Modeling the Semivariogram: New Approach, Methods Comparison, and Simulation Study

Article

3 authors, including:

Konstantin Krivoruchko
Environmental Systems Research Institute (ESRI)
34 PUBLICATIONS **1,929** CITATIONS
SEE PROFILE

Jay M. Ver Hoef
National Oceanic and Atmospheric Administration
241 PUBLICATIONS **10,339** CITATIONS
SEE PROFILE

Gribov, A., K. Krivoruchko, and J. M. Ver Hoef, 2006, Modeling the semivariogram: New approach, methods comparison, and simulation study, *in* T. C. Coburn, J. M. Yarus, and R. L. Chambers, eds., Stochastic modeling and geostatistics: Principles, methods, and case studies, volume II: AAPG Computer Applications in Geology 5, p. 45–57.

Chapter 5

Modeling the Semivariogram: New Approach, Methods Comparison, and Simulation Study

A. Gribov
K. Krivoruchko
Environmental Systems Research Institute
Redlands, California, U.S.A.

J. M. Ver Hoef
Alaska Department of Fish and Game
Fairbanks, Alaska, U.S.A.

ABSTRACT

This chapter proposes some new methods for computing empirical semivariograms and covariances and for fitting semivariogram and covariance models to empirical data. Grid-based empirical semivariograms and covariances are described, in which the grid values are smoothed using triangular kernels. A model-fitting procedure using modified iterative weighted least squares is presented. This algorithm is shown to be reliable for a wide range of data types and conditions, and its implementation in commercial software is discussed. Comparisons to restricted maximum likelihood estimation are also discussed.

INTRODUCTION

Investigation of the structure functions known as the covariance and the semivariogram has a long history. Geostatistical textbooks commonly present structural analysis and spatial interpolation in the context of geological investigations. However, both the semivariogram and the related estimation procedure known as kriging were used in other fields such as meteorology years before they were popularized by geoscientists. Early in the 20th century, Keller and Friedmann (1925) introduced a generalized form of the correlation function for hydrodynamic applications for which they assumed that moments of order greater than 2 could be ignored when characterizing average motion. Essentially, they computed the covariance in four dimensions: three-dimensional space plus time. Later, Kolmogorov (1941) presented a simplification of the theory for locally homogeneous and isotropic turbulence in two dimensions in which he showed that a velocity structure function (semivariogram) is proportional to two-thirds the power of the distance in a range for which the assumptions of local stationarity are true. After Kolmogorov's publications on structural analysis, several case studies on structure functions appeared, mostly in the former Soviet Union. Later, Gandin (1959) presented an optimal spatial interpolation procedure in two dimensions

DOI:10.1306/1063806CA53228

(kriging) that the meteorological community embraced, using semivariogram construction as a necessary step in the process. Power functions, power exponential functions, and J- and K-Bessel functions were commonly used to model the semivariogram. Soviet meteorological monitoring networks in the 1950s and early 1960s included several dozen stations, and Gandin et al. commonly used all pairs of points to estimate the parameters of the semivariogram and covariance models (Gandin, 1963). Gandin mentioned that he used a least-squares algorithm to fit the parameters of the model but did not provide any details.

Today, several reliable methods are available to estimate or fit semivariogram and (cross-)covariance models. This chapter details the commercial software implementation of one of the most popular methods: least-squares model fitting. One particular advantage of least squares is that it produces a better visual fit of the empirical semivariogram than likelihood methods. This is true simply because least squares produces the best-fitting trend through the empirical semivariogram, whereas likelihood methods more closely honor individual points on the graph.

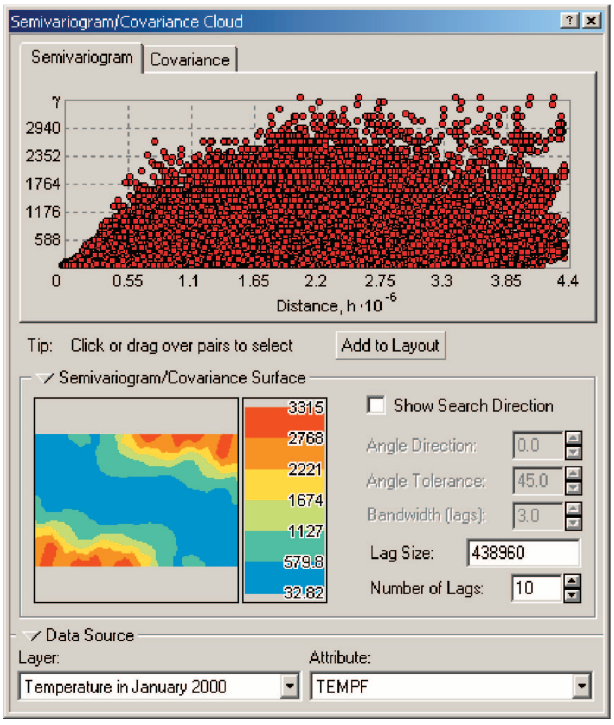


Figure 1. Computer screen shot showing an empirical semivariogram cloud. The *y*-axis of each red dot represents the empirical semivariogram of a pair of locations.

EMPIRICAL SEMIVARIOGRAM ESTIMATION

Let $z(\mathbf{s}_p)$ be a value observed at the pth location $\mathbf{s}_p$, where $\mathbf{s}_p = (x_p, y_p)$ is the vector containing the x- and y-spatial coordinates. Define the lag, $\mathbf{h} = \mathbf{s}_2 - \mathbf{s}_1$, as the vector from point $\mathbf{s}_1$ to point $\mathbf{s}_2$. The semivariogram cloud is defined as half the squared differences, $\gamma_{pq} = 0.5 \cdot [z(\mathbf{s}_p) - z(\mathbf{s}_q)]^2$ for all possible pairwise lags $\{(\mathbf{s}_p, \mathbf{s}_q); p,q = 1, 2, \ldots, \mathbf{n}\}$, plotted as a function of the distance $\|\mathbf{h}\| = \|\mathbf{s}_p - \mathbf{s}_q\| = [(x_p - x_q)^2 + (y_p - y_q)^2]^{1/2}$. Typical applications deal with hundreds and sometimes thousands of data locations, and it is commonly difficult to see any pattern because so many pairwise lags exist (Figure 1).

Not only is it difficult to see any pattern, but fitting a semivariogram model to so many values is difficult even for modern computers. Instead, the squared differences γ_{ij} are commonly binned into those distances and directions that are similar. Thus, the following empirical semivariogram formula

$$\hat{\gamma}(\mathbf{h}_k) \equiv \frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} [z(\mathbf{s}_p) - z(\mathbf{s}_q)]^2$$

is commonly used, where $z(\cdot)$ are observed values, $N(\mathbf{h})$ is the set of pairs of values at $\mathbf{s}_p$ and $\mathbf{s}_q$ that have a similar lag $\mathbf{h} \in T(\mathbf{h}_k)$, with $T(\mathbf{h}_k)$ defined to be a tolerance region around $\mathbf{h}_k$ and $|N(\mathbf{h})|$ defined to be the number of distance pairs in the set $N(\mathbf{h})$. The question remains how to best bin the data, which is equivalent to choosing the sets $N(\mathbf{h})$.

Most commonly, the binning is computed on a tolerance region $T(\mathbf{h}_k)$ representing a radial sector. Figure 2 shows the method most commonly used in software packages, including Geostatistical Software Library (Deutsch and Journel, 1998), Splus (Kaluzny et al., 1998), and SAS (SAS Institute, Inc., 1996).

A rule of thumb has been to set the maximum lag for binning to be half the maximum distance between pairs in the data set and choose the number of bins so that there are more than 30 pairs per bin (Journel and Huijbregts, 1978). However, this choice seems to be based on tradition alone, and scant theoretical evidence exists to recommend its use over other schemes.

Instead of the radial sector method, the procedure described here employs a binning approach based on tolerance regions $T(\mathbf{h}_k)$ that are rectangles distributed uniformly on a grid (Figure 3). After averaging

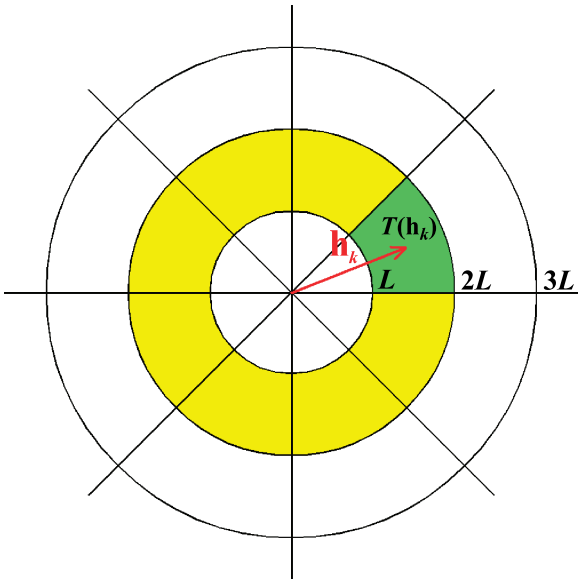


Figure 2. Radial sector tolerance regions for binning empirical semivariograms. The red vector $\mathbf{h}_k$ is the center of the bin, which has a radial tolerance region $T(\mathbf{h}_k)$ shown in green. Any vector that falls within the green area is assigned to the bin for $\mathbf{h}_k$.

the values in each bin, average values are plotted for each $\mathbf{h}_k$. This is done in two ways: constructing a semivariogram graph by plotting the average bin values versus distance and creating a semivariogram surface by plotting the average bin values directly on their grid, preserving the spatial orientation of the lags and using colors to indicate the lag values. The semivariogram surface uses cool colors like blue and green for low values and warm colors like red and yellow for high values. The color scale next to the semivariogram graph links it to both the graph and the semivariogram surface; that is, each point in the graph corresponds to a cell on the surface (Figure 4).

The grid method is more understandable because the semivariogram surface and semivariogram cloud are linked in an obvious manner (Figure 4). The grid method also helps to ensure that the number of pairs in each cell is approximately the same, which is advantageous for visually assessing the fit of the model.

If the data are uniformly distributed on rectangular domain having dimensions A by B, the distance between points has the following distribution:

$$f(\Delta x, \Delta y) = \frac{(A - |\Delta x|) \cdot (B - |\Delta y|)}{A^2 \cdot B^2} = \frac{\left(1 - \frac{|\Delta x|}{A}\right) \cdot \left(1 - \frac{|\Delta y|}{B}\right)}{A \cdot B}$$

where $|\Delta x| \leq A$ and $|\Delta y| \leq B$. The most important pairs are separated by relatively short distances, $|\Delta x| << A$ and $|\Delta y| << B$, and their distance distribution is close to the uniform distribution. Hence, the number of pairs in a rectangular averaging scheme is approximately constant for small lags. In a sector averaging scheme, the number of pairs can differ significantly from sector to sector. Note that it is not necessary to explicitly define the lag interval.

A three-dimensional view of the semivariogram surface is presented in Figure 5 for a very large simulated data set with an anisotropic structure. The sector and grid methods are presented for a 90° sector for tolerance regions $T(\mathbf{h}_k)$ of fixed size (hereafter called fixed lag sizes) and for tolerance regions that increase logarithmically in size (hereafter called logarithmic lag sizes). Note that for small fixed lags, both the sector and grid methods may have bins that are too large near the origin, and hence, in this case, the logarithmic lag size may be more effective at describing variability in the data.

The requirement to have a relatively large number of points in each lag interval, say 30 or more for each bin, can be overcome using the algorithm proposed below. In addition, fewer samples per bin can be used if smoothing has been performed (e.g., by relying on values in neighboring bins).

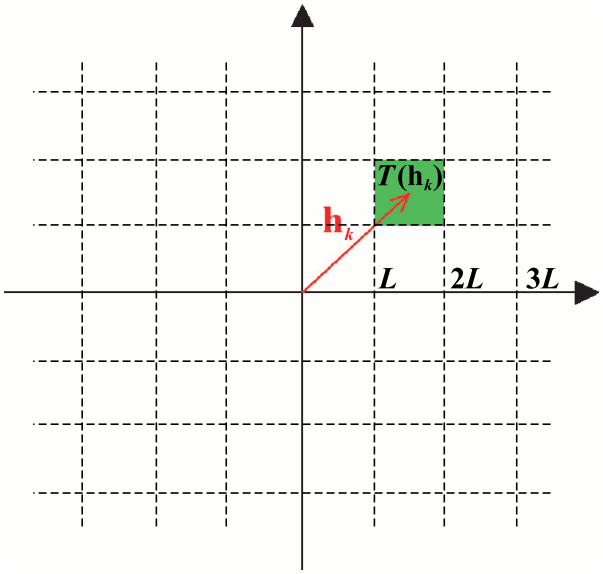


Figure 3. Grid tolerance regions for binning empirical semivariograms. The red vector $\mathbf{h}_k$ is the center of the bin, which has a radial tolerance region $T(\mathbf{h}_k)$ shown in green. Any vector that falls within the green area is assigned to the bin for $\mathbf{h}_k$.

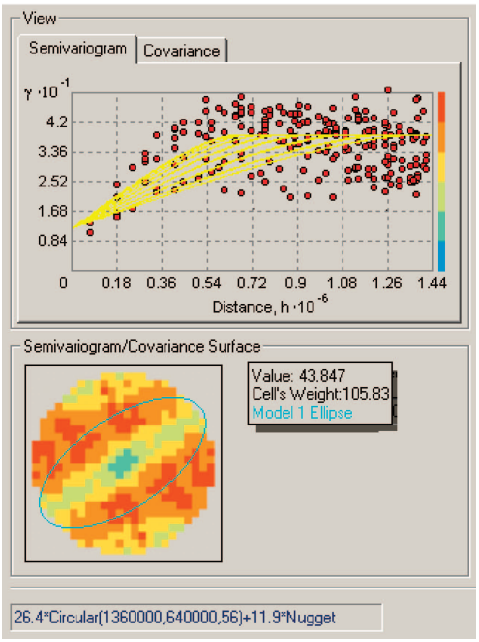


Figure 4. Computer screen shot showing a semivariogram cloud and an anisotropic model based on the semivariogram surface. Yellow lines represent models in different directions. Ellipse of anisotropy is displayed over the semivariogram surface (blue).

The algorithm proceeds as follows. A triangular kernel is used to assign weighted semivariogram values to each cell, depending on how close they are to the center of the cell (Figure 6). For example, the weight for the cell in Figure 6 containing the blue dot can be taken as the product of the two marginal profiles. The weight $w_k(\mathbf{h})$ for vector $\mathbf{h}$ for the kth bin with center $\mathbf{h}_k$ is

$$w_k(\mathbf{h}) = \left(1 - \frac{|x_k - x|}{L}\right)\left(1 - \frac{|y_k - y|}{L}\right) \times I(|x_k - x| \le L)I(|y_k - y| \le L)$$

where L is the length of the side of a square bin, x_k is the x-coordinate of $\mathbf{h}_k$, y_k is the y-coordinate of $\mathbf{h}_k$, x is the x-coordinate of $\mathbf{h}$, y is the y-coordinate of $\mathbf{h}$, and $I(\bullet)$ is the indicator function that is equal to 1 if its argument is true and zero otherwise. Modification of this method can be used for other types of averaging presented in Figure 5.

Commonly, a table of binned semivariogram values shows the number of pairs for that bin. Because smoothing with weights is used, an equivalent presentation for the proposed method is to present the sum of the weights used in each bin.

EMPIRICAL COVARIANCE AND CROSS-COVARIANCE ESTIMATION

It is also possible to estimate the empirical covariance function

$$\hat{C}(\mathbf{h}_k) \equiv \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} [z(\mathbf{s}_p) - \bar{z}][z(\mathbf{s}_q) - \bar{z}]$$

where $\bar{z}$ is the average of all the data, and all other notation is the same as for $\hat{\gamma}(\mathbf{h}_k)$. In comparison, $\hat{\gamma}(\mathbf{h}_k)$ is an unbiased estimate of the semivariogram, whereas $\hat{C}(\mathbf{h}_k)$ is a biased estimate of the population covariance. Thus, $\hat{\gamma}(\mathbf{h}_k)$ is generally preferred.

Choosing a structure function for modeling the correlation between variables is more complicated (Ver Hoef and Cressie, 1993). The traditional cross-semivariogram is defined as

$$\tau_{ij}(\mathbf{h}) \equiv \frac{1}{2} E[z_i(\mathbf{s}) - z_i(\mathbf{s} + \mathbf{h})][z_j(\mathbf{s}) - z_j(\mathbf{s} + \mathbf{h})]$$

which has, as an estimator,

$$\hat{\tau}_{ij}(\mathbf{h}_k) \equiv \frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} [z_i(\mathbf{s}_p) - z_i(\mathbf{s}_q)][z_j(\mathbf{s}_p) - z_j(\mathbf{s}_q)]$$

for the ith and jth variable types. The notation follows that of $\hat{\gamma}(\mathbf{h}_k)$ and $\hat{C}(\mathbf{h}_k)$. The cross-semivariogram, $\tau_{ij}(\mathbf{h})$, has two problems. First, it does not accommodate asymmetry; that is, the cokriging equations are only obtained by assuming that $C_{ij}(\mathbf{h}) = C_{ij}(-\mathbf{h})$. Second, both variables must be measured at the same location in order for that location to be used in the empirical estimator.

Alternatively, the cross-semivariogram can be defined as

$$\gamma_{ij}(\mathbf{h}) \equiv \frac{1}{2} \mathrm{var}[z_{\tilde{i}}(\mathbf{s}) - z_{\tilde{j}}(\mathbf{s} + \mathbf{h})]$$

which has an estimator

$$\hat{\gamma}_{ij}(\mathbf{h}_k) \equiv \frac{1}{2|N(\mathbf{h})|} \sum_{N(\mathbf{h})} [\tilde{z}_i(\mathbf{s}_p) - \tilde{z}_j(\mathbf{s}_q)]^2$$

where $\tilde{z}_i(\mathbf{s}_p)$ is standardized by subtracting off the mean and dividing by the standard deviation. The

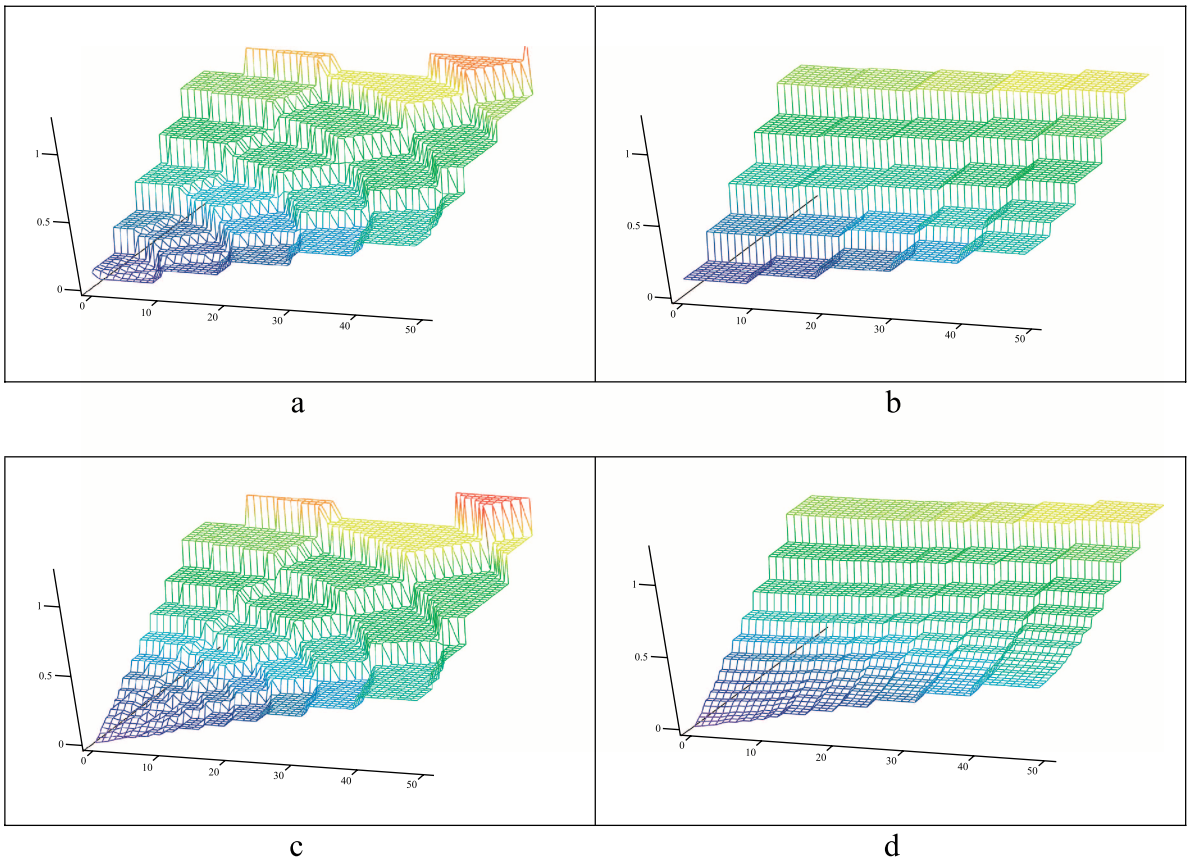


Figure 5. Empirical semivariograms: (a) sector method with fixed lag size, (b) grid method with fixed lag size, (c) sector method with logarithmic lag size, and (d) grid method with logarithmic lag size.

cross-semivariogram, $\gamma_{ij}(\mathbf{h})$, is asymmetric and allows the cokriging equations to be obtained more generally than $\tau_{ij}(\mathbf{h})$. However, because of standardization,

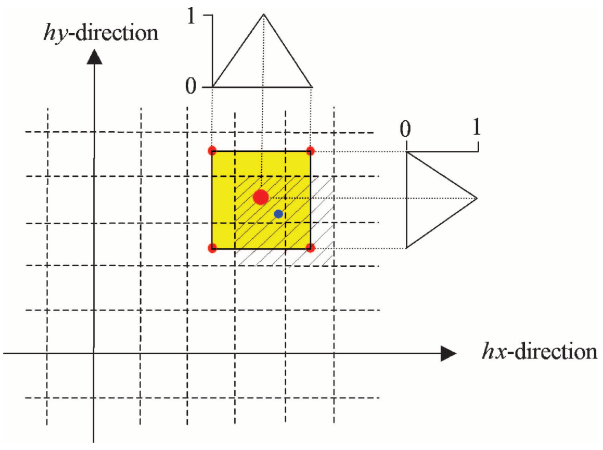


Figure 6. Smoothing the empirical semivariogram. The yellow region is the area in which lags have some positive weighting for the bin with the large red circle. The lag indicated by the blue dot will have an influence in all bins with diagonal hatching.

$\hat{\gamma}_{ij}(\mathbf{h}_k)$ is no longer unbiased, so there appears to be no advantage in using $\gamma_{ij}(\mathbf{h})$ over the cross-covariance. The first complete explanation of cokriging (Gandin, 1963) used the cross-covariance only. Hence, the present work concentrates on the cross-covariance, which is estimated by

$$\hat{C}_{ij}(\mathbf{h}_k) \equiv \frac{1}{|N(\mathbf{h})|} \sum_{N(\mathbf{h})} [z_i(\mathbf{s}_p) - \overline{z}_i][z_j(\mathbf{s}_q) - \overline{z}_j]$$

A spatial shift is an example of asymmetry, as illustrated in Figure 7. A shift that is present when the highest cross-covariance occurs for two variables at different locations represents the distance and direction where the calculated cross-covariance is at its maximum value. For example, in Figure 7, variable 1 at some location will have the highest correlation with variable 2 at a location to the west. Shifted cross-correlation is common in environmental applications (Krivoruchko, 2002).

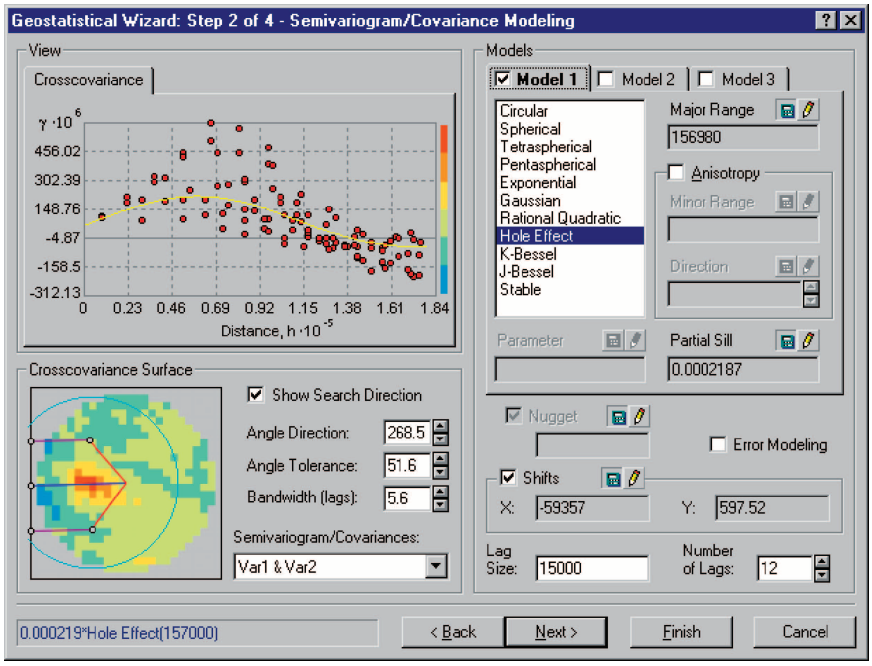


Figure 7. Computer screen shot showing how cross-covariance modeling is accomplished using the shift parameter. For the cross-covariance surface, the highest cross-covariance (red and orange colors) is shifted toward the west.

FITTING SEMIVARIOGRAM AND COVARIANCE MODELS USING A MODIFIED ITERATIVE WEIGHTED LEAST-SQUARES ALGORITHM

After calculating the empirical semivariogram, the next step is to fit a theoretical model (e.g., spherical or exponential) to the values. Three main approaches exist for estimating the parameters of the semivariogram model: visual, (weighted) least-squares, and likelihood methods. In practice, a true semivariogram is almost never known, and consequently, no method exists to determine it exactly. Before modern computers and software became available, semivariograms were commonly fitted visually. More recently, statisticians have commonly used variants of maximum likelihood estimation to fit semivariograms. Researchers in disciplines such as meteorology, geology, and environmental sciences more commonly use variants of a least-squares approach. All three techniques provide useful results, especially if the user has experience in spatial data processing. Several good comparisons of these approaches have been published, including those of Cressie (1993) and Zimmerman and Zimmerman (1991). A modification of the weighted least-squares (WLS) approach (Cressie, 1985) is described below. As noted previously, one important reason to prefer WLS in commercial software implementations is that the results can be easily visualized.

The first step in the modified WLS approach is to select a reasonable lag size for the empirical semivariogram averaging by fitting a preliminary semivariogram using the sector method with logarithmic lags (stage 1 of the algorithm) and then using the range of the fitted model to define the lag used in the second stage, when the final semivariogram model is fitted using a grid method with constant lag size. Before detailing the algorithm, it is instructive to consider an extreme situation where the method gives very good results. Figure 8 presents a simulated data set based on a spherical covariance model further described below. The data locations consist of two clusters separated by a relatively long distance. An automatically derived semivariogram is shown, which looks very reasonable when compared to the true semivariogram given in Figure 9. A screenshot produced with ArcInfo 8.2 (geographic information software developed and marketed by the Environmental Systems Research Institute [ESRI]) shows a map produced using ordinary kriging with the default semivariogram.

Stage 1

The algorithm first scales each data set, $\tilde{z}_j(\mathbf{s}_i) = \left(z_j(\mathbf{s}_i) - \overline{z}_j\right)/\sigma_j$, where σ_j is the sample standard deviation for the jth variable type. Stage 1 begins by assuming an isotropic model. The empirical semivariogram (or covariance) is computed on the scaled data $\tilde{z}_j(\mathbf{s}_i)$, using the sector method with logarithmic

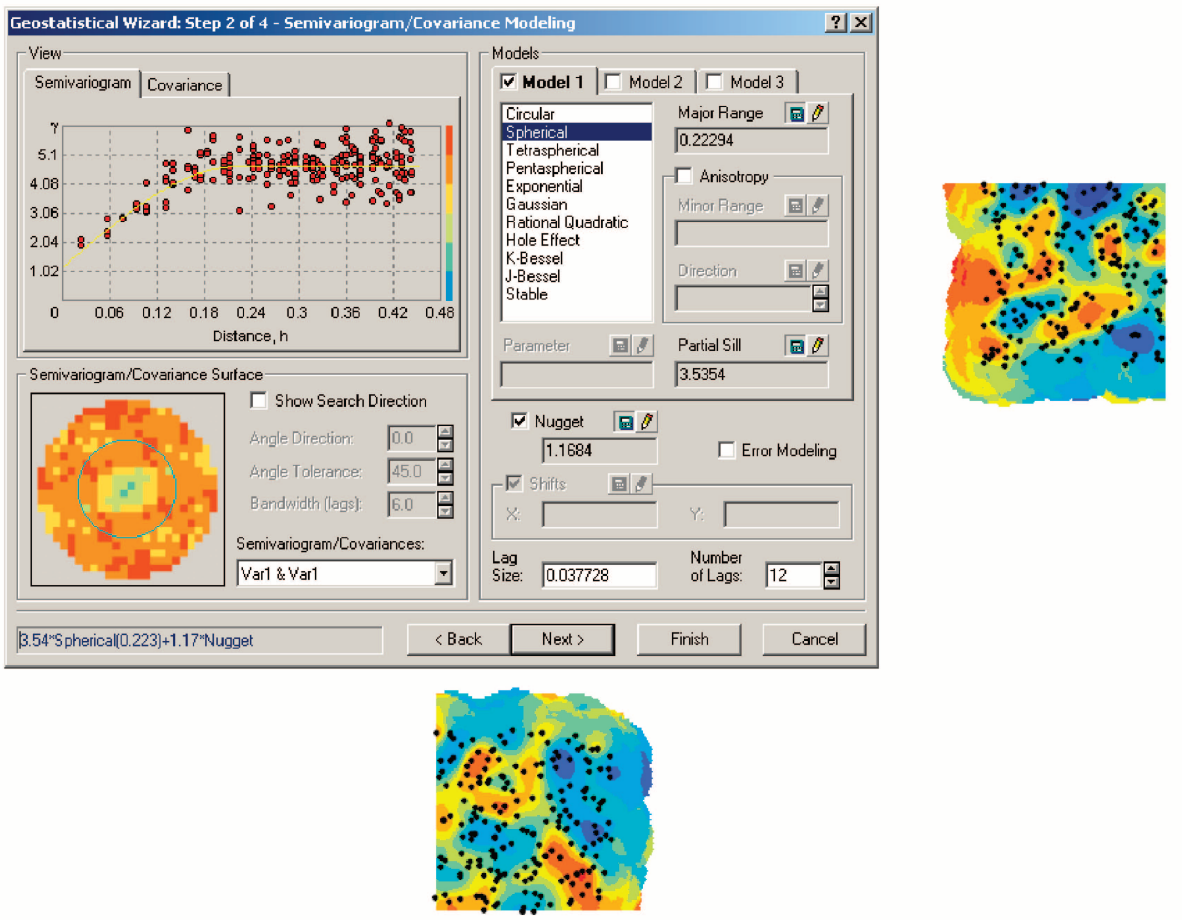


Figure 8. Computer screen shot showing the estimated semivariogram obtained using an automatic model-fitting algorithm (default). The data locations are black dots in two clusters separated by a relatively long distance. The predicted surface using kriging with the estimated semivariogram is also given. For the kriging surface, the cool colors (blue and green) represent low values, and the warm colors (red and orange) represent high values.

lag size (Figure 5c) across a large range of lag classes that progress in a geometric series. The lag classes are formed from intervals $\lfloor d^{k-1/2}, d^{k+1/2})$, where the empirical parameter d equals 1.25, and k ranges from large negative to large positive integers. Part of the series can be seen in Figure 5c. The center of each lag class is taken to be $(d^{k-1/2} + d^{k+1/2})/2$. Only lag classes that have data are used.

The empirical (cross-)covariance can be given by $\hat{C}_{ij}(\mathbf{h}_k)$, where i indicates the ith variable, j indicates the jth variable type, and k indicates the kth lag class. The first iteration of covariance parameter estimates is obtained by minimizing

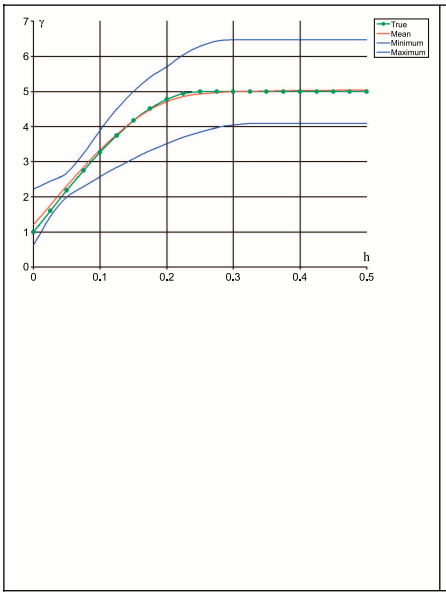
$$\sum_{i=1}^{T}\sum_{j=i}^{T}\sum_{k=1}^{n_{ij}} w_{ij}(\mathbf{h}_k)\left(\tilde{C}_{ij}(\mathbf{h}_k;\boldsymbol{\theta}) - \hat{C}_{ij}(\mathbf{h}_k)\right)^2 \qquad (1)$$

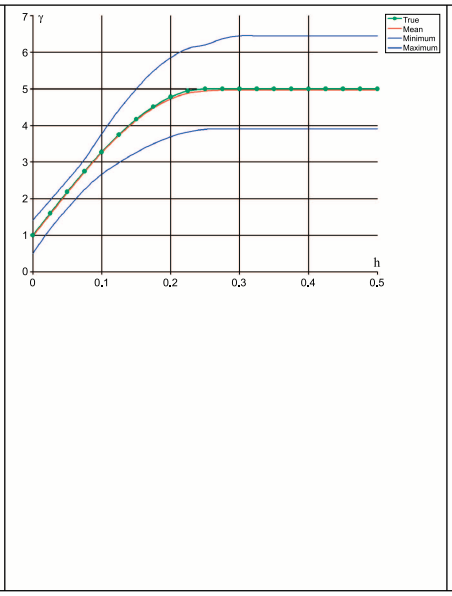
where $\boldsymbol{\theta}$ is the vector of parameters for the i,jth covariance function, and

$$w_{ij}(\mathbf{h}_k) = N_{ij}(\mathbf{h}_k)/\sum_{m=1}^{n_{ij}} N_{ij}(\mathbf{h}_m) \qquad (2)$$

where $N_{ij}(\mathbf{h}_k)$ is the number of pairs in the empirical (cross-)covariance function for variables i and j in lag class k. Let this estimate be specified as $\boldsymbol{\theta}^{(1)}$. In the next iteration, a modification of the WLS algorithm given by Cressie (1985) is used to minimize equation 1 again. In this case,

$$\varpi_{ij}(\mathbf{h}_k;\boldsymbol{\theta}^{(1)}) = \frac{N_{ij}(\mathbf{h}_k)}{\tilde{C}_{ii}(0;\boldsymbol{\theta}^{(1)})\tilde{C}_{jj}(0;\boldsymbol{\theta}^{(1)}) + \tilde{C}^2_{ij}(\mathbf{h}_k;\boldsymbol{\theta}^{(1)})} \qquad (3)$$





	a	b	MISE
REML	0	0.1	0.981946
WLS	0	0.1	1.764688
REML	0.1	0.2	4.407449
WLS	0.1	0.2	5.96006
REML	0.2	0.3	8.224324
WLS	0.2	0.3	7.892238
REML	0.3	0.5	9.090794
WLS	0.3	0.5	6.945768

a b c

Figure 9. Comparison of true and fitted semivariogram models. The green line is the true semivariogram, the red line is the average estimated semivariogram across the 22 simulations, and the blue lines are the minimum and maximum semivariogram values from the simulations. Shown in the figure are results from (a) modified WLS and (b) REML, as well as (c) a quantitative comparison for different ranges.

These weights are normalized so that each (cross-) covariance gets equal weight, that is

$$w_{ij}(\mathbf{h}_k) = \varpi_{ij}(\mathbf{h}_k;\boldsymbol{\theta}^{(1)}) / \sum_{m=1}^{n_{ij}} \varpi_{ij}(\mathbf{h}_m;\boldsymbol{\theta}^{(1)}) \qquad (4)$$

This estimate is specified as $\boldsymbol{\theta}^{(2)}$. If a semivariogram is used instead of the covariance, the estimate $\boldsymbol{\theta}^{(1)}$ is given by minimizing

$$\sum_{i=1}^{T} \left(\begin{array}{l} \sum\limits_{k=1}^{n_{ii}} w_{ii}(\mathbf{h}_k)(\tilde{\gamma}_{ii}(\mathbf{h}_k;\boldsymbol{\theta}) - \hat{\gamma}_{ii}(\mathbf{h}_k))^2 + \\ \sum\limits_{j=i+1}^{T} \sum\limits_{k=1}^{n_{ij}} w_{ij}(\mathbf{h}_k)\left(\tilde{C}_{ij}(\mathbf{h}_k;\boldsymbol{\theta}) - \hat{C}_{ij}(\mathbf{h}_k)\right)^2 \end{array} \right) \qquad (5)$$

where $w_{ii}(\mathbf{h}_k)$ is given by equation 2. $\boldsymbol{\theta}^{(2)}$ is then obtained from equation 5, with weights determined by equation 4, except that

$$\varpi_{ii}(\mathbf{h}_k;\boldsymbol{\theta}^{(1)}) = \frac{N_{ii}(\mathbf{h}_k)}{\tilde{\gamma}_{ii}^2(\mathbf{h}_k;\boldsymbol{\theta}^{(1)})}$$

Determining the estimates $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ are two steps in an iteratively reweighted least-squares algorithm. It is possible to use additional iterations as described in Shapiro and Botha (1991). $\boldsymbol{\theta}^{(2)}$ provides a value for the range of autocorrelation that is used to obtain a default lag size for the grid method in estimating the empirical semivariogram or covariance (see stage 2 below). Based on the authors' experience, a dozen lags is commonly enough to estimate the parameters of the model using the grid method, so the default lag size for the grid method for the next stage of the algorithm is taken to be (2 × range)/12.

The main difference between this approach and that of Cressie (1985) is that the weights do not change during optimization. One of the consequences of this modification is that the algorithm is not sensitive to the number of lags in the grid cells. In fact, it is possible to use even one pair in the cell; its weight will be simply less than the weights of more populated cells.

A similar approach for semivariogram calculation is discussed by Shapiro and Botha (1991) for an isotropic process using one variable. Here, however, anisotropic structure functions (covariance, semivariogram, and cross-covariance) are estimated for transformed variables in two steps, using

logarithmic lag scales and then regular lag scales, with additional kernel smoothing of empirical structure functions.

Stage 2

Stage 2 essentially repeats stage 1, using an empirical semivariogram or (cross-)covariance for the scaled data, $\tilde{z}_j(\mathbf{s}_i)$, that employs the grid method (Figure 5b), where the default lag size is obtained from the range estimate in stage 1. It also allows for anisotropy and linear combinations of (cross-)covariance or semivariogram models for each data set, such as

$$\tilde{C}_{ij}(\mathbf{h};\boldsymbol{\theta}) = \sum_{u=1}^{M} B_u(i,j)\rho_u(\mathbf{h};\boldsymbol{\varphi}_u)$$

Here, $B_u(i,j)$ is the i,jth component of $\mathbf{B}_u$, a $T \times T$ positive-definite matrix, where T is the number of variables; $\boldsymbol{\theta}$ contains the parameters $B_u(i,j)$ and $\boldsymbol{\varphi}_u$ for all u; M is the number of different (cross-)covariance models used in a linear combination; and the function $\rho_u(\mathbf{h};\boldsymbol{\varphi}_u)$ is a normalized covariance model with $\rho_u(h = 0;\boldsymbol{\varphi}_u) = 1$. The third iteration of parameter estimates, $\boldsymbol{\theta}^{(3)}$, is obtained by minimizing equation 1 with weights given in equation 2 using the empirical covariance and the grid method. $\boldsymbol{\theta}^{(4)}$ is obtained by minimizing equation 1 with weights from equations 3 and 4 using the grid method and the empirical covariance. These formulas are modified when using semivariograms, as shown in stage 1 (equation 5).

At this point, it is necessary to change back to the original scale. The final (cross-)covariance model is

$$C_{ij}(\mathbf{h}) = \sigma_i\sigma_j\tilde{C}_{ij}(\mathbf{h};\boldsymbol{\theta}^{(4)})$$

or, in the case of the semivariogram,

$$\gamma_{ii}(\mathbf{h}) = \sigma_i^2\tilde{\gamma}_{ii}(\mathbf{h};\boldsymbol{\theta}^{(4)})$$

If the user changes any of the default parameters of the semivariogram and covariance models (lag size, range, nugget, and partial sill), then estimates are recalculated beginning at stage 2.

In addition to the WLS, several other good methods of fitting spatial semivariogram models exist, including maximum likelihood, restricted maximum likelihood (REML), and generalized least squares. For a review, see Cressie (1993) and Genton (2001). The next section compares the proposed modified WLS approach and REML.

If the covariance matrix is written as $\boldsymbol{\Sigma}(\boldsymbol{\theta})$, which depends on spatial parameters $\boldsymbol{\theta}$, the REML equation to be minimized for $\boldsymbol{\theta}$ is

$$L(\boldsymbol{\theta}) = (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}})'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}(\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}) + \log(|\boldsymbol{\Sigma}(\boldsymbol{\theta})|) + \log(|\mathbf{X}'\boldsymbol{\Sigma}(\boldsymbol{\theta})^{-1}\mathbf{X}|) + (n-p)\log(2\pi) \quad (6)$$

Equation 6 can be minimized by any of the computer-intensive minimization methods such as function FMINCON in MATLAB (software developed and marketed by MathWorks, Inc.).

A SIMULATION EXAMPLE

Suppose data are simulated using the spherical semivariogram model

$$\gamma(\mathbf{h};\boldsymbol{\theta}) = \theta_n + \begin{cases} \theta_s\left[\frac{3}{2}\frac{\|\mathbf{h}\|}{\theta_r} - \frac{1}{2}\left(\frac{\|\mathbf{h}\|}{\theta_r}\right)^3\right] & \text{for} \quad \|\mathbf{h}\| < \theta_r \\ \theta_s & \text{for} \quad \theta_r \le \|\mathbf{h}\| \end{cases}$$

where $\theta_n \ge 0$ is the nugget, $\theta_s \ge 0$ is the partial sill, and $\theta_r \ge 0$ is the range. For each of the simulations, $\theta_s = 4$, $\theta_r = 0.25$, and $\theta_n = 1$. For semivariogram models with sills, it is easy to go back and forth between semivariogram and covariance functions obtained using the relations

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - \mathbf{C}(\mathbf{h}) \quad \text{and} \quad \mathbf{C}(\mathbf{h}) = \gamma(\infty) - \gamma(\mathbf{h})$$

and this approach is employed here. After converting the spherical semivariogram to a covariance function, the Cholesky decomposition method is applied (Cressie, 1993) to simulate 200 values from a Gaussian distribution in the region $0 \le x \le 1$ and $0 \le y \le 1$. Twenty-two such data sets of size 200 are simulated. Then, the grid method is used to compute the empirical semivariogram for each data set, and a spherical semivariogram model is subsequently fit to the empirical semivariogram values using modified WLS as described above. Additionally, REML is used to fit a spherical model to the empirical covariance, and the result is subsequently converted to a semivariogram.

Table 1. Estimation of the semivariogram parameters by REML and modified WLS.

	REML			*WLS*		
Quantity	*Nugget*	*Sill*	*Range*	*Nugget*	*Sill*	*Range*
Average value	0.973295	4.000645	0.250886	1.199273	3.839545	0.269727
$(1/22)\sum_{i=1}^{22}(\hat{\theta}_{r,i}-\theta_r)^2$	0.051663	0.513559	0.000717	0.161678	0.583832	0.004794

The true semivariogram model is shown as the green curve in Figure 9. The average fitted semivariogram is shown as the red curve in Figure 9, with lower and upper bounds given by the blue curves. These curves are calculated using the following formulas:

$$\gamma_{\text{minimum}}(h) = \min(\gamma_i(h)),\ i = 1 \ldots 22$$
$$\gamma_{\text{maximum}}(h) = \max(\gamma_i(h)),\ i = 1 \ldots 22,\ \text{and}$$
$$\gamma_{\text{average}}(h) = \text{average}(\gamma_i(h)),\ i = 1 \ldots 22$$

Figure 9 shows that, on average, both algorithms for fitting the semivariogram worked quite well. However, REML estimates all the semivariogram parameters (θ_r, θ_n, and θ_s) more precisely, as indicated in the Table 1.

Further, the difference between true and estimated models near the origin is smaller for REML. To quantify this, the mean integrated squared error (MISE) was used:

$$\text{MISE}(a,b) = \frac{1}{22(b-a)}\sum_{i=1}^{22}\int_a^b [\hat{\gamma}_i(h) - \gamma(h)]^2 \mathrm{d}h$$

An approximation is given by

$$\text{MISE}(a,b) \cong \frac{1}{22|B|}\sum_{i=1}^{22}\sum\nolimits_{j\in B}[\hat{\gamma}_i(h_j) - \gamma(h_j)]^2$$

where $H = \{h_j : j/c\}$ for $j = 1, 2, \ldots$ and some value c, and $B \subset H$, where $B = \{h_j : a \le h_j \le b\}$, and $|B|$ is the number of elements in B. The larger the value c, the closer the approximation becomes.

It is interesting that MISE is different for different ranges. For example, MISE(0,0.1) and MISE(0.1,0.2) are smaller for REML, but MISE(0.3,0.5) is smaller for modified WLS (Figure 9c). Thus, REML yields better estimates of the semivariogram parameters for the most important short lags.

Figure 10 presents the parameters estimated for the spherical semivariogram model on each of the 22 simulated data sets using modified WLS. The parameters are scattered around the true values given in magenta. The figure also suggests that the parameters are correlated (e.g., the estimate of the nugget effect is commonly low when the partial sill is high). Table 2 contains the estimated pairwise correlations under both the REML and modified WLS model-fitting approaches.

Estimates of the correlation between semivariogram parameters obtained using the modified WLS model-fitting approach are larger because the range is estimated first, and range influences the determination of the other semivariogram parameters. However, this is theoretically undesirable, and so the REML algorithm is considered to have performed better.

SMOOTH EMPIRICAL SEMIVARIOGRAM, COVARIANCE, AND CROSS-COVARIANCE SURFACES

Having access to a smooth semivariogram and covariance surface is useful in exploratory data analysis when investigating spatial auto- and cross-correlation. An algorithm for obtaining a smooth semivariogram surface is presented below.

The sector method with logarithmic lag size (Figure 5c) is used for averaging pairs of data values. Estimation at the point $\vec{v}$ of the semivariogram surface is accomplished using the following formula:

$$\overline{\gamma}(\vec{v}) = \frac{\sum_{i=1}^{N} n_i \cdot \mathrm{e}^{-\Theta\cdot\rho(\vec{v},\vec{c}_i)} \cdot \hat{\gamma}_i}{\sum_{i=1}^{N} n_i \cdot \mathrm{e}^{-\Theta\cdot\rho(\vec{v},\vec{c}_i)}}$$

where N is the number of sectors, $\hat{\gamma}_i$ is the average semivariogram value in the sector $i \in \overline{1,N}$, n_i is the number of pairs in sector $i \in \overline{1,N}$, and $\vec{c}_i$ are

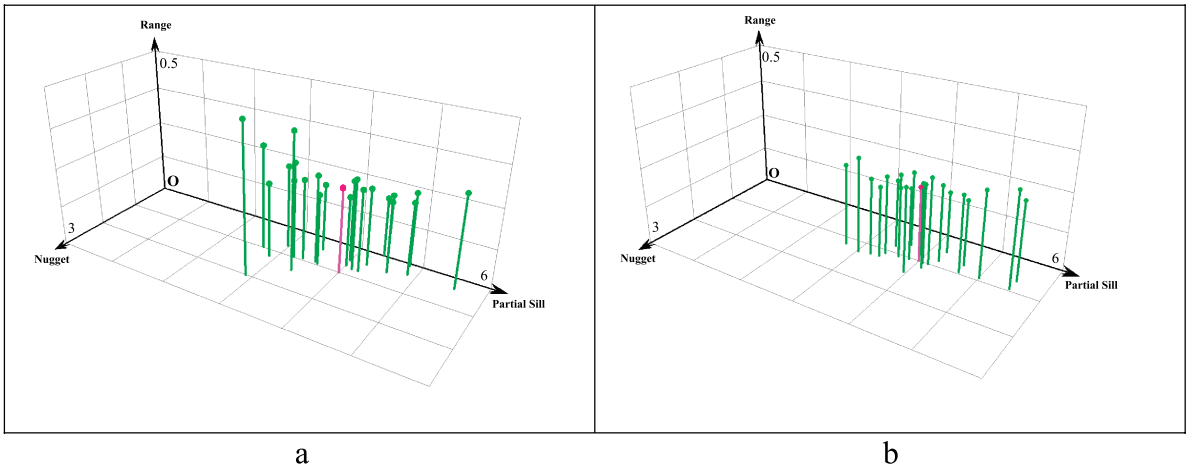


Figure 10. Three-dimensional scatter diagram of 22 sets of parameter estimates of the spherical semivariogram model. Each estimate of the nugget, range, and partial sill is given in green, and the true value is given in magenta. Results are shown for (a) WLS and (b) REML.

the coordinates of the center of sector $i \in \overline{1,N}$. Parameter Θ depends on data and data locations: $\Theta = \sqrt{\frac{8 \cdot \pi \sqrt{\sum_{i=1}^{N} n_i}}{S}}$, where S is an area of the region under investigation. Figure 11 presents examples of empirical and smooth semivariogram surfaces.

Using the smooth semivariogram and covariance surface defined above for analyses other than data exploration requires some further improvements to surface calculation. For example, $\rho(\vec{a}, \vec{b})$ can be based on a non-Euclidean distance, considering the anisotropy of the covariance surface, where the parameter Θ depends on its position on the covariance surface.

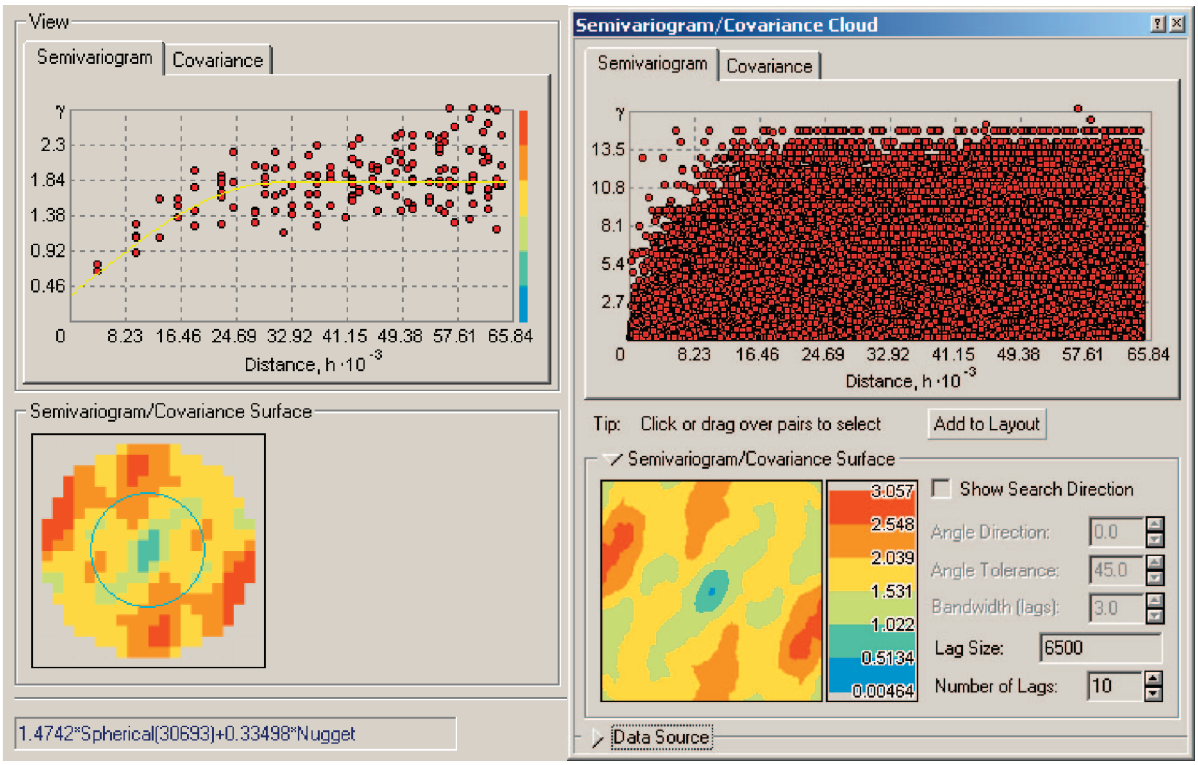
DISCUSSION AND CONCLUSION

This chapter describes grid-based empirical semivariograms and covariances, along with model fitting using modified WLS. These procedures have been implemented in commercial software packages (e.g., the Geostatistical Analyst extension to ArcGIS 8.1, a product developed and marketed by ESRI; Johnston et al., 2001). For data sets with more than 100 data points and with significant spatial dependence, the algorithm gives very good defaults, which allows users to easily update parameters using cross-validation.

However, situations exist in which modified WLS does not perform well. This is particularly true for situations in which the data set is small, a large amount of measurement error exists, and the data include clusters of high or low values (i.e., they are nonstationary). In such cases, the lag and, consequently, the range are commonly overestimated, and the default algorithm produces oversmoothed surfaces. However, the user can control this situation by selecting different lag sizes with graphical tools. Cross-validation is also an important diagnostic in these instances. If the lag size is selected properly (commonly, it is obvious how to manually improve

Table 2. Estimates of correlation between the estimated parameters of the spherical semivariogram model for 22 simulated data sets.

	REML			*WLS*		
	Nugget	***Sill***	***Range***	***Nugget***	***Sill***	***Range***
Nugget		−0.470	0.488		−0.703	0.789
Sill	−0.470		0.139	−0.703		−0.384
Range	0.488	0.139		0.789	−0.384	



a b

Figure 11. Computer screen shot of an empirical semivariogram cloud and surface using (a) the grid method and (b) all semivariogram pairs and a smooth semivariogram surface using the same data set.

default lag estimation), then other model parameters are generally well estimated.

For relatively small data sets, REML is almost always better, and this algorithm should be used as an alternative. The only serious drawback to this technique is that it requires much more computing time than modified WLS.

REFERENCES CITED

Cressie, N., 1985, Fitting variogram models by weighted least squares: Mathematical Geology, v. 17, p. 653–702.

Cressie, N., 1993, Statistics for spatial data: New York, John Wiley & Sons, 900 p.

Deutsch, C. V., and A. G. Journel, 1998, GSLIB—Geostatistical software library and user's guide: New York, Oxford University Press, 369 p.

Gandin, L. S., 1959, The problem on optimal interpolation: Leningrad, Trudy Glavnaya geofizicheskaya observatoria, v. 99, p. 67–75.

Gandin, L. S., 1963, Objective analysis of meteorological fields: Leningrad, Gidrometeorologicheskoe Izdatel'stvo, 286 p. (translated by Israel Program for Scientific Translations, Jerusalem, 1965, 242 p.).

Genton, M. G., 2001, Robustness problems in the analysis of spatial data, *in* M. Moore, ed., Spatial statistics: Methodological aspects and applications: Springer Lecture Notes in Statistics, v. 159, p. 21–37.

Johnston, K., J. Ver Hoef, K. Krivoruchko, and N. Lucas, 2001, Using ArcGIS geostatistical analyst: GIS by ESRI: ESRI Press, 300 p.

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

Kaluzny, S. P., S. C. Vega, T. P. Cardoso, and A. A. Shelly, 1998, S+ Spatial Stats: User's manual for Windows and Unix: New York, Springer-Verlag, 384 p.

Keller, L., and A. Friedmann, 1925, Differentialeichungen fur die turbulente bewegungner kompressibelen flussigkeit: Proceedings of the 1st Congress for Applied Mathematics, Delft, p. 395–405.

Kolmogorov, A., 1941, Local turbulence structure in an incompressible viscous liquid for very high Reynolds numbers: Doklady Akademii nauk SSSR, v. 30, no. 4, p. 229–232.

Krivoruchko, K., 2002, Geostatistical analysis of California air quality: ESRI Technical Report: Available at www.esri.com/software/arcgis/arcgisxtensions/geostatistical/airqualityjgra.pdf (accessed July 2004).

SAS Institute Inc., 1996, SAS/STAT technical report: Spatial prediction using the SAS system: Cary, North Carolina, SAS Institute, Inc., 80 p.

Shapiro, A., and J. D. Botha, 1991, Variogram fitting with a general class of conditionally nonnegative definite functions: Computational Statistics and Data Analysis, v. 11, p. 87–96.

Ver Hoef, J. M., and N. Cressie, 1993, Multivariable spatial prediction: Mathematical Geology, v. 25, p. 219–240.

Zimmerman, D. L., and M. B. Zimmerman, 1991, A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors: Technometrics, v. 33, p. 77–91.