

G606 Demonstration Problem Sets

Introduction

The problem sets outlined below are an integral and vital part of the course and are intended to accompany lectures. They reinforce and clarify concepts presented during the course and provide a practical introduction to the hands-on implementation of geostatistical analysis and modeling techniques. All problem sets are based on a single data set; students will start with exploratory data analysis of the raw data, working through the creation of kriged and stochastically-simulated maps of a variable, and will apply methods of quantifying uncertainty in the generated maps. The problem sets are designed around specific software in which students are expected to become proficient; students will be introduced to the software in a laboratory, hands-on setting beforehand.

Much of the software used in the course is either public domain or readily available. All Windows software is specific to Intel-based PC's, but the geostatistical software library (GSLIB) programs developed at Stanford are also available in their native FORTRAN code and could be compiled for use on Macintosh PC's. Because the GSLIB software package can be ported to multiple platforms and is so powerful, it forms the core of the software to be used during the course.

The Data Set

The problem sets are designed around a single, two-dimensional data set. The raw data are identical to those used by Isaaks and Srivastava (1989), with the coordinates rescaled to a smaller spatial domain. Thus, the study area covers approximately 1.0 x 1.2 km, with x, y coordinates given in meters relative to an arbitrary datum. A total of 470 samples are found in the data file "WalkerLakeRawUV.dat", which contains values of a continuous, primary variable and a secondary, less widely measured variable analyzed at the same sample locations; in addition, a third, discrete variable is included which represents categorical information. All values are assumed to represent averages derived from measurement of a horizontal panel of 15 x 15 meters, as well as uniformity in the third "dimension" (whether this is elevation, depth above/below surface, specific geologic formation, etc.). That is, all the data are assumed to be representative of, and drawn from, a single, 2-dimensionally distributed global population and not a mixture of different, unrelated populations from different depths, layers, etc.

Although the data incorporated in WalkerLakeRawUV.dat satisfy these conditions, it is important to understand the sampling history and strategy used to collect the data. In this data set, the first 195 samples were collected on a semi-uniform grid pattern; a second and third campaign of 150 and 125 samples, respectively, were performed in those areas in which the highest values were found during the first round of sampling. This strategy is common in practice, where the "hot " or anomalous areas command the most interest and tend to bias or cluster the data locations.

The units of the spatial variables associated with each data location are left undefined, so as to permit students to assign meaningful significance to the data depending on background and interests. Thus, a geologist might view the continuous variables as representing soil contaminant concentrations, ore grade, permeability, formation thickness, air quality or foraminifera counts; a biologist might view the values as reptile counts, species diversity, water quality variations, etc. The discrete variable could represent categorical information such as rock or soil type, presence/absence of heavy minerals, contamination, or a certain tree species, values of a variable exceeding a certain threshold, etc. These variables may or may not be correlated, but they were collected for the purposes of the study so they all carry potentially useful information about the spatial processes responsible for their variability. In what follows, units of concentration are assumed for the continuous variables wherever it is warranted for providing clarification by example.

The basic goals of working with the data set in this course are essentially the core of geostatistical analysis and modeling, in general:

1. How can we best estimate the variable at locations where data are unavailable?
2. What is the uncertainty of the value estimated at any particular location?
3. How can this uncertainty and the spatial nature of the variability be represented?

Beyond these basic goals, the applications of geostatistical techniques are as varied as the data sets and the nature of the technical problems posed by the data. For example, once these goals have been satisfactorily addressed, we might ask what is the probability that our estimated values are within a specified range; what is the probability that our estimates exceed a certain threshold; or where should additional samples be collected to maximize the increase in our knowledge of spatial variability? For example, if our sampling budget permits collection of a limited number of additional samples, where should the next samples be located to maximize the information return (in other words, to minimize prediction uncertainty).

If the estimated values of the variable are intended as inputs to a predictive model (eg: contaminant transport in ground water; faunal migration/dispersal; cost of mining, soil removal, cleanup, etc.), additional questions could be addressed: how uncertain are the predicted outcomes based on the uncertainty of our spatial estimates; or what is the probability that the results of our predictive model attain some specified performance criterion (such as acceptable ground water quality at a specified distance from the pollution source; or 75% recovery of estimated gold reserves; etc.).

Software

Almost all of the software used in this course is public domain; downloadable, zipped versions are included for these packages and can be loaded on student's home PC's; if students have access to a Macintosh Fortran compiler that meets Fortran-77 specifications, they should be able to compile executable versions for their Mac's. The version of Surfer included in the Public Domain Software directory is a demo version and may be freely copied (it is a fully functional version but cannot be used to save files or graphics).

The statistical package, StatMost, which will be used for statistical analysis, population distributions, normality tests, etc. is non-public domain software and is installed on only a limited number of workstations in the Geology Dept.'s computer labs. All students should have access to a commercial spreadsheet program, for numerical transformation of data, removal of outliers, general numerical analysis, etc. Networked or student lab versions of SPSS statistical software can be used in place of StatMost.

The directory structure on the Zip disk can be copied "as is" to the student's hard drive. It is required that students use a Zip disk to store all course work, program files, data files and problem sets; all software files are to be copied onto the Zip disk and software installed to the Zip drive for use in the campus GIS computer lab (on students' home PCs, software should be installed to the local hard drive). In addition, students are strongly urged to maintain redundant backups of all working files on your computer's hard drive - you are responsible for preventing data loss due to public access to your files, viruses, accidents, etc., by practicing safe "CIXS" (computer information exchange and storage). DO NOT keep any of your files on a computer hard disk in the Geology computer lab, Geology student lab or the campus GIS lab! You may use PCs in either computer lab during daytime hours (8am-5pm) for course work, but only if they are not in use for research purposes; because all non-research data files and programs will be expunged periodically, it is your responsibility to ensure that any work you perform in the computer labs is backed up.

Problem Sets:

I. Statistical Summarization

A. StatMost Software

Familiarize yourself with the software package. Some of the most important and most-often used routines and menu paths are summarized in attachment 1.C. to week 1 lecture notes.

a. Start StatMost, create a new worksheet, and create two normal distributions (Data|Create|Distribution) in the first two columns to "play" with, using a mean of 10 and variance of 1 for each; specify $n=20$ samples. In the third column, create another "sample" from this same population with $n=200$. In the fourth column, create a distribution from 20 samples with a mean of 11 and variance of 1.

Now, analyze these "samples" to arrive at a statistical decision made as to whether they represent the same parent population.

b. Plot graphical summaries of each distribution, highlighting a column and using Quickplot|Box Plot. Graph and compare histograms for each sample distribution (try different numbers of bins to get the best[?] looking plots) and calculate summary statistics for each. Based on these results alone, how much confidence would you have in being able to say that these represent Gaussian distributions or that they represent the same/different populations?

c. Create and compare histograms (Plot|2D Special|Histogram), cumulative frequency distribution plots (Plot|2D Statistical Charts, Quantile Plot) and normal probability plots (Plot|2D Statistical Charts, Normal Probability) for each distribution to determine if you can answer the above questions more confidently (try different bin ranges or numbers of intervals for each plot). Various tests of normality can be applied (t-test, chi-squared, Kolmogorov-Smirnov, etc.) to help determine whether a population is Gaussian; however, the apparent "normal-ness" of a sample is often dictated by the luck of the draw, and too few samples is usually the main impediment to correctly inferring whether a sample is drawn from a normal population.

d. Apply Student's t-test (Statistics|t-test) to determine whether distribution #1 and #2 are drawn from the same population mean. Repeat for population #3 and #4. Assuming you have more confidence that distribution #3 is Gaussian, does the t-test allow you to decide whether #4 is from the same population as #1 or #2? Repeat the comparison using the non-parametric, two-sample Kolmogorov-Smirnov test (Statistics|Advanced Tests, Kolmogorov-Smirnov). Remember that the t-test requires the assumption of normality for comparison; the K-S test does not.

NOTE: In the case of the two-sample tests of population means, the null hypothesis is that the population means are the same; the acceptance or rejection of the null hypothesis is based on whether the calculated test statistic exceeds the critical statistic for the specified confidence level. StatMost reports a "probability" for the two-sample t-test and the two-sample Kolmogorov-Smirnov test (Statistics|Advanced, Kolmogorov-Smirnov): this is the confidence level at which the calculated test statistic (t or D) would equal the value of the critical statistic for the specified D.F. In other words, it is the confidence level at which the null hypothesis could be accepted. In the case of the Kolmogorov-Smirnov one-sample normality test (Statistics|Normality Tests, Kolmogorov-Smirnov), the reported "probability" can be interpreted as the probability that the distribution is Gaussian. Whenever possible, don't use the calculated "probability" to decide: use the calculated or tabulated test statistic such as tabulated in statistical tables or in graphical form (eg: Fig. 7.4, in Till for Kolmogorov-Smirnov critical D-values). Remember: if the calculated test statistic exceeds the critical statistic at the specified confidence level, the null hypothesis is rejected and the populations are different.

e. Now apply these techniques to analyze the V, U data in the Walker Lake data file; see the StatMost data file WalkerLakeRawUV.dmd. Compute summary statistics and plots (means, variances, ranges, histograms, box plots, and cumulative frequency distributions) for the U and V data. Based on the summary statistics and graphical representations, can you determine if V and U are normally distributed? Apply a "one-sample" t-test (Statistics|t-test) and "one-sample" Kolmogorov-Smirnov non-parametric test (Statistics|Advanced Tests) to verify your hypotheses: do this by creating two new columns of normally-distributed data (Data|Create|Distribution) with the same means and variances as U and V; use the t-test and Kolmogorov-Smirnov test to compare V's and U's means with these created normal distributions. Are the conclusions the same using the two different test approaches (parametric vs. nonparametric)? If not, why?

Finally, plot V vs. U and V vs. T; describe how they are correlated. If T represents a facies or a soil type, could it be used to help infer the spatial distributions of U and V? Report your findings.

B. Hypothesis Testing with StatMost

Parametric Student's t-test:

a. Create two normal distributions of 100 samples each, with slightly different means and variances: $m_1 = 10$, $s^2_1 = 1$; $m_2 = 10.5$, $s^2_1 = 1.5$. Compute the 99% confidence interval for the mean of each sample dataset: Choose Statistics|Confidence Intervals|One Sample, and specify the appropriate sample column number and confidence level for the calculation. These computations are made using the Student's t-statistic. Compare the 95% and 99% confidence intervals; why is the 99% interval larger than the 95% interval?

b. Perform hypothesis tests (also based on the t-test) to determine if a sample mean is from a specified population. Choose Statistics|Hypothesis Tests|One Sample, Unknown Variance. Specify the data column to be tested, the value of the mean to be tested, the significance level and specify Raw Data to test the entire sample data set. First, test whether the sample mean represents a specified population mean (two-tailed test), then test whether the mean is less than a specified value (one-tailed test). Do this on both sample distributions and compare the results. Did the hypothesis tests yield the correct results at the .05 level of significance? at the .01 level?

c. Apply an hypothesis test (based on the t-test) to determine if the two sample populations were drawn from the same global population mean. Choose Statistics|Hypothesis Tests|Two Samples, Unknown Variance. Specify Raw Data for both columns, the level of significance of the test, the value of the mean being tested (10, for both samples) and a two-tail test (ie. means are equal). (Note: you can accomplish the same result by choosing Statistics|t-test, and determining if the computed value of the test-statistic is less than the critical t-statistic)

Non-Parametric Kolmogorov-Smirnov Test:

d. The parametric t-test requires that you know the sample distributions are drawn from a normal population. One very useful application of the K-S test is to test whether a sample population could be drawn from a normal population. Because StatMost's K-S test compares the sample distribution to a standard normal distribution, first transform the distribution to an equivalent distribution with mean of zero and variance of one. Highlight the data column to be transformed and choose Data|Special|Normalize; the resulting column contains the "z-scores" of the data, transformed to an equivalent distribution of mean = 0, variance = 1.

e. Now choose Statistics|Normality Tests and specify the column(s) to be tested and the test type (K-S). Compare the result on the normalized and unnormalized data. Note that the K-S test statistic is computed, as well as the probability or confidence level at which the means would be considered equal.

f. The second important application of the K-S non-parametric test is to compare two sample distributions, and whether they are the same or different in any way (mean, variance, skewness, etc.). Choose Statistics|Advanced Tests and specify the columns being compared. The K-S statistic is computed. Does this result compare with the results of the t-test in (c) above?

II. Exploratory Data Analysis 1

A. Basic Spatial Relationships

a. Open the raw data set (WalkerLakeRawUV.dat) with a text editor (eg: Notepad) to view the standard geostatistical data file format ("GeoEas" file format). Following header information about the number and type of variables in the file, the columns specify spatial coordinates (x, y only in this 2-D case), the primary concentration variable, V, a secondary variable (U) measured at the same location, a discrete variable (T) representing categorical information about the sample location, and the sample ID number. What are the V, U concentration values for sample number 179? What range of values does T take on over the entire study area?

b. Import WalkerLakeRawUV.dat data (or copy and paste the contents of the StatMost .dmd file) into a spreadsheet so values can be manipulated later. Add column headings for the variable names. Save the spreadsheet as a ".wk1" file for later use. Sort the file on the V data column and report the range (maximum, minimum) of V values.

c. Use the program **3Plot32** or **Surfer** to create a "post plot" map of the data in order to examine data locations and ranges of values. Identify and remove lines in the data file containing spurious x,y values that are way outside of the study area. Report which samples were removed in this way. Save the cleaned version of the .dat file which will be used for all subsequent analyses.

B. Data Distributions and Global Statistics

a. Use the zoom option in 3Plot32 or Surfer to examine the areas with highest data values. In 3Plot32, double click on the right mouse button to activate tracing (in Surfer, generate a Classified Post map) and locate the highest V concentration value. What are the coordinates of this value? Does the highest U concentration value also occur at this location?

b. Use **HISTPLT.EXE** to plot the cumulative frequency distribution of V data, and compare with the plots you made with StatMost.

c. Use **PROBPLT.EXE** to plot the V data on normal probability coordinates, and visually decide whether the data are distributed normally, log-normally or neither. Compare your GSLIB plots with those produced in StatMost.

d. Compare post plots of the U and V data; in a spreadsheet or StatMost, plot the U data vs. the V data. Are these variables correlated? Summarize what you see from the post plot of T values compared with U and V.

e. Use 3Plot32 to create a moving window average of the V data (use 60 x 60 m² window); save the windowed calculations to a .dat file and import into a spreadsheet. Sort according to number of samples per window, exclude windows with n<3; plot the variance of the remaining windows against the mean. Is there an obvious proportional effect?

III. Exploratory Data Analysis 2

C. Data Declustering and Transformation

- a. From c. above, are the data uniformly distributed over the study area or are they clustered? Use **DECLUS.EXE** to identify the optimal data spacing at which to decluster the data with a moving average technique and to compute the data declustering weights. Examine the .sum output file created by DECLUS to see the effect of averaging over different window sizes; the optimum window size is used by the declustering algorithm to assign clustering weights to the data points.
- b. Rerun HISTPLT and PROBPLT to compute mean, median, variance, minimum/maximum and lower/upper quartiles of the declustered V data. Compare the clustered and declustered statistics; what do you conclude from this comparison?
- c. Run **NSCORE.EXE** on the Walker Lake data to perform a normal-score transformation of the V data, and save it in a new file WalkerLakeFinal.dat. You will use this data file for all subsequent work.
- d. With a text editor, open the file with extension ".sum" created by NSCORE. Import this data into StatMost or a spreadsheet and plot the histogram to view the nature of the data transformation; how has the distribution changed?

Introduction to GSLIB Parameter File Structure

All GSLIB programs use an input file (.par extension) to specify the data, parameters and conditions for their associated executable file (.exe) to process. Parameters and file structure are specific to each program, but all look similar in their overall layout. Look up the documentation in Deutsch and Journal (1997; p.204, 206) for each program before attempting to run it! The .par files below are for the histogram-plotting routine HISTPLT.EXE and the cumulative frequency distribution graphing routine PROBPLT. EXE; open each with Notepad (a simple Windows text editor), find and understand the changes that need to be made (*italicized*) and key in the changes so that the final files look like those below. Then run HISTPLT and PROBPLT using these input files to create plots of Walker Lake data.

Parameters for HISTPLT

START OF PARAMETERS:

```

WalkerLakeClean.dat    \file with data
3 0                    \columns for variable and weight
0 2000                 \trimming limits
histogram.ps         \file for PostScript output
1 0                    \min and max values of variable to plot (automatic if max<min)
-1.0                   \frequency maximum (<0 for automatic)
25                      \number of bins or classes
0                       \0=arithmetic scaling, 1=log scaling
0                       \0=frequency (histogram), 1=cumulative frequency distribution
0                       \number of cfd quantiles (<0 for all)
2                       \number of decimal places (<0 for auto.)
Walker Lake Data     \title
1                       \positioning of summary statistics on plot (L to R: -1 to 1)
1                    \reference value for box plot (outside tmin,tmax)

```

Parameters for PROBPLT

START OF PARAMETERS:

```

WalkerLakeClean.dat    \file with data
3 0                    \columns for variable and weight
0 2000                 \trimming limits (set lower=min label incr.for log plot)
probplot.ps         \file for PostScript output
-1                     \number of points to plot (<0 for all)
0                       \0=arithmetic scaling, 1=log scaling
0 1600 200.0          \min,max,increment for labeling
Walker Lake Data     \title

```

The parameter files for performing data declustering (DECLUS.EXE) and normal-score transformation (NSCORE.EXE) are given below. Review parameter usage in Deutsch and Journal (1997, p.213 and 223). Italicized entries are parameters you will probably need to alter for your own data set.

Parameters for DECLUS

START OF PARAMETERS:

<i>WalkerLakeClean.dat</i>	\file with data
<i>1 2 0 3</i>	\columns for X, Y, Z, and variable
<i>-1.0e21 1.0e21</i>	\trimming limits
<i>declus.sum</i>	\file for summary output
<i>WalkerLakeDeclus.dat</i>	\file for output with data & weights
<i>1.0 1.0</i>	\Y and Z cell anisotropy (Ysize=size*Yanis)
<i>0</i>	\0=look for minimum declustered mean (1=max)
<i>24 50.0 500.0</i>	\number of cell sizes, min size, max size
<i>4</i>	\number of origin offsets (4 in 2D; 8 in 3D)

Parameters for NSCORE

START OF PARAMETERS:

<i>WalkerLakeDeclus.dat</i>	\file with data
<i>3 7</i>	\columns for variable and weight
<i>-1.0e21 1.0e21</i>	\trimming limits
<i>0</i>	\1=transform according to specified reference distribution
<i>nodata.txt</i>	\file with reference distribution
<i>1 2</i>	\columns for variable and weight in reference distribution
<i>WalkerLakeFinal.dat</i>	\file for output of transformed data
<i>nscore.trn</i>	\file for output of transformation table

IV / V. Spatial Correlation 1 - Experimental Variography

a. Plot a one-dimensional set of five data points, equally spaced along a line with sample spacing of 1 unit. Determine how many total data pairs are available in this data set for a potential variogram calculation. The data points have values of 1, 2, 2, 3, 1 along the sequence; determine the value of the semivariogram statistic at a lag spacing of 1 unit *with a manual calculation*.

b. Using Windows Notepad, create a standard format .dat file with the values in (a) as the regionalized variable z_x at locations of $x = 1, 2, \dots, 5$ and $y = 0, 0, \dots, 0$ and save it as "test.dat" (since VarioWin is written for 2D spatial data, a "dummy" y coordinate column must be included, with constant value of y for all data points).

c. Open **PreVar2D** to create a pair comparison (.pcf) file from "test.dat". Under File|Open, navigate to the directory where test.dat is saved, and read in the data. Then under Settings|XYCoordinates specify X and Y as the appropriate coordinates in the data file. The total number of pairs produced from the data file are shown on screen; does this number match your answer in (a)? A file of all possible pairs of values in the data file has been created and saved as "test.pcf".

d. Close PreVar2D and open **Vario2D**. In the File Open dialog box, specify the directory and file name for test.pcf and read it in. Under Calculate|Variogram Cloud, identify the variable column in the data file, specify a maximum distance (pair separation) of 4 units, a direction of 0 (east-west) and an angular tolerance of 45 degrees, and click on OK. A plot of all pair separations and their individual $(z_x - z_{x+h})$ values is produced. To change the x-axis scale to clearly see all the points, click on Graph|Axis and change the minimum x value to 0.0 (note that some points plot on top of one another, so less than 10 points are visible). With the graph window active, go to File|Save As, save the graph data as "test.cld". Read the contents of test.cld with Notepad; this file lists all the pairs of values and their $(z_x - z_{x+h})$ values. This type of file is very useful for creating histogram plots of the available lag spacings in the data, in order to identify lags with few data or to choose the appropriate lag intervals for subsequent variogram calculations.

e. In Vario2D, calculate the variogram under Calculate|Directional Variogram. In the dialog box, identify the variable column in the data file, specify a lag of 1, four lags, an angular tolerance of 90 degrees, and click OK. The variogram is presented, showing the semivariogram statistics for only two lags (for our small data set, lags greater than 2 units have too few points to estimate the semivariogram statistic). Compare the semivariogram value to that calculated in (a.).

Click on the first variogram point and a second graph is displayed, showing an h-scatterplot (or h-scattergram) of the data that define that variogram point. Click on the point at coordinates (1, 3) in this graph; this is a potential outlier point because it sits far from the h-scattergram's 1:1 line. A dialog box is displayed allowing you to identify which data pairs are responsible for this point, and optionally to mask or remove the pair in question to evaluate its impact on the calculated variogram statistic. Note the data pair numbers and the marked effect on the recalculated variogram after the pair is masked. Do not close the h-scattergram window. Repeat this procedure for the point in the h-scattergram at coordinates (3, 1). What single data point do all these potential outliers have in common? Its value imparts a skewed nature to the distribution of values in the data set. Thus, experimental variography becomes an extension of EDA.

f. Now work through the VarioWin tutorial example (p. 11-19 in VarioWin handout) to familiarize yourself with the capabilities and use of the program for calculating experimental variograms and discerning correlation structure. The goal in working with real data is to identify optimal lag classes, overall correlation structure and anisotropy, and to identify possible data outliers and their impacts on variogram statistics.

To aid in identifying the best lag classes to use in determining spatial continuity, use Vario2D's Calculate|Variogram Cloud option to create a variogram cloud and from File|Save As, save the pair data in a .cld file. Import this file into a spreadsheet or StatMost and plot a histogram of the distribution of |h| values. The optimum estimate of a variogram's lag interval size is one with the largest possible number of histogram bins (ie. smallest bin interval size) while maintaining as uniform a number of pairs in all bins as possible (note that variograms should NOT be calculated with |h| values larger than about half the maximum pair spacing in the data set so disregard the shape of the righthand side of the histogram). This optimum bin interval size estimate is the starting point for lag class size in your experimental variograms; you will then iteratively adjust the lag size and lag tolerance to obtain the cleanest representation of variogram structure.

The general steps in experimental variography are:

- determine best initial lag classes based on the spacing of the available sample data
- construct the best isotropic variogram (90° angular tolerance), refining the lag classes and determining the best variogram estimator to use
- examine h-scatterplots of the lag intervals, identify outlier or troublesome data points and, if necessary, remove some data points to see spatial correlation structure more clearly
- construct a variogram surface and anisotropic variograms (ideally, using angular tolerances of <30 degrees, but as large as necessary to adequately define anisotropy), and define the directional variograms for maximum and minimum correlation ranges
- reexamine alternative measures of correlation structure for the anisotropic variograms and look for internal consistency in their correlation parameters
- see Week 6 (section 3) lecture notes for more details on working with real data and dealing with the inevitable noise that creeps into variogram statistics from small and/or skewed data sets.

g. When you are confident that you can use **Vario2D** to deal with real data, convert the WalkerLakeFinal.dat file to a .pcf file using **PreVar2D**. Be sure to read Isaaks and Srivastava (1994, pp.50-64; 140-163) for help in understanding the concepts behind spatial correlation analysis. The final goal is to determine the correlation structure of the Walker Lake V data; don't rush this until you are familiar with the concepts and with the mechanics of VarioWin. When you are satisfied that you can identify the variogram structure of the Walker Lake data (including principal directions of anisotropy), save the experimental variograms to a single .var file (File|Save As); highlight the window containing the experimental variogram with the maximum correlation range and save it. Next, highlight the window with the variogram of shortest range (this must correspond to a direction perpendicular to the direction of maximum range) and Append this variogram to the same file. Do the same for the nondirectional (isotropic) variogram of V. You should now have a .var file containing three experimental variograms, which will be used subsequently for modeling the correlation structure.

VI. Spatial Correlation 2 - Variogram Modeling

- a. Now that you have analyzed the spatial correlation structure of the Walker Lake V data, it must be described by a correlation function model to be of use in kriging or simulation. Open VarioWin's modeling module, Model.exe and read in the Walker Lake .var file you created in the last exercise. In the dialog box listing the experimental variograms that are included in the .var file, select both the variograms with the maximum and the minimum ranges by holding the shift key down while clicking on each. The modeling screen appears, showing both experimental variograms in the right-hand window, and the modeling parameters window on the left. At this point, it's good practice to show the number of data pairs associated with each variogram point: click on the right-hand window, then in the menu bar, select Graph|Pairs. The number of pairs indicates the relative confidence level of each plotted variogram point, and whether your correlation function model should try to fit to points of low confidence.
- b. The first parameter, the nugget, is set in the left-hand window to an appropriate value as shown by the horizontal lines on the two variograms in the right-hand window. Note that any change you make to the model parameters is immediately updated in the variogram window.
- c. Next, set the direction of maximum spatial continuity in the top slider bar of the 1st Structure dialog box; this is the direction of the variogram with the largest range.
- d. Specify the type of correlation function Model from the drop-down list. Its Range and Sill are set with the next two slider bars below. Note that you can set any slider bar value by clicking on the value shown above the bar and directly typing in a value; this procedure is sometimes necessary when the value you want to enter is beyond the default range of the slider bar.
- e. After adjusting the 1st Structure's Range and Sill until you have the best fit possible for the variogram of maximum range, you need to fit the appropriate range for the other variogram. To do so, adjust the Anisotropy slider bar until the correlation function model best fits the second variogram. Anisotropy is defined as the ratio of the range of the two directional variograms.
- f. At this point, when you are satisfied this model fit is the best you can do, click on the Store button so that you can Restore these settings later if you play with them further. Also note that the Best Fit Found button will bring you back to the previous best fit that has not yet been stored.
- g. If additional model structures are required to improve the model fit to the variograms, repeat steps c. through f. for each additional structure in the 2nd Structure and 3rd Structure dialog boxes. Note that every time you add an additional structure, the sill and range parameters in previous structures will need to be adjusted because the various model structures are additive. This is where the Restore and Best Fit Found buttons can come in very handy.
- h. Save the model parameters to a file, using File|Save As, and specifying a .mod file name. Print out this file as well as the right-hand window showing the model fits to your experimental variograms. Describe how/why you determined these models to be the best fits to your data. You will be using these models in future kriging and simulation runs.

Procedure for Modeling Zonal Nugget Effects and Zonal Anisotropy

If the nugget is different in two directional variograms (beyond what can be construed to be an artifact of the sample data), but the variograms have the same sill (geometric anisotropy), the effect must be modeled with a set of at least three nested structures. Two of these structures are fictitious "nugget" components that are represented as correlation functions with a very small range:

1. Nugget:

- set the value of the minimum nugget (the variogram with the lowest y-intercept)

2. Structure 1: (additional zonal nugget)

- specify the direction in which the maximum nugget value occurs
- specify a fictitious correlation function e.g. spherical model
- specify a small, non-zero range (eg: 1) and large anisotropy ratio (1000 max)
- adjust sill value to approximately the difference between the two directional nuggets
- see Figure (a) for typical screen view at this point; note that an additional "nugget" has been added in one direction but not the other (a zonal nugget is probably physically impossible and is more likely a manifestation of just such a nested set of ranges)

3. Structure 2: (add an additional sill component to the variogram with the minimum nugget to compensate for the addition of Structure 1 to the other variogram)

- specify the variogram direction with the minimum nugget and a function model
- set the anisotropy ratio high (1000 max), and a small, non-zero range (eg: 1)
- specify a sill equal to the sill in Structure 1
- see Figure (b) for typical screen view at this point; note that now the cumulative sills in both directions are now equal (don't worry if the "sills" aren't horizontal)

4. Structure 3: (main correlation structure to be modeled; note that if this requires more than one structure, GSLIB or GeoEas software will have to be used - VarioWin allows 3 max)

- specify the direction of the variogram with maximum range (largest transition region)
- specify the appropriate correlation function model
- set approximate range in this direction and adjust sill and anisotropy ratio for an approximate model fit to both directional variograms
- see Figure (c) for typical screen view at this point
- at this point, the ranges and sills of all three structures can be adjusted to fine-tune the fit of the two variograms: this leads into the concept of zonal anisotropy (unequal sills)
- see Figure (d) for a screen view example of the effect of adjusting all three structures

- in performing this exercise, note how specifying a large anisotropy ratio affects only the sill of the variogram in the specified direction; conversely, specifying a very small anisotropy ratio affects the variogram in the other direction

For modeling zonal anisotropy (different sill heights, same nugget, in the two directional variograms), the procedure is very similar, except that Structure 2 is not required, and Structure 1 adds the additional sill height required by the variogram with the higher sill. Assuming equal or no nugget:

1. Nugget:
 - set the approximate value of the nugget

2. Structure 1: (additional sill increment for the variogram with the higher sill)
 - specify the direction of the variogram with the higher sill
 - set the range > 0, and a very large anisotropy ratio (1000)
 - adjust the sill to equal the difference between sill heights in the two variograms
 - see Figure (e) for typical screen view at this point; note that an additional sill increment has been added only to the variogram with the higher sill

3. Structure 2: (main correlation structure to be modeled; note that if this requires more than two structures, GSLIB or GeoEas software will have to be used)
 - specify direction of the variogram with the maximum range
 - set approximate range in this direction and adjust sill and anisotropy ratio for an approximate model fit to both directional variograms
 - adjust ranges and sills of both structures, and the anisotropy ratio of the second structure to fine-tune the model variogram fits
 - see Figure (f) for typical screen view at this point; note that zonal anisotropy is actually simpler to model than a zonal nugget, and unless a zonal nugget can be justified, directional variograms should be modeled with equal nugget values in both directions

 - note that a third structure can be incorporated, if necessary, to improve the model fit; however, do not "overmodel" a variogram, because your ability to defend the choices of a number of nested structures will almost always decline

NOTE: $Anis. = \frac{\text{Range of Complimentary Directio}}{\text{Range of Specified Direction}}$ (in VarioWin)

VII. Introduction to Kriging with GeoEas

1. Like VarioWin, GeoEas is a collection of programs with which to perform a number of different but related tasks. To open GeoEas, double-click on the file "geoeas.exe". Press any key to go to the main menu screen.
2. On the main menu screen are shown all the modules or programs that GeoEas contains. Note that there are programs that are similar to functions in VarioWin, like Prevar (similar to VarioWin, for creating pair comparison files for variogram analysis) and Vario (for calculation and modeling of variograms). Other programs are for simple univariate and bivariate statistical analysis (Stat1, Scatter), simple data transformations (Trans) and various plotting routines (Conrec, HPplot, View). For a complete description of all program functionality see the GeoEas user manual or the postscript file "geoeas.pdf".
3. In order to use GeoEas to perform 2D kriging, we will have to learn how to use the modules Krige and Xvalid (to produce gridded data files), Conrec (to convert the gridded files to contour plots, and HPplot (to translate the contour plots into a file format that can be read by Windows graphics programs, for printing). At the main menu screen, use the keyboard arrow keys to highlight the program Krige, and then press <CR>.
4. Go to the first menu screen of Krige; the menu choices are listed at the bottom of the screen. This first screen is for saving or reading previously saved kriging parameters; since we don't have any saved, highlight Option/Execute and press <CR> to go to the Kriging Options screen. This is the main screen where almost all the options for the kriging process are set.
5. Highlight Data and press<CR>; key in the exact name of the .dat file to perform kriging on. Note that the .dat file has to reside in the same directory as geoeas.exe.
6. Highlight Type, press <CR> and with the Space bar toggle to Ordinary kriging, and then to Point kriging. Use block kriging if the kriged estimates are to represent statistical averages within blocks.
7. Select Grid next, to set the kriging grid node spacing. Set the variables to the correct x, y columns in the .dat file by toggling with the Space bar; accept the selection with <CR>. The program will specify default values of xmin, ymin (Origin); dx, dy (Spacing); and nx, ny (Number). Choose these at first, then go back and change them if needed later. Note that the coarser the estimation grid spacing, the more smoothed the kriged map will look.
8. Select Search, to set the parameters controlling the nearest neighbors to be used by the kriging equations. Here again, the program provides default values which are based on the data spacing only.
 - a. Set the Major and Minor Radius and Ellipse Angle to match the maximum and minimum variogram ranges, and the direction of maximum variogram range, respectively. Choose Variogram or Euclidean distance, depending on whether nearest neighbors are to be determined

by distance or similarity according to their variogram value relative to the center of the search ellipse.

b. The #Sectors can be set to 1 for most situations; Max in Sector should be set to less than 10 for zero nugget variance models and greater than 20 for nugget variance that is half or more of the total variogram sill height. The value for Min in Sector is set to the smallest number of samples in a search neighborhood that will produce a kriged estimate (no estimate will be made if fewer than this number of points are found). If #Sectors was set to more than 1, set the number of Empty Sectors that can be tolerated before no estimate is made, otherwise it is zero.

9. Select Variables/Models to go to the variogram model parameter selection screen. On the menu, select New Variable and toggle to the variable in the .dat file being kriged. Press <CR> to select the specified variable. Enter the variogram model parameters in the remainder of the screen. Once entries have been made, changes can be entered via menu selection Edit. When done, select Quit from the menu to return to the kriging parameter screen.

10. Select Execute to perform the kriging. The screen switches to full-screen graphics mode and as the kriging system of equations is solved at each grid node, the kriging estimate, kriging standard deviation and number of local neighbors used in the kriging estimate for each grid node are displayed at the bottom of the screen. When kriging is completed, press the "Q" key to return to the kriging parameter screen. At this point, the kriged estimates have been written to the specified .grd file.

11. Exit the kriging program via Quit, back through the hierarchy of menu screens. At the main menu screen, select Conrec and <CR> to contour the .grd file. At the first Conrec menu screen, select Options/Execute, then Data and key in the name of the .grd file to contour. Specify the correct x, y variables and whether to contour the kriged estimates (*variable_name) or the kriging standard deviation (KSDvariable_name). When asked to accept defaults for contour plot parameters, choose Yes. At this point, Graph Options or Contour Options can be selected to modify the appearance of the graph or the contour interval, etc. Then select Execute and <CR> to create the contour map.

12. At this point, a DOS graphics metafile (.met suffix) has been written to disk, but this file must be translated into a form that can be read by Windows graphics programs. Quit out of Conrec and at the main menu screen choose HPplot. At the first menu, select File and key in the name of the .met file created by Conrec. Then select Execute to translate the file; specify a file name with a ".hgl" suffix. This file can be read by various Windows graphics programs such as Freelance and Powerpoint, from which the kriged map can be modified or printed out.

VIII. Kriging Estimation: the Neighborhood Search Process

GeoEas' program Krige produces a regular grid of interpolated point or block estimates using either Ordinary or Simple kriging. The default option, ordinary block kriging, is recommended for most applications. Point kriging usually provides estimates very similar to those from block kriging. However, if a point being estimated happens to exactly coincide with a sample location, the estimate is set equal to the sample value; ie. estimation does not occur, and the contour map would probably contain a "bull's eye" at that location. This is not appropriate for contour mapping, which implicitly requires a spatial estimator. Ordinary kriging estimates the point or block values with a weighted average of the sample values within a local search neighborhood, or ellipse, centered on the point or block.

The program computes a 10x10 grid by default, which we will usually want to override. The default search ellipse is a circle with a radius about one fourth the maximum x or y dimension of the site, which should be adequate for most cases. The purpose of the search is to reduce computation time by eliminating from the kriging system of equations those samples which are unlikely to get "significant weights". The default search strategy is to treat the search circle as a single "sector", to examine all samples within it and use at least one, but not more than the closest eight.

The number of neighboring data required for kriging is related to the value of the nugget term in the variogram compared to the maximum variogram value possible within the search area. The higher the nugget, the more likely that more distant samples will get significant weights. A rough rule of thumb would be to use eight samples when the nugget is near zero, increasing to twenty when the nugget is more than 50% of the maximum value. The more complex sector search options may be useful when you have unusual patterns or clusters of data.

Setting the Search Parameters in GeoEas

Rules of thumb:

- a) start with the simplest search strategy and the fewest parameters possible to obtain stable kriging solutions, then refine the search parameters
- b) start with conservative settings for the nearest neighbor criteria, obtaining a large number of null grid estimates (grid points where no kriging estimate is possible because of too few nearest neighbors), then refine the criteria to minimize or eliminate the number of null estimates
- c) initial parameter estimates should be set to (in order of priority):
 - circular search with radius about 5-6 times the kriging grid spacing used (but no larger than the average of max. and min. variogram ranges)
 - one sector; Max. Points/Sector = 12
 - a large number of Min. Points eg: 4-8 (which will identify areas where a confident estimate cannot be made because of too few data)
 - the Euclidean Distance calculation option

Then refine the search parameters to obtain (in order of priority):

- elliptical search with radii equal to max and min variogram ranges
- if pronounced zonal anisotropy exists, use the Variogram Distance calculation option
- if the variogram nugget is small or zero (relative to the sill height), use eight or fewer Max. Points/Sector; if the nugget is half or more of the total sill height, use up to twenty Max. Points/Sector (ie. at high nugget values, distant neighbors in the search neighborhood will receive a higher weight, because the estimate relies on local averaging of more data)
- use the Debug keys (<Caps Lock> shows the search neighborhood and number of nearest neighbors for each grid location; <Num Lock> shows the distances of each neighbor from the grid location and their calculated weights); adjust the number of Max. Points/Sector until the most distant neighbors consistently receive a weight of 0.05 or less
- reduce the number of Min. Points to Use until areas with null estimates disappear or are restricted to those portions of the map area where very few data exist

To enhance interactivity with the kriging process and to more easily see the effects of changes made during the parameter fine-tuning, use 3Plot32 to read the GeoEas kriging output .grd file and view the kriged values with the postplot option. To do this, open 3Plot32 and under File|Open specify the directory and file name of the .grd file (GeoEas creates the .grd file in exactly the same format as a .dat file) to read in the file, then create a post plot, using a suitably edited .lvs file to color-code the kriging estimates. It is easiest to create several .grd files of different names, representing different parameter combinations, then to open multiple versions of 3Plot32 and toggle between each to see the differences in the kriged maps.

Once a kriged .grd file has been created, it can be analyzed statistically just like the original data. One of the goals of post-kriging analysis is to evaluate the estimates obtained through kriging by comparing their statistics with the original data's statistics (this is a measure of the "unbiasedness" aspect of estimation). Another goal is to evaluate which kriging strategies produced the "best" estimates (remember, that for every set of search parameters, solving the kriging matrices always produces "best" estimates for the neighboring data utilized, but that these estimates vary with the kriging strategy employed; therefore, several different kriging scenarios/strategies must be compared to evaluate whether significant differences exist in the estimates produced, and whether any one stands out above the others as a superior "best" estimate; this is the process known as cross-validation).

Krige the Walker Lake V data, using your WalkerLakeFinal.dat file; perform the following steps:

1. Start GeoEas's kriging module, specify the .dat file and the output .grd file names.
2. Start with conservative search parameter values as discussed above. For the grid parameters, use $n_x = 70$, $n_y = 80$, $dx, dy = 15$ and $x_0, y_0 = 7.5$. Make sure you understand how the grid origin (x_0, y_0) is defined. Execute the kriging process to ensure that a stable solution is possible.
3. With CapsLock and NumLock OFF, start the kriging process. After a few seconds, press <CapsLock> to see the search neighborhood; press <CapsLock> again to turn it off then press <NumLock> to see a list of the neighboring data points used at the current grid location, with their distances and kriging weights. How many nearest neighbors have weights greater than 0.01? Turn off NumLock and press Q again to resume kriging. Repeat the above process five or six times and record the numbers of nearest neighbors at each grid location with kriging weights greater than about 0.01.
4. Use an average of these numbers to refine your starting estimate of Max. Points/Sector; also modify your search neighborhood to an elliptical area and repeat the kriging; again, view the kriging weights at five or six grid locations and report the minimum values at each. If the average of these values is greater than about 0.05, increase your estimate of Max. Points/Sector slightly and repeat the kriging. Ideally, the ideal number of neighboring data points should be small enough to avoid many with very small weights, yet large enough to ensure that kriging estimates are representative of their local areas.
5. View the resulting kriged grid with 3Plot32, using an .lvs file appropriately edited to show concentration levels of -100, 0, 100, 200, 1500. If there are numerous and/or large areas with no estimated values (null estimates), reduce the Min. Points to Use parameter and repeat the kriging.
6. Produce a plot of the kriged output, by capturing a 3Plot32 screen with PaintShop or Canvas and printing it.
7. Create your "best" kriged map, plus another that is quite different-looking (by specifying a nugget equal to the variogram sill height). Produce a plot of this kriged map and compare. Why do they differ because a higher nugget was used?
8. Read the two .grd files into StatMost and compare the statistics of the estimated V values against those in the original .dat file; compare the kriged means and variances with those of the clustered and declustered data. What do you conclude about your kriging estimates?

IX. Kriging Cross-Validation, Introduction to GSLIB's Kriging and Simulation Programs

A. Cross-Validation

1. Open GeoEas' "XValid" module. At this point, your Walker Lake .dat file should reside in the GeoEas directory; space over to the Data menu item and enter the 8-character.dat name of your Walker Lake .dat file. Specify the X, Y and V variables and Log Option off (ie. do not use a lognormal transform).
2. Go to the Options/Execute screen; set the kriging type to Ordinary Kriging, and set the Search and Model parameters the same as you used in one of your kriging runs in Problem Set XIII, Question 7. Execute to perform the cross-validation calculations.
3. When the program has completed the calculations, press Q to return to the calculation results menu screen. The screen summarizes the cross-validation results. How do the minimum and maximum values of the estimates compare with those of the V data? What is the mean difference between the estimated and observed values? Is the median difference similar to the mean difference? If it is, what would this indicate about the distribution of estimation errors?
4. Select Error Map to see a map of the relative differences between estimated and observed values; X symbols represent estimates that were less than the observed data value, + symbols represent estimates that were greater than the observed value, and the size of the symbol is a coarse indication of the magnitude of the difference. Look at the distribution of X's and +'s in the clustered (hot spot) areas: can you visually tell if there is a difference in the proportion of X's and +'s between the hot spots and the sparsely sampled (background) areas? If the answer were yes, it would indicate that kriging may have introduced a bias.
5. Press Q to return to the menu screen, and choose Scatter Plot, and Observed vs. Estimated. Does the plot show any systematic bias away from the 1:1 line? if so, this would indicate that kriged estimates of high or low values may be biased. Look at the box plots on the top and right sides of the plot; do the plotted differences appear to be normally distributed or not?
6. Return to the menu screen and choose Histogram, and Frequency vs. (z-z*). Does the distribution of estimation errors look normal? if not, kriging may have produced biased estimates. By choosing Write from the menu screen, the cross-validation differences at each data point location are written to a .dat file, from which the normality of the distribution of kriging errors can be analyzed with StatMost.
7. Repeat questions 2. through 6. for the second kriging run you produced in Problem Set XIII, Question 7. Based on the normality of the cross-validation error distributions between these two runs and the mean differences calculated in the two cross-validation runs, are you able to distinguish which of the two kriging runs is "best"? if so, which one (you will use this variogram model in subsequent simulation exercises).

B. Introduction to GSLIB Kriging Program KT3D

KT3D is a powerful and very flexible kriging program that can be used for 2D and 3D data sets. Its power lies in its ability to do things that GeoEas cannot (such as kriging data sets that are characterized by a regional trend, and being capable of a form of cokriging known as "kriging with a drift". The parameter file below is designed to work with the Walker Lake data; it is set up for a grid of 70 x 80 nodes and generic variogram parameters (note that unlike VarioWin and GeoEas, GSLIB variogram directional parameters are specified like compass bearings (0o is north, 90o is east, etc.). If you use this parameter file to kriging your Walker Lake data, you will need to change the variogram parameters to match your model variogram values. As before, parameter values you will likely need to adjust are in *italics*; non-italicized parameter values probably don't need to be altered.

Refer to Deutsch and Journel (1997, p. 96-100) for detailed description of parameter file structure.

Parameters for KT3D

START OF PARAMETERS:

```

WalkerLakeFinal.dat      \file with data
1 2 0 3 0                \columns for X,Y,Z, primary variable, external drift variable(=nonstationary mean)
0      1.0e21            \ trimming limits
0                          \option: 0=std.kriging, 1=cross-valid'n, 2=jackknife
xvk.dat                  \file name for jackknife data
1 2 0 3 0                \columns in xvk.dat for X,Y,Z, primary variable and secondary variable
0                          \debugging level: 0,1,2,3
kt3d.dbg                 \file for debugging output
kt3d.out                 \file for kriged output, viewed with PIXELPLT.EXE
70 7.5 15.0              \nx,xmn,xsiz (kriging grid parameters)
80 7.5 15.0              \ny,ymn,ysiz
1 0.5 1.0                \nz,zmn,zsiz
2 2 2                    \block discretization(1,1,1 for point kriging; 2,2,2 for 2x2x2 block kriging, etc.)
4 8                       \min, max data in search neighborhood octants
0                          \max number of data points per octant (if 0, an octant search is not used)
450.0 450.0 0.0         \maximum search ellipse radii in principal horiz. directions and vertical direction
0.0 0.0 0.0             \search ellipsoid angle parameters (see Deutsch and Journel, 1997, p.28)
1 2.30                   \0=SK + SK mean;1=OK; 2=SK  $\bar{w}$  nonstationary mean in extdrift.dat; 3=OK  $\bar{w}$  external drift
0 0 0 0 0 0 0 0         \drift: x,y,z,xx,yy,zz,xy,xz,zy (see Deutsch and Journel, 1997, p.99)
0                          \0=estimate variable;1=est. trend (or use local search, OK and all drift terms=0, to kriging trend)
extdrift.dat             \gridded file with drift variable or nonstationary mean
4                          \column number of external drift variable or nonstationary mean in extdrift.dat
2 0.02                   \number of nested variogram structures, variogram nugget
1 0.183 165.0 0.0 0.0   \model type (see p.25), sill, anis.ellipsoid horiz.angle, vertical angle, rotation angle
      87.0 44.0 0.0       \max horiz. variogram range, min horiz. range, vertical range
1 0.04 165.0 0.0 0.0    \parameters for variogram structure #2 (if needed), as above
      345.0 38.0 0.0      \ " " "

```

In the parameter file for KT3D above, the kriged output is written to the file kt3d.out; to view this file, run the GSLIB plotting program PIXELPLT.EXE, with the following parameter file. This plotting program is a general-purpose routine that is also used for viewing the output of simulation programs. PIXELPLT creates a post-script (.ps) file; in Windows Explorer, double-clicking on this .ps file will automatically open and allow you to view the kriging results in a color-graduated display that can be directly plotted. As before, parameter values you will likely need to adjust are in *italics*; non-italicized parameter values probably don't need to be altered.

Refer to Deutsch and Journal (1997, p.202-204) for detailed description of the parameter file.

Parameters for PIXELPLT

START OF PARAMETERS:

```

file.out          \output file with gridded data, from kriging, simulation or post-processing of simulations
1                \column number for variable in file.out
0  1.0e21        \data trimming limits
pixlpltE.ps      \file for PostScript output
1                \realization number (=1 for viewing kriged output; =n for the nth of multiple simulations)
70  7.5  15.0    \nx,xmn,xsiz (has to be exactly the same as the kriging grid or the simulation grid)
80  7.5  15.0    \ny,ymn,ysiz  "   "   "
1  0.5  1.0      \nz,zmn,zsiz  "   "   "
1                \slice orientation: 1=XY, 2=XZ, 3=YZ (for viewing cross-sections of a 3D grid)
1                \slice number (the nth layer in the direction perpendicular to the slice)
Walker Lake Kriging Estimate  \Title
East              \X label
North            \Y label
0                \0=arithmetic, 1=log scaling
1                \0=gray scale, 1=color scale
0                \0=continuous, 1=categorical
0.0 1100.0 110.0 \continuous data: min, max of data values, color/grey-scale increment
4                \categorical data: number of categories (used for indicator kriging or indicator simulations)
1  3  Code_One   \category number for categorical variable plot(), colorcode(), legend name()
2  1  Code_Two   "   "   "
3  6  Code_Three "   "   "
4  10 Code_Four  "   "   "

```

Color Codes for Categorical Variable Plotting:

1=red, 2=orange, 3=yellow, 4=light green, 5=green, 6=light blue,
7=dark blue, 8=violet, 9=white, 10=black, 11=purple, 12=brown,
13=pink, 14=intermediate green, 15=gray

C. Introduction to GSLIB Gaussian Simulation Program SGSIM

SGSIM performs stochastic simulations using sequential Gaussian simulation, either conditional to observed data or unconditional (only informed by variogram structure and global histogram). Note that if data clustering exists, the program requires a .dat file with the declustering weights (which you calculated in Problem Set III). As before, parameter values you will likely need to adjust are in *italics*; non-italicized parameter values probably don't need to be altered.

For a detailed description of SGSIM's parameter file, see Deutsch and Journel (1997, p. 170-174).

Parameters for SGSIM *****

START OF PARAMETERS:

```

WalkerLakeFinal.dat  \file with data (if data file not found, simulations are unconditional)
1 2 0 3 7 0         \columns for X,Y,Z, primary variable, declus.weights, secondary variable(for ext.drift)
-1.0e21  1.0e21     \trimming limits (beyond which values in the .dat file are ignored)
1                  \perform normal-score transform?0=no,1=yes, including automatic back-transformation
sgsim.trn          \file to which normal-score transform table is output
0                  \use a distribution other than the data to perform normal-score transform? (0=no, 1=yes)
histsmth.out      \file containing a reference distribution for normal-score transform
1 2               \columns containing variable and weights in histsmth.out
0.0 2000.0        \zmin,zmax(tail extrapolation)for back-trnfrm
1 0.0             \lower tail option, + parameter
2 0.5             \upper tail option, + parameter
0                 \debugging level: 0,1,2,3
sgsim.dbg         \file for debugging output
sgsim.out         \file for simulation output, to be read by POSTSIM.EXE or PIXLPLT.EXE
5                 \number of realizations to generate
70 7.5 15.0       \nx,xmn,xsiz (simulation grid parameters)
80 7.5 15.0       \ny,ymn,ysiz
1 0.5 1.0         \nz,zmn,zsiz
69069            \random number seed
0 12             \min and max number of data points in search neighborhood for simulation
12              \maximum number of previously simulated nodes in search neighborhood to use
1               \relocate data to nodes (0=no, 1=yes) ?
0 3             \multiple grid search (0=no;std spiral search), (1=yes,+ number of grids)
0              \maximum data per octant (0=not used,otherwise overrides ndmax, performs oct.search)
150.0 150.0 10.0 \maximum search ellipse radii (hmax,hmin,hvert)
165.0 0.0 0.0    \angles for search ellipsoid
0 0.60 1.0       \0=SK (default);1=OK; 2=SK, locally varying mean(LVM);3=ext.drift(ExDr); 4=CoCK
ydata.dat        \file with LVM, ExDr or CoCK variable (must be gridded same as primary variable)
4               \column in ydat.dat containing LVM, EXDR or CoCK variable
2 0.12          \number of nested variogram structures, variogram nugget
3 0.88 165.0 0.0 0.0 \model type (see p.25), sill, anis.ellipsoid's horiz.angle, vertical angle, rotation angle
262.0 132.0 10.0 \max horiz. variogram range, min horiz. range, vertical range
1 0.04 165.0 0.0 0.0 \parameters for variogram structure #2 (if needed), as above
345.0 38.0 0.0  \ " " "

```

Individual realizations in the output file from SGSIM (sgsim.out) can be plotted with PIXLPLT to view and compare the results of various individual simulations. However, in a single run, SGSIM can create a huge output file containing the results of hundreds of simulations, each of which may contain thousands of simulated values. To be able to summarize and evaluate the results of multiple simulations, the program POSTSIM is used to post-process the results of n simulations generated by SGSIM that have been written to the file sgsim.out. For the results to be most useful, no fewer than 20-30 simulations should be post-processed with POSTSIM.

Five options are provided to analyze the results of multiple simulations:

1. at each grid node calculate the mean of n simulations (the expected value or E-type estimate);
2. compute the probability of exceeding a specified threshold and the means above and below the specified threshold at each grid node;
3. for a specified cdf quantile level (probability), compute the value of the simulated variable at each grid node;
4. at each grid node, compute the range of simulated values that define an interval of specified probability about the median (assuming a symmetric cdf); or
5. calculate the variance of the n simulations at each grid node.

Note that the postsim.out file will contain 1, 2 or 3 columns of gridded data, depending on which option is chosen.

For a detailed description of the .par file parameters, see Deutsch and Journel (1997, p.239).

Parameters for POSTSIM

START OF PARAMETERS:

sgsim.out	\output file containing simulated realizations
30	\number of realizations in sgsim.out
-1.0e21 1.0e21	\trimming limits
70 80 1	\nx, ny, nz
postsim.out	\file for output array(s), to be read by PIXLPLT.EXE
2 900.0	\1 = expected value, or 5 = variance, at each grid point
	\2 = probability of exceeding a specified threshold and the mean above/below threshold
	\3 = the value at each grid node at a specified cdf quantile level
	\4 = the range of simulated values in interval of specified (assumed normal) probability

X. Description of the SGSIM parameter file

SGSIM's parameter file contains numerous options and capabilities whose use needs to be understood before using the program for routine simulations. In the description and user instructions that follow, line numbers refer to the bolded digits at the beginning of each line in the file listing given in IX. C. of the Demonstration Problem Set. For more information, see Deutsch and Journel, 1997, p. 170-174.

Parameters for SGSIM

START OF PARAMETERS:

- 1.** *WalkerLakeFinal.dat* \file with data (if data file not found, simulations are unconditional)
- 2.** 1 2 0 3 7 0 \columns for X,Y,Z, primary variable, declus.weights, secondary variable(for ext.drift)
- 3.** -1.0e21 1.0e21 \trimming limits (beyond which values in the .dat file are ignored)

The first three lines of the data file describe the data file name, if it exists (a filename has to be given, but if it is not found, then unconditional simulations are performed), and associated data parameters. As in most GSLIB programs, the column list (line #2) and the trimming limits (line #3) are standard; if Y and/or Z column numbers are set to 0, then the simulation will be 2D or 1D. Note, however, that the simulation program expects to find declustering weights (calculated by DECLUS) in a separate column; optionally, a secondary variable can be specified to be used as a drift variable.

- 4.** 1 \perform normal-score transform?0=no,1=yes, including automatic back-transformation
- 5.** *sgsim.trn* \file to which normal-score transform table is output
- 6.** 0 \use a distribution other than the data to perform normal-score transform? (0=no, 1=yes)
- 7.** *histsmth.out* \file containing a reference distribution for normal-score transform
- 8.** 1 2 \columns containing variable and weights in *histsmth.out*
- 9.** 0.0 2000.0 \zmin,zmax(tail extrapolation)for back-trnfrm
- 10.** 1 0.0 \lower tail option, + parameter
- 11.** 2 0.5 \upper tail option, + parameter

Lines #4 - 11 are used to specify normal score-transform parameters, and parameters to be used for the back-transformation after simulation is completed. Line #4, if set to 1, automatically transforms the data and back-transforms the simulated values afterward. Line#6 specifies whether a data distribution other than the data in the .dat file are to be used to define the normal-score transform; if so, that distribution is specified in file *histsmth.out* (created by GSLIB's HISTSMTH program; see Deutsch and Journel, 1997, p.214-218). The parameters listed in lines #9 - 11 specify the manner in which the back-transformation is to be made; in most simulation problems, only the maximum and minimum expected simulated values need be changed in line #9.

- 12.** 0 \debugging level: 0,1,2,3
- 14.** *sgsim.dbg* \file for debugging output
- 15.** *sgsim.out* \file for simulation output, to be read by POSTSIM.EXE or PIXLPLT.EXE

Lines #12 - 15 simply specify output options, and are self-explanatory (however, note that a debugging level of 3 will result in a HUGE .dbg file!).

16. 5 \number of realizations to generate
17. 70 7.5 15.0 \nx,xmn,xsiz (simulation grid parameters)
18. 80 7.5 15.0 \ny,ymn,ysiz
19. 1 0.5 1.0 \nz,zmn,zsiz
20. 69069 \random number seed

Lines #16 - 20 specify parameters for the simulation grid, number of simulations to be created during a single run, and the random number seed used to create the random-walk through the simulation grids.

21. 0 12 \min and max number of data points in search neighborhood for simulation
22. 12 \maximum number of previously simulated nodes in search neighborhood to use
23. 1 \relocate data to nodes (0=no, 1=yes) ?
24. 0 3 \multiple grid search (0=no;std spiral search), (1=yes,+ number of grids)
25. 0 \maximum data per octant (0=not used,otherwise overrides ndmax, performs oct.search)
26. 150.0 150.0 10.0 \maximum search ellipse radii (hmax,hmin,hvert)
27. 165.0 0.0 0.0 \angles for search ellipsoid
28. 0 0.60 1.0 \0=SK (default);1=OK; 2=SK, locally varying mean(LVM);3=ext.drift(ExDr); 4=CoCK
29. ydata.dat \file with LVM, ExDr or CoCK variable (must be gridded same as primary variable)
30. 4 \column in ydat.dat containing LVM, EXDR or CoCK variable

Lines #21 - 30 contain the kriging parameters to be used in estimating the Gaussian pdf at each grid location, and should be self-explanatory (refer to GeoEas kriging notes and problem sets for information on search strategy). A small number for previously simulated nodes (line #22) will force the pdf conditioning to be affected more by the global histogram; a large value will give more weight to the conditioning by the variogram. The option to relocate data to grid nodes (line #23) speeds execution time, by sacrificing local accuracy by moving data points away from their actual locations to grid points; for large simulations ($>10^6$ nodes) or very fine grids, data should be relocated. Line #24 is for a more complex nearest-neighbor search option and can be left as is. Lines #26, 27 specify the search ellipsoid radii (maximum and minimum horizontal length, vertical length) and the ellipsoid orientation angles (see Deutsch and Journel, 1997, p.28 for angular conventions). For most situations, simple kriging (line #28) should be specified; if a locally varying mean (LVM) or an external drift (ExDr) needs to be specified, or if co-located cokriging (CoCK) is to be utilized, additional information will need to be provided in file ydata.dat (line #29, 30). See Deutsch and Journel for details.

31. 2 0.12 \number of nested variogram structures, variogram nugget
32. 3 0.88 165.0 0.0 0.0 \model type (see p.25), sill, anis.ellipsoid's horiz.angle, vertical angle, rotation angle
33. 262.0 132.0 10.0 \max horiz. variogram range, min horiz. range, vertical range
34. 1 0.04 165.0 0.0 0.0 \parameters for variogram structure #2 (if needed), as above
35. 345.0 38.0 0.0 \ " " "

The remaining lines are used to provide the variogram correlation function model specifications. Line #31 stipulates the number of nested variogram structures in the correlation function model and the variogram nugget (note that, as in all GSLIB programs, a nugget cannot be anisotropic). As many successive pairs of lines (eg: #32 - 33; 34 - 35) as number of nested structures must follow line #31. The first number in the first line of each pair is the type of variogram model structure (1=spherical, 2=exponential, etc.; see Deutsch and Journel, 1997, p.25 for the

numbering convention). Also, refer to Deutsch and Journel (1997, p. 28) for the conventions on specifying angular orientations of variogram anisotropy.

The results of SGSIM need to post-processed with program POSTSIM, and its .par file's grid size must match the nx, ny, nz values specified in lines #17 - 19 exactly. To view the results of POSTSIM, or to view individual simulations created by SGSIM, use the following PIXELPLT .par file (note that the values specified for nx, ny, nz in this file must also exactly match those specified in SGSIM's). To view different realizations created by SGSIM, the realization number in the fifth parameter line must be changed each time PIXELPLT is run. Each time PIXELPLT is executed, it writes to the postscript file (.ps suffix) specified in the fourth parameter line, so it is recommended that this filename also be changed at each run to avoid overwriting of previously generated files.

Parameters for PIXELPLT

START OF PARAMETERS:

```

file.out          \output file with gridded data from simulation or post-processing of simulations
I                \column number for variable in file.out
0  1.0e21        \data trimming limits
pixlplt.ps       \file for PostScript output
I                \realization number (1 for viewing POSTSIM output; =n for the nth of multiple simulations)
70  7.5  15.0    \nx,xmn,xsiz (has to be exactly the same as the kriging grid or the simulation grid)
80  7.5  15.0    \ny,ymn,ysiz  "   "   "
1  0.5  1.0      \nz,zmn,zsiz  "   "   "
1                \slice orientation: 1=XY, 2=XZ, 3=YZ (for viewing cross-sections of a 3D grid)
1                \slice number (the nth layer in the direction perpendicular to the slice)
Plot Title Here (40 characters max)          \Title
East                                           \X label
North                                          \Y label
0                                               \0=arithmetic, 1=log scaling
1                                               \0=gray scale, 1=color scale
0                                               \0=continuous data (SGSIM output), 1=categorical data (indicator simulation output)
0.0 2000.0 200.0 \continuous data: min, max of data values, color/grey-scale increment
4 ----- \categorical data: number of categories (used for indicator kriging or indicator simulations)
1 3 Code_One \category number for categorical variable plot(), colorcode(), legend name()
2 1 Code_Two  " " "
3 6 Code_Three " " "
4 10 Code_Four " " "

```

Color Codes for Categorical Variable Plotting:

*— 1=red, 2=orange, 3=yellow, 4=light green, 5=green, 6=light blue,
 — 7=dark blue, 8=violet, 9=white, 10=black, 11=purple, 12=brown,
 — 13=pink, 14=intermediate green, 15=gray*

Note: the last five lines and color code list shown in italics above are not needed for viewing SGSIM (continuous variable) output.

Walker Lake Gaussian Simulation

1. Set up and run SGSIM using a .par file similar to that above, to create 10 realizations from the Walker Lake V data. Modify the PIXELPLT .par file to read SGSIM's simulation output file (sgsim.out) and to create a grey-scale (not color) plot. View the plot with GSView (double-click on the .ps file in Windows Explorer to open and view the plot), and print. Print three of the realizations and describe in what map areas they differ most; is this where you would expect a lot of variation? why?
2. Modify POSTSIM's .par file from last week's lab notes and create the E-type estimate (expected value) based on your five simulated realizations. Modify PIXELPLT's .par file to read POSTSIM's output and create a postscript plot file; print it out. Does this map differ from your earlier kriged map? explain how.
3. Repeat (2) to create a plot of the probability of exceeding the 500 ppm concentration level. Do the areas of highest probability occur where you expected them to occur?

XI. Introduction to Sequential Indicator Simulation of Categorical Variables

SISIMPDF is a stochastic simulation program that implements sequential indicator simulation, either conditional to observed data or unconditional, for categorical data. The program requires integer-coded categorical variables, their global proportions, and their indicator variograms. As before, parameter values you will likely need to adjust are in *italics*; non-italicized parameter values probably don't need to be altered.

A detailed description of SISIMPDF's parameter file is given below. This program was not included in the second edition of GSLIB but is much simpler to use than the sequential indicator simulator included in that edition. For details on the program's input file, see the first edition of Deutsch and Journel (1992, p.168-170).

Parameters for SISIMPDF

START OF PARAMETERS:

1. *WalkerLakeFinal.dat* \data file
2. *1 2 0 4* \column: x,y,z,variable
3. *-1.0e21 1.0e21* \data trimming limits
4. *sisimpdf.out* \output file for simulation results
5. *0* \debugging level: 0,1,2,3
6. *sisimpdf.dbg* \output file for debugging
7. *69069* \random number seed
8. *10* \number of simulations
9. *70 7.5 15.0* \nx,xmn,xsiz
10. *80 7.5 15.0* \ny,ymn,ysiz
11. *1 0.5 1.0* \nz,zmn,zsiz
12. *1* \0=two part search, 1=data relocated to grid nodes
14. *0* \max neighborhood data per octant (if 0, not used)
15. *400.0* \maximum radius for the neighborhood search ellipsoid
16. *15.0 0.0 0.0 0.5 1.0* \ellipsoid's horiz, vertical, tilt orientation angles, and horiz, vert. anisotropy ratios
17. *0 12* \min, max number of neighborhood data points to use for simulation
18. *12* \max number of simulated nodes in the search neighborhood to use for simulation
19. *0 2* \0=use individual indicator variograms for each category; 1=median approx(+category#)
20. *0* \0=use simple kriging (default), 1=use ordinary kriging (only if sufficient data exist)
21. *2* \number of categories to be simulated
22. *0 0.75 1 0.00* \category # , global proportion, number of nested variogram structures, nugget
23. *1 100.0 1.00* \variogram model type, range, sill
24. *90. 0.0 0.0 0.5 1.0* \variogram anisotropy orientation angles: horizontal, vertical, tilt, and anisotropy ratios
25. *1 0.25 1 0.00* \category # , global proportion, number of nested variogram structures, nugget
26. *1 100.0 1.00* \variogram model type, range, sill
27. *90. 0.0 0.0 0.5 1.0* \variogram anisotropy orientation angles: horizontal, vertical, tilt, and anisotropy ratios

Lines 1 - 6 are the standard input and output specifications common to most GSLIB programs. Lines 7 - 11 specify the random walk, number of simulations and simulation grid parameters. Lines 12-18 specify the local neighborhood search parameters; Line 20 specifies the kriging type. If Line 19 is set to 1, the category corresponding to the population median is specified and only one variogram (that for the median indicator) is specified in Lines 22-24; otherwise, N categories (Line 21) must have individual indicator variograms specified in Lines 22 and following.

Run SISMPDF on the Walker Lake T data, for the parameter values listed above. Fit a correlation model to the indicator variogram of T, using a) a spherical model and b) a gaussian model. Run SISIMpdf with the model you consider most defensible, and create an E-type estimate of 10 realizations. Compare this to the post plot of the T categorical data and your subjective view of what the T categories should look like if they represent outcrop exposure of two geologic units.