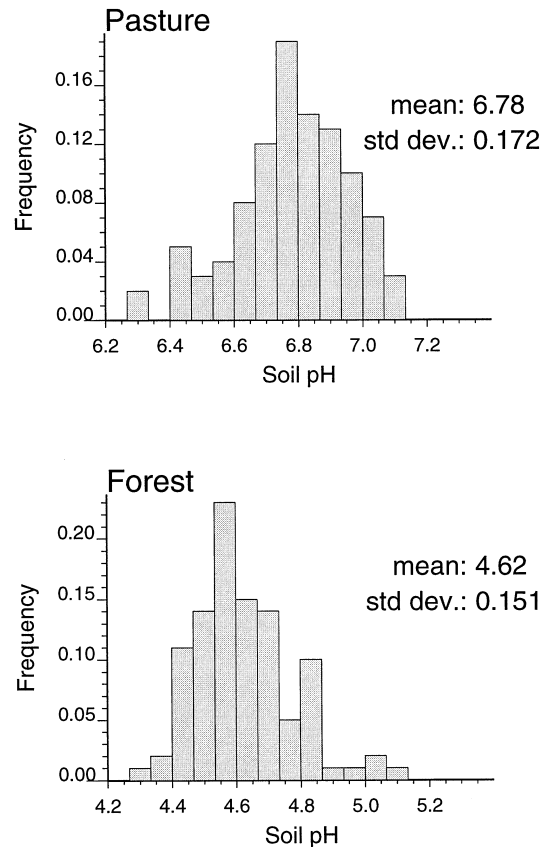
This paper aims at providing soil scientists with an applied introduction to basic descriptive and predictive geostatistical tools such as the semivariogram and kriging, while presenting recent developments in multivariate geostatistics and the modeling of local and spatial uncertainty. The different tools will be used for a step-by-step analysis of a one-dimensional set of soil data, which consists of two transects of 100 topsoil pH and electrical conductivity values measured every meter in pasture and forest.

### Description of spatial patterns

Figure 1 shows two series of 100 topsoil pH values measured every meter along transects, one in a pasture and the other in a forest (Goovaerts et al. 1989). Too often, analysis of such data amounts to plotting the histogram and computing summary statistics such as mean and standard deviation (Fig. 2). By so doing, one ignores critical information, that is the spatial location of pH measurements. For example, the comparison of



**Fig. 1** Transects of 100 pH values measured in the topsoil of a pasture and a forest plot



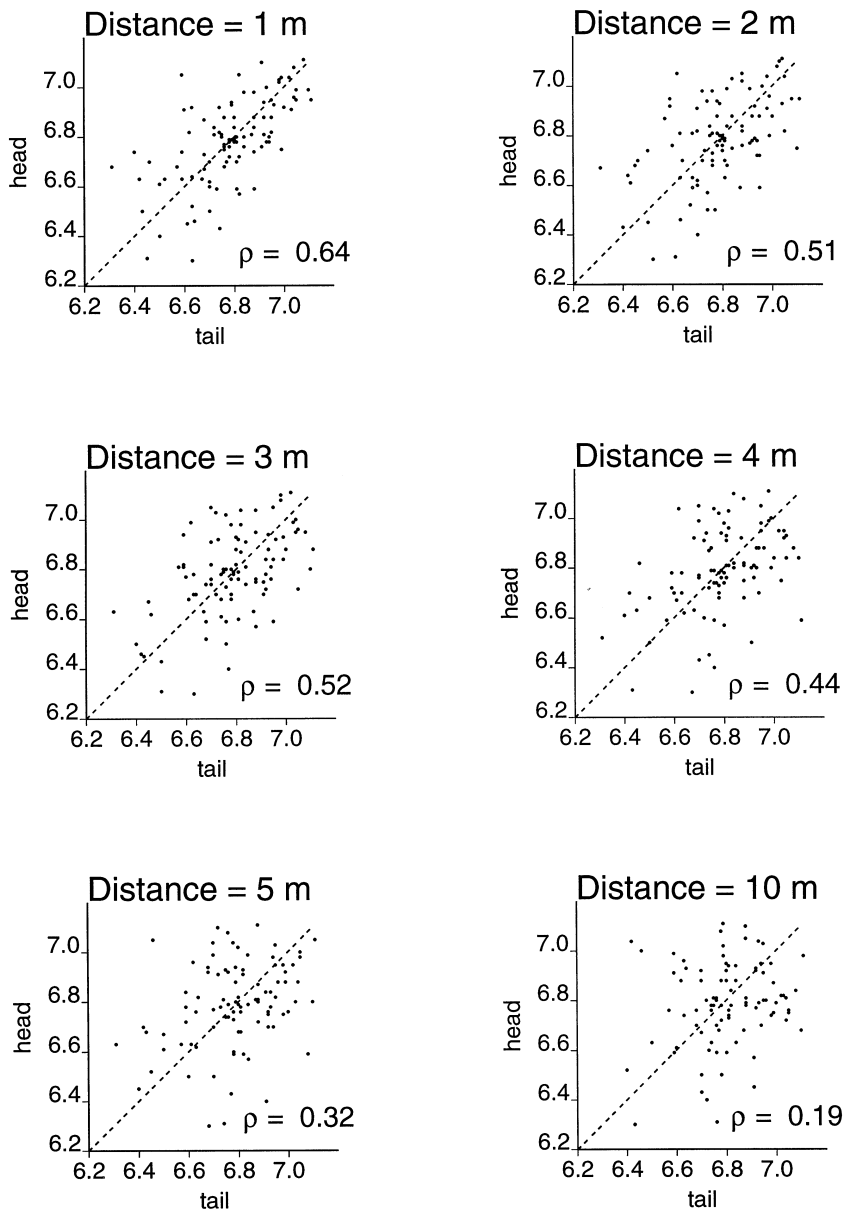
**Fig. 2** Histograms of soil pH values measured in a pasture and a forest plot

standard deviations suggests that the variability of pH is similar for both land uses. The shape of pH transects indicates, however, that pH values vary over shorter distances in the forest than in the pasture! This spatial feature, which may be important for interpretation, is not captured if one ignores the spatial information. Geostatistics provides a set of statistical tools for detecting and quantifying the major scales of spatial variability.

### The correlogram

Let  $z(\mathbf{u}_\alpha)$ ,  $\alpha=1, 2, \dots, n$  denote the set of  $n=100$  pH values measured in a pasture, where  $\mathbf{u}_\alpha$  is the vector of spatial coordinates of the  $\alpha$ th observation. Figure 1 (top graph) shows that the distribution of pH values along the transect is not fully random in that observations that are close to each other on the ground tend to be more alike than those further apart. The similarity between adjacent pH values can be depicted by plotting each observation  $z(\mathbf{u}_\alpha)$  versus the one measured 1 m away,  $z(\mathbf{u}_\alpha + \mathbf{h})$  with  $|\mathbf{h}| = 1$  m. In this example, 99 pairs of pH measurements ( $z(\mathbf{u}_\alpha)$ ,  $z(\mathbf{u}_\alpha + \mathbf{h})$ ) can be formed from the initial set of 100 values, and the resulting plot is called an **h-scattergram** (Fig. 3; left top graph). Readers are likely to be familiar with the scattergram, which

**Fig. 3** Scattergrams between soil pH values separated by a distance of 1 m, 2 m, ..., 10 m. Note how the correlation between observations decreases as the separation distance increases

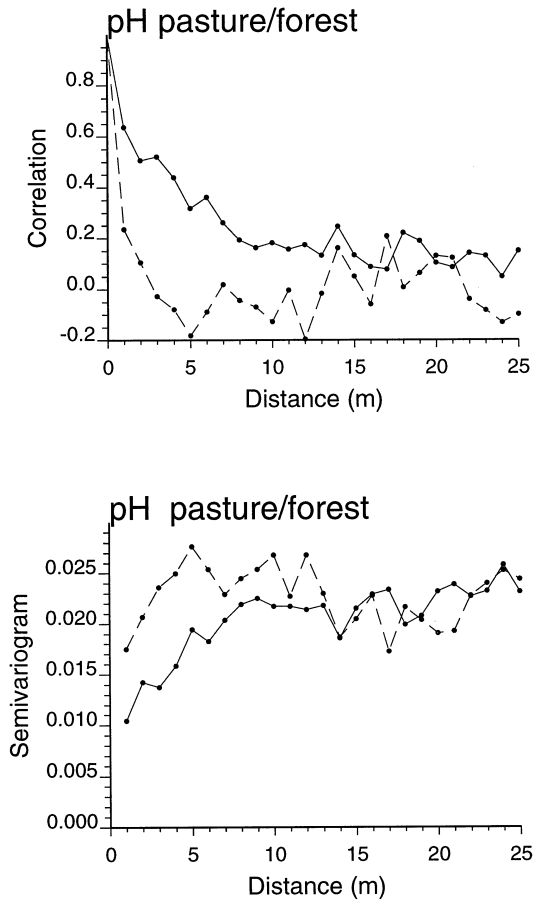


is a plot of all pairs of values related to two different attributes measured at the same locations. By analogy, the **h**-scattergram is a plot of all pairs of measurements of the same attribute  $z$  (i.e., pH) at locations separated by a given distance  $|\mathbf{h}|$  in a particular direction. By convention, the value at the start of the vector  $\mathbf{h}$ ,  $z(\mathbf{u}_\alpha)$ , is called the tail value, whereas the value at the end,  $z(\mathbf{u}_\alpha + \mathbf{h})$ , is the head value.

The shape of the cloud of points on the **h**-scattergram indicates that there is some correlation between adjacent pH values, and this can be measured using the linear correlation coefficient which is traditionally used to assess correlation between different attributes. The correlation is 0.64 and agrees with our visual impression that the pH value at any location is related to the pH measured 1 m away. Intuitively, one would expect that the relation between pH values weakens for a separation distance of 2 m, which is confirmed by the sec-

ond **h**-scattergram of Fig. 3 (right top graph): the correlation drops from 0.64 to 0.51. The increasing inflation of the cloud of points with increasing distance reflects the decreasing similarity of measurements farther apart: the correlation at 10 m is only 0.19!

The plot of the estimated correlation coefficients as a function of the separation distance is called the experimental correlogram. Figure 4 (top graph) shows the correlograms for pH measured in a pasture (solid line) and in a forest (dashed line), respectively. For the pasture, the correlation becomes negligible at a separation distance of about 12 m, which is referred to as the range and is interpreted as the distance beyond which two pH values can be considered as statistically independent. The decline in correlation is much sharper in the forest in that any two pH values 3 m apart can already be considered as independent. The correlogram thus allows one to quantify in terms of correlation the visual im-



**Fig. 4** Correlograms and semivariograms for the transects of 100 soil pH values measured in pasture (*solid line*) and forest (*dashed line*)

pression that small-scale fluctuations prevail for the transect of forest pH values displayed at the bottom of Fig. 1.

#### The semivariogram

In the geostatistical literature, spatial patterns are usually described in terms of dissimilarity (instead of similarity) between observations as a function of the separation distance. The average dissimilarity between data separated by a vector  $\mathbf{h}$  is measured by the experimental semivariogram  $\hat{\gamma}(\mathbf{h})$ , which is computed as half the average squared difference between the components of every data pair:

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

where  $N(\mathbf{h})$  is the number of data pairs for a given distance, e.g.,  $N(\mathbf{h})=99$  for  $|\mathbf{h}|=1$  m for the transects of Fig. 1.

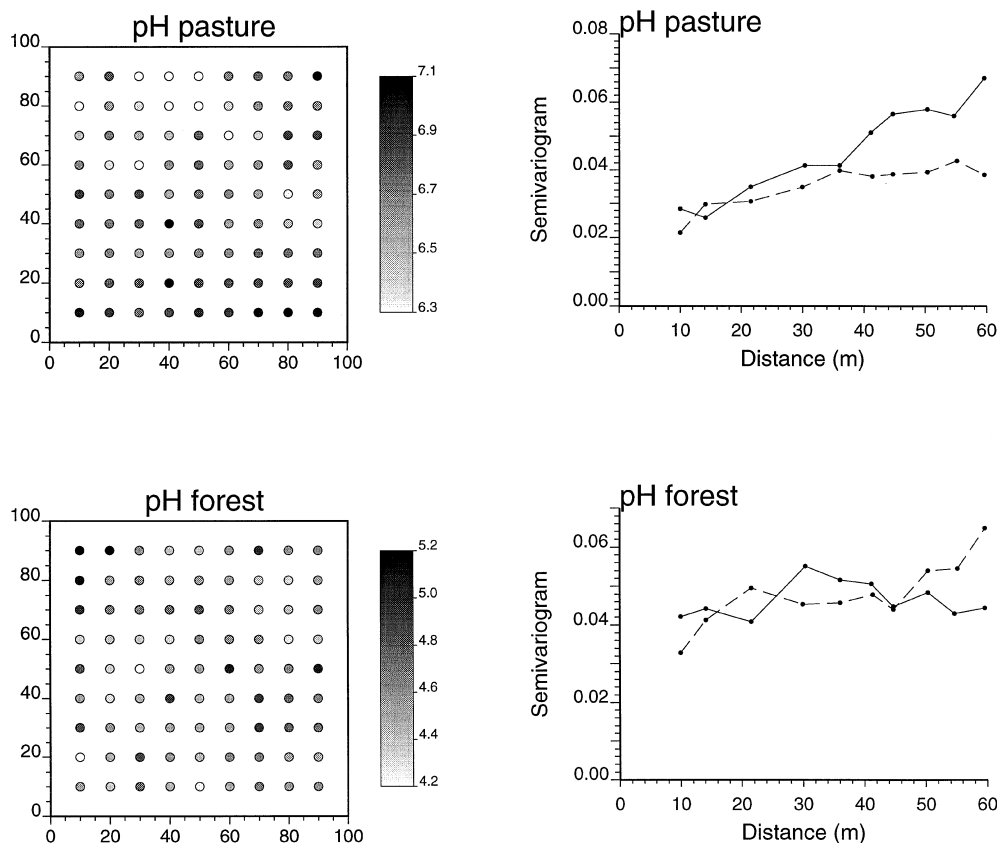
Figure 4 (bottom graph) shows the semivariograms of pH measured in pasture (solid line) and forest (dashed line), respectively. In both cases, semivario-

gram values increase with the separation distance, reflecting our intuitive feeling that two pH values close to each other on the ground are more alike, and thus their squared difference is smaller, than those further apart. The two semivariograms stop increasing at a given distance, called the range, which corresponds to the range identified on the correlogram, and again it can be interpreted as the distance of dependence or zone of influence of pH measurements. Beyond the range, the two semivariograms fluctuate around the same sill value, which is roughly equal to the sample variance computed on Fig. 2. In other words, the overall variation of pH measurements across the transect is similar for both land uses, yet the larger values of the forest semivariogram at short distances indicate that most of this variability is present over a few meters in forest. The discontinuity at the origin of the semivariogram (i.e., zero separation distance) is called the nugget effect and arises from measurement errors or spatial sources of variation at distances smaller than the shortest sampling interval or both (Journel and Huijbregts 1978, p. 39). Depending on the sampling scale, a semivariogram can appear as a pure nugget effect even if the variation is locally spatially structured, and so it might be useful to conduct a preliminary survey to approximate the major scales of spatial variation (Oliver and Webster 1986a).

In combination with a good knowledge about the phenomenon and the study area, such a spatial description can improve our understanding of the physical underlying mechanisms controlling spatial patterns. For example, McBratney and Webster (1981) related the periodicity displayed by the correlogram of wheat yield measured in a 0.4 ha area to an earlier ploughing. Robertson et al. (1988) found nitrogen availability in a 0.5 ha old field to be highly patterned at scales 20–40 m, which corresponds to the scales over which plant community composition varies during early succession, suggesting that the community structure is strongly influenced by the spatial heterogeneity of available nutrients. Goovaerts (1994a) found a remarkable similarity between the shapes of semivariograms for eight soil chemical properties and six banana leaf contents measured across a 100 km<sup>2</sup> area, and he was able to relate this common spatial pattern to the geographical distribution of soil types in the study area. Other examples in earth sciences can be found in Oliver and Webster (1986b). In the present study, the short-range variability of forest pH values probably reflects the local heterogeneity of ground vegetation compared with the uniformity of grass cover.

The information collected during soil surveys rarely reduces to a single transect, but a two-dimensional domain is typically sampled. It is thus necessary to characterize the spatial variability along different directions. For example, Fig. 5 (left column) shows two grids of pH values which were measured in the same plots as the two transects of Fig. 1. Semivariograms have been computed along the N-S (solid line) and E-W (dashed line) directions. Data pairs have been grouped into classes of

**Fig. 5** Grids of soil pH values measured on 1 ha pasture and forest plots, and the corresponding semivariograms computed in the NS (*solid line*) and EW (*dashed line*) directions. For pasture, the variability is said to be anisotropic in that it depends on the direction



distances (lags) and angles,  $[h \pm \Delta h]$  and  $[\theta \pm \Delta \theta]$ , so that each semivariogram value is computed from a sufficient number of data pairs (i.e.,  $N(\mathbf{h}) > 30$ ). In the example of Fig. 5, the two classes of angles are  $0^\circ \pm 45^\circ$  and  $90^\circ \pm 45^\circ$ , whereas the lag tolerance is 5 m. These graphs show distinct spatial patterns of acidity on both land uses: pH appears to vary more continuously in pasture as illustrated by the smaller nugget effect and larger range of the corresponding semivariogram. Another important feature is that the variability of pH values in pasture is anisotropic, that is it depends on the direction. The continuous increase of the N-S semivariogram reflects a trend in pH along that direction; the large pH values in the southern part of the plot are in fact due to the liming of the neighboring plot a few hours before sampling.

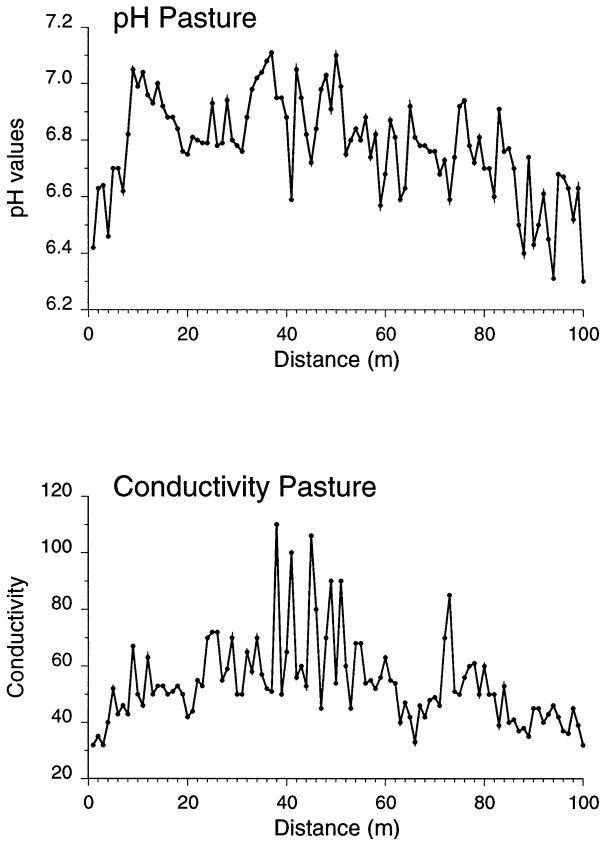
#### The cross correlogram

In most situations, several properties are measured on each soil sample, and geostatistics is increasingly used to process such multivariate spatial soil information (Wackernagel 1988; Goovaerts 1992). For example, in addition to pH, electrical conductivity was measured at each of the 100 locations along the transect in pasture (Fig. 6). Typically, the relationship between two soil properties is assessed by plotting the scattergram of the two variables and computing the corresponding corre-

lation coefficient (Fig. 7; left top graph). Once again, such an approach ignores the spatial coordinates of the measurements.

Instead of looking at the relation between pH and conductivity measured at the same location, one may assess the correlation between adjacent observations of both attributes by plotting each pH value  $z(\mathbf{u}_\alpha)$  versus the conductivity value measured 1 m away,  $y(\mathbf{u}_\alpha + \mathbf{h})$  with  $|\mathbf{h}| = 1$  m. The so-called cross  $\mathbf{h}$ -scattergram is displayed at the top of Fig. 7 (right graph). Surprisingly, the correlation between the two attributes increases with the separation distance, an effect referred to as the lag effect (Journel and Huijbregts 1978, p. 48). Such an effect can be observed in geochemistry, where different rates of precipitation may cause enrichment in some minerals to lag behind that of others along the direction of hydrothermal flow. In this particular example, the lag effect is not backed by any physical interpretation and most likely reflects sample fluctuations.

Figure 7 shows cross  $\mathbf{h}$ -scattergrams for a separation distance up to 10 m. The plot of the estimated correlation coefficient as a function of the separation distance is called the experimental cross correlogram (Fig. 8, top graph). Note that two cross correlograms can be computed, depending on whether a pH value is compared with the conductivity value east of it (easterly direction) or west of it (westerly direction). In both cases, the correlation between pH and electrical conductivity tends to decrease with the separation distance, and becomes negligible around 15 m.



**Fig. 6** Transect of 100 pH and electrical conductivity values measured in the topsoil of a pasture

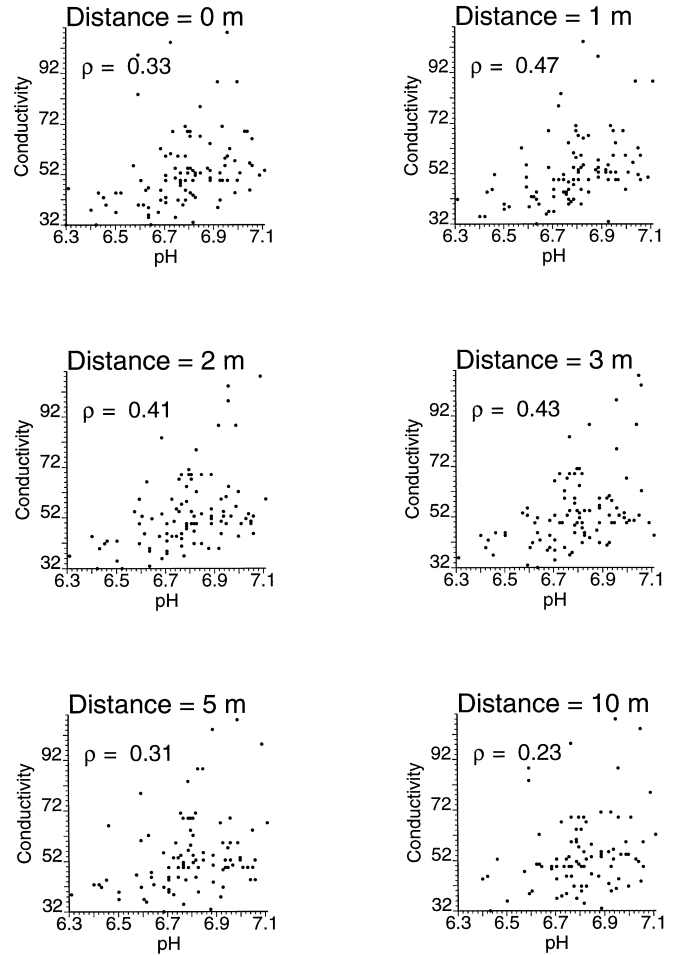
### The cross semivariogram

Instead of looking at the similarity between pH and conductivity values separated by a vector  $\mathbf{h}$  say  $z(\mathbf{u}_\alpha)$  and  $y(\mathbf{u}_\alpha + \mathbf{h})$ , one may consider how these two attributes jointly vary over a vector  $\mathbf{h}$ , that is the relation between the increments  $[z(\mathbf{u}_\alpha) - z(\mathbf{u}_\alpha + \mathbf{h})]$  and  $[y(\mathbf{u}_\alpha) - y(\mathbf{u}_\alpha + \mathbf{h})]$ . If both attributes are positively correlated, an increase (decrease) in pH values from  $\mathbf{u}_\alpha$  to  $\mathbf{u}_\alpha + \mathbf{h}$  tends to be associated with an increase (decrease) in conductivity values. Conversely, a negative correlation between attributes would entail that an increase (decrease) in pH values tends to be associated with a decrease (increase) in conductivity values.

A measure of the joint variability of two continuous attributes  $z$  and  $y$  is the experimental cross semivariogram which is computed as:

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

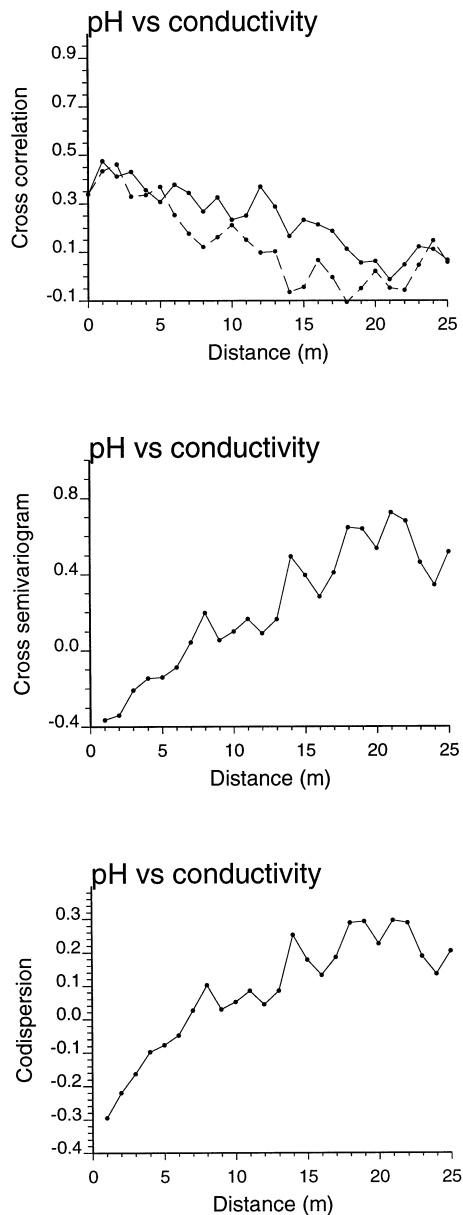
The rescaling of cross semivariogram values by the corresponding direct semivariogram values yields the codispersion coefficient  $\hat{\nu}_{ZY}(\mathbf{h})$  which can be interpreted as a linear correlation coefficient between the spatial increments of both attributes (Goovaerts 1997a):



**Fig. 7** Scattergrams between soil pH and electrical conductivity values separated by a distance of 0 m, 1 m, 2 m, ..., 10 m. Note how the correlation between observations decreases as the separation distance increases

$$\hat{\nu}_{ZY}(\mathbf{h}) = \frac{\hat{\gamma}_{ZY}(\mathbf{h})}{\sqrt{\hat{\gamma}_{ZZ}(\mathbf{h}) \cdot \hat{\gamma}_{YY}(\mathbf{h})}} \in [-1, +1] \quad (3)$$

Figure 8 shows the experimental cross semivariogram (middle graph) and codispersion function (bottom graph) between pH and electrical conductivity in pasture. Note that unlike the cross correlogram, the cross semivariogram is identical in the easterly and westerly directions. An interesting feature is the change in sign of cross semivariogram values, which indicates that the two attributes vary in opposite ways at short distances (small scale), whereas at larger distances their increments are positively related. Such scale-dependent relations may reflect the existence of different physico-chemical processes controlling the fluctuations in pH and electrical conductivity values at different scales. Changes in the correlation between soil attributes as a function of the spatial scale have been reported in several studies, e.g., see Wackernagel (1988), Goulard and Voltz (1992), Goovaerts (1992, 1994a), Goovaerts and Webster (1994), and Dobermann et al. (1995, 1997).



**Fig. 8** Experimental cross correlogram and semivariogram for the transect of 100 soil pH and electrical conductivity values measured in pasture

### The indicator semivariogram

The different tools introduced so far allow the quantification of the spatial variability over the full range of attribute values. In most applications, such description should be supplemented by the characterization of the spatial distribution of specific classes of values. For example, knowledge of whether large values are clustered or scattered in space can improve our interpretation of underlying physical processes. In many environmental studies, a few random hot spots of large pollutant concentrations caused by local human activities coexist with a natural background of small values (Leonte and

Schofield 1996; Mohammadi et al. 1997). Similarly, many microbiological processes in soil, such as denitrification, show hot spots of high activity which may be related to the patchy distribution of organic material in soil (Parkin 1987; Robertson et al. 1988).

The characterization of the spatial distribution of  $z$ -values above or below a given threshold value  $z_k$  requires a prior coding of each observation  $z(\mathbf{u}_\alpha)$  into an indicator datum  $i(\mathbf{u}_\alpha; z_k)$ , defined as:

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{u}_\alpha) \leq z_k \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

Indicator semivariograms can then be computed by substituting indicator data  $i(\mathbf{u}_\alpha; z_k)$  for  $z$ -data  $z(\mathbf{u}_\alpha)$  in Eq. (1):

$$\hat{\gamma}_I(\mathbf{h}; z_k) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [i(\mathbf{u}_\alpha; z_k) - i(\mathbf{u}_\alpha + \mathbf{h}; z_k)]^2 \quad (5)$$

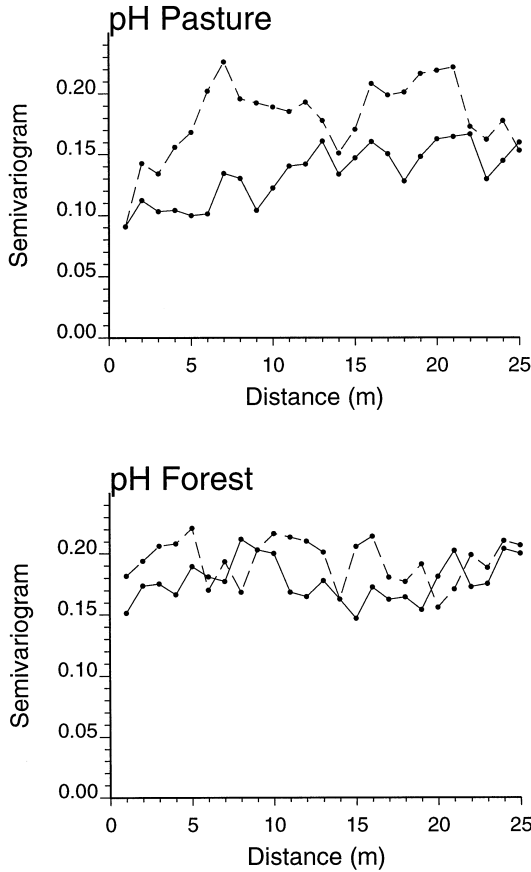
The indicator variogram value  $2\hat{\gamma}_I(\mathbf{h}; z_k)$  measures how often two  $z$ -values separated by a vector  $\mathbf{h}$  are on opposite sides of the threshold value  $z_k$ . In other words,  $2\hat{\gamma}_I(\mathbf{h}; z_k)$  measures the transition frequency between two classes of  $z$ -values as a function of  $\mathbf{h}$ . The greater  $\hat{\gamma}_I(\mathbf{h}; z_k)$ , the less connected in space are the small or large values.

Figure 9 shows the indicator semivariograms computed for the lower and upper quartiles of the distributions of pH values in pasture and forest. For forest soils, the two indicator semivariograms have a very large nugget, which indicates that small and large pH values are almost randomly distributed along the transect. Indicator semivariograms for pasture measurements are clearly different: the shorter range of the upper quartile semivariogram reflects the existence of short high-valued (pH larger than 6.92) segments along the transect, whereas small pH values are better connected in space as indicated by smaller values of the lower quartile semivariogram.

Indicator cross semivariogram  $\hat{\gamma}_I(\mathbf{h}; z_k, z_{k'})$  can be computed by substituting indicator data defined for two different threshold values  $z_k$  and  $z_{k'}$  for  $z$  and  $y$ -data in the equation (2):

$$\hat{\gamma}_I(\mathbf{h}; z_k, z_{k'}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [i(\mathbf{u}_\alpha; z_k) - i(\mathbf{u}_\alpha + \mathbf{h}; z_k)] \cdot [i(\mathbf{u}_\alpha; z_{k'}) - i(\mathbf{u}_\alpha + \mathbf{h}; z_{k'})] \quad (6)$$

Webster and Boag (1992) used a combination of indicator direct and cross semivariograms to characterize the spatial distribution of the number of cyst nematodes in topsoil. They found that as the threshold value  $z_k$  increased, the range of indicator direct semivariograms shortens progressively, whereas the nugget effect of the cross semivariograms decreased, which indicates a border effect with infestation increasing gradually from the outside of patches towards their centres. Biologically, the distribution can be explained by the nematode being spread by cultivation from initial fortuitous foci of infestation or control by nematophagous fungi or both.



**Fig. 9** Experimental indicator semivariograms for soil pH values measured in pasture and forest. The two threshold values correspond to the lower (*solid line*) and upper (*dashed line*) quartiles of the sample distributions of Figure 2

Indicator semivariograms can also be used to characterize the spatial patterns of categorical variables, such as texture or water table classes, which take only a limited number of states which might be ordered or not. Let  $S$  be a categorical attribute with possible states  $s_k$ ,  $k=1,2,\dots,K$ . The  $K$  states are exhaustive and mutually exclusive in the sense that one and only one state  $s_k$  occurs at each data location  $\mathbf{u}_\alpha$ . The pattern of variability of a category  $s_k$  can be characterized by semivariograms of type (5) defined on an indicator coding of the presence of that category:

$$i(\mathbf{u}_\alpha; s_k) = \begin{cases} 1 & \text{if } s(\mathbf{u}_\alpha) = s_k \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

The indicator variogram value  $2\hat{\gamma}_I(\mathbf{h}; s_k)$  measures how often two locations a vector  $\mathbf{h}$  apart belong to different categories  $s_{k'} \neq s_k$ . The smaller  $2\hat{\gamma}_I(\mathbf{h}; s_k)$ , the greater the spatial connectivity of category  $s_k$ . For example, categorical indicator semivariograms have been used to characterize the spatial distribution of water table classes (Bierkens and Burrough 1993a, b) and the spatial arrangement of soil types (Goovaerts 1994b).

### Semivariogram modeling

One of the main applications of geostatistics is the prediction of attribute values at unsampled locations. Prediction is made possible by the existence of spatial dependence between observations as assessed by the correlogram or semivariogram. A key step between description and prediction is the modeling of the spatial distribution of attribute values. Most of geostatistics is based on the concept of random function, whereby the set of unknown values is regarded as a set of spatially dependent random variables. Each measurement  $z(\mathbf{u}_\alpha)$  is thus interpreted as a particular realization of a random variable  $Z(\mathbf{u}_\alpha)$ . Interested readers should refer to textbooks such as Isaaks and Srivastava (1989, pp. 196–236) or Goovaerts (1997a, pp. 59–74) for a detailed presentation of the theory of random functions. An important characteristic of the random function is its semivariogram which must be modeled from the experimental values.

#### The univariate case

Let  $[\hat{\gamma}(\mathbf{h}_k), k=1,\dots,K]$  be the set of experimental semivariogram values computed for a finite number of lags, say the 25 semivariogram values displayed at the bottom of Fig. 4 for either pasture or forest pH measurements. A continuous function must be fitted to these values so as to deduce semivariogram values for any possible lag  $\mathbf{h}$  required by prediction algorithms and also to smooth out sample fluctuations. The difficulty is that only functions that are conditionally negative definite can be considered as semivariogram models, in order to ensure the non-negativity of the variance of the prediction error, see later expression (21). In practice, only a few models are known to be permissible, and Fig. 10 shows the ones that are included in most geostatistical software. The three top models are bounded in that they reach a sill either at a given range value (spherical model) or asymptotically (exponential, Gaussian model):

- Spherical model with range  $a$

$$g(h) = \text{Sph}\left(\frac{h}{a}\right) = \begin{cases} 1.5 \cdot \frac{h}{a} - 0.5 \cdot \left(\frac{h}{a}\right)^3 & \text{if } h < a \\ 1 & \text{otherwise} \end{cases} \quad (8)$$

- Exponential model with distance parameter  $b$

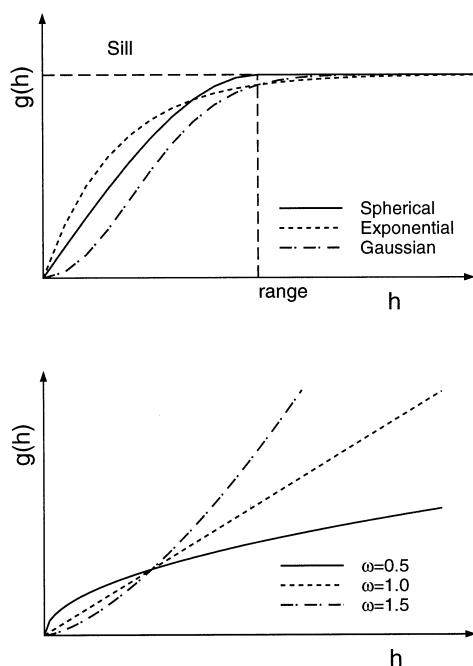
$$g(h) = 1 - \exp\left(\frac{-h}{b}\right) \quad (9)$$

- Gaussian model with distance parameter  $b$

$$g(h) = 1 - \exp\left(\frac{-h^2}{b^2}\right) \quad (10)$$

The practical range is defined as the distance at which the model value is at 95% of the sill, that is  $3b$  for the exponential model and  $\sqrt{3}b$  for the Gaussian model.





**Fig. 10** Bounded semivariogram models with the same practical range (top graph), and power models for different values of the parameter  $\omega$  (bottom graph)

The bottom graph shows power models for different values of the parameter  $\omega$ :

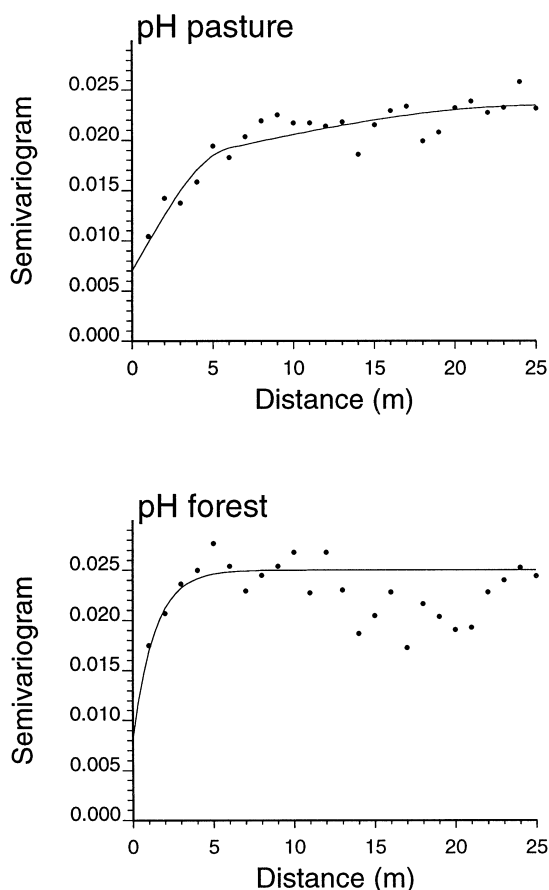
$$g(h) = h^\omega \text{ with } 0 < \omega < 2 \quad (11)$$

These functions allow one to model different behaviours at infinity (bounded, unbounded) and at the origin (linear, quadratic) of the semivariogram. Several authors (McBratney and Webster 1986; Wackernagel 1995; Goovaerts 1997a) have warned users of the risk of numerical instability when using the Gaussian model for spatial interpolation (kriging). A common recommendation is to add a small nugget effect to the Gaussian model to destroy its extreme extrapolative properties and avoid the generation of artifacts in interpolated maps. A better solution is never to use the Gaussian model. If the experimental semivariogram appears to have both an asymptote and an inflexion, it can be modeled using the following function:

$$g(h) = 1 - \exp\left(\frac{-h^\omega}{b}\right) \text{ with } 0 < \omega < 2 \quad (12)$$

where the parameter  $\omega$  can be adjusted (Wackernagel 1995, p. 110; Webster 1997). Note that for  $\omega=2$ , one retrieves the Gaussian model (10). A theoretically sounder alternative is to use Whittle's elementary correlation (McBratney and Webster 1986).

Typically, two or more permissible models must be combined to fit the shape of the experimental semivariogram. For example, the semivariogram of pH on pasture has been modeled using a combination of a nugget effect and two spherical models of range 6 m and 26 m



**Fig. 11** Experimental semivariograms for soil pH values measured along the pasture and forest transects with the model fitted

(Fig. 11; top graph). The semivariogram of forest pH has been modeled using a nugget effect and an exponential model of practical range 4 m. Combinations of permissible models are permissible as long as the contribution of each basic model is positive, that is the nested model is written as:

$$\gamma(\mathbf{h}) = \sum_{l=0}^L b^l g_l(\mathbf{h}) \text{ with } b^l \geq 0 \quad (13)$$

where  $b^l$  is the positive sill or slope of the corresponding basic semivariogram model  $g_l(\mathbf{h})$ .

The way in which these permissible models are chosen and their parameters (range, sill) are estimated is still controversial (McBratney and Webster 1986; Goovaerts 1997a, pp. 97–107). Several methods have been proposed, ranging from full blackbox procedures in which the choice of model and its fitting are automatic to visual approaches where the model is selected so that the fit is satisfactory from a graphical point of view, as in the examples of Fig. 11. An intermediate approach consists of an automatic (least-squares) estimation of parameters of models chosen by the user. In this semi-automatic procedure, the objective is usually to minimize a weighted sum of squares (WSS) of differ-

ences between experimental  $\hat{\gamma}(\mathbf{h}_k)$  and model  $\gamma(\mathbf{h}_k)$  semivariogram values:

$$\text{WSS} = \sum_{k=1}^K \omega(\mathbf{h}_k) \cdot [\hat{\gamma}(\mathbf{h}_k) - \gamma(\mathbf{h}_k)]^2 \quad (14)$$

The weight  $\omega(\mathbf{h}_k)$  given to each lag  $h_k$  is often taken proportional to the number  $N(\mathbf{h}_k)$  of data pairs that contribute to the estimate  $\hat{\gamma}(\mathbf{h}_k)$ . The implicit assumption is that the reliability of an experimental semivariogram value increases with statistical mass. An alternative that gives more weight to the first lags consists of dividing the number of data pairs by the squared model value:  $N(\mathbf{h}_k)/[\gamma(\mathbf{h}_k)]^2$  (Cressie 1985).

The value of the WSS criterion is frequently used to rank alternative models, say to compare the goodness-of-fit of a spherical model versus an exponential one. In many situations, the user tends to increase the number of basic structures in the nested model (13) to lower the WSS value as much as possible. However, one should avoid overfitting experimental semivariograms: the more complicated model generally does not lead to more accurate estimates. McBratney and Webster (1986) proposed using the Akaike Information Criterion to achieve a satisfactory compromise between goodness-of-fit and parsimony of the model.

Semivariogram modeling is rarely a goal per se. The ultimate objective is usually to estimate the value of a soil property at an unsampled place. Cross validation allows one to compare the impact of different semivariogram models on interpolation results (McBratney and Webster 1986; Isaaks and Srivastava 1989, pp. 351–368). The idea consists of removing one datum at a time from the data set and re-estimating this value from remaining data using different semivariogram models. Interpolated and actual values are compared, and the model that yields the most accurate predictions is retained. A measure of accuracy is the mean absolute error of prediction (MAE), which is computed as the average absolute difference between actual and estimated values at the  $n$  sampled locations:

$$\text{MAE} = \frac{1}{n} \sum_{\alpha=1}^n |z(\mathbf{u}_\alpha) - z^*(\mathbf{u}_\alpha)| \quad (15)$$

In the presence of anisotropic (direction-dependent) variation, one must build a model that describes how the semivariogram changes as the distance and direction change. Two types of anisotropy are usually distinguished: geometric and zonal. In the first case, the directional semivariograms have the same shape and sill but different range values that plot as an ellipse as a function of the azimuth of the direction. An anisotropy that involves sill values varying with direction is said to be zonal. These two models of spatial anisotropy are presented in textbooks such as Isaaks and Srivastava (1989, pp. 377–390) and Goovaerts (1997a, pp. 90–95).

Regardless of the fitting procedure adopted, one must keep in mind that the objective is to build a permissible semivariogram model that captures the major

spatial features of the attribute under study. The use of least-squares criteria amounts to reducing semivariogram modeling to an exercise in fitting a curve to experimental values, which I think is too restrictive. Although experimental semivariogram values play an important role in this process, ancillary information such as provided by physical knowledge of the area and phenomenon may be of great interest. For example, strong prior qualitative information may lead one to adopt an anisotropic model even if data sparsity prevents seeing anisotropy from the experimental semivariograms computed in different directions.

### Bivariate case

Modeling the coregionalization between two variables  $Z$  and  $Y$  involves choosing and fitting functions to the two direct semivariograms  $\gamma_{ZZ}(\mathbf{h})$  and  $\gamma_{YY}(\mathbf{h})$  plus the cross semivariogram  $\gamma_{ZY}(\mathbf{h})$ . The difficulty lies in the fact that the three models cannot be built independently from one another. The easiest approach consists of modeling the three semivariograms as linear combinations of the same set of basic semivariogram models  $g_l(\mathbf{h})$ :

$$\gamma_{ZZ}(\mathbf{h}) = \sum_{l=0}^L b'_{ZZ} g_l(\mathbf{h})$$

$$\gamma_{YY}(\mathbf{h}) = \sum_{l=0}^L b'_{YY} g_l(\mathbf{h})$$

$$\gamma_{ZY}(\mathbf{h}) = \sum_{l=0}^L b'_{ZY} g_l(\mathbf{h})$$

The so-called linear model of coregionalization (LMC) is permissible if: (1) the  $(L+1)$  basic models  $g_l(\mathbf{h})$  are permissible, and (2) the coefficients satisfy the following necessary and sufficient conditions:

$$b'_{ZZ} \geq 0, b'_{YY} \geq 0 \quad \forall l \quad (16)$$

$$|b'_{ZY}| \leq \sqrt{b'_{ZZ} b'_{YY}} \quad \forall l \quad (17)$$

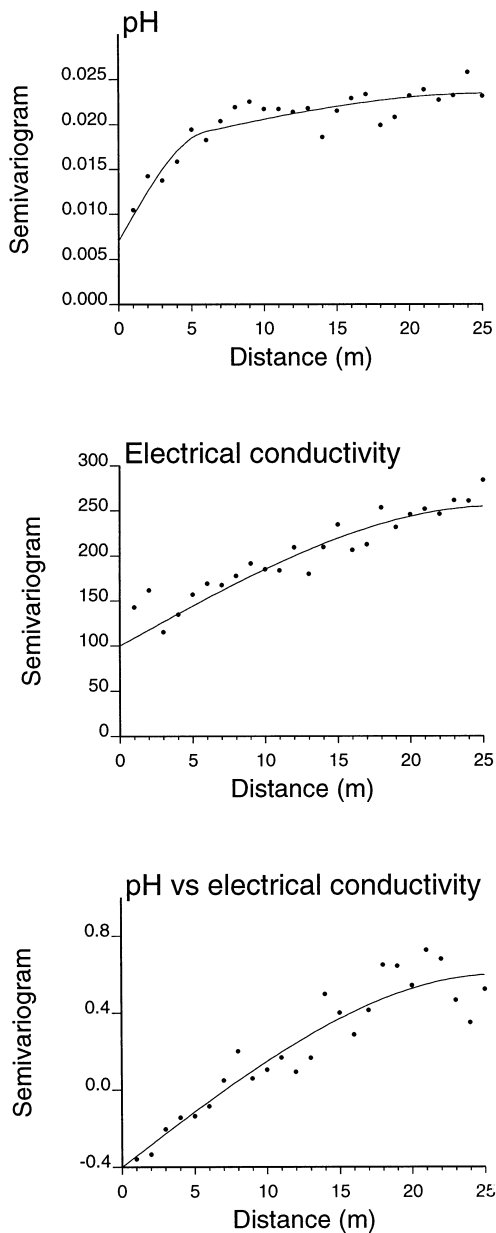
In practice, the modeling is performed in two steps:

1. Both direct semivariograms are first modeled as linear combinations of selected basic structures  $g_l(\mathbf{h})$ .
2. The same basic structures are then fitted to the cross semivariogram under the constraint (17).

This approach was used to fit visually the following model to the (cross) semivariograms of pH and electrical conductivity displayed in Fig. 12:

$$\begin{aligned} \gamma_{\text{PH}}(\mathbf{h}) &= 0.007 g_0(\mathbf{h}) + 0.01 \text{ Sph}(h/6 \text{ m}) \\ &\quad + 0.0065 \text{ Sph}(h/26 \text{ m}) \\ \gamma_{\text{EC}}(\mathbf{h}) &= 100 g_0(\mathbf{h}) + 155 \text{ Sph}(h/26 \text{ m}) \\ \gamma_{\text{PH-EC}}(\mathbf{h}) &= -0.4 g_0(\mathbf{h}) + 1.0 \text{ Sph}(h/26 \text{ m}) \end{aligned} \quad (18)$$

where  $g_0(\mathbf{h})$  is a nugget effect model, and  $\text{Sph}(h/6 \text{ m})$  is the spherical model (8) with range  $a=6 \text{ m}$ . For each of the three basic models, the coefficients of the direct semivariograms are positive and the inequality (17) is



**Fig. 12** Experimental direct and cross semivariograms for soil pH and electrical conductivity along the pasture transect, with the linear model of coregionalization fitted

satisfied:  $|-0.4| \leq \sqrt{0.007 \cdot 100}$ ,  $|0| \leq \sqrt{0.01 \cdot 0}$ , and  $|1| \leq \sqrt{0.0065 \cdot 155}$ .

The requirement that all semivariograms must share the same set of basic structures may sound like a severe limitation on the linear model of coregionalization. Variables that are well cross-correlated are, however, likely to show similar patterns of spatial variability. In addition, there is no need for the direct and cross semivariograms to include all the basic structures; for example, the semivariogram of electrical conductivity and the cross semivariogram do not include the short-range (6 m) spherical structure.

**Multivariate case**

Soil studies typically involve more than two attributes, and it is often necessary to model the coregionalization for a large number of variables. Checking the permissibility of the LMC becomes cumbersome because the simple constraints (16) and (17) do not suffice for more than two variables. Mathematically speaking, one must check that, for each structure  $g_l(\mathbf{h})$ , the matrix of  $b$ -coefficients is positive semi-definite. A symmetric matrix is positive semi-definite if its determinant and all its principal minor determinants are non-negative. Consider, for example, the linear model of coregionalization for three variables  $Z$ ,  $Y$  and  $X$ :

$$\begin{bmatrix} \gamma_{ZZ}(\mathbf{h}) & \gamma_{ZY}(\mathbf{h}) & \gamma_{ZX}(\mathbf{h}) \\ \gamma_{YZ}(\mathbf{h}) & \gamma_{YY}(\mathbf{h}) & \gamma_{YX}(\mathbf{h}) \\ \gamma_{XZ}(\mathbf{h}) & \gamma_{XY}(\mathbf{h}) & \gamma_{XX}(\mathbf{h}) \end{bmatrix} = \sum_{l=0}^L \begin{bmatrix} b_{ZZ}^l & b_{ZY}^l & b_{ZX}^l \\ b_{YZ}^l & b_{YY}^l & b_{YX}^l \\ b_{XZ}^l & b_{XY}^l & b_{XX}^l \end{bmatrix} g_l(\mathbf{h})$$

For each structure  $g_l(\mathbf{h})$ , the matrix of coefficients must satisfy the following seven inequalities:

- All diagonal elements are non-negative:

$$b_{ZZ}^l \geq 0 \quad b_{YY}^l \geq 0 \quad b_{XX}^l \geq 0$$

- All principal minor determinants of order 2 are non-negative:

$$\begin{vmatrix} b_{ZZ}^l & b_{ZY}^l \\ b_{YZ}^l & b_{YY}^l \end{vmatrix} = b_{ZZ}^l b_{YY}^l - [b_{YZ}^l]^2 \geq 0$$

$$\begin{vmatrix} b_{YY}^l & b_{YX}^l \\ b_{XY}^l & b_{XX}^l \end{vmatrix} = b_{YY}^l b_{XX}^l - [b_{YX}^l]^2 \geq 0$$

$$\begin{vmatrix} b_{ZZ}^l & b_{ZX}^l \\ b_{XZ}^l & b_{XX}^l \end{vmatrix} = b_{ZZ}^l b_{XX}^l - [b_{XZ}^l]^2 \geq 0$$

- The determinant of order 3 is non-negative:

$$\begin{vmatrix} b_{ZZ}^l & b_{ZY}^l & b_{ZX}^l \\ b_{YZ}^l & b_{YY}^l & b_{YX}^l \\ b_{XZ}^l & b_{XY}^l & b_{XX}^l \end{vmatrix} = b_{ZZ}^l (b_{YY}^l b_{XX}^l - [b_{YX}^l]^2) - b_{ZY}^l (b_{YZ}^l b_{XX}^l - b_{XZ}^l b_{YX}^l) + b_{ZX}^l (b_{YZ}^l b_{XY}^l - b_{XZ}^l b_{YY}^l) \geq 0$$

An iterative procedure that fits the linear model of coregionalization directly under such a constraint of positive semi-definiteness has been developed by Goulard (1989). This technique is increasingly used to model the coregionalization of soil properties (Goovaerts 1992; Goulard and Voltz 1992; Voltz and Goulard 1994; Webster et al. 1994; Dobermann et al. 1995, 1997).

**Spatial prediction**

Most geostatistical studies in soil science aim at estimating soil properties at unsampled places and mapping them. Kriging is a generic name adopted by the geostatisticians for a family of generalized least-squares regression algorithms (Webster 1996). There are many different kriging algorithms, and most of them are reviewed in Goovaerts (1998) with references to soil ap-

plications. In this paper, only the most common ordinary kriging and cokriging are presented.

### Ordinary kriging

Consider the problem of estimating the value of a continuous soil attribute  $z$  at an unsampled location  $\mathbf{u}$  using only data on this attribute  $[z(\mathbf{u}_\alpha), \alpha=1, \dots, n]$ . Ordinary kriging (OK) estimates this value as a linear combination of neighboring observations:

$$z_{OK}^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) z(\mathbf{u}_\alpha) \quad (19)$$

Typically, the 10–20 observations closest to  $\mathbf{u}$  being estimated are retained because they screen the influence of those farther away, that is  $n(\mathbf{u}) \ll n$ . The key issue is the determination of the weight  $\lambda_\alpha(\mathbf{u})$  assigned to each observation.

As for other linear regression procedures, ordinary kriging weights are chosen so as to minimize the estimation or error variance  $\sigma_E^2(\mathbf{u}) = \text{Var} [Z^*(\mathbf{u}) - Z(\mathbf{u})]$  under the constraint of unbiasedness of the estimator. These weights are obtained by solving a system of linear equations which is known as “ordinary kriging system”:

$$\begin{cases} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta(\mathbf{u}) \gamma(\mathbf{u}_\alpha - \mathbf{u}_\beta) - \mu(\mathbf{u}) = \gamma(\mathbf{u}_\alpha - \mathbf{u}) \\ \alpha = 1, \dots, n(\mathbf{u}) \\ \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta(\mathbf{u}) = 1 \end{cases} \quad (20)$$

Unbiasedness of the estimator is ensured by constraining the weights to sum to one, which requires the definition of the Lagrange parameter  $\mu(\mathbf{u})$ . The only information required by the kriging system are semivariogram values for different lags, and these are readily derived from the semivariogram model fitted to experimental values (recall previous section).

In addition to an estimate for the unknown  $z$ -value, ordinary kriging provides an error variance which is computed as:

$$\sigma_{OK}^2(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) \gamma(\mathbf{u}_\alpha - \mathbf{u}) - \mu(\mathbf{u}) \quad (21)$$

Under stringent hypotheses of normality and homoscedasticity, the kriging variance can be combined with the estimated value to derive a confidence interval; for example, the 95% confidence interval is taken as:

$$\text{Prob}\{Z(\mathbf{u}) \in [z_{OK}^*(\mathbf{u}) - 2\sigma_{OK}(\mathbf{u}), z_{OK}^*(\mathbf{u}) + 2\sigma_{OK}(\mathbf{u})]\} = 0.95 \quad (22)$$

Figure 13 (left column) shows the ordinary kriging estimates and error variances for pH along the transect in pasture. Only 20 pH values out of the 100 original data were considered here so that estimated values can be later compared with actual measurements to assess the

prediction performances of kriging. The estimation was performed every meter using the ten closest pH values and the semivariogram model displayed at the top of Fig. 11. The ordinary kriging estimator is exact in that it honors observations at data locations, and the kriging variance is zero there. The error variance increases away from the data and reaches a maximum value beyond the extreme right datum (extrapolation situation).

Besides providing an error variance, kriging has several advantages over other interpolation techniques such as the inverse distances algorithm: (1) it accounts for the pattern of spatial variability (range, anisotropy) through the semivariogram model, and (2) it allows estimation of the target attribute on a support that is different from the data support, e.g., pH values measured on soil cores can be used to predict average pH values on larger surfaces, such as 1 ha plots, which are better suited for land management (Burgess and Webster 1980b).

### Ordinary cokriging

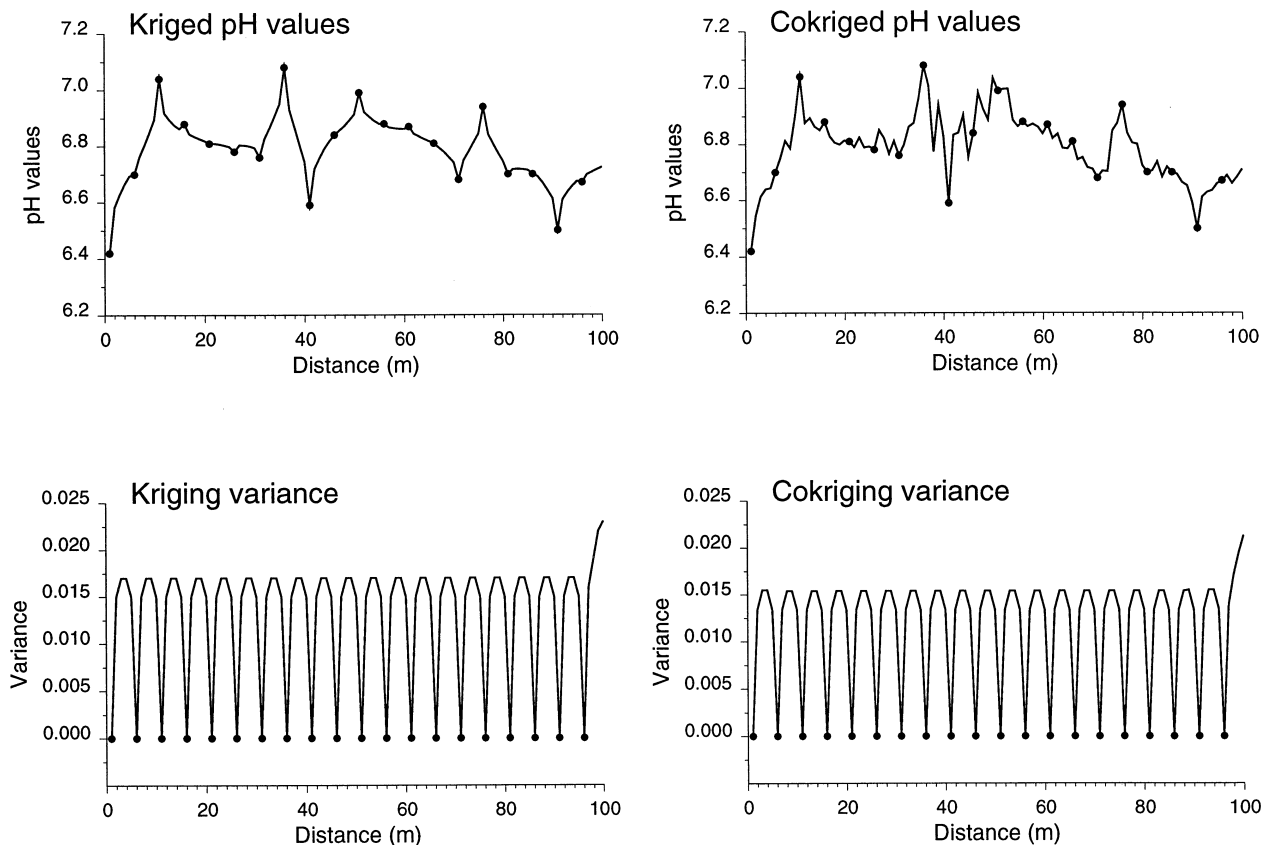
When measurements are sparse or poorly correlated in space, the estimation of the primary attribute of interest is generally improved by taking into account secondary information originating from other correlated continuous attributes. The ordinary kriging estimate (19) is readily extended to incorporate that additional information. In the simplest case of a single secondary attribute  $Y$ , the so-called ordinary cokriging (OCK) estimate is written as a linear combination of both neighboring primary and secondary data:

$$z_{OCK}^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}) z(\mathbf{u}_\alpha) + \sum_{\alpha'=1}^{n'(\mathbf{u})} \nu_{\alpha'}(\mathbf{u}) y(\mathbf{u}_{\alpha'}) \quad (23)$$

where secondary data have been measured at possibly different locations  $\mathbf{u}_{\alpha'}$ . Like ordinary kriging, the objective is to minimize the error variance under unbiasedness constraint, which yields a very complex system of  $(n(\mathbf{u}) + n'(\mathbf{u}) + 2)$  linear equations:

$$\begin{cases} \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta(\mathbf{u}) \gamma_{ZZ}(\mathbf{u}_\alpha - \mathbf{u}_\beta) + \sum_{\beta'=1}^{n'(\mathbf{u})} \nu_{\beta'}(\mathbf{u}) \gamma_{ZY}(\mathbf{u}_\alpha - \mathbf{u}_{\beta'}) \\ \quad - \mu_Z(\mathbf{u}) = \gamma_{ZZ}(\mathbf{u}_\alpha - \mathbf{u}) \quad \alpha = 1, \dots, n(\mathbf{u}) \\ \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta(\mathbf{u}) \gamma_{YZ}(\mathbf{u}_{\alpha'} - \mathbf{u}_\beta) + \sum_{\beta'=1}^{n'(\mathbf{u})} \nu_{\beta'}(\mathbf{u}) \gamma_{YY}(\mathbf{u}_{\alpha'} - \mathbf{u}_{\beta'}) \\ \quad - \mu_Y(\mathbf{u}) = \gamma_{YZ}(\mathbf{u}_{\alpha'} - \mathbf{u}) \quad \alpha' = 1, \dots, n'(\mathbf{u}) \\ \sum_{\beta=1}^{n(\mathbf{u})} \lambda_\beta(\mathbf{u}) = 1 \\ \sum_{\beta'=1}^{n'(\mathbf{u})} \nu_{\beta'}(\mathbf{u}) = 0 \end{cases} \quad (24)$$

There are now two Lagrange parameters to account for the constraints on primary and secondary data weights, and the input information comprises the values of direct and cross semivariograms for different lags. The cokriging variance is computed as:



**Fig. 13** Ordinary kriging and cokriging estimates of soil pH along the pasture transect, with the corresponding error variances. The information available consists of 20 pH values (*black dots*) and 100 electrical conductivity values

$$\sigma_{OCK}^2(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_{\alpha}(\mathbf{u}) \gamma_{ZZ}(\mathbf{u}_{\alpha} - \mathbf{u}) + \sum_{\alpha'=1}^{n'(\mathbf{u})} \nu_{\alpha'}(\mathbf{u}) \gamma_{YZ}(\mathbf{u}_{\alpha'} - \mathbf{u}) - \mu_Z(\mathbf{u}) \quad (25)$$

Figure 13 (right column) shows the ordinary cokriging estimates and the corresponding error variances for pH along the transect in pasture. The information available consists of 20 pH values plus the 100 electrical conductivity data shown at the bottom of Fig. 6. The estimation was performed every meter using the ten closest pH values and the ten closest electrical conductivity values. The linear model of coregionalization of Fig. 12 was used to create the cokriging system. Accounting for this additional information yields a more detailed estimated transect and smaller error variances than kriging does. The absolute estimation error was also computed at the 80 locations where the pH value was ignored, and results show that cokriging yields a 10% reduction of the error: 0.0978 instead of 0.1031 for kriging. The small magnitude of the reduction is explained mainly by the weak correlation ( $r=0.33$ ) between pH and electrical conductivity. The benefit of cokriging over kriging increases as the correlation between the two attributes increases and as the secondary

information becomes more abundant than the primary information. Cokriging is thus particularly useful when a cheap-to-measure attribute is strongly correlated with the attribute of interest.

Many examples of cokriging can be found in the soil literature. For example, Yates and Warrick (1987) estimated gravimetric moisture content in a 1 ha plot using the bare soil surface temperature and the percent sand content as secondary variables. On a much larger scale, Gallichand and Marcotte (1993) mapped the clay percentage for a study area of 33 500 ha using 485 clay content values and 3488 values of saturated hydraulic conductivity. Several authors (Leenaers et al. 1990; Odeh et al. 1997) have shown how the readily available digital elevation model can be used to improve the prediction of soil attributes through cokriging. Finally, Gotway and Hartford (1996) demonstrated that accounting for corn yield measurements through cokriging improved the prediction of residual nitrate in the field, leading to better recommendations for variable-rate N fertilizer application. Beware that several applications of cokriging in the soil literature do not use a linear model of coregionalization, and even worse, the permissibility of the model fitted is rarely checked!

### Factorial kriging

Any soil property measured at a certain location in space is the outcome of several physical, chemical or

biological processes. Some of them operate only at microscopic scales (e.g., biological activities), whereas others such as weathering of parent material, erosion or solute transport may act over larger distances. The combined effect of different sources of spatial variation that operate at distinct scales yields nested semivariograms, such as the semivariogram of pH on pasture (Fig. 11, top graph) which was modeled as the sum of a nugget effect and two spherical models of range 6 m and 26 m:

$$\gamma_{\text{pH}}(\mathbf{h}) = 0.007 g_0(\mathbf{h}) + 0.01 \text{Sph}(h/6 \text{ m}) + 0.0065 \text{Sph}(h/26 \text{ m}) \quad (26)$$

Factorial kriging or kriging analysis (Matheron 1982; Wackernagel 1988, 1995; Goovaerts 1992) is a variant of kriging which aims at estimating and mapping the different sources of spatial variability identified on the experimental semivariogram. For example, on the basis of the semivariogram model (26), the transect of pH values measured in pasture can be viewed as the sum of different microscopic (nugget effect), short-range (6 m), and long-range (26 m) spatial components:

$$z(\mathbf{u}_\alpha) = z_0(\mathbf{u}_\alpha) + z_1(\mathbf{u}_\alpha) + z_2(\mathbf{u}_\alpha) \quad \forall \alpha$$

Assuming that these components are independent, the transect of original pH values is decomposed into local (microscopic + short-range) and regional (long-range) components using the approach described in Goovaerts (1992) (Fig. 14, left column). A similar decomposition is performed for the transect of electrical conductivity values on the basis of the semivariogram model of Fig. 12 (middle graph).

Although spatial components are mathematical constructions with no a priori physical meaning, they can help to identify the main sources of spatial variation and improve our understanding of underlying physical processes in soil. Goovaerts (1994a) used factorial kriging to separate local variation in soil and vegetation properties due to field-to-field differences from regional variation related to the presence of different soil types. Webster et al. (1994) used a similar approach to map local variation in heavy metal concentrations which are partly due to anthropic activities and regional changes caused by the geology. Dobermann et al. (1995) investigated field-scale variation of soil fertility in the humid tropics, and factorial kriging allowed the discrimination of short-range variation in P, Ca, Mg and K topsoil concentrations, which is due to patchy inputs of nutrients from buffalo excrement, from long-range variation related to vertical and lateral soil water flow along the topographic slope.

### Scale-dependent relations

An interesting application of factorial kriging is the detection of scale-dependent correlations between soil attributes, i.e. correlations that change as a function of the spatial scale. For example, the scattergram of pH

versus electrical conductivity values indicates that these two soil attributes are weakly correlated along the transect in pasture (Fig. 15, top graph). Original measurements, however, result from a combination of different factors operating at different spatial scales; some of them such as measurement errors or microscale variation in biological activities likely reduce artificially the overall correlation between attributes. This effect is made clear by looking at the relation between the local or regional components of pH and electrical conductivity shown in Fig. 14. Whereas the two attributes are slightly negatively correlated at a local scale, they display a strong positive correlation at the regional scale (Fig. 15, bottom graph), which agrees with the change of sign of the cross semivariogram noticed on Figure 8.

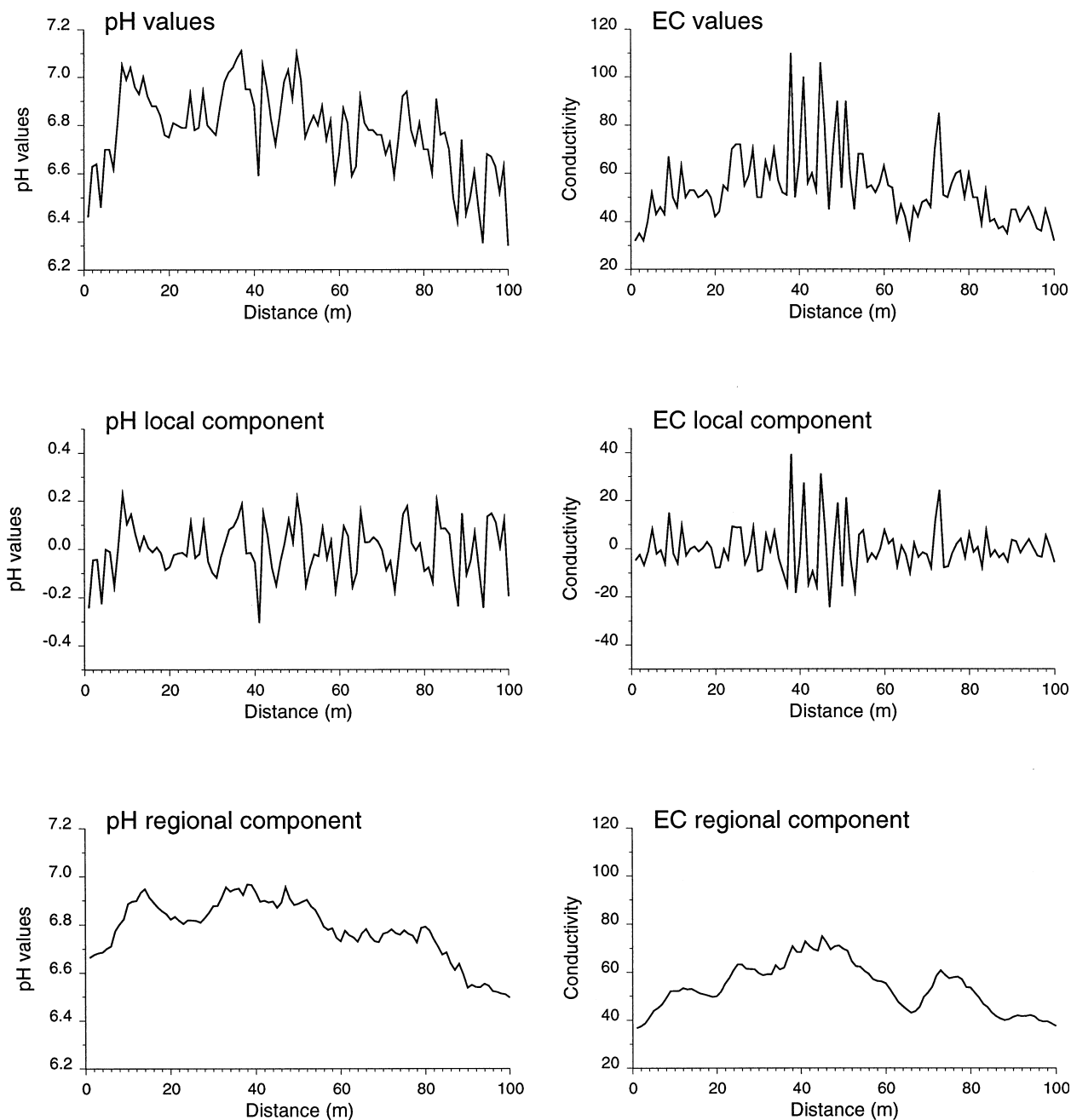
Such scale-dependent relations are frequent in soil science where many properties are controlled by the same physical processes that operate at different spatial scales and influence these properties in different ways. Goovaerts and Webster (1994) found a strong positive correlation between topsoil cobalt and copper concentrations once large micro-scale variations due to measurement and procedural errors were filtered from the original data. They also showed that the classification by soil association accounts for a large proportion of the variance at the regional scale, suggesting that the parent material contributes substantially more to the trace element content of the soil than had been thought earlier on the basis of a classical analysis where all different sources of variation were mixed. Accounting for the spatial scale in the study of correlation can thus improve our understanding of complex relations between soil attributes and their environment.

Instead of computing the correlation coefficient between previously estimated spatial components, the correlation between two attributes  $z$  and  $y$  at a particular scale  $l$  can be assessed directly from the parameters of the linear model of coregionalization as:

$$l_{ZY} = \text{Corr}[Z_l(\mathbf{u}), Y_l(\mathbf{u})] = \frac{b'_{ZY}}{\sqrt{b'_{ZZ} \cdot b'_{YY}}} \quad (27)$$

Application of expression (27) to the linear model of coregionalization (18) yields the following structural correlation coefficients  $l_{ZY}$  for the pair pH-electrical conductivity:  $-0.31$  and  $0.99$  at the local (up to 6 m) and regional (26 m) scales, respectively. These values are somewhat larger than those of Fig. 15, which are based on estimated values.

Multivariate factorial kriging is a variant of cokriging which aims at estimating and mapping the different sources of spatial variability which are common to the set of attributes. The maps of the so-called regionalized factors have been used to summarize major features of multivariate soil data sets at different spatial scales identified on the experimental semivariogram (Goovaerts 1992; Dobermann et al. 1995). More recently, Monestiez et al. (1997) used factorial kriging to map



**Fig. 14** Decomposition of the pasture transect of pH and electrical conductivity values into local and regional components on the basis of the nested semivariogram models of Fig. 12

the short-range (120 km) and long-range (300 km) spatial genetic structures of wild populations of perennial ryegrass in France.

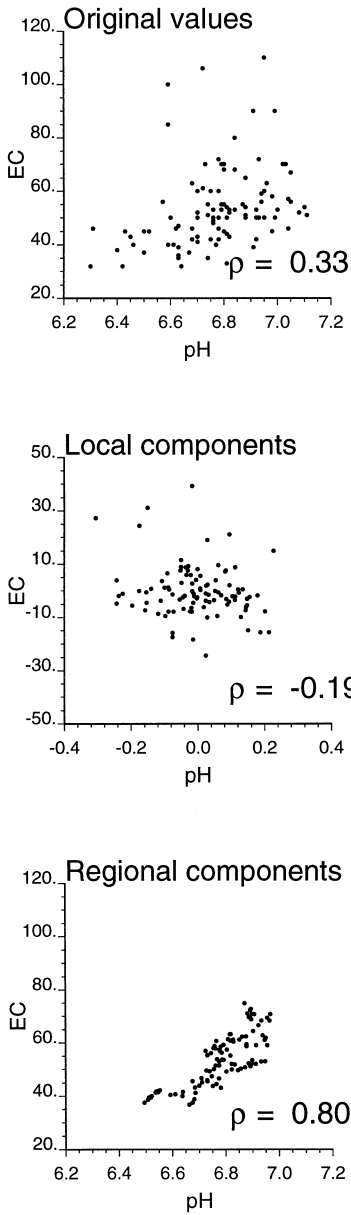
### Advanced geostatistical methods

Although geostatistics is still mainly used for description of spatial patterns and prediction of unsampled soil attributes, the last 5 years have seen the develop-

ment of new techniques which allow soil scientists to tackle advanced issues such as the assessment of the uncertainty about soil quality parameters or soil pollutant concentrations, or the stochastic simulation of the spatial distribution of attribute values. These topics are briefly addressed in this section, and the reader should refer to Goovaerts (1997a, pp. 259–436) or Goovaerts (1998) for a more detailed presentation.

### Assessing the local uncertainty

Mapping metal concentrations or other soil properties is often a preliminary step towards decision making such as the delineation of polluted areas or the identi-



**Fig. 15** Scattergrams of pH vs electrical conductivity values: original concentrations and spatial components displayed on Fig. 14

fication of zones that are suitable for crop growth. In many applications, the focus is on the exceedence of particular values such as regulatory thresholds for heavy metal contamination or critical values of limiting factors (e.g., depth to parent material, soil acidity) for soil quality. There is necessarily some error attached to the kriging estimate, and this must be accounted for in the decision rule. Ignoring this uncertainty may, for example, lead one to declare safe a contaminated location on the basis of a wrong estimate of pollutant concentration which is slightly below the regulatory threshold.

Geostatistics is increasingly used to estimate and map the risk of exceeding specific threshold values in soil science (Webster and Oliver 1989; Smith et al. 1993; Goovaerts and Journel 1995; Goovaerts et al. 1997; Mohammadi et al. 1997). The most straightforward approach is based on a coding of each observation  $z(\mathbf{u}_\alpha)$  into an indicator of non-exceedence of the target threshold value  $z_k$ ; recall expression (4)

$$i(\mathbf{u}_\alpha; z_k) = \begin{cases} 1 & \text{if } z(\mathbf{u}_\alpha) \leq z_k \\ 0 & \text{otherwise} \end{cases}$$

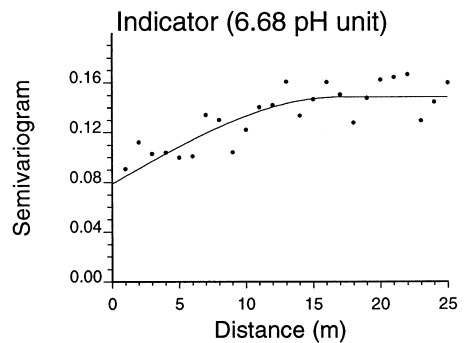
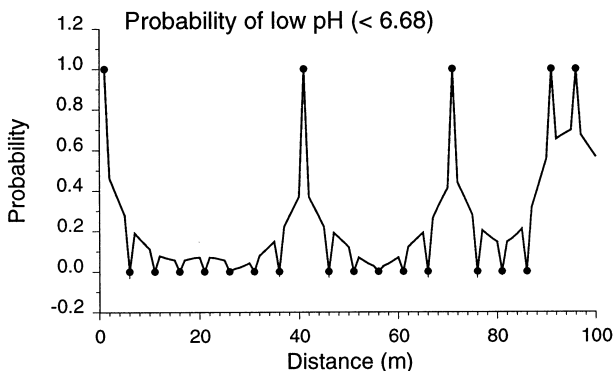
The probability that the attribute value  $z$  does not exceed  $z_k$  at the unsampled location  $\mathbf{u}$  is then estimated from the indicator transforms using a kriging estimator similar to the one developed for continuous attributes. For example, ordinary indicator kriging builds the probability estimate as a linear combination of neighboring indicator data:

$$[\text{Prob}\{Z(\mathbf{u}) \leq z_k | (n)\}]^* = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u}; z_k) i(\mathbf{u}_\alpha; z_k) \quad (28)$$

where the weights  $\lambda_\alpha(\mathbf{u}; z_k)$  are obtained by solving a system of linear equations identical to system (20) except that semivariogram values are now derived from the model  $\gamma_I(\mathbf{h}; z_k)$  fitted to the experimental indicator semivariogram (5).

Figure 16 shows the probability that the soil pH is no greater than 6.68 along the transect in pasture. These probabilities were estimated every meter using ordina-

**Fig. 16** Kriged estimates of the probability for the soil pH along the pasture transect to be smaller than 6.68. The information available consists of 20 pH values (black dots) plus the indicator semivariogram model





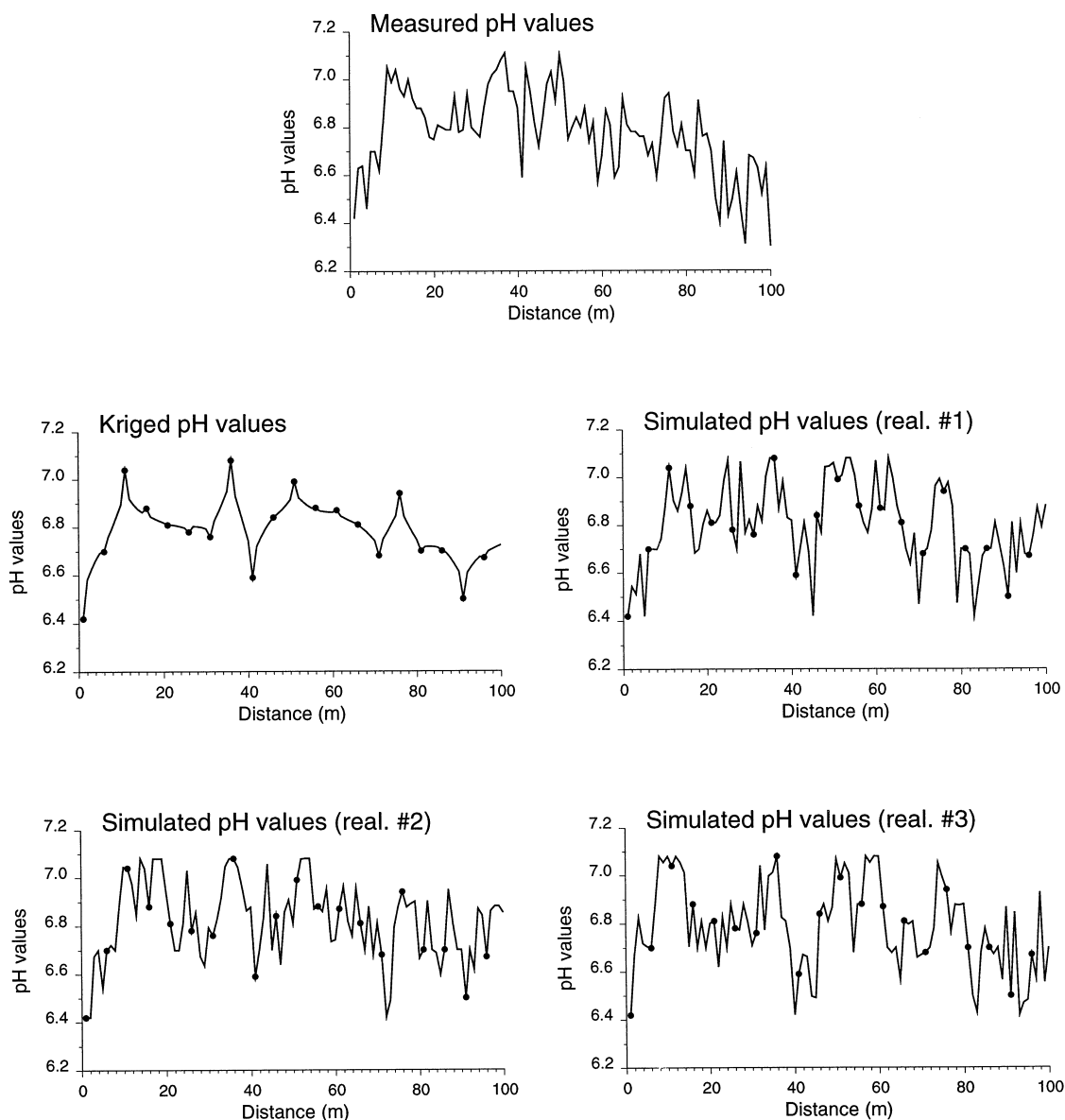
ry indicator kriging and the indicator semivariogram model displayed in Fig. 16. Note that the modeling of indicator semivariograms follows the same rules as that of semivariograms of continuous attributes.

Prediction of probabilities is generally improved by incorporating additional information, e.g., by cokriging of indicator transforms of both primary and secondary information. Goovaerts and Journel (1995) showed how probabilities of deficiency in copper and cobalt can be derived from the calibration of a soil map and combined with precise measurements of metal concentration to map the risk of deficiency of these metals in the soil. They have also introduced the concept of loss function which, in combination with probability maps, allowed one to map the economical impact of declaring wrongly that a location is deficient or sufficient in these metals. Similar applications to the delineation of contaminated soils are presented in Colin et al. (1996) and Goovaerts et al. (1997).

## Stochastic simulation

The comparison of the transects of measured and estimated pH values displayed in Fig. 17 indicates that kriging smooths out local details of the spatial variation of pH values in pasture. Such smoothing results from the least-squares criterion of the kriging algorithm and leads to an unfortunate overestimation of small values and underestimation of large values. In stochastic simulation, the aim is not the minimization of the error variance but to generate a set of values that reproduces statistics such as the sample histogram or the semivario-

**Fig. 17** Transects of 100 pH values measured in the topsoil of a pasture (top graph), and the transects of estimated or simulated values constructed from only 20 observations depicted by *black dots*. Three different realizations were generated using stochastic simulation



gram model while honoring data at their locations. Mathematically speaking, each set of values is viewed as a realization of the random function.

For example, Fig. 17 shows three realizations of the spatial distribution of pH values generated using the same information as in kriging, that is, 20 pH values depicted by black dots and the semivariogram model of Fig. 11 (top graph). The transects of simulated values look more “realistic” than the transect of kriging estimates because they reproduce the spatial variability modeled from the sample information. Stochastic simulation is thus increasingly preferred to kriging for all applications where the spatial variability of the measured field must be preserved, such as the delineation of contaminated areas (Desbarats 1996; Goovaerts 1997b) or the modeling of solute transport in the vadoze zone (Vanderborght et al. 1997). Like estimation, simulation can be accomplished using a growing variety of techniques, and most of them are described in Goovaerts (1997a, pp. 376–424).

Each of the three realizations shown in Fig. 17 is a plausible representation of the unique and unknown distribution of pH values along the transect in that each simulated transect honors the 20 data and reproduces approximately the sample histogram and semivariogram model. Differences between the three realizations thus provide a measure of *spatial* uncertainty: features such as segments of high pH are deemed certain if seen on most of the realizations, and their probability of occurrence (i.e., probability that a given threshold is jointly exceeded at a series of locations) can be computed as long as the realizations are equiprobable. Unlike stochastic simulation, indicator kriging provides only a measure of *local* uncertainty in that it estimates the probability that the threshold is exceeded at a single location; recall expression (28).

The impact of a given scenario, such as application of a particular amount of fertilizer, can be investigated from a simulated map that reproduces aspects of the pattern of spatial dependence or other statistics deemed consequential for the problem at hand (e.g., connectivity of large values, spatial correlation with secondary attributes). Moreover, the availability of many equiprobable realizations allows one to assess the uncertainty about the consequences of this particular scenario, such as leaching of fertilizer and contamination of groundwater, which results from our imperfect knowledge of the spatial distribution of soil attribute values.

---

## Conclusions

Any quantitative processing of soil information is incomplete as long as it takes no account of the spatial coordinates of observations. Geostatistical characterization of the spatial variability through semivariograms or correlograms generally brings new insight into the way soil attributes are influenced by the environment,

such as geographical distribution of soil types or topography. Indicator semivariograms provide additional information about the spatial distribution of specific classes of values of continuous attributes as well as categorical soil attributes. Cross semivariograms and cross correlograms complete the description by assessing spatial relations between pairs of attributes.

A key step in any geostatistical analysis is the fitting of permissible models to the experimental semivariogram values. In this paper, basic rules for univariate and multivariate semivariogram modeling have been recalled, with the hope that it will reduce the use of unstable (i.e., Gaussian) or non-permissible models in soil studies. Permissible models are, however, one ingredient of the modeling process, and whenever physical knowledge of the area and phenomenon under study is available it should be accounted for in the building of a model of spatial variability.

The existence of a model of spatial variability (dependence) allows one to tackle the problem of estimating soil attribute values at unsampled locations. Geostatistics offers a palette of linear least-squares (kriging) algorithms to incorporate different types of information in the estimation process. Kriging algorithms can also be modified to decompose original observations into specific spatial components, enabling the spatial fluctuations and interactions between soil attributes to be studied at different scales.

Recent developments in indicator geostatistics allow a straightforward assessment of the probability to exceed critical values, such as regulatory thresholds in soil pollution or criteria for soil quality. Another way to model uncertainty is to generate many images (realizations) that all honor the data and reproduce aspects of the patterns of spatial dependence or other statistics deemed consequential for the problem at hand. In the future, more attention should be devoted to the use of these various measures of local or spatial uncertainty in decision making, such as determination of the amount of fertilizer to be applied or the delineation of polluted areas targeted for remediation.

Processes such as weathering of parent material, soil erosion or solute transport are dynamic, leading to variation of soil attribute values in both space and time. Temporal fluctuations are particularly important for microbiological processes and cannot be ignored in statistical modeling of soil variability. Although not introduced in the present paper, spatial geostatistical tools such as semivariogram or kriging can be extended to the space-time domain.

---

## References

- Bierkens MFP, Burrough PA (1993a) The indicator approach to categorical soil data. I. Theory. *J Soil Sci* 44:361–368
- Bierkens MFP, Burrough PA (1993b) The indicator approach to categorical soil data. II. Application to mapping and land use suitability analysis. *J Soil Sci* 44:369–381

- Burgess TM, Webster R (1980a) Optimal interpolation and isarithmic mapping of soil properties. I. The variogram and punctual kriging. *J Soil Sci* 31:315–331
- Burgess TM, Webster R (1980b) Optimal interpolation and isarithmic mapping of soil properties. II. Block kriging. *J Soil Sci* 31:333–341
- Burgess TM, Webster R, McBratney AB (1981) Optimal interpolation and isarithmic mapping of soil properties. VI. Sampling strategy. *J Soil Sci* 31:643–659
- Colin P, Froidevaux R, Garcia M, Nicoletis S (1996) Integrating geophysical data for mapping the contamination of industrial sites by polycyclic aromatic hydrocarbons: a geostatistical approach. In: Rouhani S, Srivastava RM, Desbarats AJ, Cromer MV, Johnson AI (eds) *Geostatistics for environmental and geotechnical applications*. American Society for Testing and Materials STP 1283, Philadelphia, pp 69–87
- Cressie N (1985) Fitting variogram models by weighted least squares. *Math Geol* 17:563–586
- Desbarats AJ (1996) Modeling spatial variability using geostatistical simulation. In: Rouhani S, Srivastava RM, Desbarats AJ, Cromer MV, Johnson AI (eds) *Geostatistics for environmental and geotechnical applications*. American Society for Testing and Materials STP 1283, Philadelphia, pp 32–48
- Deutsch CV, Journel AG (1998) *GSLIB: geostatistical software library and user's guide*, 2nd edn. Oxford University Press, New York
- Dobermann A, Goovaerts P, George T (1995) Sources of soil variation in an acid Ultisol of the Philippines. *Geoderma* 68:173–191
- Dobermann A, Goovaerts P, Neue HU (1997) Scale-dependent correlations among soil properties in two tropical lowland rice fields. *Soil Sci Soc Am J* 61:1483–1496
- Englund E, Sparks A (1991) *Geo-EAS 1.2.1 user's guide*. EPA report # 600/8-91/008, EPA-EMSL, Las Vegas, Nev
- Gallichand J, Marcotte D (1993) Mapping clay content for subsurface drainage in the Nile Delta. *Geoderma* 58:165–179
- Goovaerts P (1992) Factorial kriging analysis: a useful tool for exploring the structure of multivariate spatial soil information. *J Soil Sci* 43:597–619
- Goovaerts P (1994a) Study of spatial relationships between two sets of variables using multivariate geostatistics. *Geoderma* 62:93–107
- Goovaerts P (1994b) Comparison of coIK, IK, and mIK performances for modeling conditional probabilities of categorical variables. In: Dimitrakopoulos R (ed) *Geostatistics for the next century*. Kluwer, Dordrecht, pp 18–29
- Goovaerts P (1997a) *Geostatistics for natural resources evaluation*. Oxford University Press, New York
- Goovaerts P (1997b) Kriging vs stochastic simulation for risk analysis in soil contamination. In: Soares A, Gómez-Hernández J, Froidevaux R (eds) *geoENV I – geostatistics for environmental applications*. Kluwer, Dordrecht, pp 247–258
- Goovaerts P (1998) *Geostatistics in soil science: state-of-the-art and perspectives*. *Geoderma* (in press)
- Goovaerts P, Journel AG (1995) Integrating soil map information in modelling the spatial variation of continuous soil properties. *Eur J Soil Sci* 46:397–414
- Goovaerts P, Webster R (1994) Scale-dependent correlation between topsoil copper and cobalt concentrations in Scotland. *Eur J Soil Sci* 45:79–95
- Goovaerts P, Gérard G, Frankart R (1989) Etude de la variabilité spatiale de quelques propriétés chimiques du sol en Fagne de Chimay, Belgique. *Pédologie* 39:191–207
- Goovaerts P, Webster R, Dubois J-P (1997) Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environ Ecol Stat* 4:31–48
- Gotway CA, Hartford AH (1996) Geostatistical methods for incorporating auxiliary information in the prediction of spatial variables. *J Agric Biol Environ Stat* 1:17–39
- Goulard M (1989) Inference in a coregionalization model. In: Armstrong M (ed) *Geostatistics*. Kluwer, Dordrecht, pp 397–408
- Goulard M, Voltz M (1992) Linear coregionalization model: tools for estimation and choice of cross-variogram matrix. *Math Geol* 24:269–286
- Isaaks EH, Srivastava RM (1989) *An introduction to applied geostatistics*. Oxford University Press, New York
- Journel AG, Huijbregts CJ (1978) *Mining geostatistics*. Academic Press, New York
- Leenaers H, Okx JP, Burrough PA (1990) Employing elevation data for efficient mapping of soil pollution on floodplains. *Soil Use Manag* 6:105–114
- Leonte D, Schofield N (1996) Evaluation of a soil contaminated site and clean-up criteria: a geostatistical approach. In: Rouhani S, Srivastava RM, Desbarats AJ, Cromer MV, Johnson AI (eds) *Geostatistics for environmental and geotechnical applications*. American Society for Testing and Materials STP 1283, Philadelphia, pp 133–145
- Matheron G (1982) *Pour une analyse krigeante de données régionalisées*. Centre de Géostatistique, Report N-732, Fontainebleau
- McBratney AB, Webster R (1981) Detection of ridge and furrow pattern by spectral analysis of crop yield. *Int Stat Rev* 49:45–52
- McBratney AB, Webster R (1986) Choosing functions for semi-variograms of soil properties and fitting them to sampling estimates. *Soil Sci* 37:617–639
- Mohammadi J, Van Meirvenne M, Goovaerts P (1997) Mapping cadmium concentration and the risk of exceeding a local sanitation threshold using indicator geostatistics. In: Soares A, Gómez-Hernández J, Froidevaux R (eds) *geoENV I – geostatistics for environmental applications*. Kluwer, Dordrecht, pp 327–337
- Monestiez P, Goulard M, Charmet G, Balfourier F (1997) Analysing spatial genetic structures by multivariate geostatistics: study of wild populations of perennial ryegrass. In: Baafi EY, Schofield NA (eds) *Geostatistics Wollongong '96*. Kluwer, Dordrecht, pp 1197–1208
- Odeh IOA, McBratney AB, Slater BK (1997) Predicting soil properties from ancillary information: non-spatial models compared with geostatistical and combined methods. In: Baafi EY, Schofield NA (eds) *Geostatistics Wollongong '96*. Kluwer, Dordrecht, pp 1008–1019
- Oliver MA, Webster R (1986a) Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. *Geogr Anal* 18:227–242
- Oliver MA, Webster R (1986b) Semi-variograms for modelling the spatial pattern of landform and soil properties. *Earth Surface Proc Landforms* 11:491–504
- Pannatier Y (1996) *Variowin: software for spatial data analysis in 2-D*. Springer, New York Berlin Heidelberg
- Parkin TB (1987) Soil microsites as a source of denitrification variability. *Soil Sci Soc Am J* 51:1194–1199
- Robertson GP, Huston MA, Evans FC, Tiedje JM (1988) Spatial variability in a successional plant community: patterns of nitrogen availability. *Ecology* 69:1517–1524
- Rossi RE, Mulla DJ, Journel AG, Franz EH (1992) Geostatistical tools for modeling and interpreting ecological spatial dependence. *Ecol Monogr* 62:277–314
- Smith JL, Halvorson JJ, Papendick RI (1993) Using multiple-variable indicator kriging for evaluating soil quality. *Soil Sci Soc Am J* 57:743–749
- Sutherland RA, Kessel C van, Pennock D (1991) Spatial variability of nitrogen-15 natural abundance. *Soil Sci Soc Am J* 55:1339–1347
- Trangmar BB, Yost RS, Uehara G (1985) Application of geostatistics to spatial studies of soil properties. *Adv Agron* 38:45–94
- Vanderborcht J, Jacques D, Mallants D, Tseng PH, 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, Dordrecht, pp 283–295

- Vieira SR, Hatfield JL, Nielsen DR, Biggar JW (1983) Geostatistical theory and application to variability of some agronomical properties. *Hilgardia* 5:1–75
- Voltz M, Goulard M (1994) Spatial interpolation of soil moisture retention curves. *Geoderma* 62:109–123
- Wackernagel H (1988) Geostatistical techniques for interpreting multivariate spatial information. In: Chung CF, Fabbri AG, Sinding-Larsen R (eds) *Quantitative analysis of mineral and energy resources*. Reidel, Dordrecht, pp 393–409
- Wackernagel H (1995) *Multivariate geostatistics: an introduction with applications*. Springer, Berlin Heidelberg New York
- Warrick AW, Myers DE, Nielsen DR (1986) Geostatistical methods applied to soil science. In: *Methods of soil analysis, part 1, 2nd edn Physical and mineralogical methods*. Agronomy Monograph 9, pp 53–82
- Webster R (1996) What is kriging? *Aspects Appl Biol* 46:57–66
- Webster R (1997) Software review. *Eur J Soil Sci* 48:173–175
- Webster R, Boag B (1992) Geostatistical analysis of cyst nematodes in soil. *J Soil Sci* 43:583–595
- Webster R, Burgess TM (1980) Optimal interpolation and isarithmic mapping of soil properties. III. Changing drift and universal kriging. *J Soil Sci* 31:505–524
- Webster R, Oliver MA (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
- Wollum AG, Cassel DK (1984) Spatial variability of *Rhizobium japonicum* in two North Carolina soils. *Soil Sci Soc Am J* 48:1082–1086
- Yates SR, Warrick AW (1987) Estimating soil water content using cokriging. *Soil Sci Soc Am J* 51:23–30