

**ROBUST SEMIVARIOGRAM ESTIMATION  
IN THE PRESENCE OF INFLUENTIAL  
SPATIAL DATA VALUES**

Richard F. Gunst and Molly I. Hartfield  
Department of Statistical Science  
Southern Methodist University  
Dallas, TX 75275-0332

**SMU / DS / TR - 286**  
**September 1996**

# **Robust Semivariogram Estimation in the Presence of Influential Spatial Data Values**

**Richard F. Gunst and Molly I. Hartfield**

**Department of Statistical Science**

**Southern Methodist University**

**Dallas, TX 75275-0332**

## **Abstract**

Recent investigations have stimulated renewed interest in accommodating the effects of influential observations in the estimation of semivariogram values. Because of the increasing size and complexity of spatial data sets, it is not sufficient to rely solely on graphical methods for detecting aberrant data values. Influence diagnostics have long been used to identify influential observations in regression analyses and have recently been adapted to the fitting of variogram and kriging models. Less emphasis has been placed on the identification and accommodation of influential data in the estimation of the sample semivariogram values that are used to fit these models. In this paper prior work on the identification of influential observations is extended to robust estimation of semivariogram values. We concentrate on robust semivariogram estimators that can be readily implemented by modest adaptations of existing computer software. The robust estimators evaluated in this paper are compared to classical semivariogram estimators and to one another in terms of both accuracy and precision.

## **1. Introduction**

Semivariogram modeling is central to the prediction of point values and areal averages of geostatistical random fields. Additionally, estimates of semivariogram model parameters themselves are of intrinsic interest because of the information they provide about spatial dependence across a region. The fitting of semivariogram models is in turn critically dependent on the shape of sample semivariogram plots. While this dependence has been studied extensively (see Cressie 1991 for numerous citations), recent work by Basu et al. (1995) documents that the presence of influential spatial data values can seriously distort sample semivariogram plots, in some cases even when a robust estimator is used. These distortions can critically affect both the choice of models that are fit to sample semivariograms and the calculation of parameters for the chosen models. Basu et al. (1995) also present graphical and numerical diagnostics that are highly effective for identifying influential spatial data values. In this paper, we discuss several robust estimators of semivariogram values that aid in the identification of influential spatial data values and that can be used as alternatives to the deletion of the observations from the data base.

The need for further research into robust alternatives to the sample semivariogram stems in part from the application of spatial methods to very large data sets and in part from the

observation that current robust methods are not highly effective in accommodating influential data values. One application of interest in this work is to the modeling of temperature anomaly trends over the last century. The data files used in these investigations consist of monthly temperature readings from several thousand reporting stations from approximately 1850 to 1991. In terms of resources and the time expended, it would be prohibitive to graphically investigate the presence of influential observations for each station in each of a number of regions of the globe for each month of each year included in the data base. Thus, there is a well defined need both for influence diagnostics and effective robust methods that can accommodate these large spatial data sets. Moreover, the application of some of the more popular robust methods to these data files did not satisfactorily accommodate some obvious influential observations. This latter finding will be documented below, following a brief review of semivariogram estimation.

We begin with the usual definition of semivariogram values:

$$\gamma(\mathbf{s}_i, \mathbf{s}_j) = \text{var}\{z(\mathbf{s}_i) - z(\mathbf{s}_j)\} / 2 ,$$

where  $z(\mathbf{s})$  denotes a spatial variate measured at location  $\mathbf{s}$ . Semivariogram values quantify the spatial covariance structure of the random variables in a fashion similar to covariances and correlations. Under the second-order spatial stationarity assumptions stated below,  $\text{cov}\{z(\mathbf{s}_i), z(\mathbf{s}_j)\} = \text{var}\{z(\mathbf{s})\} - \gamma\{z(\mathbf{s}_i) - z(\mathbf{s}_j)\}$ , where  $\text{var}\{z(\mathbf{s})\}$  is the common variance of spatial variates. Many spatial measurements are characterized by variation that is a function of distance but not direction. These *isotropic* spatial variates have variograms defined as

$$\gamma(d) = \text{var}\{z(\mathbf{s}_i) - z(\mathbf{s}_j)\} / 2 , \quad (1)$$

where  $d = \|\mathbf{s}_i - \mathbf{s}_j\|$  is the (Euclidean or great circle, as appropriate) separation distance between the spatial locations. With irregularly spaced locations, pairs of locations are ordinarily assigned to bins that are multiples of a nominal distance. In this case,  $d$  represents the midpoint of the range of distances in a bin. Under *intrinsic stationarity* assumptions,  $E\{z(\mathbf{s}_i) - z(\mathbf{s}_j)\} = 0$  and  $\text{cov}\{z(\mathbf{s}_i), z(\mathbf{s}_j)\}$  is a function of only  $\mathbf{s}_i - \mathbf{s}_j$ , so that for isotropic spatial variates (1) can also be expressed as

$$\gamma(d) = E\{z(\mathbf{s}_i) - z(\mathbf{s}_j)\}^2 / 2 , \quad (2)$$

*Second-order spatial stationarity* is slightly stronger than intrinsic stationarity since the first assumption is replaced by  $E\{z(\mathbf{s}_i)\} = \mu$ , a constant for all locations in the region of interest, and the variance of individual spatial variates is assumed to be finite. With intrinsic stationarity assumptions, the variance of differences of spatial variates is assumed to be finite, even though the variance of an individual spatial variate need not be so. Except where explicitly stated otherwise, we assume isotropic, second-order stationary spatial variates in the remainder of this paper.

Although stationarity is assumed in this work, if the random field does not satisfy stationarity assumptions one ordinarily either (1) fits low-order polynomial models to the data using the spatial locations as predictors, (2) fits local linear or quadratic models using the spatial locations, (3) performs median polish, or (4) fits more complex models to the spatial variables using both the spatial locations and other spatial covariates. The residuals from such fits are then assumed to satisfy stationarity assumptions and are used in the calculation of semivariogram values and in the fitting of semivariogram models. Although the following discussion is framed in terms of the spatial variates  $z(s_i)$ , the methods discussed may in practice be applied to residuals.

From (2), the classical method-of-moments semivariogram estimator (Matheron 1962) for pairs of locations binned a distance  $d$  apart is

$$\hat{\gamma}(d) = \sum_{N(d)} \{z(s_i) - z(s_j)\}^2 / 2N_d, \quad (3)$$

where  $N(d)$  denotes the set of all pairs of locations binned together at (nominal) separation distance  $d$  and  $N_d$  is the number of such pairs of locations. For locations a fixed distance  $d$  apart, this sample semivariogram estimator is unbiased for the semivariogram value  $\gamma(d)$  under either second-order or intrinsic stationarity.

Cressie and Hawkins (1980) introduced a robust estimator of semivariogram values that is less susceptible to influential data values than the sample semivariogram estimator. Their robust estimator is

$$\hat{\gamma}_{CH}(d) = \frac{\left\{ \sum_{N(d)} |z(s_i) - z(s_j)|^{1/2} / N_d \right\}^4}{2(0.457 + 0.494 / N_d)}. \quad (4)$$

This estimator was derived under the assumption that the differences  $z(s_i) - z(s_j)$  are normally distributed for all station pairs  $(s_i, s_j)$ . The square root transformation of the differences was shown to have moments close to those of a normal distribution and the denominator in (4) is a bias correction.

Basu, et al. (1995, Figure 1) demonstrate a variety of effects that can be induced on semivariogram plots because of the presence of influential observations. The examples presented in that paper include some in which a single influential observation causes dramatic spikes in the semivariogram plot and others in which a small number of influential observations cause a mound-shaped *excitation crest*. In the following example, we expose yet another way in which influential observations can affect semivariogram plots, the inducement of *anisotropy*.

The semivariogram plots in Figure 1 are of soil nitrate concentrations from core samples taken from a 3 ha field. The sample locations were on a regular 9x7 grid with 25m spacing

between north-south and east-west grid locations. Bins for this semivariogram plot were chosen to be multiples of 25m. The first bin includes all core sample locations that are between 0 and 25m; the second, those between 25m and 50m; etc. Each symbol denotes a different direction between pairs of locations on the grid. For example, the north-south semivariogram values, identified by the triangles in Figure 1, include all grid locations within each bin that are  $90^\circ \pm 22.5^\circ$  from one another. Prior to calculating the semivariogram values, median polish (e.g., Cressie 1984, 1986) was performed on the raw nitrate concentrations in order to reduce the effects of any nonstationarity in directions along the north-south and east-west grid lines. The median-polish residual pairs were then binned and directional semivariogram values were calculated using the robust estimator (4) with the median polish residuals inserted as the  $z(s_i)$ .

Consider first the semivariogram plots indicated by the various symbols connected with solid lines. The semivariogram values for each of the four directions generally increase throughout the range of distances plotted. However, the semivariogram plots for each direction do not increase at the same rate. The least change in variability across the range of distances occurs with pairs of grid locations that are in the northwest-southeast direction from one another. The most striking change in variability occurs for location pairs that are in the northeast-southwest direction from one another. When spatial variability patterns change with direction, the spatial variability is called *anisotropic*. It is highly desirable when fitting semivariogram models that the semivariogram plots be isotropic and that the semivariogram values remain basically constant or increase smoothly to a plateau, referred to as the *sill* of the semivariogram. This desirability is not simply computational; in many instances the nature of the measurements strongly suggests that variability should be isotropic. For example, the homogeneity of the soil treatments applied to this field does lead one to expect isotropic spatial variability. The spatial variability indicated by the symbols connected with solid lines in Figure 1 is clearly not isotropic and the sills, if they exist, are quite different for the four directional semivariogram plots.

Contrast the semivariogram values for the solid lines with those indicated by the dashed lines in Figure 1. The latter semivariogram values are much smaller than the former ones. They appear to be isotropic and the semivariogram plots are relatively flat, suggesting white noise errors and a common sill. Figure 2 highlights the reason for the differences in the two sets of semivariogram values. Figure 2 is a three-dimensional smoothed plot (S-PLUS, 1991) of the data. Strikingly evident in the plot are several pockets of very large nitrate concentrations. The semivariogram values identified by the solid lines in Figure 1 were calculated using all the nitrate values while those indicated by the dashed lines were calculated from a reduced data set in which the 8 largest nitrate values were removed.

It may be that one considers the large nitrate values the most important features of the data. Nevertheless, they were unexpected and they cause severe problems when one attempts to fit semivariogram models to the sample and robust semivariogram values. We are not concluding that these data values are necessarily anomalous nor that they should necessarily be eliminated from the data base. However, this example illustrates the

dramatic effect a small number of data values can have even on some robust estimators. If these values are determined to be anomalous, neither the sample nor the robust semivariogram values plotted in Figure 1 would properly characterize the spatial variability of nitrate concentrations in this field. In addition, with data sets of the size of global temperature data sets, it is not likely that sufficient time and resources could be relegated to construct perspective plots for each spatial variate of interest over all regions of the globe for each month and each year contained in the data bases. Hence, there is a need for more effective robust semivariogram estimators

## 2. Influence Functions

Influence Functions (e.g., Hampel et al. 1986) quantify the effects of influential data values on an estimator. One can think of an influence function as a measure of the change in an estimator due to the presence of a small number of data values from a specified distribution that is different from the distribution of the bulk of the data. If  $F$  represents the distribution (e.g., normal) generating the bulk of the data and  $G$  represents the distribution of the influential observations (e.g., a very skewed distribution), then the distribution of all the data can be represented as a mixture  $F^\alpha$  of these two distributions:  $F^\alpha = (1 - \alpha)F + \alpha G$ , for  $0 < \alpha < 1$ . For a fixed separation distance  $d$ , the semivariogram parameter one wishes to estimate can be represented as  $\gamma(F) \equiv \gamma(d) = \text{var}\{z(s) - z(s + d)\}/2$ . The influence of outliers from the distribution  $G$  can then be quantified as  $IF_\gamma(G) = \partial\gamma(F^\alpha) / \partial\alpha|_{\alpha=0}$ .

For illustration purposes, let  $F$  represent a second-order stationary spatial random field with variance  $\sigma_{zz}$ . If  $G$  represents the presence of a single influential outlier of magnitude  $\mu + \delta_m$  at location  $s_m$  in the domain  $D$  of the random field, and if location  $s_m$  is one of the locations in the sampled field, one can show that the influence function for the sample semivariogram estimator (3) is

$$IF_s(s_m) = \frac{N_m}{2N_d} \delta_m^2 \quad (5)$$

Basu et al. (1995) demonstrate that influential spatial data values typically result in overestimation by both semivariogram estimators. In particular, they show that for the sample semivariogram (3), if there is an influential spatial variate at location  $s_m$  then  $E\{\hat{\gamma}(d)\} = \gamma(d) + (N_m/2N_d)\delta_m^2$ . Note that the second term of this expression is the value of the influence function (5).

Two features of this influence function are important in the assessment of the effect of extreme data values. First, as one might expect from the form of the estimator (3), the influence of an extreme data value is proportional to the square of its magnitude. Second, the influence of an extreme data value is attenuated by the proportion of pairs  $N_m/N_d$  in the bin that include the outlier. The full effect of the extreme data value is realized only if all the pairs in a bin include the outlier, whereas the extreme data value has no effect if none

of the pairs include the outlier. The proportionality factor  $N_m/N_d$  explains the spiking often seen in semivariogram plots when anomalous data are present. There is a relatively high proportion of outlier pairs in the bins for which spiking is evident but relatively few in the bins immediately adjacent to them.

The influence function for the robust estimator (4) is expressible as

$$\begin{aligned} \text{IF}_R(\mathbf{s}_m) = \gamma(d) & \left[ \left\{ (1-f_m)^4 \left( g_1^4 + 6 \frac{g_1^2 g_2}{N_d} + 3 \frac{g_2^2}{N_d^2} \right) \right. \right. \\ & + 4(1-f_m)^3 f_m \left( g_1^3 + 3 \frac{g_1 g_2}{N_d(1-f_m)} \right) h_1 \left( 1 + \frac{\lambda_m}{1+\lambda_m} \right)^{1/4} \\ & + 6(1-f_m)^2 f_m^2 \left( g_1^2 + \frac{g_2}{N_d(1-f_m)} \right) \left( h_1^2 + \frac{h_2}{N_m} \right) \left( 1 + \frac{\lambda_m}{1+\lambda_m} \right)^{1/2} \\ & + 4(1-f_m) f_m^3 g_1 \left( h_1^3 + 3 \frac{h_1 h_2}{N_m} \right) \left( 1 + \frac{\lambda_m}{1+\lambda_m} \right)^{3/4} \\ & \left. \left. + f_m^4 \left( h_1^3 + 6 \frac{h_1 h_2}{N_m} + 3 \frac{h_2^2}{N_m^2} \right) \left( 1 + \frac{\lambda_m}{1+\lambda_m} \right) \right\} / \left( g_1^4 + 6 \frac{g_1^2 g_2}{N_d} + 3 \frac{g_2^2}{N_d^2} \right) - 1 \right] \quad (6) \end{aligned}$$

where  $f_m = N_m/N_d$ ,  $\lambda_m = \delta_m^2 / 2\{2\gamma(d)\}$ ,  $g_j = h_j(0)$ ,  $h_j = h_j(\lambda_m)$ ,

$$\begin{aligned} h_1(t) &= \Gamma\left(\frac{3}{4} + \frac{t^2}{2+4t}\right) / \Gamma\left(\frac{1}{2} + \frac{t^2}{2+4t}\right) \\ h_2(t) &= \left\{ \Gamma\left(1 + \frac{t^2}{2+4t}\right) / \Gamma\left(\frac{1}{2} + \frac{t^2}{2+4t}\right) \right\} - \{h_1(t)\}^2 \end{aligned}$$

and  $\Gamma(x)$  is the gamma function  $\Gamma(x) = \int_0^\infty w^{x-1} e^{-w} dw$ . The value of the influence function (6) is defined to be zero if  $N_m = 0$ .

While this expression is far more complicated than the influence function for the sample semivariogram, it, like the one for the sample semivariogram, enables one to quantitatively evaluate the effect of extreme data values. To illustrate the effectiveness of these influence functions on assessing the effects of influential observations, we calculated influence function values for one of the more interesting semivariogram plots in Basu et al. (1995). Figure 1(c) in Basu et al. (1995) contains a mound-shaped semivariogram plot of nitrate calculations from the same field as those used in the previous but from a different depth and using nitrate concentrations from both the 25 meter grid and a smaller 6x6 5m grid in the center of the larger grid.

The parameters needed to calculate the influence functions (5) and (6) were estimated from the large- and small-grid data sets after an influential observation that was identified in Basu et al (1995) in the large-grid data was removed. The variance  $\sigma_{zz}$  was estimated to be approximately 40, using the average of variance estimates from the large-grid data and the small-grid data. The influential observation's  $\delta_m$  effect was estimated from its median-polish residual to be approximately 41. The values of  $N_m$  and  $N_d$  for each bin were determined from the location distances in the data file. Figure 3 is a display of the resulting calculations for both the sample and the robust semivariogram estimators.

The horizontal line at a semivariogram value of 40 represents the theoretical value of the sample semivariogram estimator (3) or the robust estimator (4) if the random field were a white-noise process with variance  $\sigma_{zz} = 40$  and there were no influential data values. The sample and robust semivariogram plots in Figure 3 with the effect of the influential observation added are obtained by adding the influence function values calculated from (5) and (6) to the white noise semivariogram value. The excitation crests in each of these latter semivariograms are evident, as is the lesser effect of the influential data value on the robust semivariogram values than on the sample semivariogram values.

The simple white-noise model is likely not the most appropriate model for the combined large- and small-grid semivariogram model. In practice, each of these grids would be fit separately and the resulting fitted semivariogram models would be combined to represent the spatial variability of the field. The spatial variation in nitrate concentrations can be modeled as separate white-noise processes with large-grid variance  $\sigma_{zz} = 50$  (with the influential data value removed) and small-grid variance  $\sigma_{zz} = 30$ . Reconstruction of Figure 3 with these estimates and separate estimates of the semivariogram values for each bin with the influential data value removed produced crests more similar to those in Basu et al. (1995) Figure 1(c), but did not meaningfully change the conclusions so clearly evident from Figure 5. The importance of this illustration is that even under such a simple white-noise model the influence functions adequately characterize the gross effects of the influential data. One can perform similar computations from the European temperature anomaly data to reproduce the spiking in Basu et al. (1995) Figure 1(a) from influence function calculations, as well as the anisotropic behavior in Figure 1 of this paper.

### 3. Robust Estimators

McBratney and Webster (1986) concluded that neither the sample semivariogram estimator (3) nor the robust semivariogram estimator (4) could always be preferred to the other. Our analyses of a variety of data sets and a number of simulations lead to the conclusion that the robust semivariogram is less sensitive to the effects of influential observations than the sample semivariogram but it can still be seriously affected, as is indicated in Figure 1. These analyses and simulations support the general conclusions drawn from a comparison of the influence functions (5) and (6) and they reemphasize the need for other robust alternatives to the sample semivariogram.

Prior to discussing possible alternative robust semivariogram estimators, it is important to recognize difficulties that arise as a result of unavoidable data reuse in the estimation of semivariogram values. Regardless of the estimation scheme selected, each data value is used many times in the calculation of sample and robust semivariogram values. For a simple example of how this can exacerbate the influence of extreme data values, consider a transect of length  $2m+1$ . Suppose a single unusually large spatial variate occurs at the middle location on the transect. If the bin distances are one unit, there will be  $2m$  pairs of locations in the first bin, of which 2 will include the influential spatial variate. In the second bin, there will be  $2m-1$  pairs of locations, of which 2 again will include the influential spatial variate. One can continue in this fashion and show that up until bin  $m$  there are fewer and fewer pairs of locations, of which 2 always include the influential spatial variate. Thus, the percentage of differences in a bin which include an influential spatial data value increases with the bin number and can far exceed the percentage of such influential data values in a data set. The percentage can increase even more rapidly if there is more than one influential data value in the data set.

One can easily construct examples with irregularly spaced data for which a disproportionate percentage of differences in a bin will contain influential observations but many fewer or none of the differences in neighboring bins contain influential observations, leading to spikes in the sample semivariogram. Depending on the proportions of pairs in a bin that involve influential data, the excitation crests discussed previously are another effect of data reuse. In addition, it is not necessarily true that robust methods operating on differences of the pairs in a bin will operate only on the differences that contain influential data values. Indeed, in bins that do not contain outliers, robust methods will often downweight or eliminate differences involving pairs of valid data, thereby tending to underestimate the variability for distances represented by those bins. Robust estimators must accommodate all of these difficulties caused by data reuse. It is not at all clear that robust estimators that operate on pairs of data values in each bin can satisfactorily do so. This concern is part of the motivation for the following comparisons of robust estimators

In the course of this research, several robust alternatives to (4) were evaluated. The following estimators are typical of those investigated and include those found most effective:

- trim a fixed percentage of raw data values and then bin and apply the sample semivariogram estimator (3) to the remainder of the data;
- trim a fixed percentage of the differences in each bin and then apply (3) to the remaining differences;
- apply a robust  $m$ -estimator to the differences in each bin;
- perform a preliminary test to identify and eliminate unusual data values in the data set, then bin and apply (3) to the remaining data values;
- perform a preliminary test on the differences in each bin to identify and eliminate unusual differences, then apply (3) to the remaining differences in each bin.

Consideration was given to applying the robust estimator (4) rather than the classical estimator (3) to the trimmed data. It was feared that trimming followed by the use of a robust estimator would cause serious underestimation of the spatial variability, especially when there were no extreme data values. Spot checks were made by rerunning some of the simulations using (4) with the trimmed data. These simulations confirmed that substantial underestimation often occurred. Consequently, the trimmed estimators used in the simulations all use the classical estimator (3) with the trimmed data.

Because of the data reuse issues raised above, trimming a *fixed* percentage of the raw data values or trimming the differences in bins is likely to be problematic. In the investigations discussed below, we include for comparison purposes one robust estimator that trims 5% of the largest raw data values in a data set. Denote this estimator  $\hat{\gamma}_T(d)$ . For the reasons cited above, trimming a fixed percentage of the binned differences was not regarded as a viable estimation strategy and no such estimator was included in the simulations.

Two versions of a robust estimator which tests for influential observations are included in the simulations reported below. The preliminary test is used to determine whether any of the raw data values or, alternatively, any of the differences in a bin are influential. If the preliminary test determines that some of the raw data values or some of the differences are influential, then the robust estimator (4) is applied to the remaining data values or differences. This robust estimation scheme is included so that, unlike trimming a fixed percentage of the data values, the natural variability in the data can assist in determining the number of possible influential observations. For the raw data values, the preliminary test is a robust version of a t-test. Outliers are deleted if  $|t_i| > 3$ , where

$$t_i = \frac{z(s_i) - M}{S}, \quad (7)$$

$M = \text{median}\{z(s_i)\}$ , and  $S$  is the median absolute deviation of the spatial variates:  $S = \text{median}\{|z(s_i) - M|\}/0.6745$ . Denote this estimator  $\hat{\gamma}_P(d)$ . Differences in each bin were also tested for their influence using (7) with obvious modifications. Denote this estimator  $\hat{\gamma}_B(d)$ .

An alternative to trimming that does not require a preliminary test is a robust M-estimator of location. Both from efficiency and bias considerations, Cressie and Hawkins (1980) found M-estimators applied to the square root differences  $\{|z(s_i) - z(s_j)|\}^{1/2}$  to be among the preferred robust alternatives in their simulations. The M-estimates used in our work are calculated separately for each bin, similar to Cressie and Hawkins (1980). Let  $N_d$  denote the number of pairs of locations binned a nominal separation distance  $d$  apart. Let  $M_d$  denote the median of the square-root differences in the bin with nominal separation distance  $d$  and let  $S_d$  be the corresponding median absolute deviation:

$$M_d = \text{median}_{\text{bin } d} \{|z(s_i) - z(s_j)|\}^{1/2}, \quad S_d = \text{median}_{\text{bin } d} \{|z(s_i) - z(s_j)|\}^{1/2} - M_d / 0.6745$$

Although the M-estimator is ordinarily calculated iteratively until convergence, a one-step iteration often is used. The one-step iterative estimate has been shown to give an excellent approximation to the fully iterated estimate and to be computationally expedient when computational requirements are large. For our work the one-step robust M-estimator of location was calculated as:

$$\hat{\gamma}_M(d) = \frac{\left\{ M_d + S_d \frac{\sum \psi(y_{ij})}{\sum \dot{\psi}(y_{ij})} \right\}^4}{2(0.457 + 0.494 / N_d)} \quad (8)$$

where  $y_{ij} = \{[|z(s_i) - z(s_j)|\}^{1/2} - M_d\}/S_d$ , and  $\dot{\psi}$  is the derivative of Tukey's biweight  $\psi$  (Hampel et al. 1986, Staudte and Sheather 1990):

$$\psi(t) = \begin{cases} t(4^2 - t^2)^2 & |t| \leq 4 \\ 0 & |t| > 4 \end{cases}$$

The bias correction in the denominator of (8) is the same as that in the robust estimator (4) since both are based on estimating the center of the distribution of  $\{|z(s_i) - z(s_j)|\}^{1/2}$ .

### 3.1 Simulations

To assess the relative merits of these six sample and robust semivariogram estimators, both simulations and analyses of actual data were conducted. The simulations consisted of generating random field data using Gaussian and other random processes. Spatial correlations were induced by specifying covariance matrices calculated from Gaussian and spherical, and white noise semivariogram models. Influential data were then added with stipulated mixture probabilities so that the random field was a known mixture of a specified random field process and influential data values. Very consistent results were obtained over a wide range of models, model parameters, sample sizes and mixture proportions on both transects and two-dimensional grids. A summary of the key results and illustrative plots are now given.

Plotted on Figure 4 are the results from a simulation of 200 realizations of an ordinary kriging model  $z(s) = \mu + e(s)$  with mean  $\mu = 0$  and spatially correlated errors  $e(s)$  generated by a spherical semivariogram model:

$$\gamma(d) = \begin{cases} \theta_1 + \theta_2 \{1.5(d/\theta_3) - 0.5(d/\theta_3)^3\} & d \leq \theta_3 \\ \theta_1 + \theta_2 & d > \theta_3 \end{cases}$$

Each realization consisted of data along a transect of length 100. The spherical model had a nugget  $\theta_1 = 2$ , a range  $\theta_3 = 10$ , and a sill  $\theta_1 + \theta_2 = 12$ . At location 20, an influential observation of magnitude  $\delta_m = 25$  or 50 was inserted in place of the randomly generated data value with probability 0.1 or 1 as shown over each panel in the figure. The solid curve in the figure is the theoretical semivariogram. The calculated semivariogram values averaged over the 200 realizations are indicated by the following symbols: sample (S), Cressie-Hawkins robust (R), M-Estimator (M), sample estimate following a 5% trimming of the raw data (T), sample estimate following a preliminary test for trimming anomalous raw data values (P), and sample estimate following a preliminary test for trimming anomalous binned differences (B).

The strong positive bias of the sample semivariogram is evident in three of the four panels of Figure 4. The sample semivariogram does not appear to be much affected in the first panel, but in most of the simulations the positive bias was clearly evident. The Cressie-Hawkins robust estimator does provide protection against the effects of the influential observations, as is evident in Figure 4, but it too is generally biased upward. Typically the Cressie-Hawkins robust estimator had less bias than the sample semivariogram but more than some of the other robust estimators, notably the preliminary test estimators and the M-estimator. The shift in the semivariogram plots for the sample and robust estimators occurs because the influential data value was paired with two other observations for all separation distances up to 20 and with only one thereafter. The 5% trimmed estimator consistently had large negative bias in the simulations. This negative bias is evident in Figure 4. The major result suggested in the figure and consistent throughout all the simulations was that the preliminary test estimators  $\hat{\gamma}_P(d)$  and  $\hat{\gamma}_B(d)$  and the M-estimator  $\hat{\gamma}_M(d)$  suffered the least from the effects of the influential observations.

Figure 5 displays the sample standard deviations of the semivariogram estimates across the 200 realizations for the same model that was used in Figure 4. Across the many simulation models examined, the sample semivariogram tended to be the most variable. This is especially true when the probability of an outlier was less than 1, as in the two left-hand panels of Figure 5. Thus, the sample semivariogram tends to have strong positive bias and be highly variable. At the other extreme of variability is the fixed 5% trimmed mean. This semivariogram estimator tends to have the least variability, although often by only a small amount. In spite of its relatively smaller variation, the reduction in variability does not offset its large negative bias. The preliminary test estimators not only have small bias, as discussed above, they generally have the least variability of the remaining robust estimators. The M-estimator usually has somewhat greater variability but it is not substantially worse. The Cressie-Hawkins robust estimator often has substantially more variability than the preliminary test estimators and the M-estimator.

Figures 6 and 7 show comparable results to Figures 4 and 5 for data that were generated on a two-dimensional grid using the same spherical semivariogram model, but with the range changed from 10 to 4. The influential data value was placed at the center of the 10 x 10 grid. The conclusions drawn from Figures 6 and 7 for the transect data remain basically the same for the two-dimensional grid data.

Figure 8 displays average semivariogram values from a simulation in which no influential observations were added to a random field. Data were generated from a normal random field and a highly skewed random field. The normal field data were generated using the same spherical semivariogram model as in Figure 4. The skewed data were generated by squaring data values from this same normal field. To make the theoretical semivariograms comparable in expectation, the squared normal values were location- and scale-adjusted. As is suggested in Figure 8, the preliminary test estimators  $\hat{\gamma}_P(d)$  and  $\hat{\gamma}_B(d)$  performed comparably to the sample semivariogram estimator. Overall, these three estimators had the least bias and small variability, but this finding was not consistent throughout a series of simulations of skewed data with and without influential observations. The simulations with influential observations showed that the robust estimators had far less bias than the sample semivariogram estimator, but none of the robust estimators were uniformly superior to the others. All the robust estimators were inferior to the sample semivariogram estimator when the data were highly skewed but contained no influential observations.

For spatial data generated from normal random fields, both the preliminary-test estimators  $\hat{\gamma}_P(d)$  and  $\hat{\gamma}_B(d)$  and the M-estimator  $\hat{\gamma}_M(d)$  were much closer over a wide range of models to the theoretical semivariogram than the other estimators studied in this simulation. All three provided excellent protection against serious bias due to influential observations. In addition, both provided informative diagnostics for influential observations: the weights  $\psi(t)/t$  for the M-estimator and the  $t$  values for the preliminary-test estimators. It was surprising to find that the preliminary test estimators performed as well as is indicated in Figures 4-8. They were studied because other robust estimators, including several that were not reported in this paper, did not accommodate the reuse of influential observations in the binned differences used to estimate semivariogram values.

In addition to these robust location estimators, robust scale estimators were investigated, including trimmed standard deviations (e.g. Bickel and Lehman 1976). None that were examined in preliminary simulations performed better than the location estimators that were included in the main simulations.

### 3.2 Spatial Data Analyses

The various semivariogram estimators were applied to the data sets discussed in Section 1. The 5% trimmed estimator  $\hat{\gamma}_T(d)$  will not be discussed further because of its poor performance in the simulations.

Figures 9 and 10 display directional semivariogram plots for the nitrate soil data discussed in Section 1 for the M-estimator  $\hat{\gamma}_M(d)$  and the preliminary test estimator  $\hat{\gamma}_P(d)$ , respectively. The M-estimator maintains much of the directional differences noted in the Cressie-Hawkins robust estimates shown in Figure 1. On the other hand, the preliminary test estimates  $\hat{\gamma}_P(d)$  plotted in Figure 10 are directionally consistent and they are

comparable in magnitude to the sample semivariogram values obtained by the subjective elimination in Figure 2 of the 8 largest nitrate values. The preliminary test identified 14 of the 63 data values as having unusually large values, with the  $t$  values (7) ranging from 3.1 to 23.0. Nine of the 14  $t$  values exceeded 5.0. Thus, the preliminary test estimator is achieving one of the goals of this work: the robust estimation of semivariogram values without the labor-intensive investigation of graphical and numerical summaries of the data. Note too that the  $t$  values (7) identify the influential observations and would thereby flag them to an investigator who then could choose to graphically investigate the causes of the large diagnostics.

Figure 11 compares semivariogram estimates for the European temperature anomalies discussed in Basu et al. (1995). All the sample and robust estimates except  $\hat{\gamma}_T(d)$  are plotted. The M-estimator again performs similarly to the Cressie-Hawkins robust estimator. The binned preliminary test estimator  $\hat{\gamma}_B(d)$  tracks the sample semivariogram estimator and is virtually identical with it in most of the bins. The preliminary test estimator  $\hat{\gamma}_P(d)$  identifies ( $t = 5.9$ ) and removes Copenhagen, as well as 6 other influential data values that have  $t$  values ranging from 3.1 to 4.0. Copenhagen's  $t$  value would certainly cause one to investigate it further. The 6 other temperature stations are all clustered nearby one another just north of the Caspian Sea and might reflect some regional anomaly. Perhaps the most striking feature of Figure 11 is that the preliminary test estimator is the only estimator that plateaus, reaching a sill of about 0.8 at a range of approximately 1,200-1,400 km. None of the other estimators in the figure show evidence of having reached a sill.

Figure 12 is included only for completeness. The preliminary test estimator clearly identifies the influential observation ( $t = 8.2$ ) and as a result eliminates most of the excitation crest that is so prominent in the sample and the other robust semivariograms. The remaining portion of the excitation crest is due to the different magnitudes of variability in the large and the small grid portions of the data. In practice, these would be separately estimated.

## Conclusion

The preliminary test estimator P was most effective overall in identifying and accommodating highly influential spatial data values. The M-estimator and the preliminary test estimator B also were generally effective, but occasionally, as in Figure 11, they did not satisfactorily accommodate one or more influential data values. Perhaps the most important conclusions from this work are that robust estimators are preferable to the classical estimator and that estimators such as the preliminary test estimators and the M-estimator are well-suited to large data sets because they efficiently identify highly influential data.

## References

- Basu, S., Gunst, R.F., Guertal, E.A., and Hartfield, M.I. (1995). "The Effects of Influential Observations on Sample Semivariograms," submitted to *the Journal of Agricultural, Biological, and Environmental Statistics*.
- Bickel, P. and Lehmann, E.L. (1976). "Descriptive Statistics for Nonparametric Models: III. Dispersion," *The Annals of Statistics*, 4, 1139-1158.
- Cressie, N. (1984). "Toward Resistant Geostatistics," in G. Verly, M. David, A.G. Journel, and A. Marechal (eds.), *Geostatistics for Natural Resource Characterization*. Dordrecht, Holland: D. Reidel Publishing Co.
- Cressie, N. (1986). "Kriging Nonstationary Data," *Journal of the American Statistical Association*, 81, 625-634.
- Cressie, N. (1991). *Statistics for Spatial Data*. New York: John Wiley and Sons, Inc.
- Cressie, N. and Hawkins, D.M. (1980). "Robust Estimation of the Variogram: I," *Mathematical Geology*, 12, 115-125.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics*, New York: John Wiley and Sons, Inc.
- Matheron, G. (1962). *Traite de Geostatistique Appliquee, Tome I. Memoires du Bureau de Recherches Geologiques et Minieres, No. 14*, Paris: Editions Technip.
- McBratney, A.B. and Webster, R. (1986). "Choosing Functions of Semivariograms of Soil Properties and Fitting Them to Sampling Estimates," *Journal of Soil Science*, 37, 617-639.
- Staudte, R.G. and Sheather, S.J. (1990). *Robust Estimation and Testing*. New York: John Wiley and Sons, Inc.
- S-PLUS (1991). *S-PLUS Reference Manual, Version 3.0*. Seattle, WA: Statistical Sciences, Inc.

Fig. 1. Robust Nitrate Semivariogram Values.

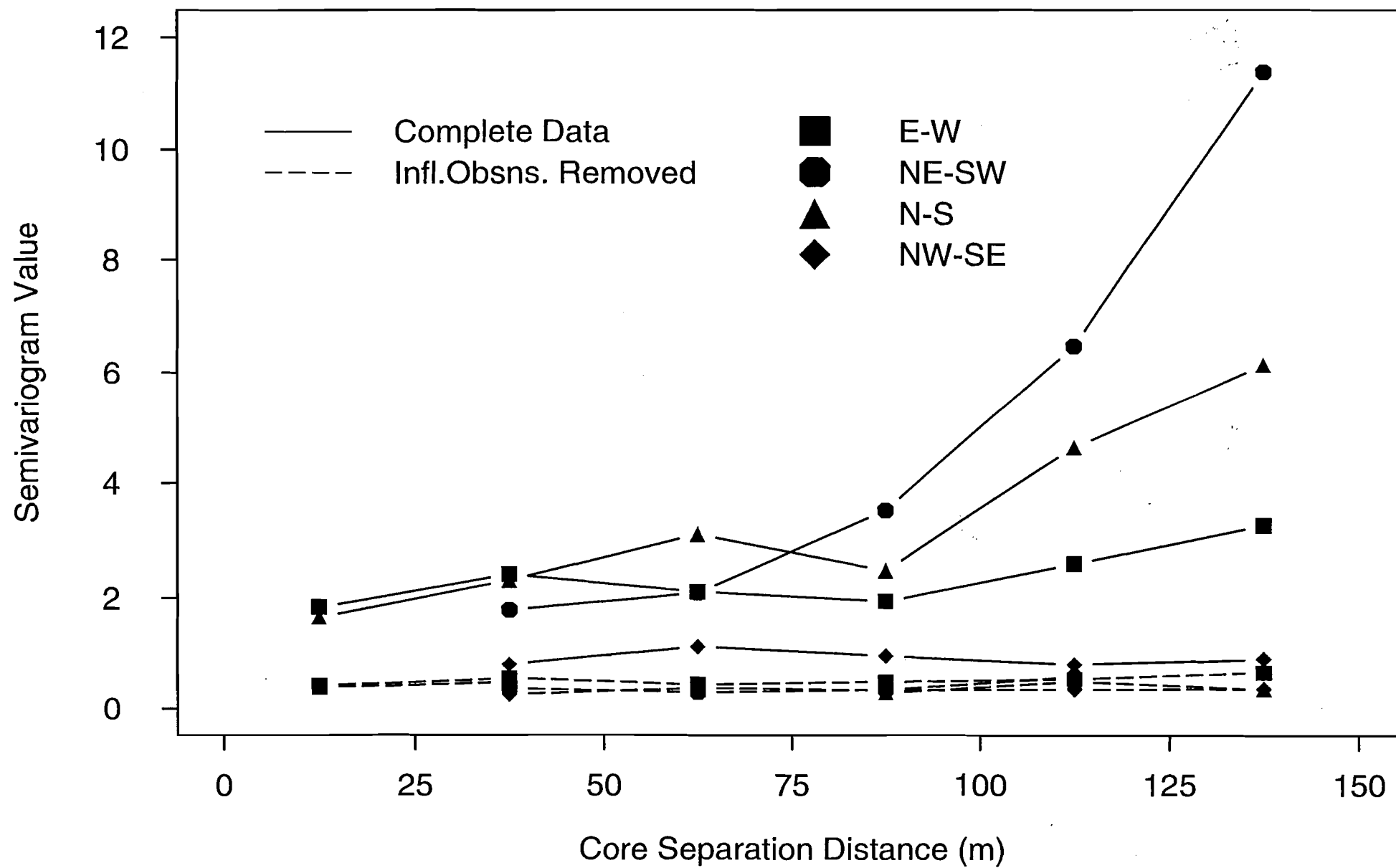


Fig. 2. Three-Dimensional Plot of Median-Polish Nitrate Residuals.

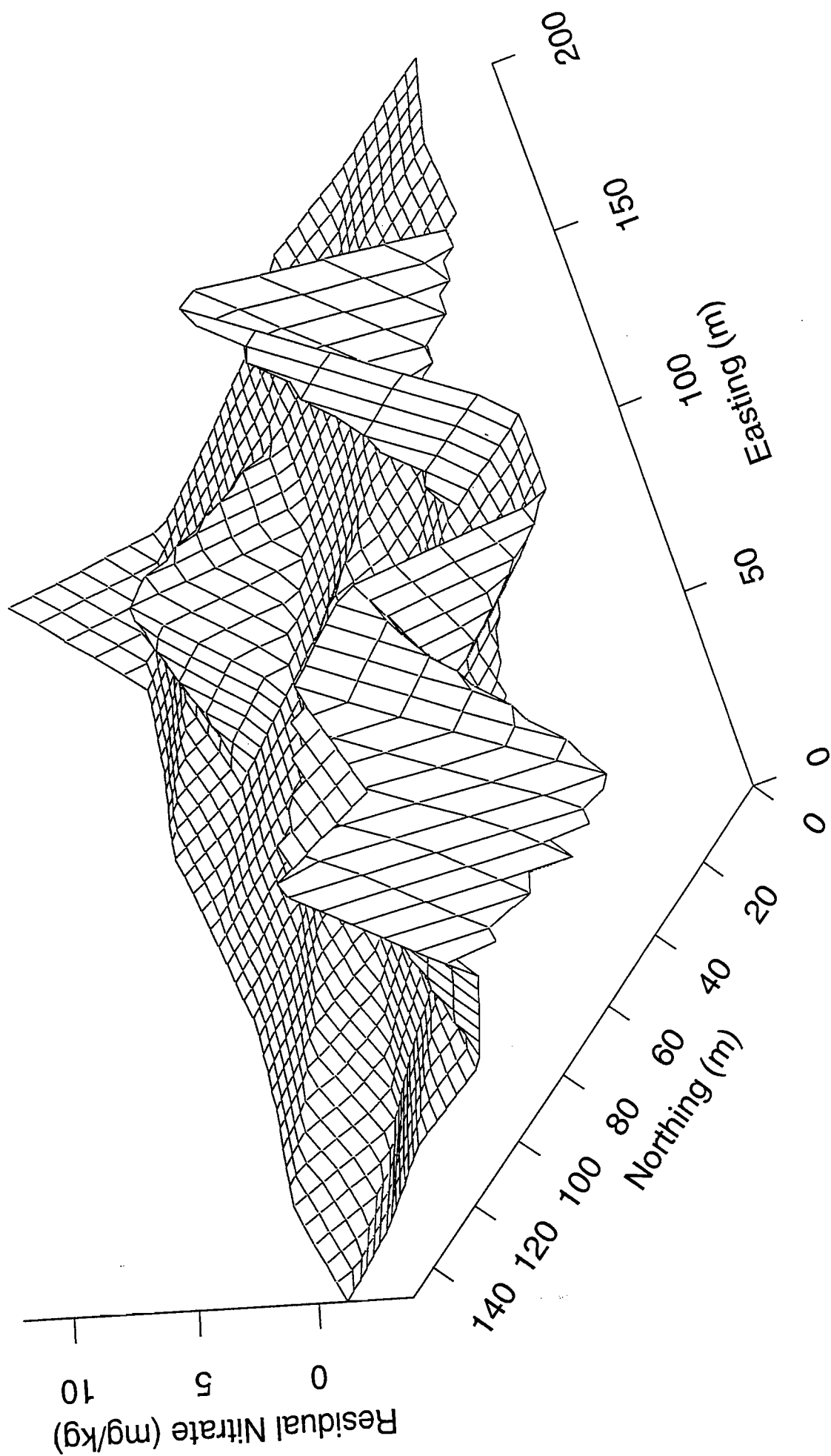


Fig. 3. Semivariogram Approximation Using Influence Functions.

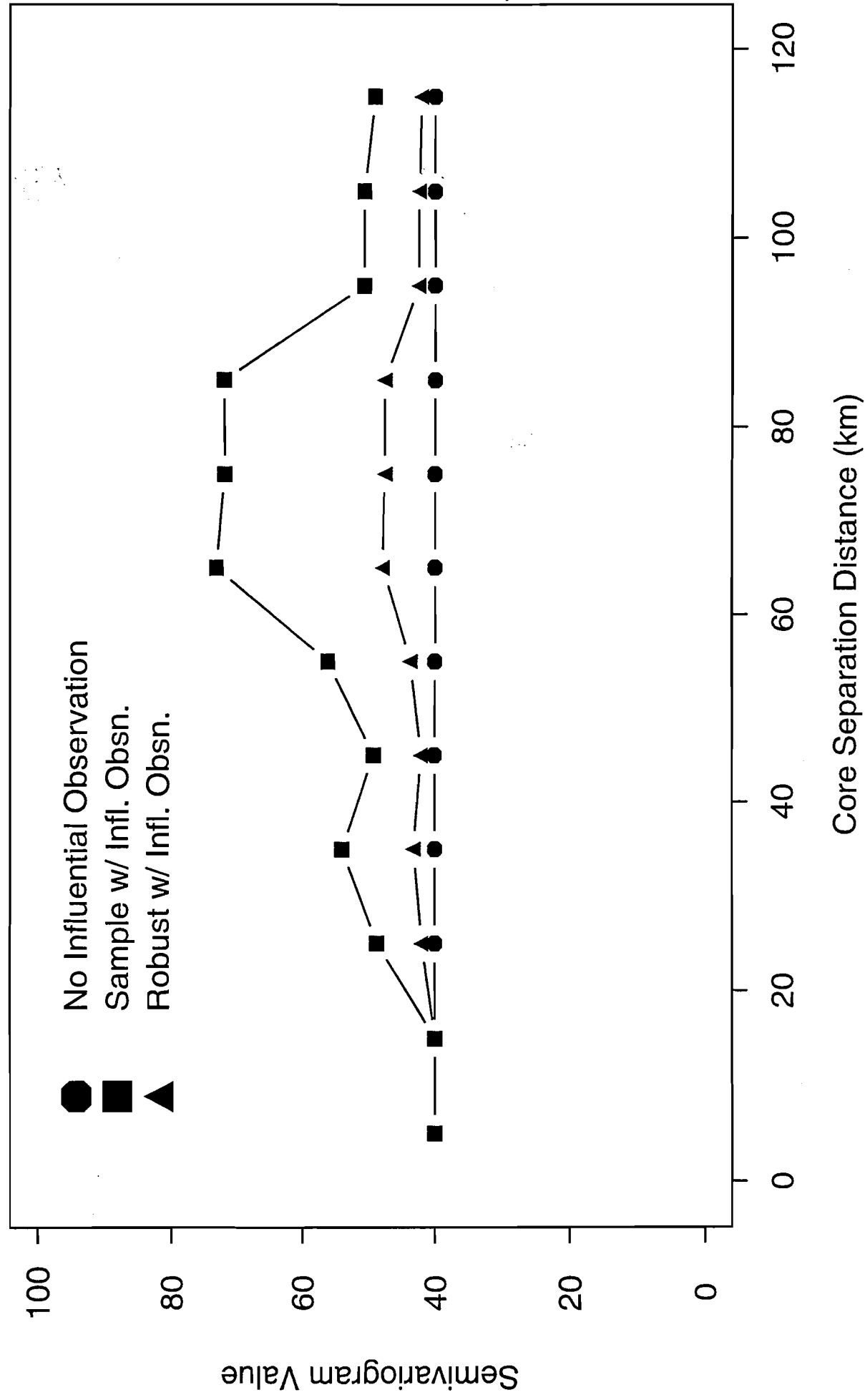


Fig. 4. Average Semivariogram Estimates, Transect Data

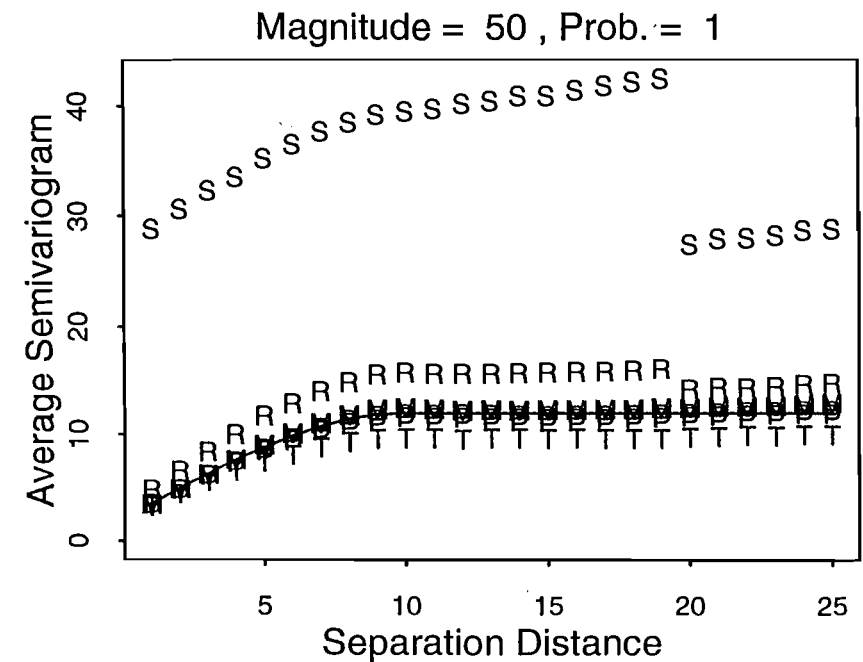
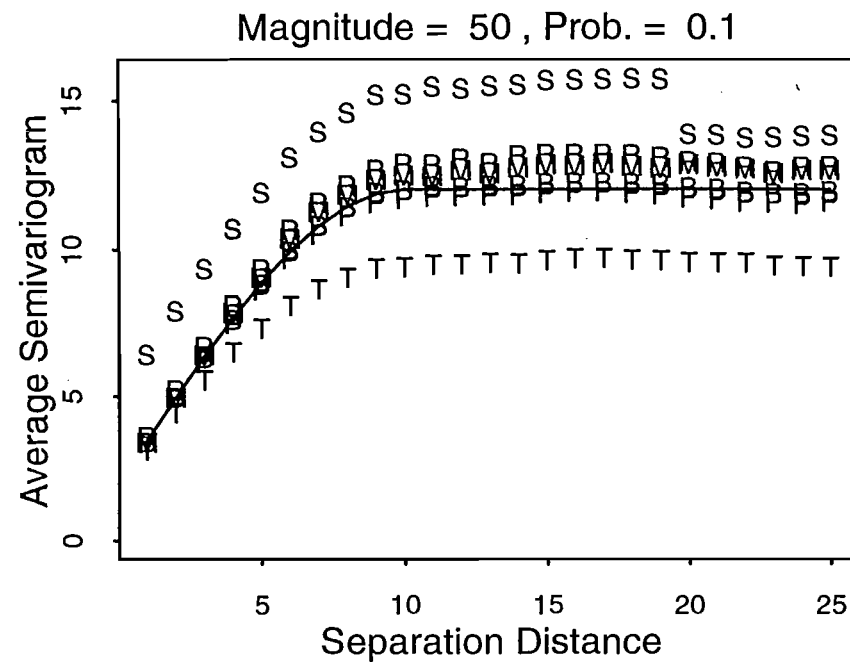
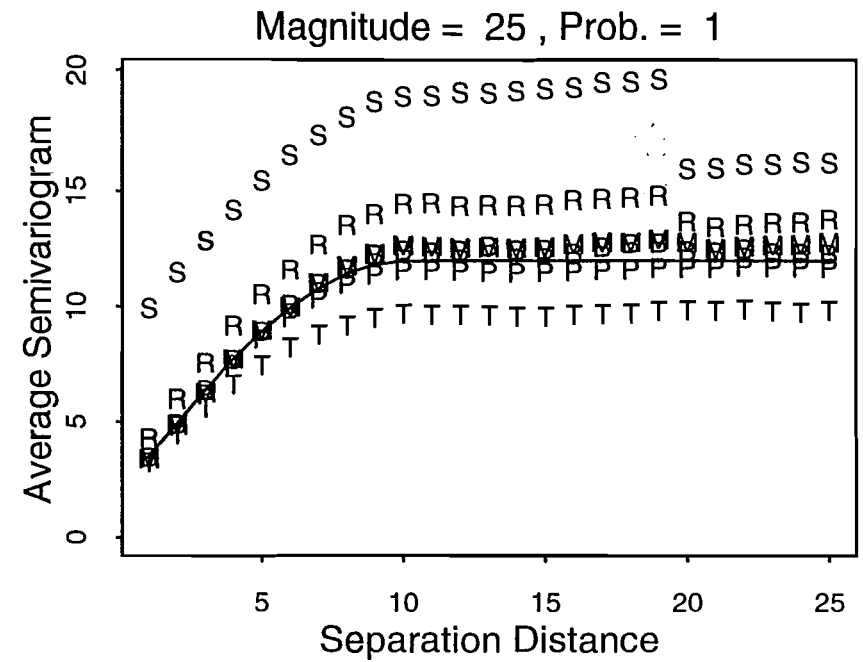
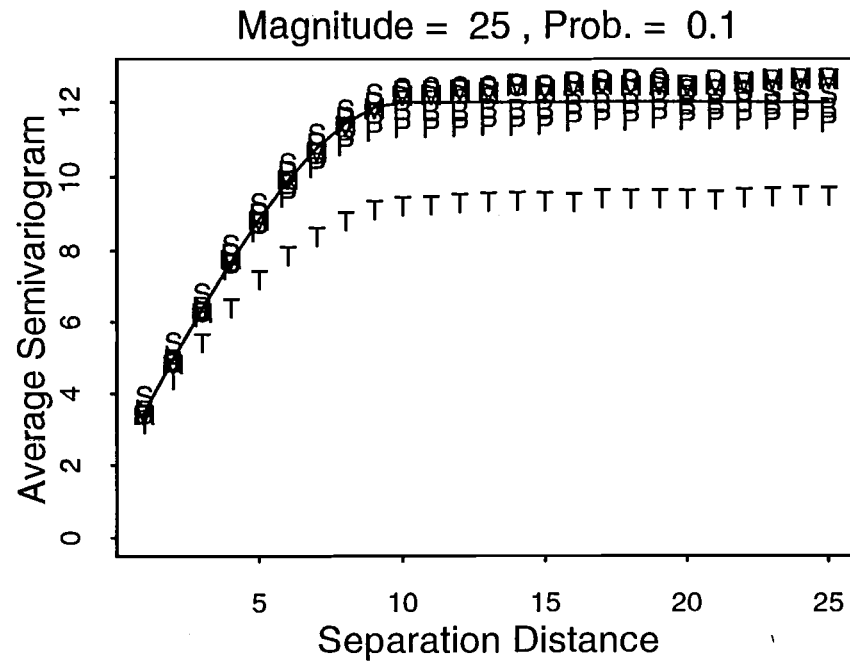
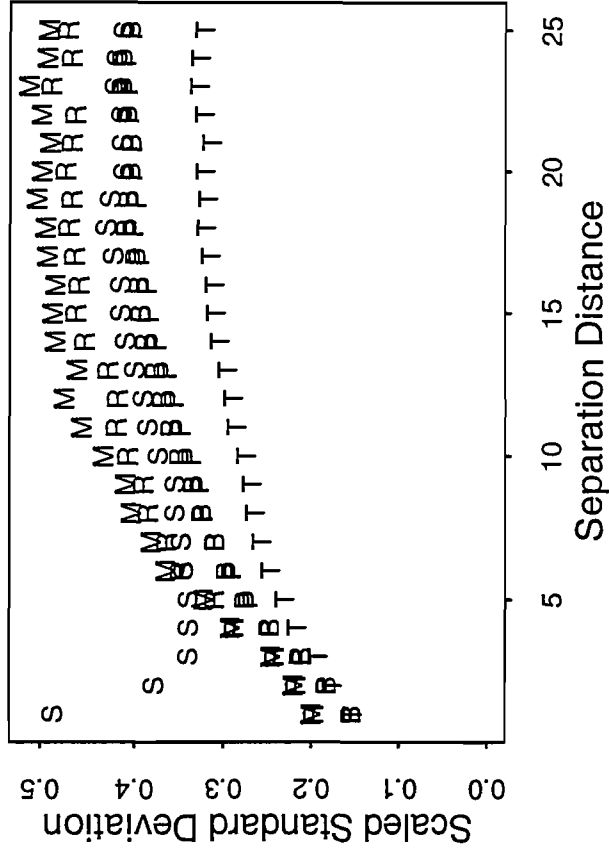
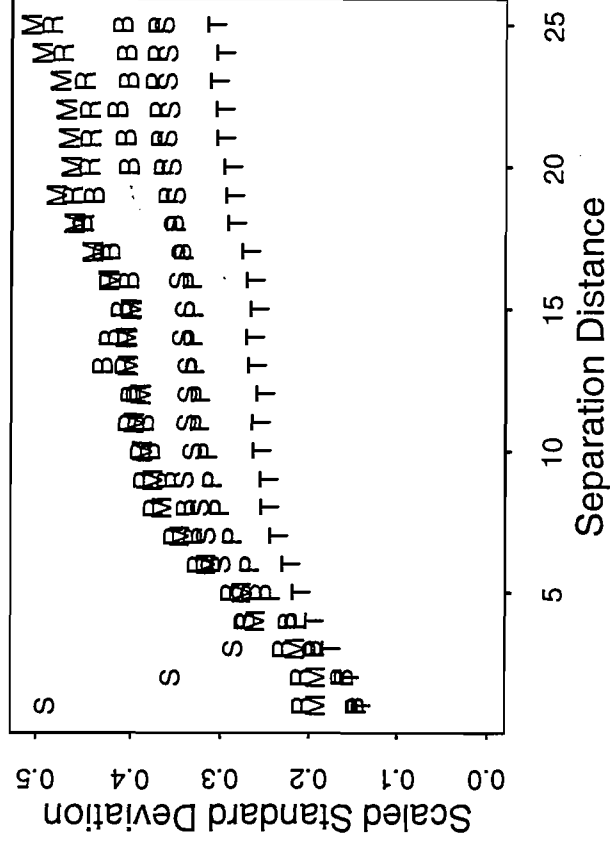


Fig. 5. Variability of Semivariogram Estimators, Transect Data

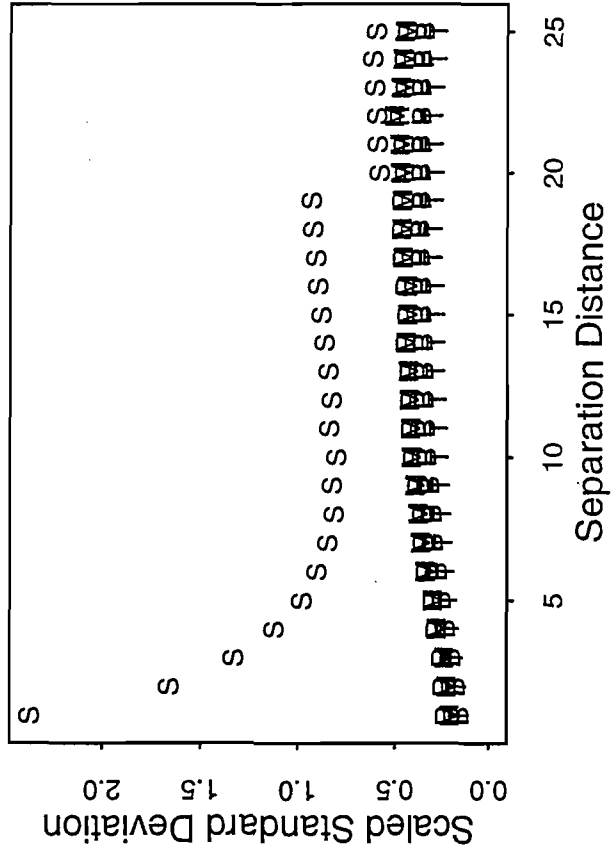
Magnitude = 25 , Prob. = 0.1



Magnitude = 25 , Prob. = 1



Magnitude = 50 , Prob. = 0.1



Magnitude = 50 , Prob. = 1

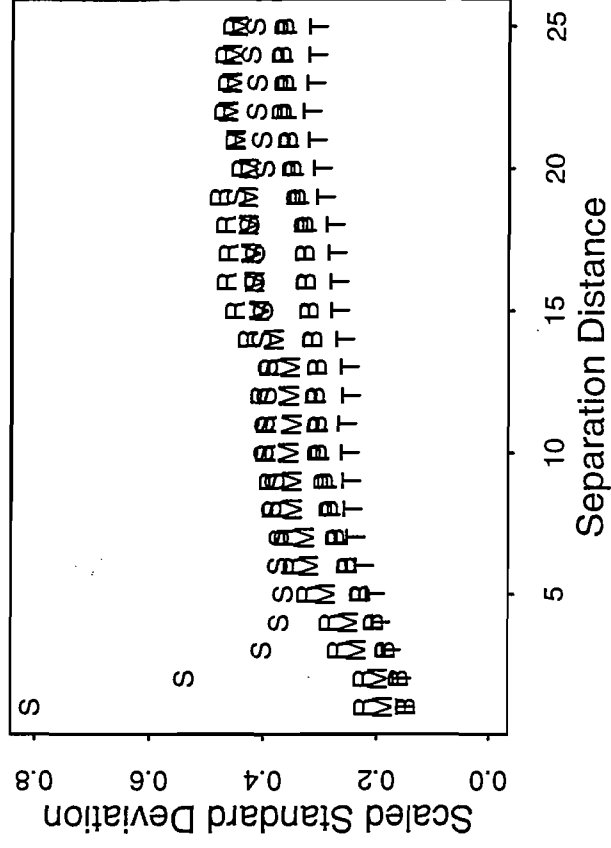


Fig. 6. Average Semivariogram Estimates, Gridded Data

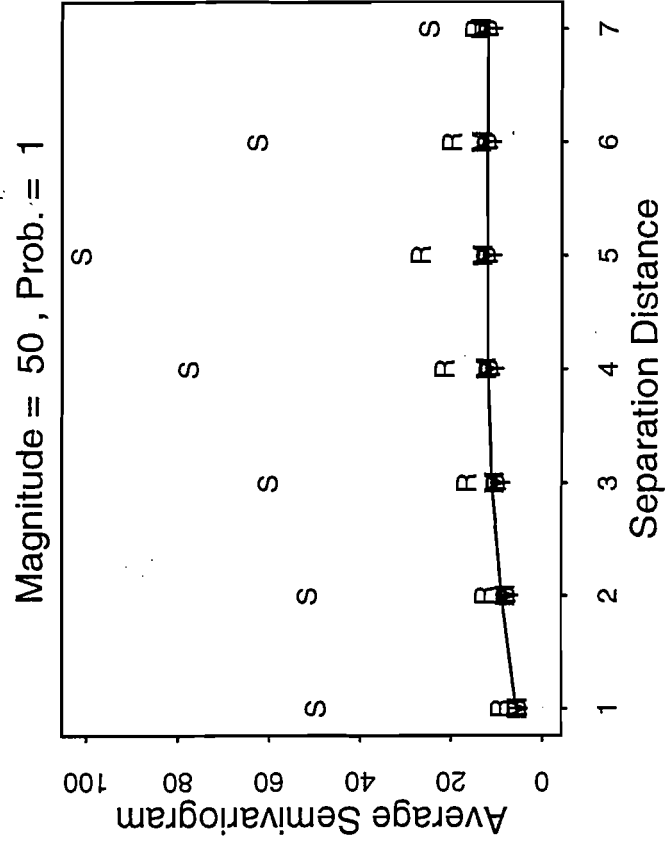
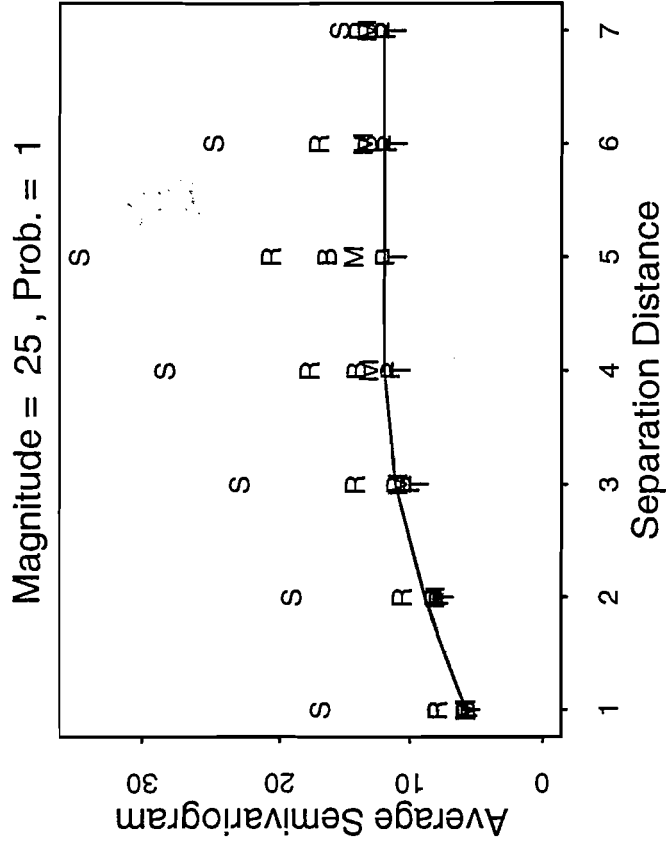
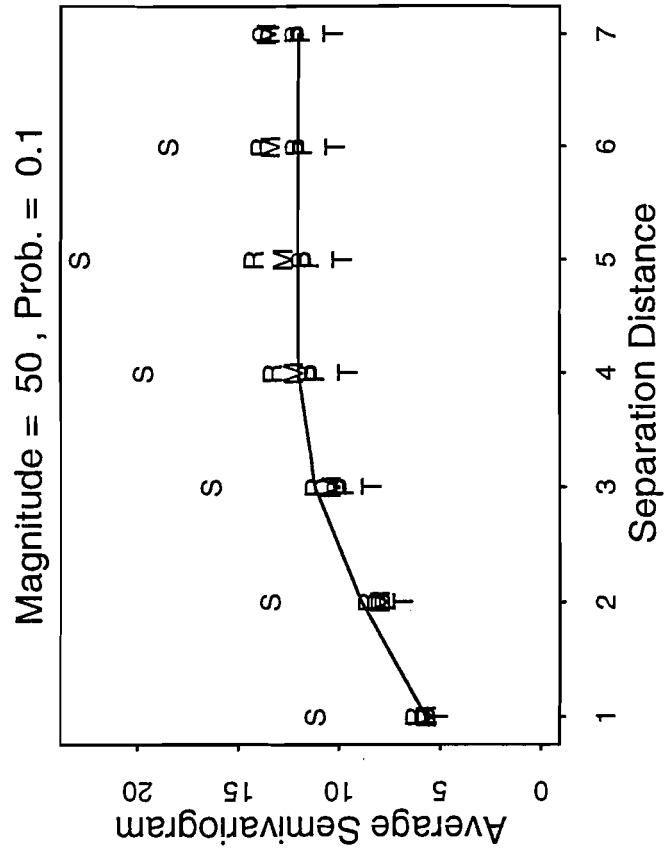
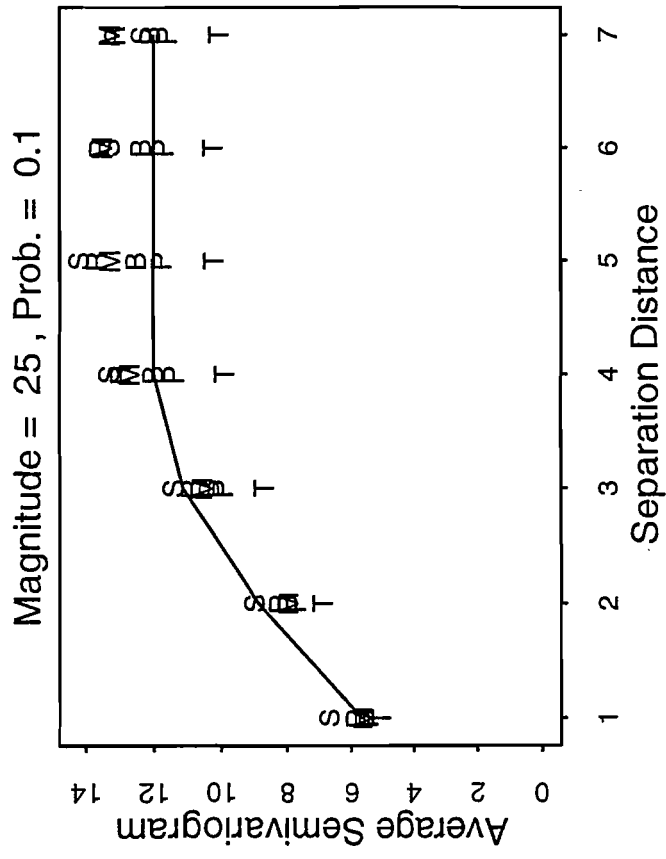


Fig. 7. Variability of Semivariogram Estimators, Gridded Data

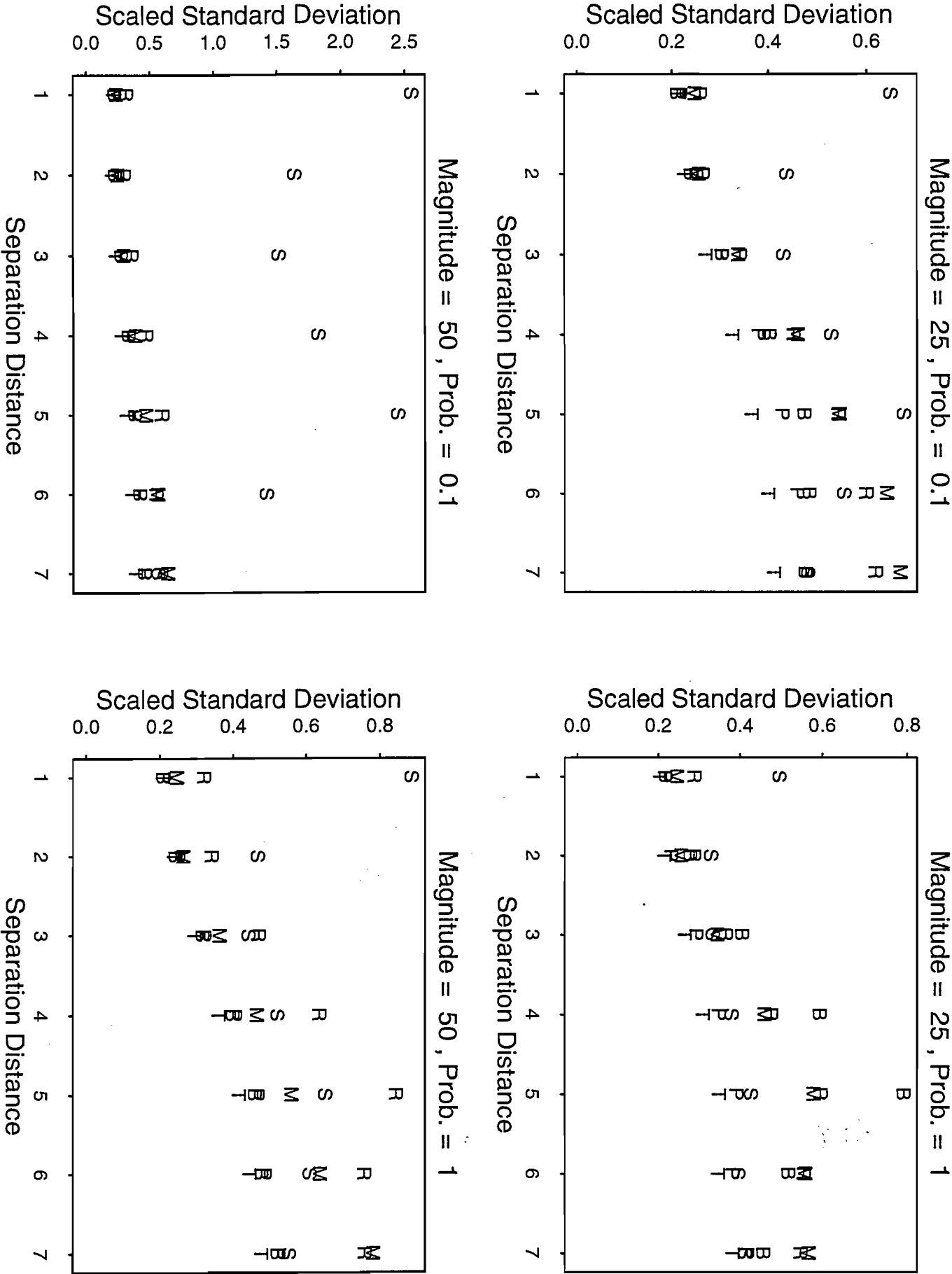


Fig. 8. Comparison of Semivariogram Estimators for Uncontaminated Distributions

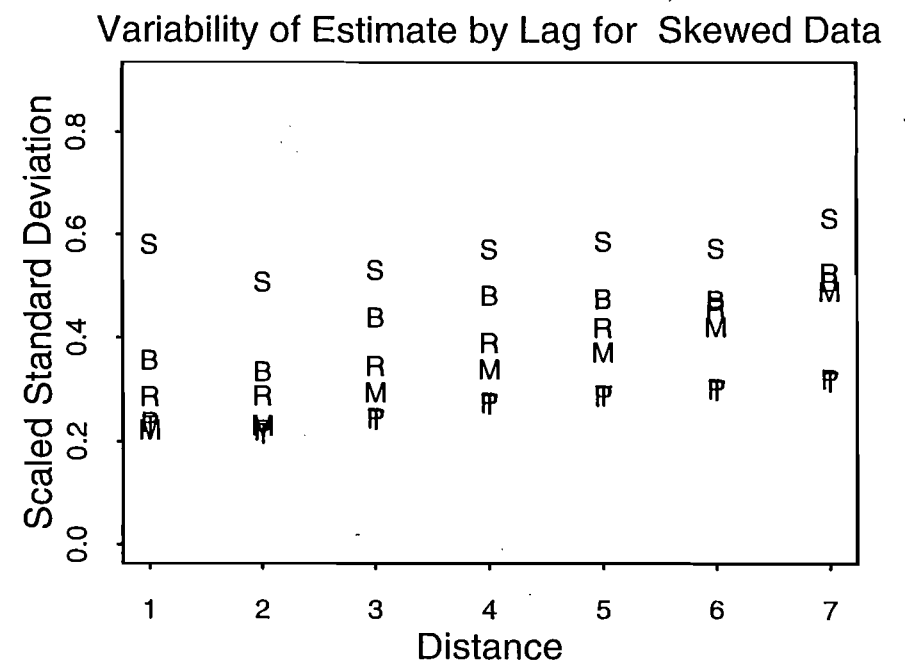
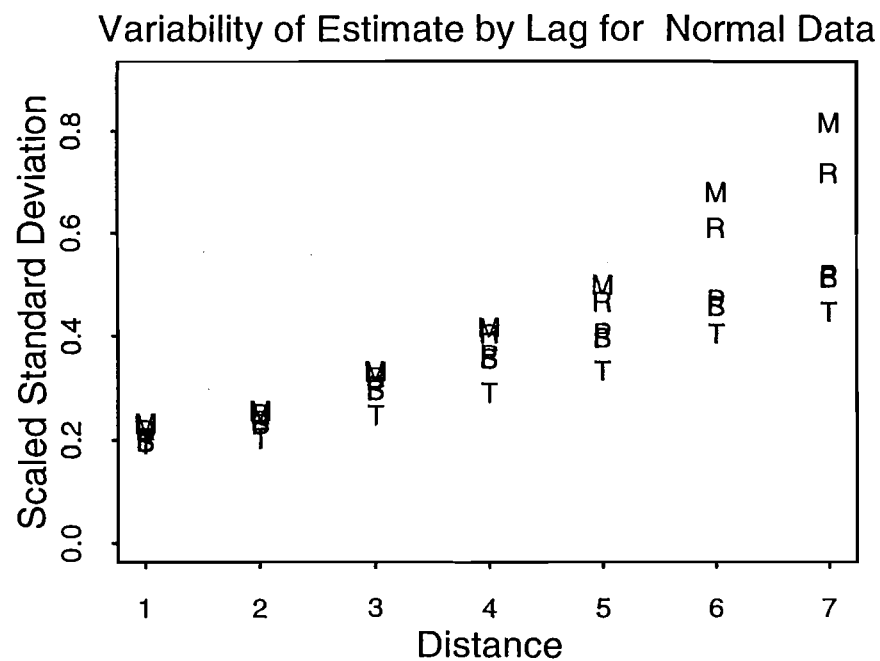
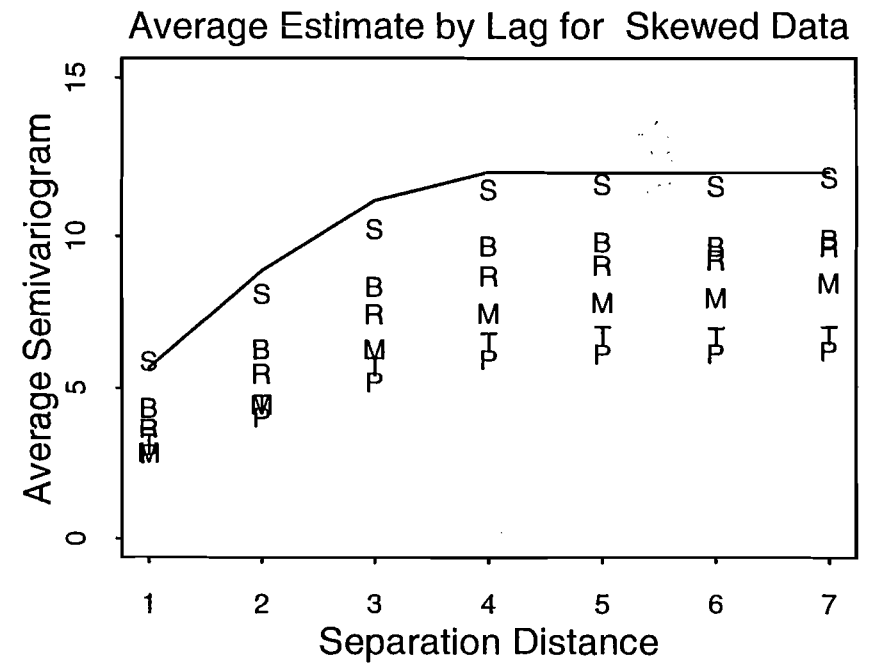
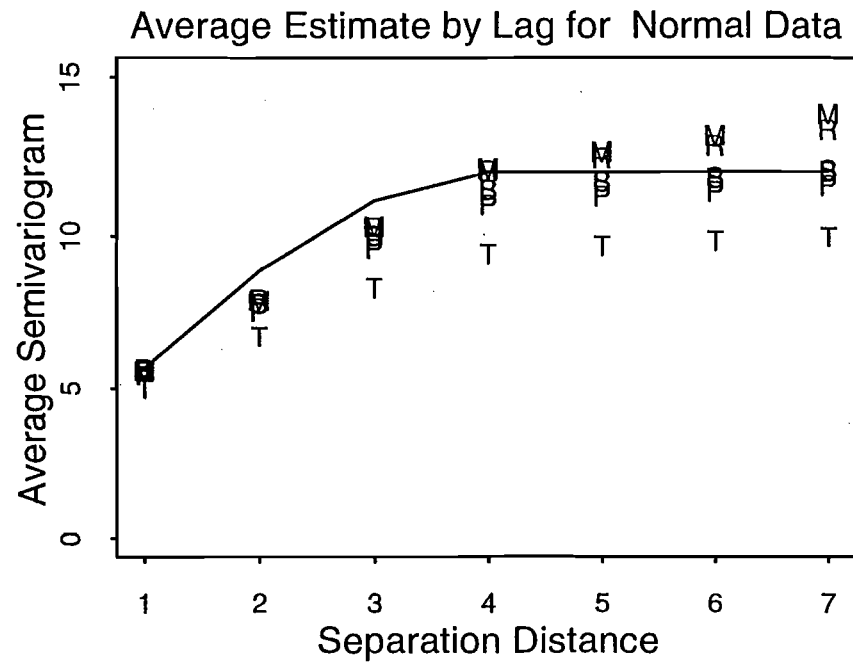


Fig. 9. M-Estimator Nitrate Semivariogram Values.

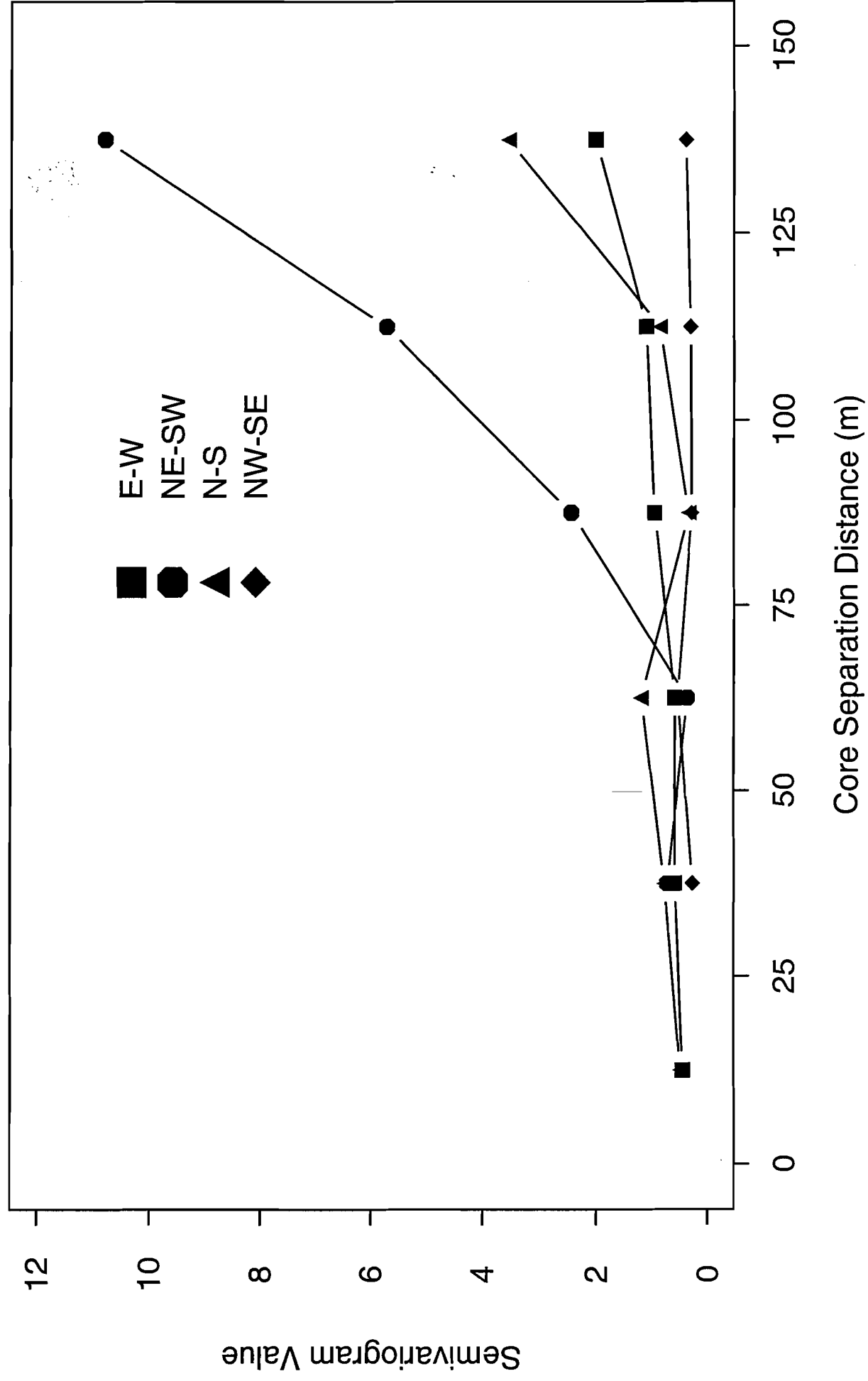


Fig. 10. Preliminary Test Trimmed Nitrate Semivariogram Values.

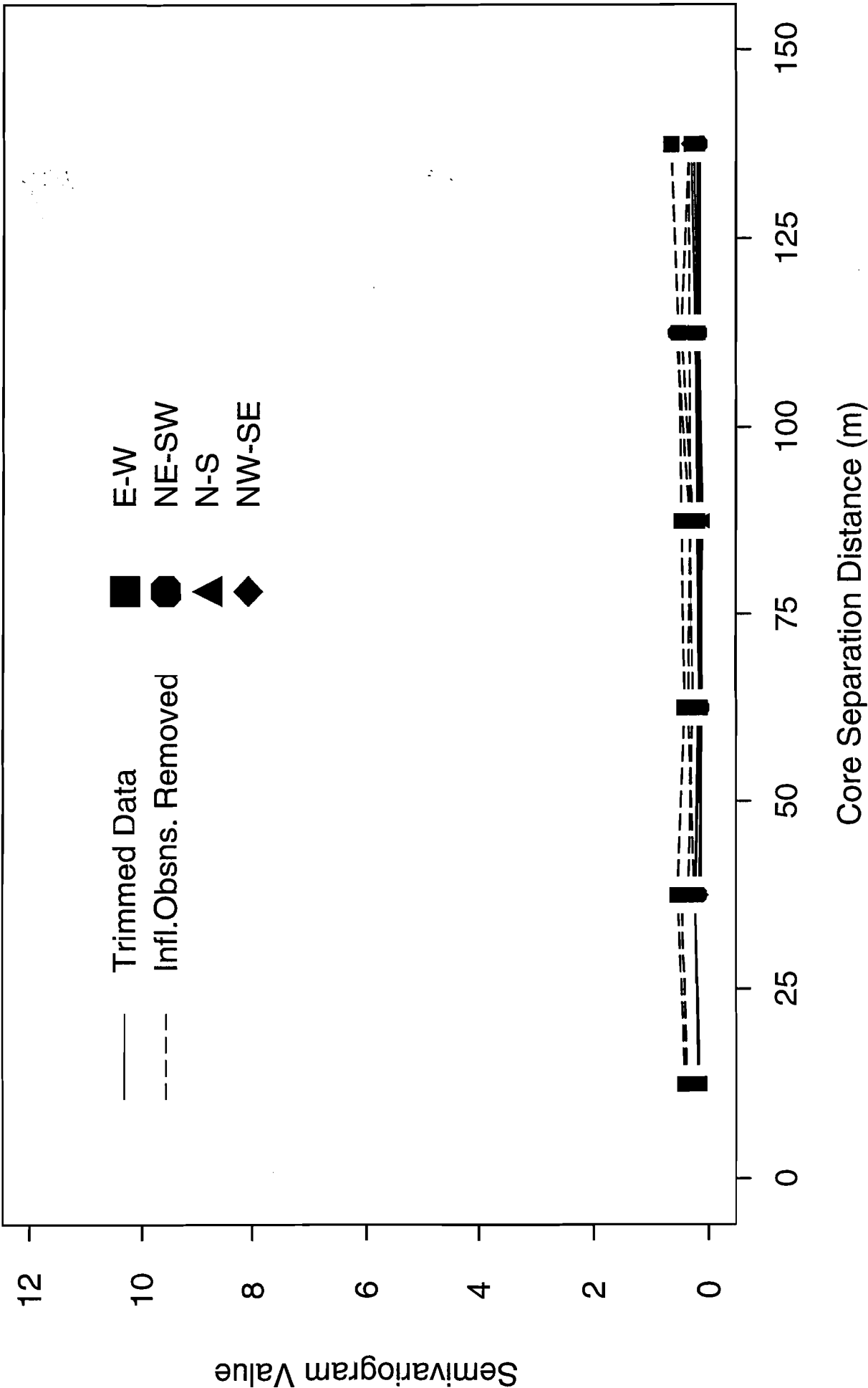


Fig. 11. Comparison of Semivariogram Estimates for European Temperature Anomalies.

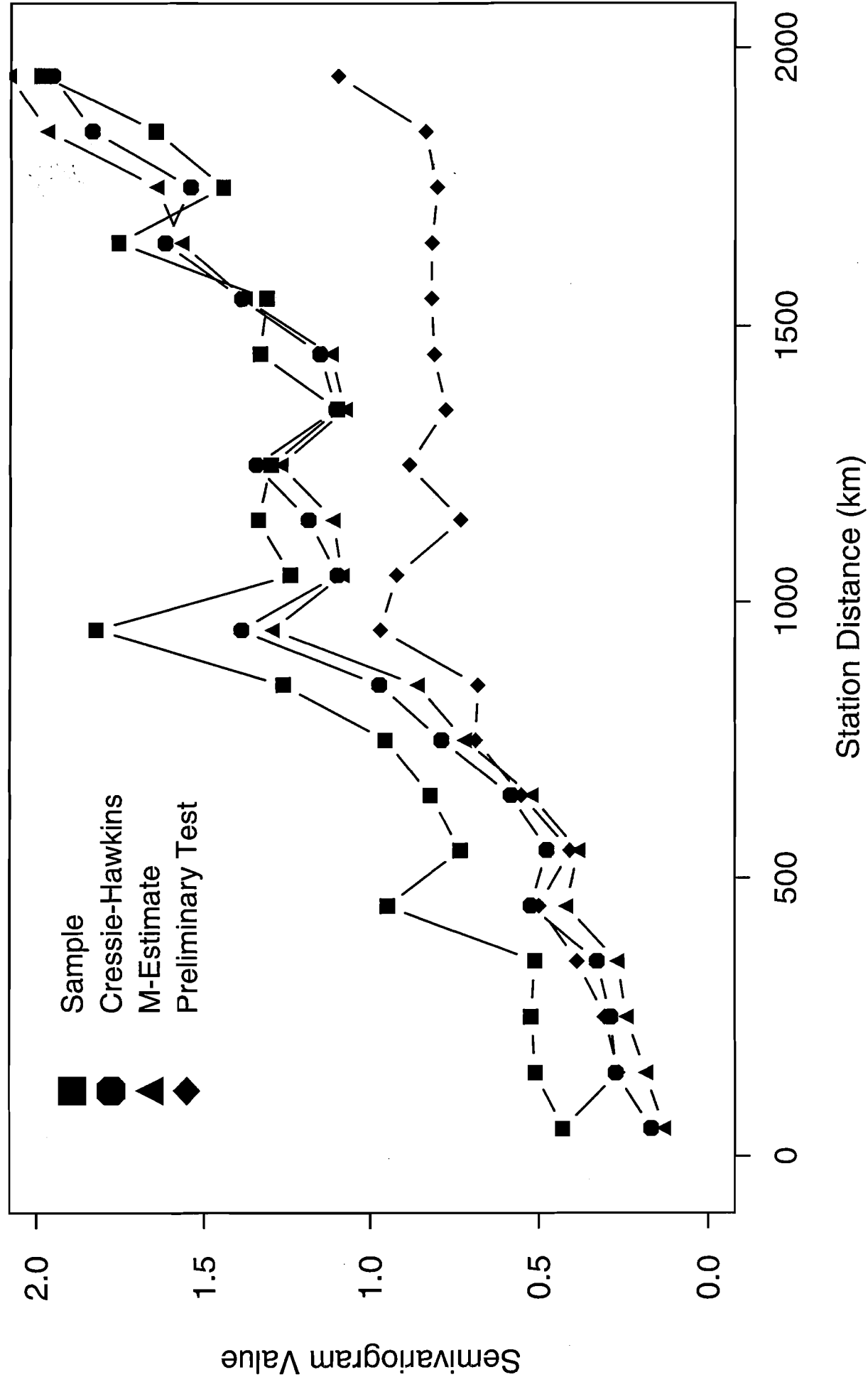




Fig.12. Comparison of Semivariogram Estimates for Nitrate Residuals.

