



**The Abdus Salam
International Centre for Theoretical Physics**



1867-60

College of Soil Physics

22 October - 9 November, 2007

**AN EXAMPLE OF APPLICATION OF THE GEOSTATISTICS APPROACH
USING THE GS+ SOFTWARE**

Luis Carlos Timm
*ICTP Regular Associate, University of Pelotas
Brazil*

AN EXAMPLE OF APPLICATION OF THE GEOSTATISTICS

APPROACH USING THE GS+ SOFTWARE

TIMM, L.C.^{1*}; RECKZIEGEL, N.L.²; AQUINO, L.S.²; REICHARDT, K.³;

TAVARES, V.E.Q.²; BAMBERG, A.L.⁴

¹Regular Associate of the ICTP, Rural Engineering Department, Faculty of Agronomy, Federal University of Pelotas, CP 354, 96001-970, Capão do Leão-RS, Brazil.

*lctimm@ufpel.edu.br; lcartimm@yahoo.com.br.

²Rural Engineering Department, Faculty of Agronomy, Federal University of Pelotas, CP 354, 96001-970, Capão do Leão-RS, Brazil.

³Soil Physics Laboratory, Center for Nuclear Energy in Agriculture, University of São Paulo, CP 96, 13400-970, Piracicaba-SP, Brazil.

⁴PhD Student, Soil Science Department, Faculty of Agronomy, Federal University of Pelotas, CP 354, 96001-970, Capão do Leão-RS, Brazil.

Introduction

This text has the objective to illustrate a complete example of the Geostatistics approach from the GS+ software (Geostatistics for the Environmental Sciences software) developed by Gamma Design Software, LLC. It is based on GS+ User's Guide Version 7.0 (Gamma Design Software, 2004) and Guimarães (2004).

A soil water content data set (% basis of weight), extracted from Guimarães (2004) is used as an example of spatial data analysis from the GS+ software. Soil water content observations were measured 20 m apart from each other in an experimental field grid (9 x 7 points), totalizing 63 observations.

The GS+ implementation starts bringing the data into the GS+ Data Worksheet (Figure 1). Observed values can be entered directly into the

worksheet or can be imported from a spreadsheet, for example, from an Excel worksheet. In this case, the easiest way to import data is to cut-and-past or to copy-and-past from the source spreadsheet (Figure 1). The sample locations can be organized in a cartesian (X,Y) coordinate system when they were sampled on a grid (two dimensions). They can also be measured along a spatial line which is called transect (one dimension).

The screenshot shows the GS+ Geostatistics for the Environmental Sciences software window. The title bar reads "GS+ Geostatistics for the Environmental Sciences (tabela 4 apostila.par)". The menu bar includes "File", "Edit", "Data", "Autocorrelation", "Interpolate", "Map", "Window", and "Help". The toolbar contains various icons for file operations, data manipulation, and visualization. The main window is divided into sections: "Base Input File" (empty), "Data Title / Description" (empty), and "Data Records". The "Data Records" section contains a table with 11 rows and 11 columns. The columns are labeled 1 through 10, with the first column being an index. The first three columns are labeled "X Coord", "Y Coord", and "Z". The data values are as follows:

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|---------|---------|-------|---|---|---|---|---|---|----|
| | X Coord | Y Coord | Z | | | | | | | |
| 1 | 20.00 | 20.00 | 28.16 | | | | | | | |
| 2 | 40.00 | 20.00 | 27.16 | | | | | | | |
| 3 | 60.00 | 20.00 | 26.09 | | | | | | | |
| 4 | 80.00 | 20.00 | 27.27 | | | | | | | |
| 5 | 100.00 | 20.00 | 27.61 | | | | | | | |
| 6 | 120.00 | 20.00 | 26.61 | | | | | | | |
| 7 | 140.00 | 20.00 | 26.13 | | | | | | | |
| 8 | 160.00 | 20.00 | 29.73 | | | | | | | |
| 9 | 180.00 | 20.00 | 31.12 | | | | | | | |
| 10 | 20.00 | 40.00 | 27.52 | | | | | | | |
| 11 | 40.00 | 40.00 | 26.54 | | | | | | | |

Figure 1: GS+ data worksheet window illustrating an input data file. (X Coord,YCoord) columns are the cartesian coordinates of the sample locations; Z column contains the observed values of the soil water content.

First of all, it was performed an exploratory data analysis in which was calculated data position and dispersion measures as well as the data probability

distribution analysis. Figure 2 shows the icons that represent short-cuts of these descriptive statistics.

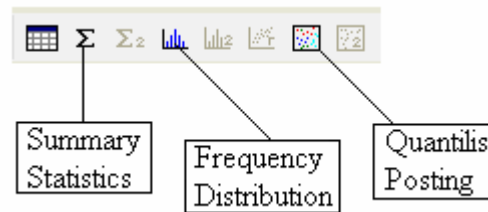


Figure 2: GS+ short-cuts to calculate descriptive statistics.

The no-available icons are used for cross-variogram analysis which is not presented here. Figure 3 provides the calculated descriptive statistics for the soil water content data set defined in the data worksheet window (Figure 1): the mean value (mean), the standard deviation (std deviation), the sample variance, the minimum and maximum values, the number of observations (n) [and between brackets (n missing or excluded)], the frequency distribution, the skewness coefficient and its standard error (se) and the kurtosis coefficient and its standard error (se). For the Z variable it is also possible to make a transformation of the data set in order to better normalize the variable distribution previously to perform the Geostatistics analysis, such as: Scale to 0-1, Log Transform and Square-Root Transform. It is also possible to backtransform the values to the original measurement domain, however the output is customarily (but not necessarily) backtransformed to the original data domain (Gamma Design Software, 2004).

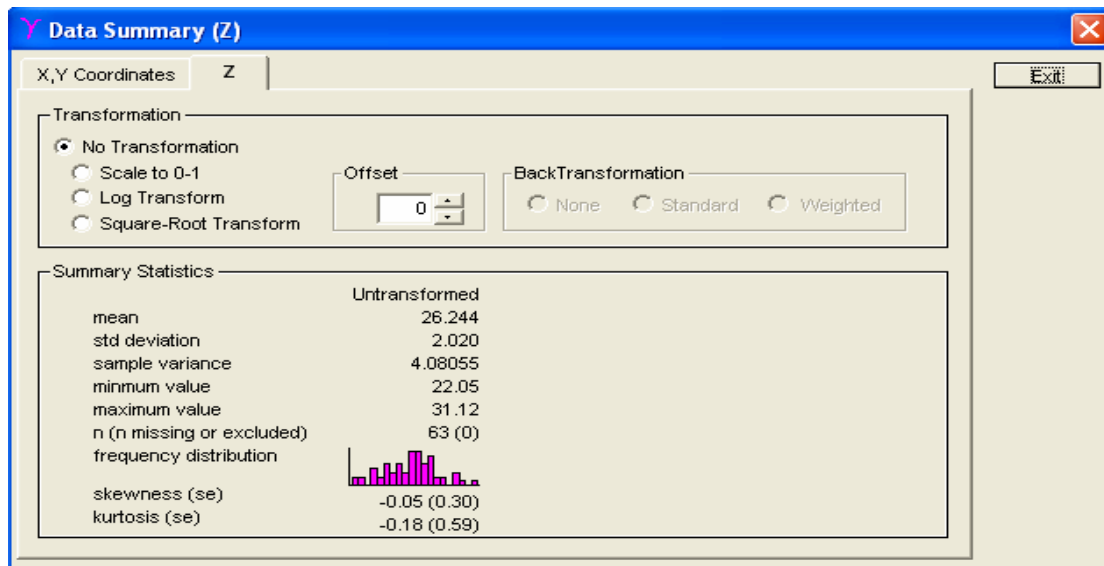


Figure 3: GS+ Data Summary window illustrating the calculated descriptive statistics of the soil water content data set (Z column).

From the Figure 3 the user can also access a full-window frequency distribution which allows it to analyze in more details the data frequency distribution (histogram graph), the Cumulative Frequency and the Normal Probability graphs (Figure 4).

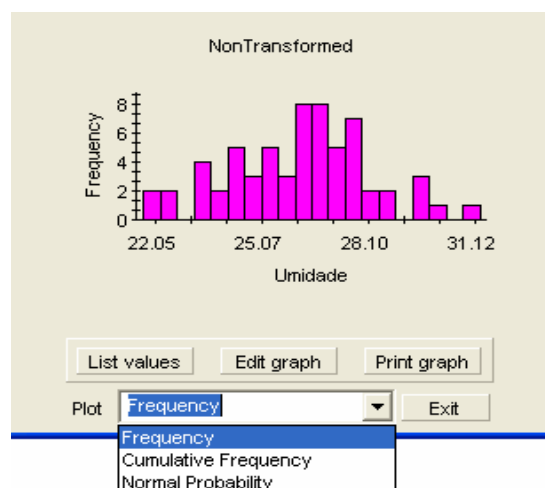



Figure 4: GS+ window to analyze in more details the spatial data distribution in the experimental field.

The data spatial distribution is shown from a GS+ coordinate posting window (Figure 5). This graph is obtained when the  icon is selected in Figure 2. On the right side of this window there are many options for presenting the data spatial.

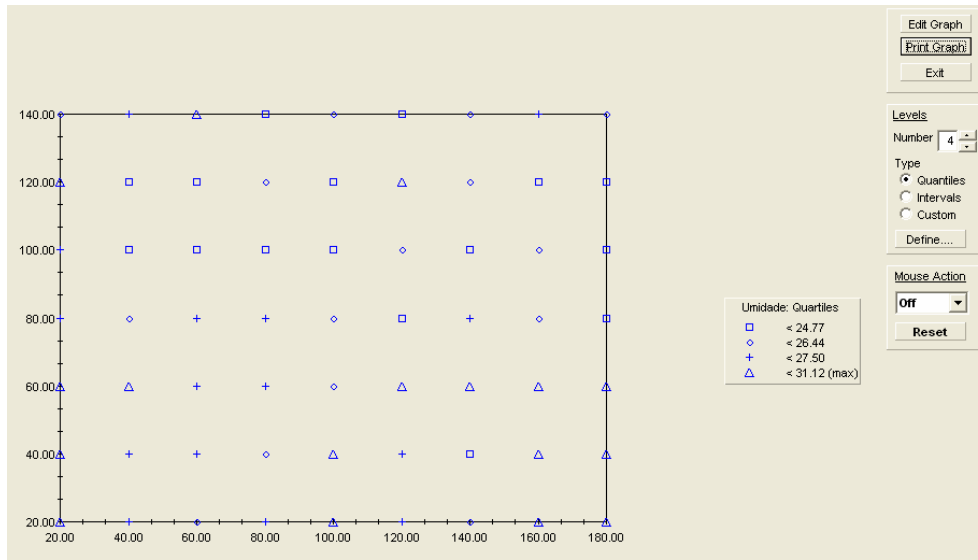



Figure 5: GS+ coordinate posting window which is shown the spatial distribution of the soil water content data set in the experimental area.

The next step will be to study the spatial dependence structure of the soil water content data set applying Geostatistical tools which are available in the GS+ software. In this way, it is possible to calculate the experimental semivariance values for pairs of points separated by a lag class distance interval, to construct the experimental semivariogram graph and to adjust the best mathematical model (which is here called theoretical semivariogram graph) to experimental semivariogram. The GS+ software runs this procedure when the  icon is selected (Figure 6).

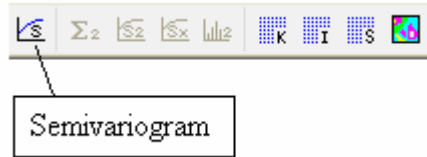



Figure 6: GS+ short-cuts for studying the structure of a data spatial dependence of a selected variable.

Clicking on the  icon, a window with entry parameters which should be given by the user in order to GS+ routine to proceed the calculation of the structure of the data spatial dependence of the selected variable, is available (Figure 7).

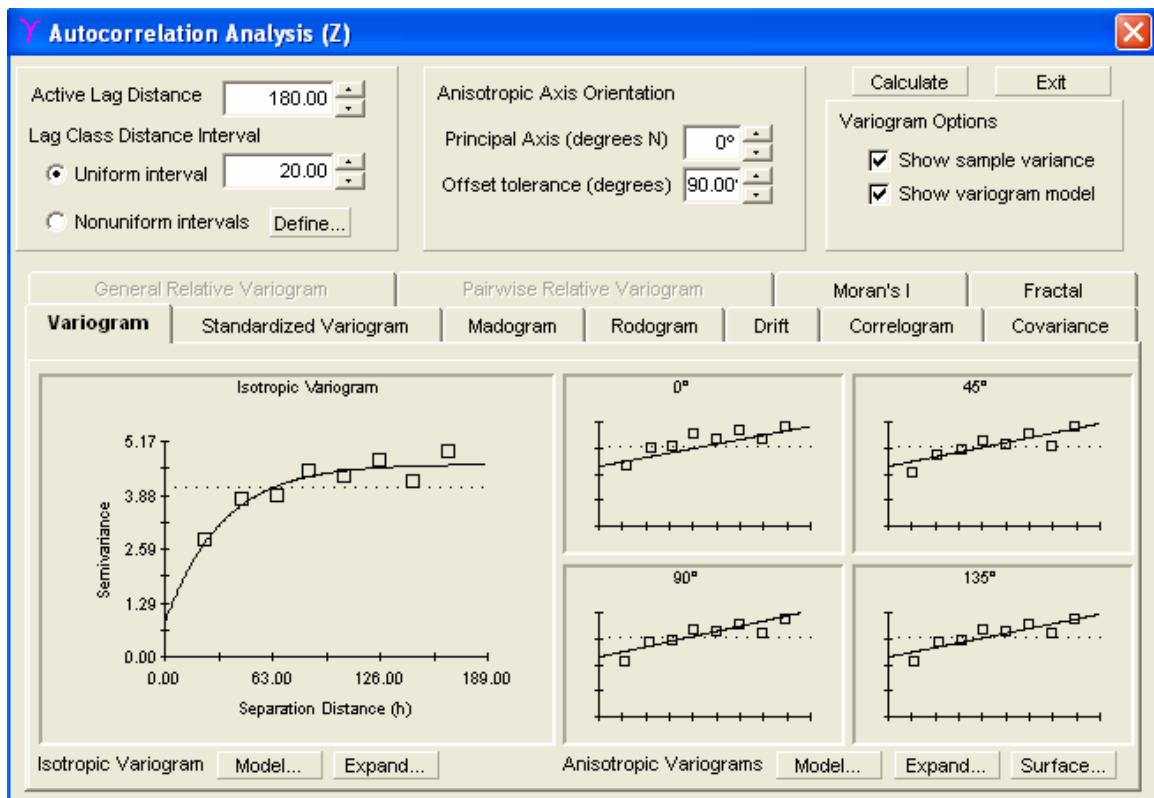


Figure 7: GS+ window with entry parameters which should be given by the user so that the software routine runs the calculation of the structure of the data spatial dependence of the selected variable.

In Figure 7, the Active Lag Distance and Lag Class Distance Interval entry parameters should be given by the user. The GS+ software allows studying the spatial dependence structure of a selected variable in which it was sampled in Uniform or in Nonuniform intervals. In the last case, the intervals are defined by the user clicking on the Define option (Figure 7).

The Active Lag Distance value which is the distance for calculating the experimental semivariance values is a function of the sampling maximum distance in relation to the point coordinates taken as a reference. The Lag Class Distance Interval value defines how the calculated experimental semivariance values are grouped into lag classes. The bigger this value, the smaller is the number of points for constructing the experimental semivariogram graph. However, if the distance between the lag classes is very small there will be lag class intervals without pairs for calculating experimental semivariance values.

Isotropic and anisotropic analyses can be performed by the GS+ software. For the isotropic analysis, the Offset tolerance angle (degrees) (Figure 7) should be 90° and, in this case, all experimental semivariogram (0° , 45° , 90° and 135°) will be the same as the isotropic experimental semivariogram.

The GS+ window shown in Figure 7, has also other Variogram options, like: Show sample variance which is the data sample variance in which is represented by a straight dashed line parallel to the X axis [separation distance (h)]. Selecting the Show variogram model option, the adjusted mathematical model is shown in the same experimental semivariogram graph.

After inputting entry parameters, it is possible to check how the adjusted model was fitted to the experimental semivariogram. This GS+ procedure is based on the lowest Residual Sum of Squares (Residual SS).

Clicking on the Model... option (Figure 7), the GS+ shows the parameters of the adjusted mathematical model indicating the Variogram model type giving the Nugget variance (C_0), the Structural Variance Sill (C_0+C) and the Range (A) (Figure 8). However, the user has the possibility of checking the performance of other adjusted models to the experimental semivariogram. For this, clicking on the Variogram model type option and choose, among the other available models (linear model, Spherical model, Exponential model and Gaussian model) the more appropriate one. At the bottom of Figure 8 are the following statistics measures which indicate the performance of the GS+ automatically calculated model (AutoFit) and the user's selected performance model (This Fit): Residual SS, r^2 coefficient and the Proportion $[C/(C_0+C)]$ which is the dependence degree (DD) of the variable: DD is considered weak when $DD < 25\%$; moderate when $25\% < DD < 75\%$; and strong when $DD > 75\%$. The Print, Cancel and Ok options are also presented in Figure 8.

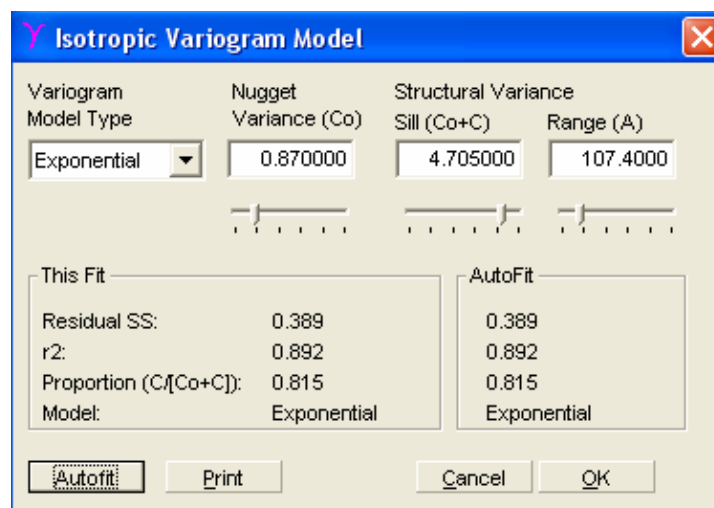


Figure 8: Window showing the automatically calculated adjusted model by the GS+ software with its parameters as well as the user's selected model and its

parameters. It also shows the statistics measures which are used for selecting a theoretical semivariogram model.

Clicking the Expand... option (Figure 7), the GS+ software shows the experimental isotropic and theoretical isotropic semivariograms with its adjusted parameters. This option also gives the calculated statistics measures from the adjustment process (Figure 9). There is also the possibility of verifying the behavior of the experimental and theoretical semivariograms in the 0° (+ 0 Degrees), 45° (+ 45 Degrees), 90° (+ 90 Degrees) and 135° (+ 135 Degrees) directions. Other options in this GS+ window are: List, Edit, Print, Cloud, Scatter and Exit.

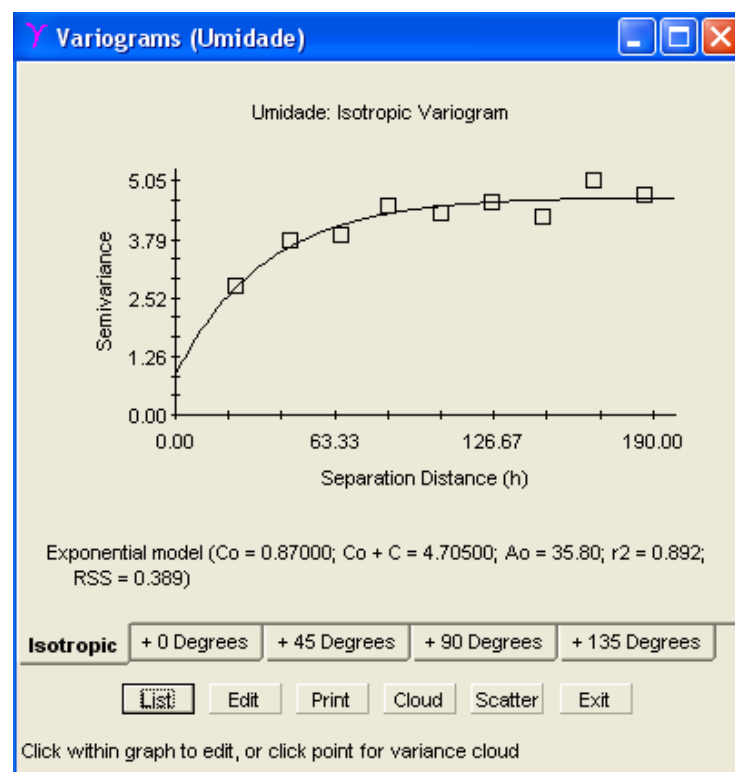


Figure 9: GS+ window showing the experimental semivariogram and the adjusted mathematical model (theoretical semivariogram) with C_0 , C_0+C and A_0 parameters and the calculated statistics measures (r^2 and RSS measures) from the adjustment process.

An important aspect from the adjustment process is that when the selected mathematical model is exponential or Gaussian, the A_0 parameter (Figure 9) should be times by 3 and by $\sqrt{3}$, respectively, to obtain the value of A range [Range (A), Figure 8), which will be the effective range in which the spatial dependence of the variable is apparent (Gamma Design Software, 2004).

Clicking on the List option (Figure 9), the semivariance values are listed for each lag class as a function of the average separation distance for pairs of points in that class, the average semivariance for those points and the number of pairs of points upon which the average distance and semivariance are based (Figure 10).

| Lag Class | Average Distance | Average Semivariance | Pairs |
|-----------|------------------|----------------------|-------|
| 1 | | | 0 |
| 2 | | | 0 |
| 3 | 23.86 | 2.7736 | 206 |
| 4 | | | 0 |
| 5 | 43.00 | 3.6402 | 258 |
| 6 | 56.57 | 4.2453 | 70 |
| 7 | 62.06 | 3.6304 | 214 |
| 8 | 72.11 | 4.3416 | 116 |
| 9 | 84.41 | 4.5041 | 310 |
| 10 | | | 0 |
| 11 | 102.53 | 4.1520 | 270 |
| 12 | 115.40 | 4.9848 | 86 |
| 13 | 124.27 | 4.3712 | 170 |
| 14 | 134.16 | 5.5675 | 36 |
| 15 | 142.81 | 4.0777 | 102 |
| 16 | 154.48 | 4.8446 | 36 |
| 17 | 162.92 | 4.6457 | 47 |
| 18 | 173.49 | 5.9037 | 22 |
| 19 | 186.54 | 4.7383 | 8 |
| 20 | | | 0 |

Figure 10: GS+ window showing the calculated average semivariance values as a function of the average separation distance for each lag class.

Clicking on the Cloud option (Figure 9), it is shown a variance cloud graph as illustrated in Figure 11. For this, the GS+ software brings up all

possible pairs of the variance values which are separated by distances lower than the selected lag class separation distance. For example, if the separation distance between the lag class intervals is 20 m, GS+ shows for the Lag Class 1 all calculated variance values from pairs of points separated by distances lower than 20 m. As the sampled soil water content data were collected 20 m apart from each other in the experimental grid, Lag Class 1 does not have any calculated variance value. For the Lag Class 2 all possible pairs of the variance values are calculated for points separated by distances lower than 40 m and higher than or equal to 20 m, and successively for the other Lag Classes.

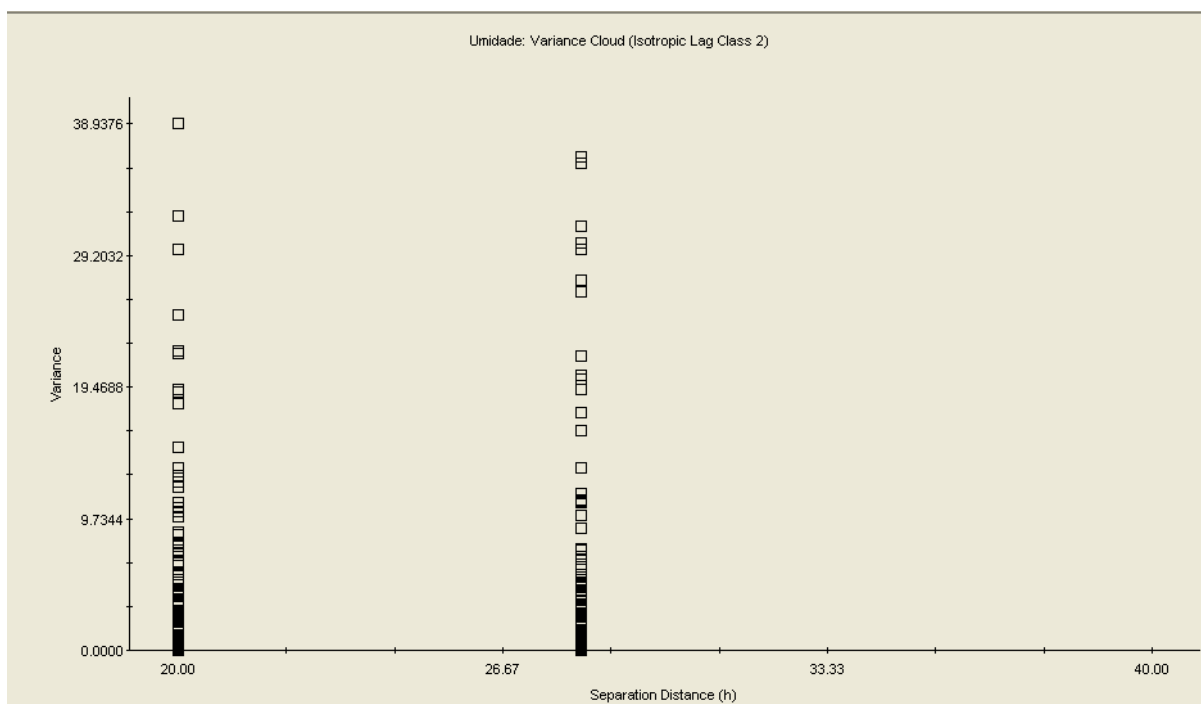


Figure 11: GS+ window showing the variance cloud graph of the soil water content data set for the Lag Class 2.

In the GS+ worksheet shown in Figure 10, the average separation distance for pairs of points for each lag class is calculated as follows: if in a grid

we have 12 pairs of points separated by a distance of 10 m and 10 pairs of points separated by a distance of 14.14 m, then the average separation distance is calculated as:


$$h = \frac{(10 \times 12) + (14.14 \times 10)}{12 + 10} = 30.02$$

i.e. the pair average distance of 30.02 m and calculated average semivariance is plotted in the experimental semivariogram (Figure 10).

From the theoretical semivariogram, the kriging interpolation method can be applied using the GS+ software. Interpolation is the estimation of values in an experimental area for points not actually sampled of the studied variable.

GS+ provides three types of interpolation: kriging, conditional simulation and inverse distance weighting (IDW) methods. Here only the kriging basic aspects are presented.

In the kriging interpolation method the estimates are based on values at neighboring locations plus knowledge about the underlying spatial relationship in a data set. Semivariograms provide knowledge about the spatial relationships. The estimated value at a given location is a weighted moving average of best estimates calculated to minimize local area variance.

The GS+ kriging interpolation method can be accessed from the  shortcut (Figure 1). After selecting this option, GS+ provides a window (Figure 12) with entry parameters which should be given by the user in order to GS+ routine to proceed the interpolation for estimating values for points not sampled. The theoretical semivariogram must be calculated before applying kriging method.

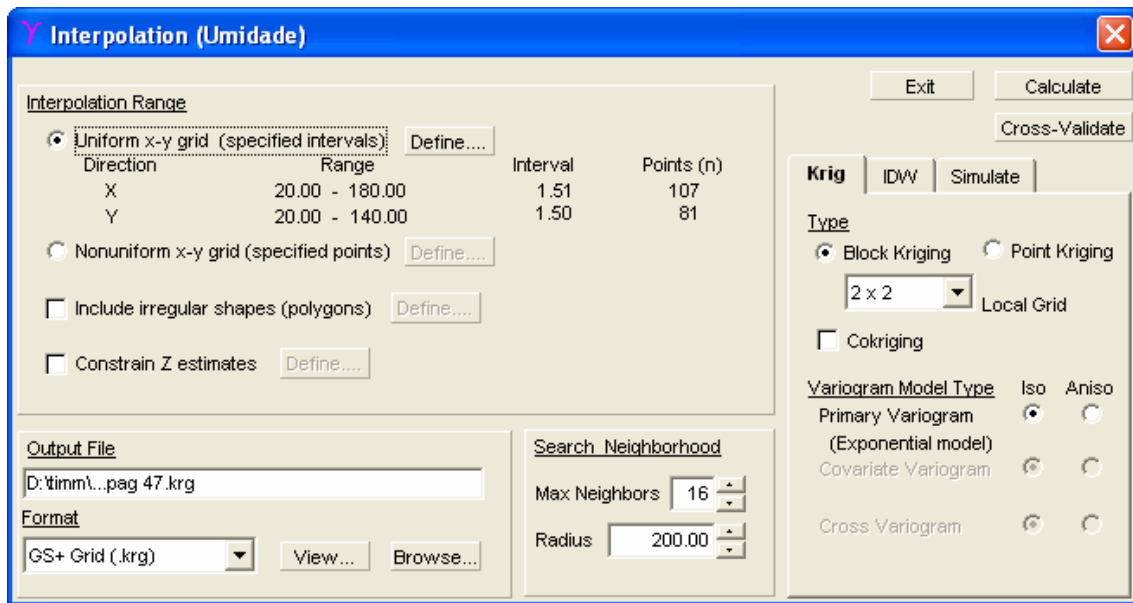


Figure 12: GS+ window with entry parameters which should be given by the user before applying kriging interpolation method.

GS+ software interpolates values for a specific location using nearest neighbor values (Search Neighborhood) weighted by distance and the degree of autocorrelation present for that distance (as defined by the semivariogram model). Searches are limited to a certain number of nearest neighbors (Max Neighbors option), and can also be restricted to a particular geographic radius (Radius option, Figure 12). The default value of 16 nearest neighbors is usually sufficient, with no restrictions placed on radius (in kriging, neighbors outside of the semivariogram range are weighted identically and, if significant structural dependence is present, weighted minimally). Specifying more than 16 neighbors can slow interpolation substantially (Gamma Design Software, 2004).

In the GS+ the user may choose either block or point kriging. Its choice should be made on the basis of sampling design and variable characteristics. If samples were taken to represent an area around the actual sample point, then

block kriging may be more appropriate than point kriging. If samples were taken to represent point values then point kriging may be more appropriate.

Defined the entry parameters by the user, the kriging interpolation analysis can be performed by clicking on the calculate option (Figure 12). GS+ produces 2D and 3D maps of spatial data following interpolation. The data to be mapped come from kriging analysis, and are thus contained in interpolation output files. Figure 13 illustrates, as an example, a 3D map of the soil water content data set constructed from kriging analysis.

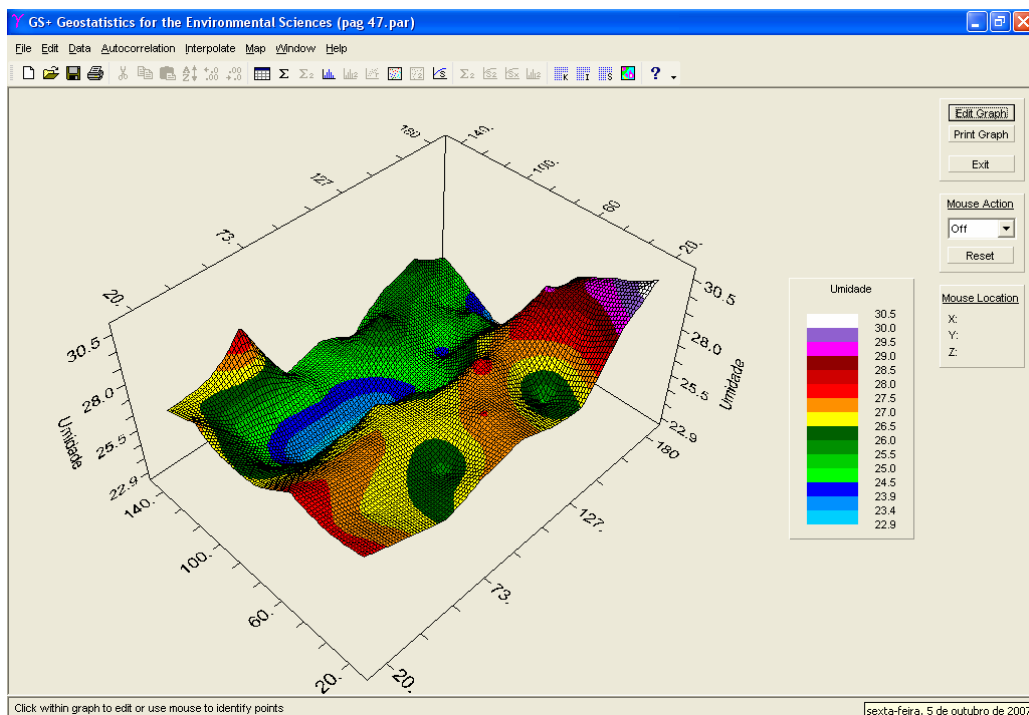


Figure 13: Illustration of a 3D map of the soil water content data set constructed from kriging analysis.

As you can see in the right side of Figure 13, there are some options for producing the 3D map image. Clicking on the Edit Graph option, the user can

change axis scales and other graph formats via the Graph Settings dialog window (Figure 14).

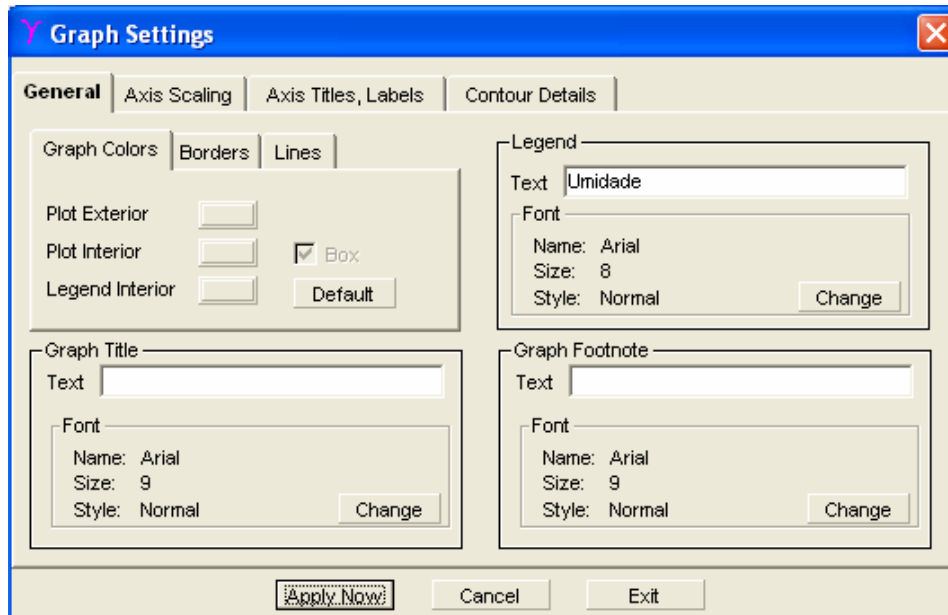


Figure 14: GS+ Graph Settings dialog window.

The GS+ interpolation window (Figure 12) allows selecting an existing or new output file to which kriging interpolation estimates will be written. The format in which GS+ will write estimates to the file can be one of the following types: GS+ krig format (.krig), Surfer grid format (.grd) and Arcview format (.asc). To examine the contents of an existing file press View option (Figure 12).

Clicking on the Print graph option (Figure 13) the user can print the active graph to a file, printer, or other device. On the other hand, clicking on the Mouse Action option the map can be rotated, moved, among other options.

Cross-validation analysis can be performed in the GS+ software. Clicking on the Cross-Validate option (Figure 12) in which each measured point in the spatial domain is individually removed from the domain and its value estimated

via kriging as though it were never there (Gamma Design Software, 2004). In this way a graph can be constructed of the estimated versus actual values for each sample location in the domain. Placing the cursor over a point provides information on the estimate for a specific location as in the example shown in Figure 15.

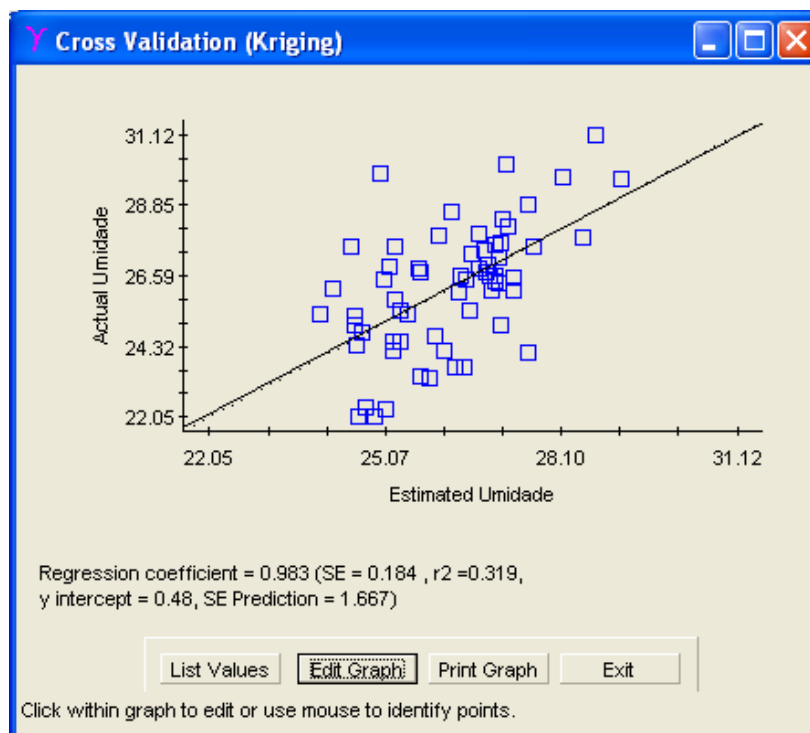


Figure 15: Illustration of a Cross-validation analysis performed by the GS+ software.

The regression coefficient at the bottom of the graph represents a measure of the goodness of fit for the least-squares model describing the linear regression equation. The standard error (SE) refers to the standard error of the regression coefficient; the r^2 value is the proportion of variation explained by the best-fit line; and the y-intercept of the best-fit line is also provided. The SE Prediction term is defined as $SD \times (1-r^2)^{0.5}$, where SD= standard deviation of the actual data (the data graphed on the y-axis).

Acknowledgments

To CNPq and FAPERGS for financial support.

References

Gamma Design Software. GS+: Geostatistics for the Environmental Sciences. Plainwell: Gamma Design Software, LLC, 2004.160p.

GUIMARÃES, E.C. Geoestatística básica e aplicada. Uberlândia: Faculdade de Matemática-Universidade Federal de Uberlândia, 2004. 77p.