Calculating Variograms for 2-D Regularly Grided Data using XLISP-STAT

Stanley S. Bentow

June 14, 1995

1 Description

This program produces variograms using algorithm's from C.V. Deutsch and A.G. Journel, "GSLIB: Geostatistical Software Library and User's Guide," Oxford University Press, New York, 1992. Features are being added regularly, and at the present time the program has a plotting option that creates a "circle-plot" of the data as well as variogram plots. Currently in the works is an exponential kriging option. The program is still in it's early stages, and any suggestions are welcome.

2 Data and Terminology

The data must be entered in the same format as that required by GSLIB. Here, the data is assumed to come from a 2-D regularly spaced grid of the form:

$$\begin{array}{ccccc}
\mathbf{Z}_{(1,n_y)} & \cdots & \mathbf{Z}_{(x,n_y)} & \cdots & \mathbf{Z}_{(n_x,n_y)} \\
\vdots & & \vdots & & \vdots \\
\mathbf{Z}_{(1,y)} & \cdots & \mathbf{Z}_{(x,y)} & \cdots & \mathbf{Z}_{(n_x,y)} \\
 & & & & \\
\vdots & & \vdots & & \vdots \\
 & & & & \\
\mathbf{Z}_{(1,1)} & \cdots & \mathbf{Z}_{(x,1)} & \cdots & \mathbf{Z}_{(n_x,1)}
\end{array} \tag{1}$$

where $\mathbf{Z}_{(x,y)}$ is a vector of continuous random variables at location (x, y), $x = 1, \ldots, n_x$; $y = 1, \ldots, n_y$. Following the geostatistical framework, x and y will be considered the "easting" and "northing" directions respectively.

Variograms may be produced for any number of lags between variables and directions. Generally, a variogram's magnitude and direction is denoted by the vector **h** where, in the 2-D grided space, **h**=(x-lag,y-lag). For example, the resulting variogram for **h**=(0,2) consists of pairs of points at a distance $|\mathbf{h}| = 2$ units apart corresponding to a lag of two in the y direction. If **h**=(1,1), the resulting variogram considers pairs of points at a distance $|\mathbf{h}| = \sqrt{2}$ units apart corresponding to a lag of one in the x and y direction (north/east). The current version of the program only considers three directions: north, east, and north/east, with **h**=(0,1), **h**=(1,0), and **h**=(1,1) respectively for lags of one. Data points that don't have corresponding values at a considered lagged distance are discarded.

Since this program is designed for regularly spaced data, the locations (x, y) need not be entered explicitly. The location of the data points within the column represents their location on the grid. The program requires that variables be entered as columns with each column vector ordered as

$$\left(z_{(1,1)}, z_{(2,1)}, \ldots, z_{(n_x,1)}, \ldots, z_{(1,n_y)}, z_{(2,n_y)}, \ldots, z_{(n_x,n_y)}\right)'. \qquad (2)$$

As the program currently stands, missing values are not allowed.

3 Running the Program

Once a data file has been created, the program may be loaded into xlisp-stat by typing *(load "variogram")*, at which point a series of dialog boxes request various basic information. Once this has been entered, the user may choose from a number of descriptive and analytical options.

3.1 The Input Dialog Box

The input dialog boxes request the user for information about data format, and the type of variogram to calculate. Two methods for viewing the data are offered as well.

- Number of nodes in the x and y direction (*xnode*, *ynode*).
- View Data: Prints the data matrix.
- Circle Plot: Draws a circle plot (e.g. a method for describing spatial variabiltiy in the data. See Appendix B for a full explanation).
- Number of lags to consider (*nlag*): The lags for which variograms are calculated. For example, if *nlag* is set to 10, the program calculates separate variogram values for points that are $1, 2, 3, \ldots, 10$ lags apart.
- Head and Tail variable (*head*, *tail*): This refers to the variables used in calculating the variogram. For example, one may use a *tail* value from one variable and a *head* value from another that are $1, \ldots,$ *nlag* apart.

- x and y direction (*xdir*, *ydir*): This refers to the direction in which the variogram is calculated.

- Type of Procedure (*ptype*): This refers to the type of variogram the user wishes to calculate (See Appendix A).

3.2 The Results Dialog Box

The results dialog box provides a method for viewing the resulting variogram as well as other relevant statistics:

- Results: The vector of variogram values requested by *ptype*.

- Head Means: The vector of *head* means $1, \ldots, nlag$ apart.

- Tail Means: The vector of *tail* means $1, \ldots, nlag$ apart.

- Variogram: A plot of the variogram values by *nlag*.

4 Example: Simulated Toxic Waste Data

This example uses the dataset *true.dat* from problem set one of Deutch and Journal[1], pg.35. These data are created by the proceess of simulated annealing of a variogram representing the distribution of a toxic substance over a 2500 square mile grid. Variograms are calculated for 20 lags. The head and tail values are taken from variable 0 (the first column variable) and the direction is set to east. Figures 1 and 2 are the resulting variograms for the four procedures.

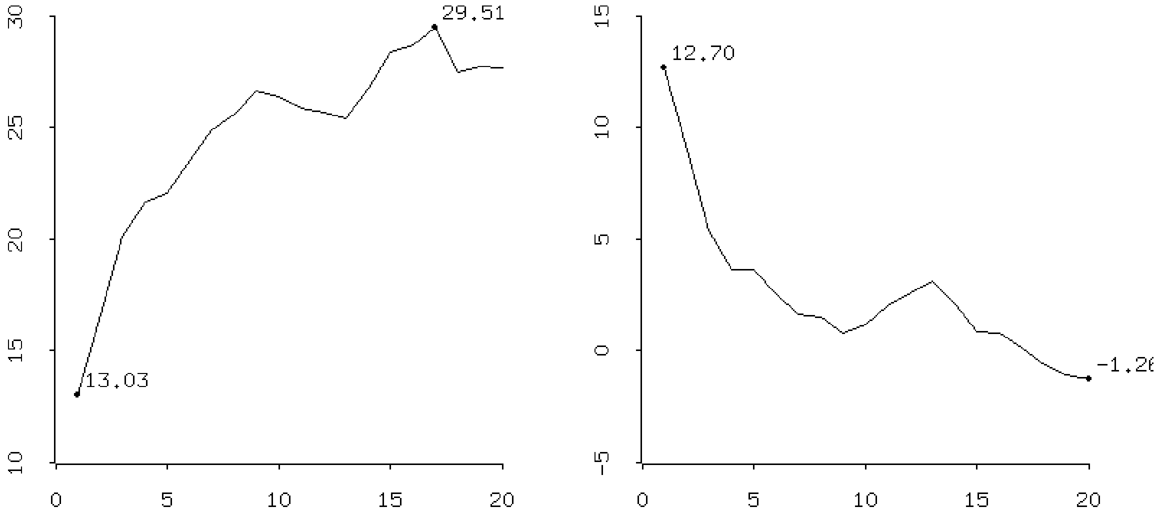

Figure 1: Semivariogram; Covariogram

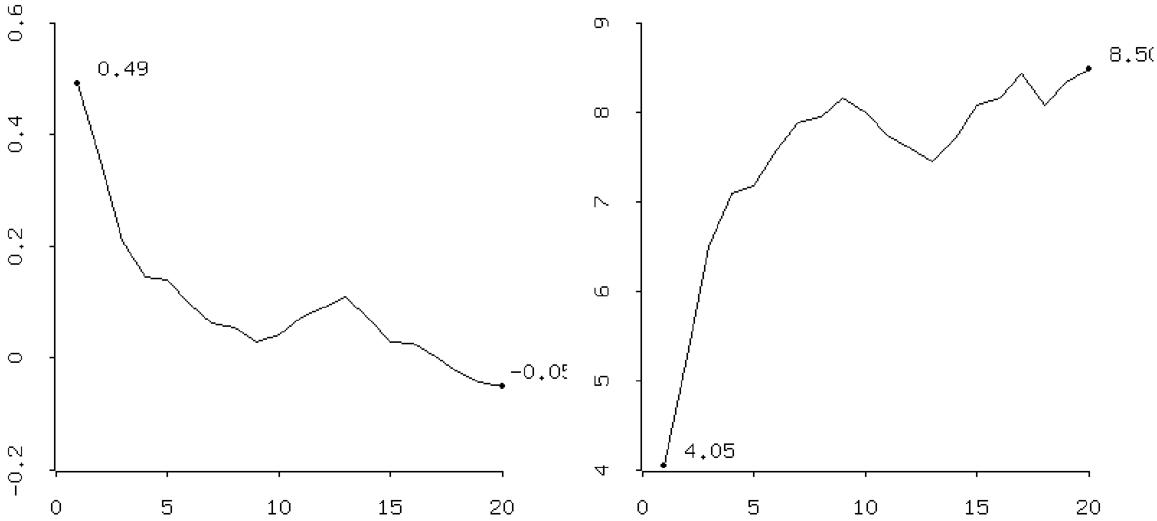

Figure 2: Correlogram; General Relative Semivariogram

Appendix A

A Variograms

A variogram is a method of describing the spatial variability between points a specified distance apart. There are many different procedures for calculating this quantity. The program offers an option for computing four of the most widely used.

A.1 Semivariogram

This measure is defined as half the average squared distance between two points a specified distance apart:

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} (x_i - y_i)^2 \tag{3}$$

where $N(\mathbf{h})$ is the number of pairs, x_i is the tail value, and y_i is the head value. It is generally recommended that tail and head values be of the same variable.

A.2 Covariogram

This measure is the traditional covariance commonly used in statistics:

$$C(\mathbf{h}) = \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} x_i y_i - m_{-\mathbf{h}} m_{+\mathbf{h}} \tag{4}$$

where $m_{-\mathbf{h}}$ is the mean of the tail values,

$$m_{-\mathbf{h}} = \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} x_i \tag{5}$$

and $m_{+\mathbf{h}}$ is the mean of the head values,

$$m_{+\mathbf{h}} = \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} y_i \tag{6}$$

If x and y are two different variables, the expression identifies the sample cross covariance.

A.3 Correlogram

This measure is the traditional correlation commonly used in statistics:

$$\rho(\mathbf{h}) = \frac{C(\mathbf{h})}{\sigma_{-\mathbf{h}}\sigma_{+\mathbf{h}}} \tag{7}$$

where $\sigma_{-\mathbf{h}}$ is the standard deviation of the tail values,

$$\sigma_{-\mathbf{h}} = \left[\frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} x_i^2 - m_{-\mathbf{h}}^2 \right]^{\frac{1}{2}} \tag{8}$$

and $\sigma_{+\mathbf{h}}$ is the standard deviation of the head values,

$$\sigma_{+\mathbf{h}} = \left[\frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} y_i^2 - m_{+\mathbf{h}}^2 \right]^{\frac{1}{2}} \tag{9}$$

A.4 General Relative Semivariogram

Here, the Semivariogram (3) is standardized by the squared mean of the data:

$$\gamma_{GR}(\mathbf{h}) = \frac{\gamma(\mathbf{h})}{(\frac{m_{-\mathbf{h}}+m_{+\mathbf{h}}}{2})^2} \tag{10}$$

Appendix B

B Circle Plot

A circle plot is an interactive and dynamic method for describing the variability of spatial data. These features make it particularly suitable for the XLISP-STAT environment since it takes advantage of XLISP-STAT's excellent dynamic graphing capabilities.

The circle plot can be thought of as a cross between a greyscale map and an indicator map ([3], pg.45). Although both maps attempt to describe spatial continuity while preserving the relative magnitude of the data, they have their advantages and disadvantages.

The greyscale map describes spatial variability by representing each data point as a shade of grey, with darker shades indicating larger values. Although this type of map is pleasing to the eye, it suffers from limitations associated with the variability in the data. For example, if the data values are not evenly distributed throughout the range, it may be difficult to evaluate the magnitude of the outlying values through grey shading alone.

On the other hand, an indicator map can be thought of as a greyscale map with only two colors; black and white. The data is split into two classes, and a series of indicator maps are presented that correspond with progressively increasing cutoff values. Here the series itself is the focus of the analysis, with each successive map describing a spatial pattern, as well as direction. This procedure is excellent at revealing outliers and spatial continuity, but masks much of the subtle variability that techniques like the greyscale map reveal.

The circle plot is an attempt to combine the pleasant appearance and detail of the greyscale map with an ability to reveal outliers and spatial continuity analogous to that of a series of indicator maps. The plotting algorithm proceeds as follow:

1. Sort the $z_{(x,y)}$ data values in ascending order. Let $n = n_x n_y$ and let $z_{(i)}$, $i = 1, \ldots, n$ denote the i^{th} order statistic. Set the largest value $z_{(n)}$ to *maxval*.

2. Generate a vector of proportional scalings $y_{(i)}$, $(0 < y_{(i)} < 1)$, where

$$y_{(i)} = \frac{z_{(i)}}{maxval}. \tag{11}$$

3. If $y_{(1)} < tol$ (where tol is a user specified tolerance level)

   then

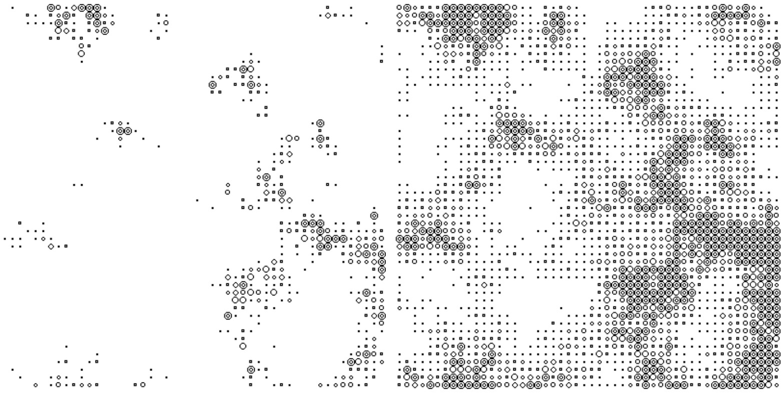

Figure 3: Tolerance=.05(default); Tolerance=.2

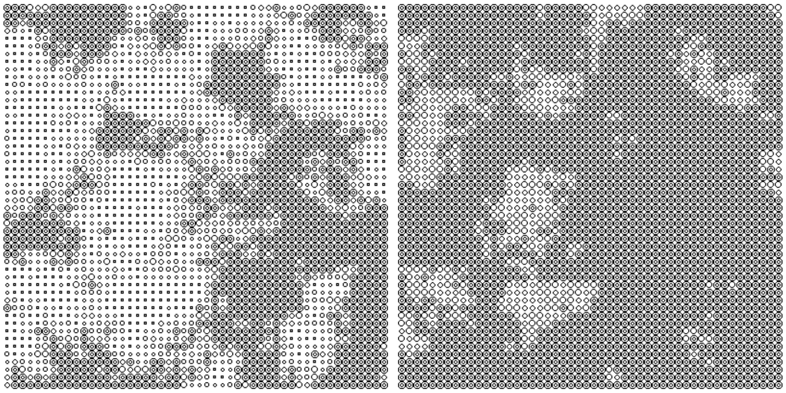

Figure 4: Tolerance=.4; Tolerance=.8

```
  fix $y_{(n)} = 1$ and set $maxval = z_{(n-1)}$. Go to step 2.
else
  Stop.
```

If $y_{(i)} = 1$, then plot it's location on the grid as a circle within a circle. Else if $y_{(i)} < 1$ then plot it's location on the grid as a circle whose area within the grid is $y_{(i)}$.

The larger the tolerance, the greater the emphasis on the variation in smaller values of $z_{(x,y)}$ and vise versa. The initial tolerance level is set at .05.

The series of circle plots in figures 3 and 4 are derived from the 2500 data values in the dataset *true.dat*[1] As expected, variability in smaller values of $z_{(x,y)}$ is more apparent for higher tolerance levels. Note: These plots are rotated 90 degrees clockwise from their original orientation.

References

[1] Deutch, CV., Journel, AG. (1992) *GSLIB: Geostatistical Software Library and User's Guide*. Oxford University Press.

[2] Englund, E., Sparks, A. (1991) *GEO-EAS 1.2.1 User's Guide*. Environmental Monitoring Systems Laboratory, EPA.

[3] Isaaks, EH., Srivastava, RM. (1989) *An Introduction to Applied Geostatistics*. Oxford University Press.