Kriging

Summary

Kriging is a procedure that is widely used to analyze geospatial data. Given a set of measurements taken on a variable at various locations within a two-dimensional region, estimates are derived for the value of that variable throughout the region. The primary output is a map of the estimated value, together with the variance of the estimate.

Sample StatFolio: *kriging.sgp*

Sample Data

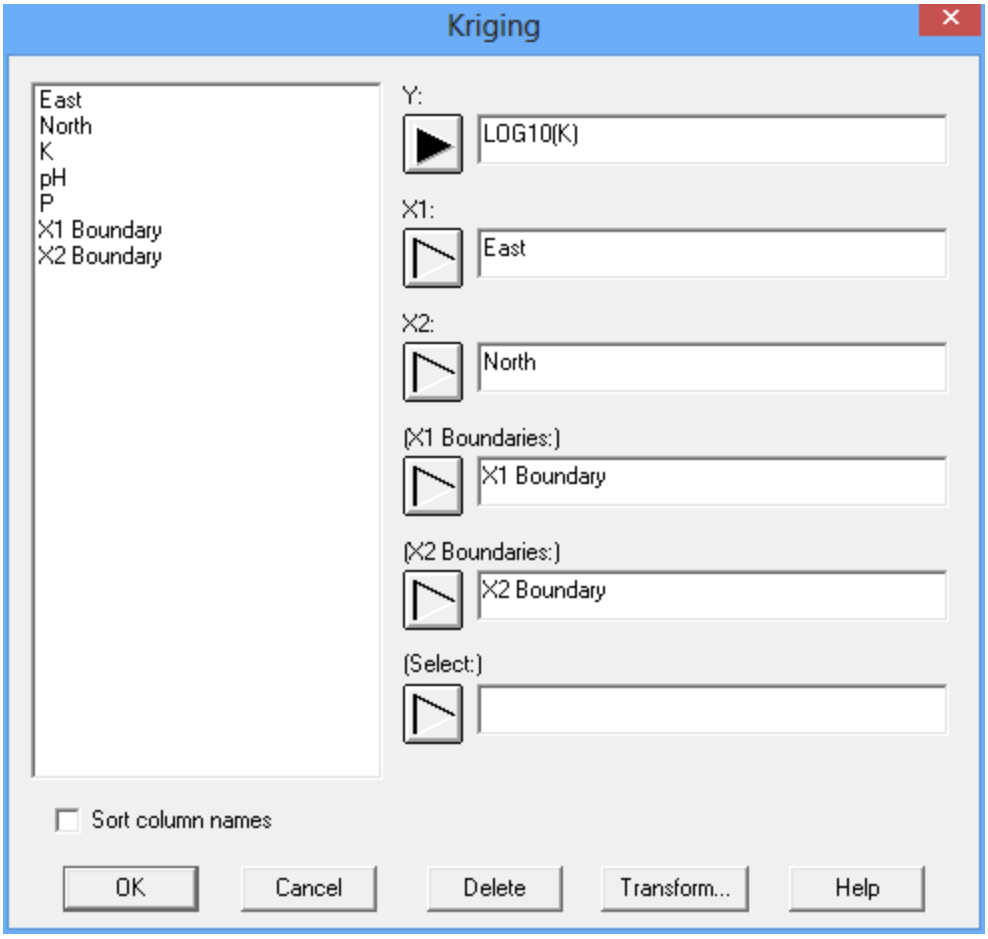
The file *BroomsBarn.sgd* contains a widely analyzed set of data obtained at the Rothamsted Agricultural Research Center in the UK. Soil samples were taken at 435 locations throughout a 77 ha field at Broom's Farm Barn. Measurements were made of the exchangeable potassium, pH, and available phosphorus of each of the samples. A section of that data is shown below:

East	North	K	pH	P	X1 Boundary	X2 Boundary
20	940	26.0	7.2	5.5	0	1240
20	980	22.0	7.2	5.2	0	920
20	1020	18.0	6.8	2.6	40	740
20	1060	19.0	6.4	1.3	80	540
20	1100	26.0	6.1	1.3	80	0
20	1140	23.0	6.6	6.9	140	0
20	1180	32.0	7.8	6.6	220	160
20	1220	28.0	8.0	7.8	260	240
60	740	55.0	6.8	2.6	300	240

The samples were taken at 40-m intervals at locations defined by the *East* and *North* columns.

Data Input

To analyze the data, enter the names of the variables into the following data input dialog box:



- **Y:** numeric column containing the data to be analyzed.
- **X1:** numeric column containing the location of the samples in dimension #1. The locations may be in any units and do not have to be regularly spaced.
- **X2:** numeric column containing the location of the samples in dimension #2. The units should be the same as *X1* so that valid distances between samples may be calculated.
- **X1 Boundaries:** numeric column containing the coordinates of the region's boundary points along the horizontal dimension.
- **X2 Boundaries:** numeric column containing the coordinates of the region's boundary points along the vertical dimension.
- **Select:** subset selection.

In the discussion below, $X_{i,j}$ represents the *i-th* value of the *j-th* variable, for $i = 1, \ldots, n$ and $j = 1$ or 2.

Note that the samples do not need to be located on a rectangular grid, although they often are.

The boundary columns define the region to be Kriged by specifying a set of coordinates that may be connected to form an enclosed polygon. For the sample data, the boundary points are:

X1 Boundary	X2 Boundary
0	1240
0	920
40	740
80	540
80	0
140	0
220	160
260	240
300	240
360	40
580	40
720	80
720	1200
180	1240
0	1240

To form a closed boundary, the last point should be identical to the first.

Analysis Options

The *Analysis Options* dialog box specifies general options for the analysis:

- **Variogram Maximum lag distance:** the largest lag distance at which the variogram will be calculated.

- **Variogram Increment:** the distance between estimates of the variogram.

- **Variogram Model:** type of model used to describe the experimental variogram. The choices are described later in this document.

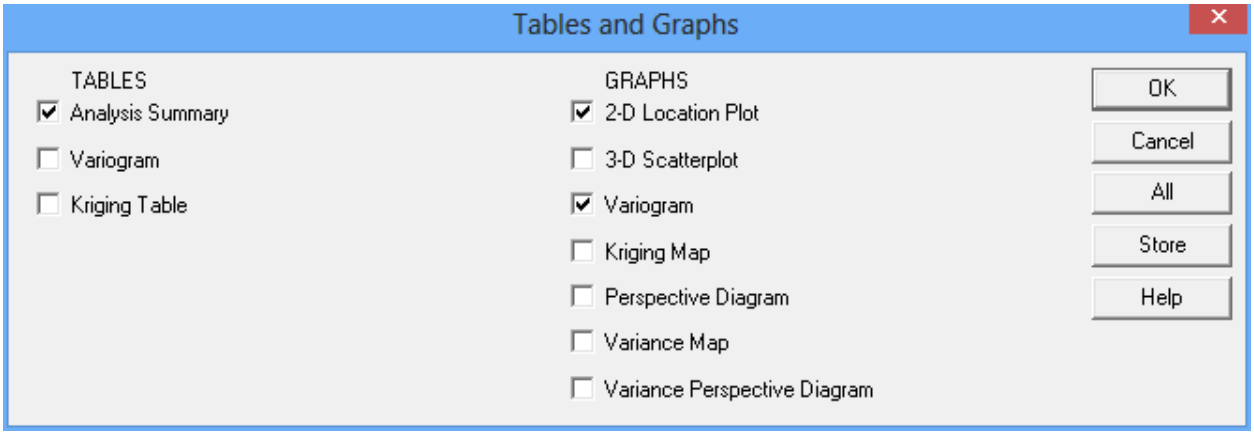
- **Include nugget:** whether to include a term in the variogram model called the *nugget*, which is the value of the variogram for a lag equal to 0. If not included, the variogram will be forced to pass through the origin.

- **Variogram Weights:** the weights used by the nonlinear least squares algorithm when fitting the variogram, as described later.

- **Kriging Parameters:** settings that affect the estimated response when kriging is performed, as described in a later section.

Tables and Graphs

The following tables and graphs are available:



Analysis Summary

The *Analysis Summary* summarizes the input data:

Kriging - LOG10(K)
Data variable: LOG10(K) ranging from 1.07918 to 1.98227
Dimension 1: East ranging from 20.0 to 700.0
Dimension 2: North ranging from 20.0 to 1220.0

Number of observations: 434
Minimum distance between observations: 40.0
Maximum distance between observations: 1310.27
Average distance between nearest neighbors: 40.0

The table shows:

1. The range of the measured variable *Y*.
2. The range of the two dimensions X_1 and X_2.
3. The number of locations where *Y* was measured.
4. The minimum distance between any 2 locations.
5. The maximum distance between any 2 locations.
6. The average distance between each point and its nearest neighbor.

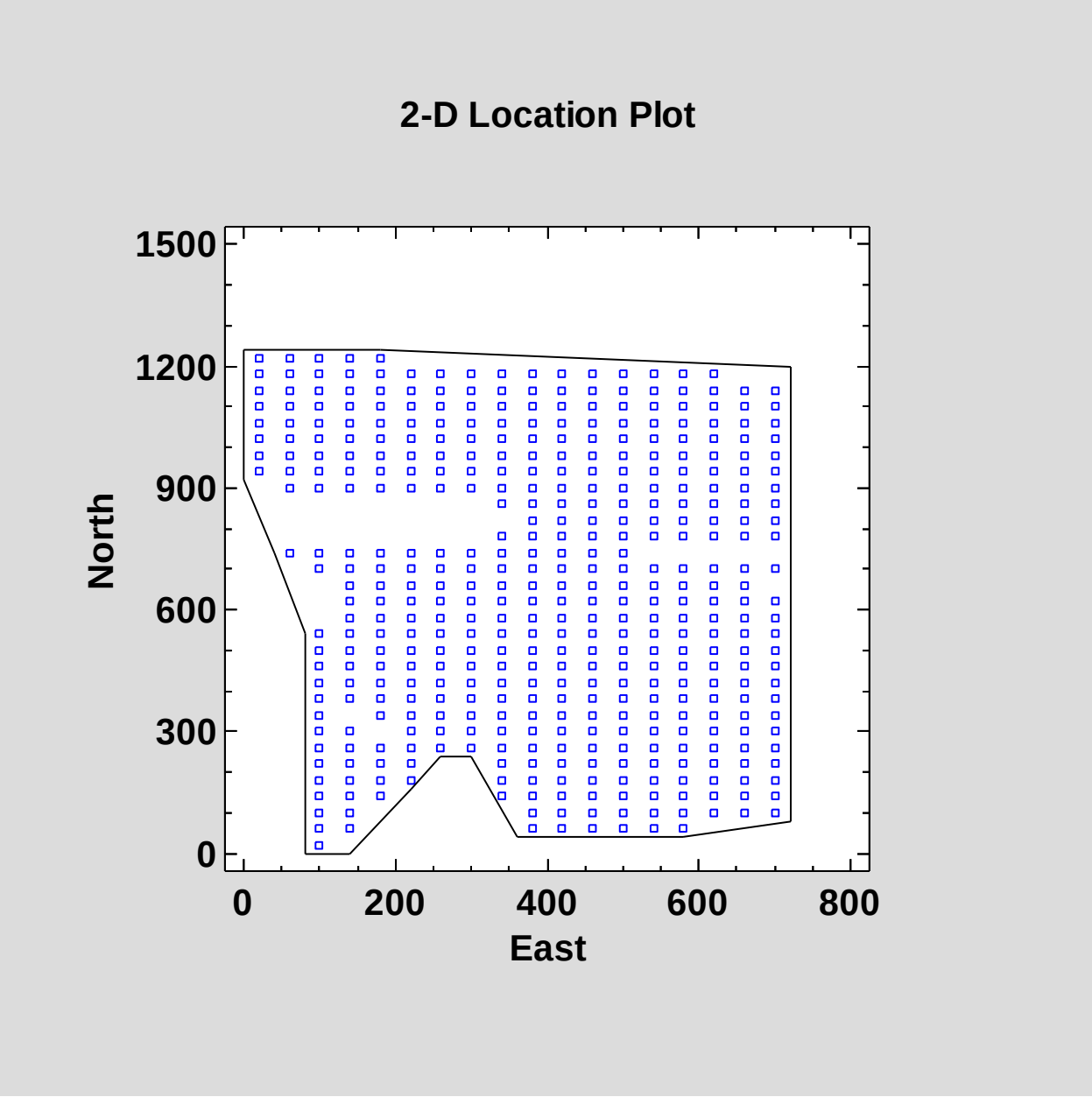
The distance between point *i* and point *j* is defined as:

$$d_{i,j} = \sqrt{\left(X_{i,1} - X_{j,1}\right)^2 + \left(X_{i,2} - X_{j,2}\right)^2} \qquad (1)$$

In the example, there were 434 valid measurements of K. (One measurement was negative and has been treated as a missing value.) The minimum distance between observations was 40 m, which was also the average distance between each point and its nearest neighbor (since the observations were taken on a rectangular grid). The maximum distance between any 2 observations was 1310.27 m.

2-Dimensional Location Plot

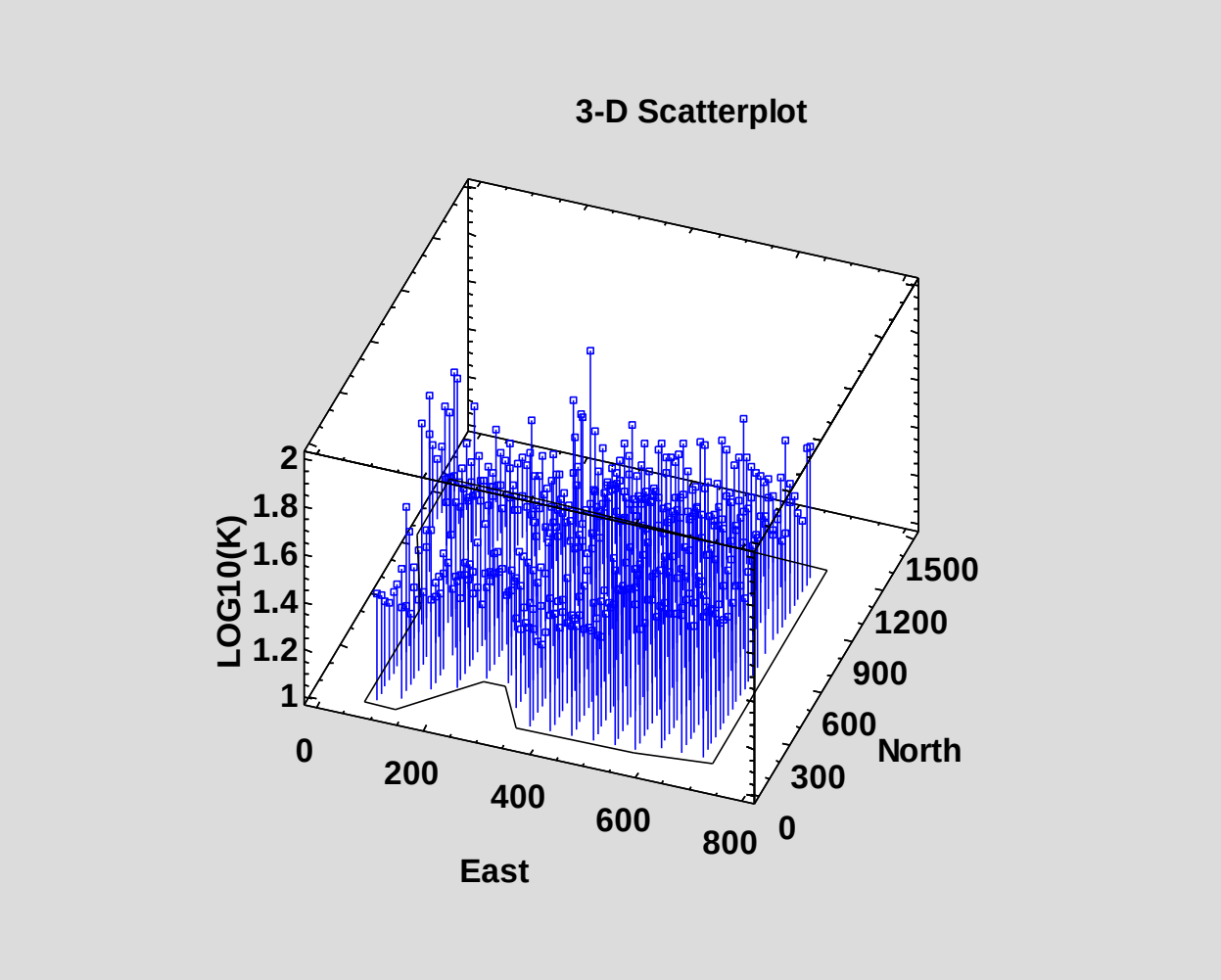
This plot shows the locations at which samples were obtained:



Any locations with missing or invalid data are not shown.

3-Dimensional Scatterplot

This plot shows the measurements taken at each location:



Variogram

An important function that must be computed before Kriging can be performed is the *variogram* or *semivariogram*. This function measures the variance between pairs of points located a distance *h* apart from each other, as a function of *h*. The value of the variogram at a lag distance *h* is calculated by

$$\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} \{y(x_i) - y(x_i + h)\}^2 \qquad (2)$$

where $y(x_i)$ is the observed value of *Y* at the location x_i, which is a two-dimensional point. The summation is taken over the *m(h)* pairs of points that are separated by a distance ranging between *h - d/2* and *h + d/2*. The variogram is evaluated at various distances h_j separated by the variogram increment *d*. Note that the definition in equation (2) has a 2 in the denominator. For this reason, the quantity calculated is sometimes referred to as the *semivariance*.

If you select *Variogram* from the list of tables and graphs, the program will show the calculated values at various lags:

Variogram for LOG10(K)
Fitted model: Spherical: nugget=0.00520676 sill=0.0198453 range=437.506 R-squared=99.77% RMSE=0.000243689
Weights: counts and model

Lag	*Sample Semivariance*	*Pairs*	*Model Semivariance*	*Residual*
40.0	0.00725882	1545	0.00720871	0.0000501052
80.0	0.00944553	2135	0.00917711	0.000268427
120.0	0.0107559	2643	0.0110784	-0.000322524
160.0	0.0126	4824	0.012879	-0.000278993
200.0	0.0145756	3916	0.0145453	0.0000303212
240.0	0.0162739	5142	0.0160438	0.000230093
280.0	0.017451	4748	0.017341	0.000110002
320.0	0.0183789	5177	0.0184032	-0.0000243233
360.0	0.0193334	6547	0.0191969	0.000136479
400.0	0.0196017	4876	0.0196886	-0.0000869128
440.0	0.0201391	5630	0.0198453	0.00029376
480.0	0.0194541	4855	0.0198453	-0.000391246
520.0	0.0198106	5468	0.0198453	-0.000034744
Total		57506		

Included in the table are:

1. The lag value *h*.
2. The *Sample Semivariance* $\hat{\gamma}(h)$.
3. The number of pairs of observations used to estimate the variance at that lag.
4. The *Model Semivariance* calculated from the fitted model, as described below.
5. The *Residual*, which equals the *Sample Semivariance* minus the *Model Semivariance*.

The rightmost column shows the predicted semivariance given from one of 6 models that may be fit to the observed variogram estimates. The models are:

Circular model: $$\gamma(h) = c_0 + \begin{cases} c\left\{1 - \frac{2}{\pi}\cos^{-1}\left(\frac{h}{a}\right) + \frac{2h}{\pi a}\sqrt{1-\frac{h^2}{a^2}}\right\} \\ c \end{cases} \text{ for } \begin{cases} h \le a \\ h > a \end{cases} \quad (3)$$

Exponential model: $$\gamma(h) = c_0 + c\left\{1 - \exp\left(-\frac{h}{r}\right)\right\}, \quad a = 3r \quad (4)$$

Gaussian model: $$\gamma(h) = c_0 + c\left(1 - \exp\left(-\frac{h^2}{r^2}\right)\right), \qquad a = \sqrt{3}r \quad (5)$$

Pentaspherical model: $$\gamma(h) = c_0 + \begin{cases} c\left\{\frac{15h}{8a} - \frac{5}{4}\left(\frac{h}{a}\right)^3 + \frac{3}{8}\left(\frac{h}{a}\right)^5\right\} \\ c \end{cases} \text{ for } \begin{cases} h \le a \\ h > a \end{cases} \quad (6)$$

Power function model: $$\gamma(h) = c_0 + wh^{\alpha} \text{ for } 0 < \alpha < 2 \quad (7)$$

Spherical model: $$\gamma(h) = c_0 + \begin{cases} c\left\{\frac{3h}{2a} - \frac{1}{2}\left(\frac{h}{a}\right)^3\right\} \\ c \end{cases} \text{ for } \begin{cases} h \le a \\ h > a \end{cases} \quad (8)$$

All of the models except the power function approach an asymptotic value of $c_0 + c$, called the *sill*, for large values of the lag h. c_0 is called the nugget variance and equals the variance at lag $h = 0$. The lag beyond which the function is effectively constant is called the *range*, which is represented by a.

The functions are estimated using nonlinear weighted least squares. The *Analysis Options* dialog box gives several choices for setting the weights:

1. *None:* the weights for each variogram estimate are set equal to 1.

2. *Counts:* the weights are set equal to *m(h)*, the number of pairs used to estimate the variance at lag *h*.

3. *Model:* Two iterations are performed. In the first iteration, the weights are all set equal to 1. In the second iteration, the weights are set equal to

$$w(h) = 1/\gamma^{*2}(h) \quad (9)$$

where $\gamma^*(h)$ is the estimated model variance at lag *h* obtained from the first iteration.

4. *Counts and model:* Two iterations are performed. In the first iteration, the weights are all set equal to *m(h)*. In the second iteration, the weights are set equal to

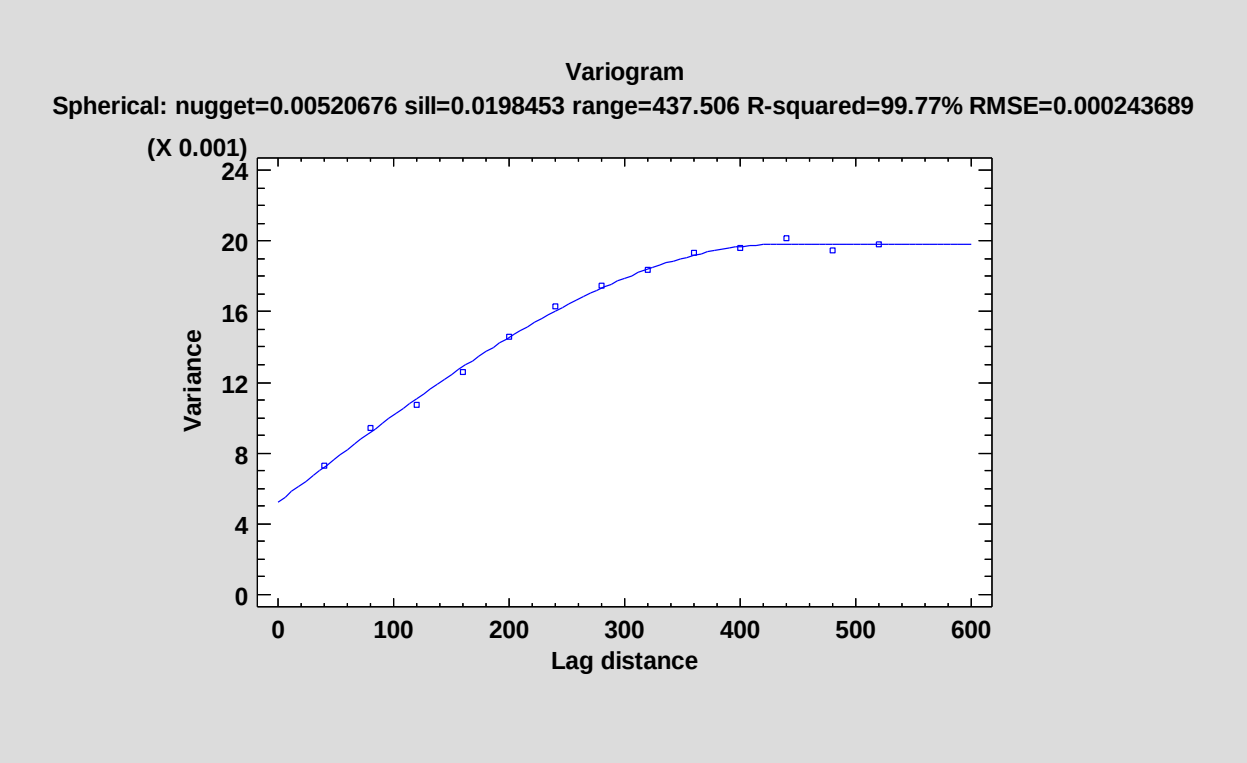
$$w(h) = m(h) / \gamma^{*2}(h) \tag{10}$$

5. *Counts, sample and model:* Two iterations are performed. In the first iteration, the weights are all set equal to *m(h)*. In the second iteration, the weights are set equal to

$$w(h) = m(h)\hat{\gamma}(h) / \gamma^{*3}(h) \tag{11}$$

Variogram Plot

This plot shows the estimated variogram together with the fitted model:



The model crosses the Y-axis at the nugget c_0 = 0.005207. It becomes horizontal beginning at the range a = 437.5, after which it equals the value of the sill c_0+c = 0.1985. As indicated by the large R-squared statistic, the fit is very good.

Kriging Table

Once a model has been built for the variogram, kriging may be performed at any target location represented by $\mathbf{x} = \{x_1, x_2\}$. The estimated response is given by

$$\hat{Y}(x) = \sum_{i=1}^{N} \lambda_i \, y(x_i) \tag{12}$$

which is a weighted average of the observed value of the random variable at the *N* nearest neighbors to **x**. The weights λ_i are made to sum to 1. The weights depend on the semivariances between the target point and its nearest neighbors, as described in detail by Webster and Oliver (2007). In addition, the variance of the estimate may be determined.

Observations are added to the estimate in equation (12) beginning in order of their proximity to the target location. The number of nearest neighbors *N* included in the estimate is determined by 3 settings on the *Analysis Options* dialog box:

- **Maximum radius:** an observation is only included in the estimate if the distance between the location of that observation and the target location does not exceed this value. This requirement is overridden if there are less observations within the specified radius than the *minimum points* setting below.

- **Minimum points:** the smallest number of points that must be included in the estimate.

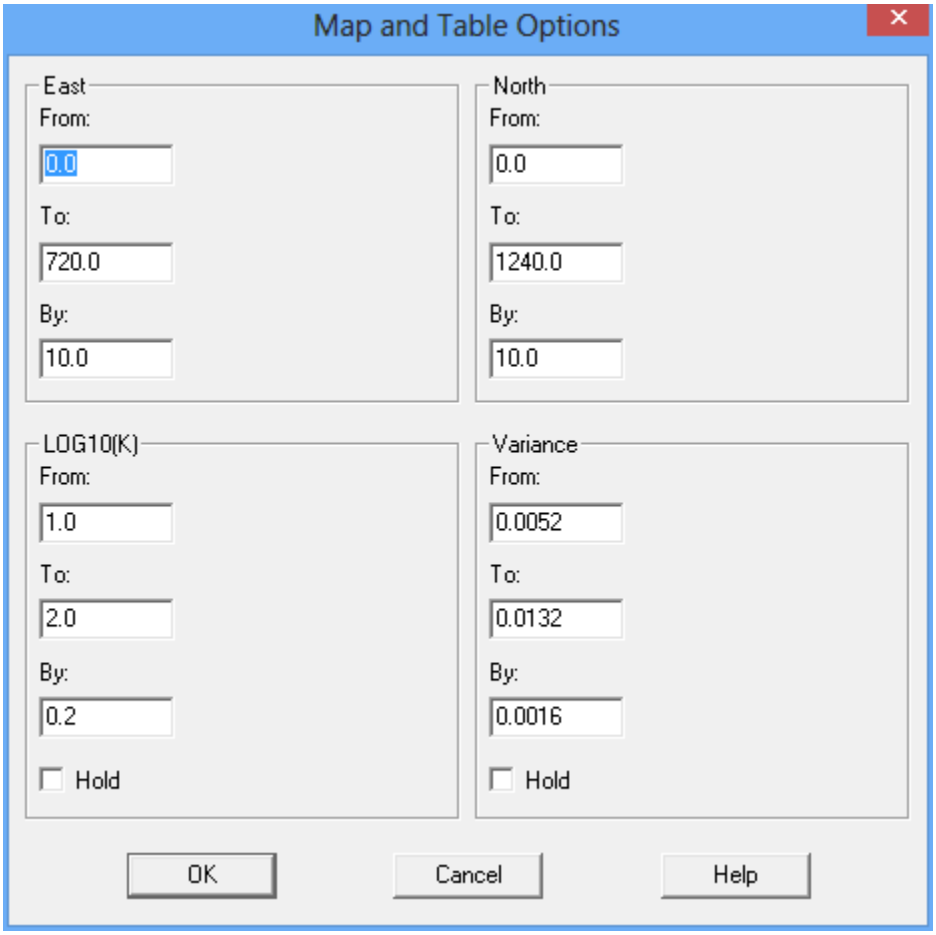
- **Maximum points:** the largest number of points that may be included in the estimate.

The estimated response and its variance are included in the *Kriging Table*, a small section of which is shown below:

Kriging Table for LOG10(K)

East	*North*	*LOG10(K)*	*Variance*
0.0	0.0	1.4608	0.0111319
0.0	10.0	1.45655	0.0109706
0.0	20.0	1.45223	0.0108416
0.0	30.0	1.44784	0.0107418
0.0	40.0	1.4434	0.0106671
0.0	50.0	1.43897	0.0106129
0.0	60.0	1.43459	0.0105746
0.0	70.0	1.43036	0.0105483
0.0	80.0	1.42638	0.0105305
0.0	90.0	1.42408	0.0105088
0.0	100.0	1.42058	0.0104999
0.0	110.0	1.4177	0.0104936
0.0	120.0	1.41556	0.0104887
0.0	130.0	1.41425	0.0104846
0.0	140.0	1.41387	0.0104808
0.0	150.0	1.39896	0.0105607
0.0	160.0	1.39948	0.0105605

The intervals at which the estimates are calculated are controlled by the settings on *Pane Options*:



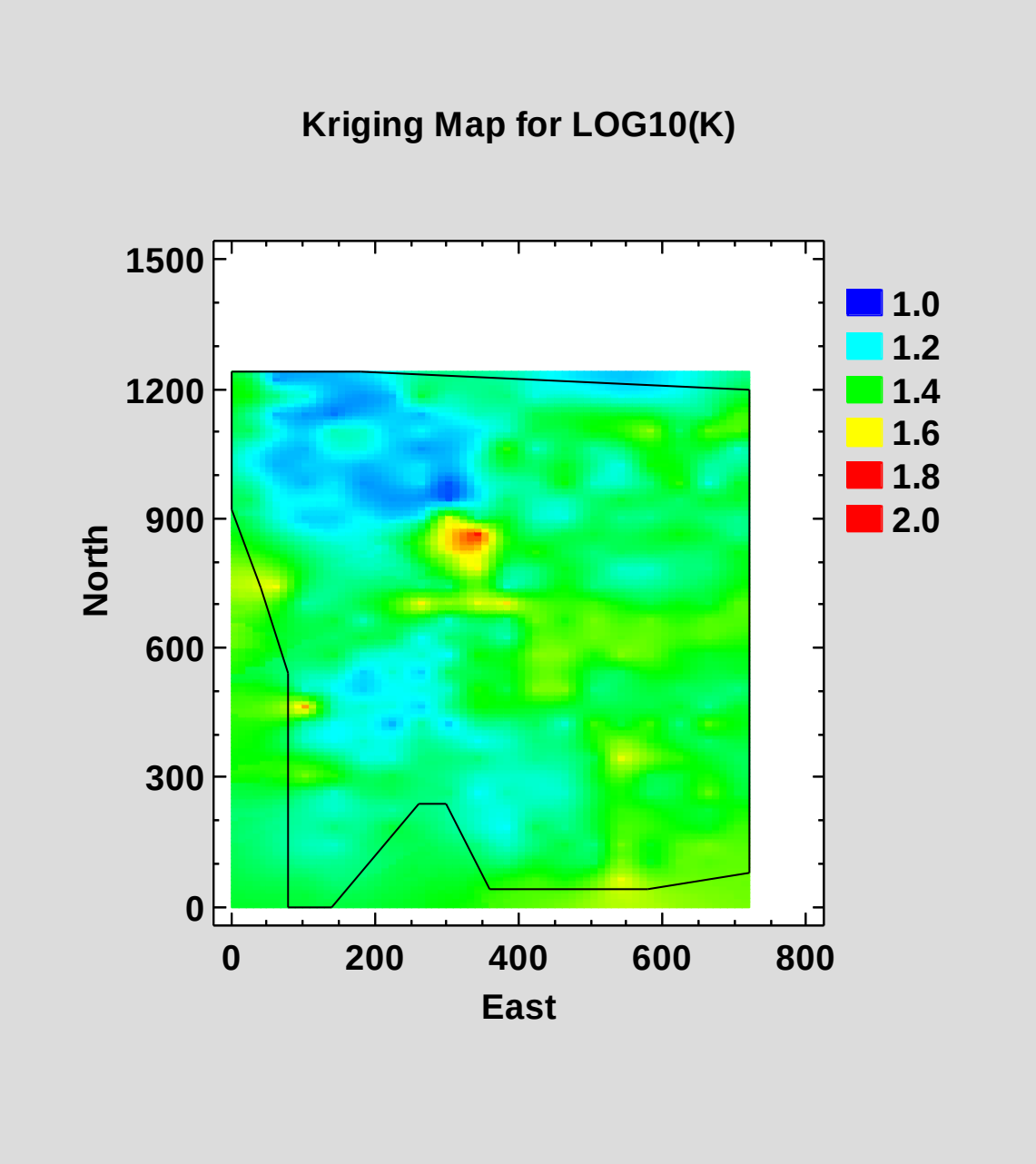
- **East:** the locations along the *East* axis at which estimates will be made.
- **West:** the locations along the *West* axis at which estimates will be made.

In the example, estimates were obtained over a 10m by 10m grid.

Note: the *LOG10(K)* and *Variance* settings apply to the kriging plots, not the table.

Kriging Map

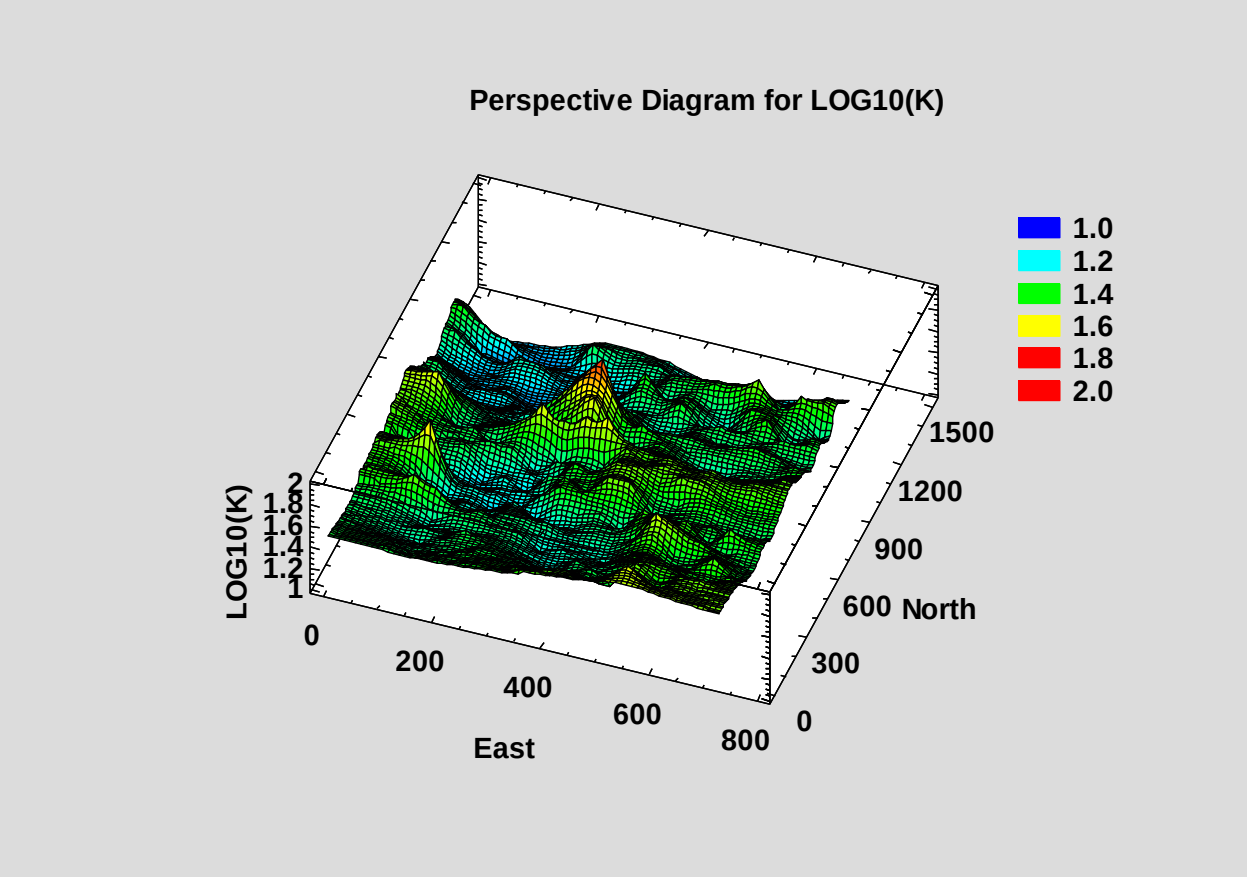
The estimates may be plotted on a two dimensional *Kriging Map*:



In the above plot, the estimated LOG10(K) is color coded from dark blue at 1.0 to dark red at 2.0. The settings under *LOG10(K)* on the *Pane Options* dialog box define the color range and increment. Checking the *Hold* box prevents these settings from changing if the Kriging parameters are changed.

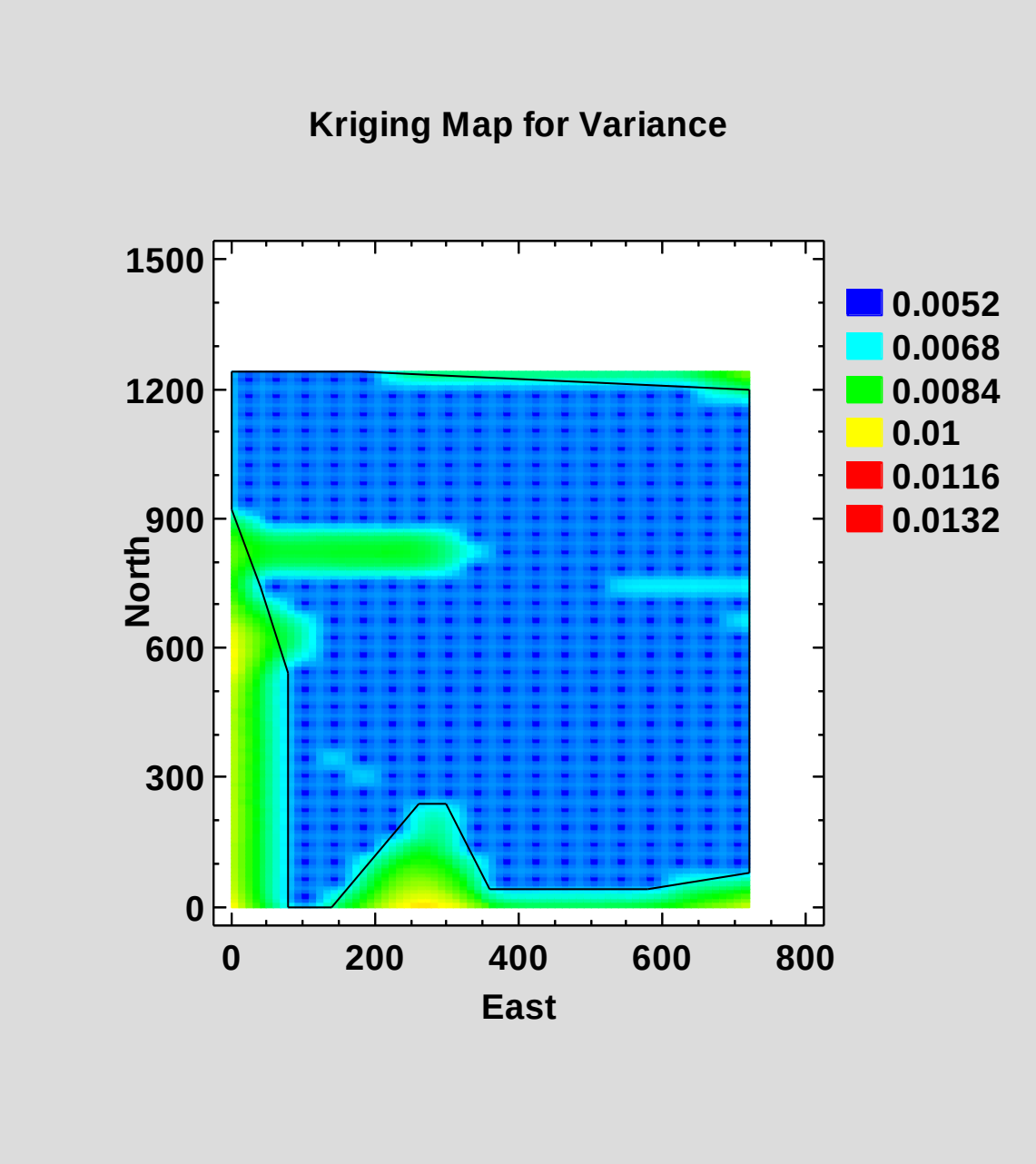
Perspective Diagram

The estimates may also be plotted on a three dimensional *Perspective Diagram:*



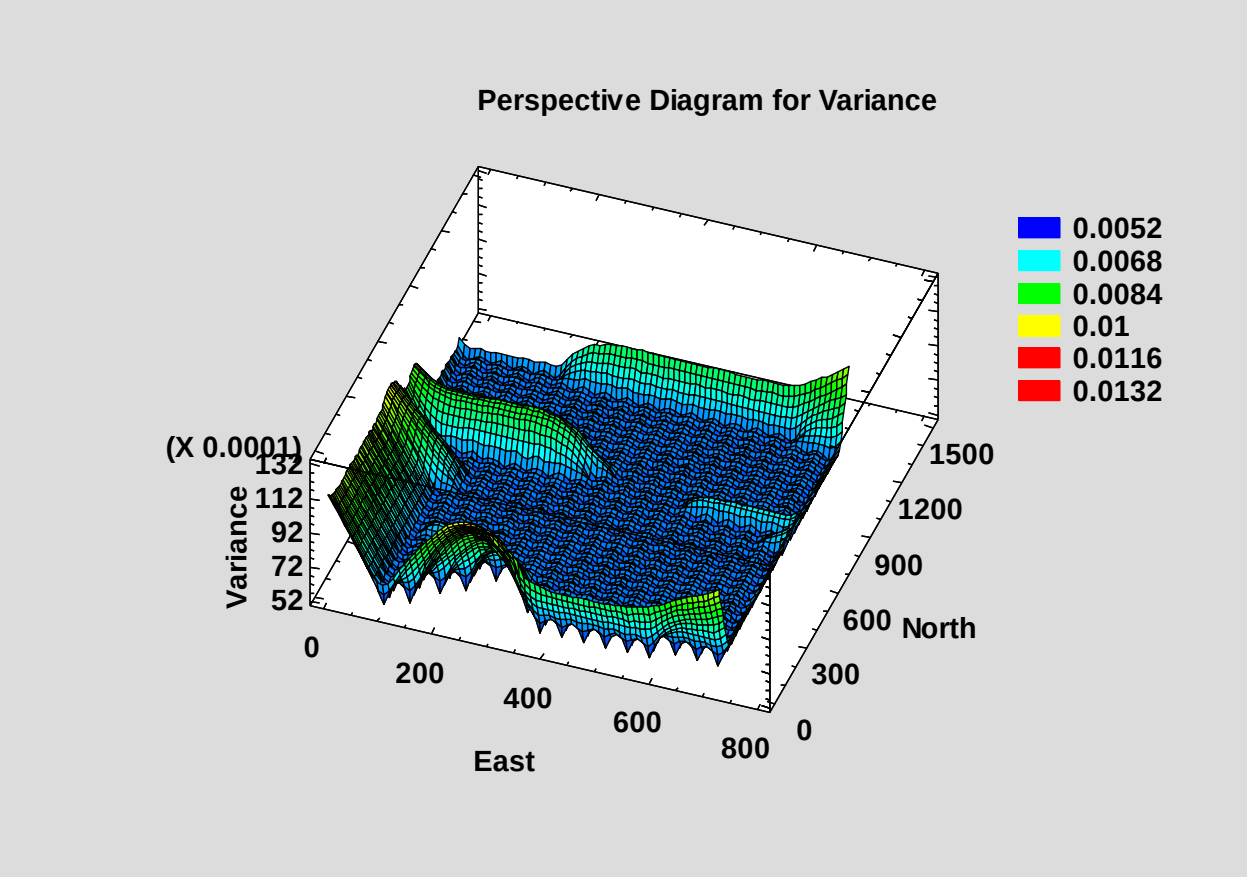
Variance Map

The variance of the estimated response may be plotted on a two dimensional *Map of Variance*:



Perspective Diagram of Variance

The variance of the estimates may be plotted on a three dimensional *Perspective Diagram:*



References

Webster and Oliver (2007). Geostatistics for Environmental Scientists. Wiley, New York.

