ASSIGNMENT 4

This study is designed as an extension of Assignment 3. Hence there is no need to repeat everything you said (or perhaps didn't say!) in that assignment. One strategy you might consider in writing your report for this assignment is to treat Assignment 3 as a "previous study" that will be referred to in this report. A common convention here is to give this previous work a designation, such as [A3], and to include it among you references (say [A3] Your Name (year) "Assignment 3: Cobalt Study for ESE 502"). Hence your introduction might include a brief review of what was done in [A3], and how that work will be extended here. Similarly, when you develop the analytical methods in this assignment, you might focus on how they extend or refine the analyses in [A3]. This will necessarily involve some degree of repetition, but should help you to concentrate on the *new* elements in this assignment.

(1) In this study you will examine **Manganese** (**Mn**) deposits in the same region of Vancouver Island as in Assignment 3. The objective is to interpolate this data using **Ordinary Kriging** in MATLAB and to compare the result with ARCMAP.

(a) First you will construct a visual representation of the data as follows:

a. In ARCMAP open the map document **Cobalt_2.mxd** constructed in your home directory, say **S:\home\Cobalt_2**, in Assignment 3. Now save this document under the new name **Manganese.mxd**. You will use this new document from now on.

b. Rename the data frame as "Manganese Region", and rename the "Cobalt Data" layer as the "Manganese Data" layer.

c. To display the **Manganese data** use the procedure in Problem 1 [Part (a).6] in Assignment 3, now applied to the layer "Manganese Data". In the "Quantities" window set **Value** = "MN". All else is the same.

d. For later use, it is convenient to set the map and display units to meters. To do so, right click on "Manganese Region", select **Properties > General** and under **Units** set **Map** = "meters" and **Display** = "meters".

2. Next you will examine the frequency distribution of Manganese:

a. Right click on "Manganese Data" layer and open the attribute table.

b. Now right click on the "MN" column and select "Statistics".

c. You will then see the summary statistics for Manganese together with its frequency distribution (histogram).

d. Notice that this distribution is skewed to the right, and appears to be very non-normal in shape. (Change the **Field** to "CO" and you will see that this data is much more skewed that the Cobalt data.) This non-normality cannot be analyzed further in ARCMAP, so you will now export this data to JMP for further analysis.

(b) The desired data can be exported to JMP as follows:

1. Open EXCEL, and click: **File > Open**, and navigate to your home directory, **S:\home\Cobalt_2**.

2. Set **Files of Type** = "dBase Files (*.dbf)", and open the file **geo_dat.dbf**. (This data base file is one component of the shape file for **geo_dat**.)

3. You should now see a data set that looks essentially like the Attribute Table in ARCMAP.

4. Save this file to your home directory as **Manganese.xlsx**.

5. Now you can open the file directly in **JMP**.

6. Save this file to your home directory as **Manganese.jmp**. You will then be able to save the edits.

   a. Rename the columns "EASTING" and "NORTHING" as "X" and "Y", respectively.

   b. Remove all columns except "X", "Y", and "MN" (**Cols > Delete Cols**).

   c. Save these edits.

(c) Now open the frequency distribution for Manganese (**Analyze > Distribution > MN**) and you will see (a vertical version of) the same frequency distribution in ARCMAP.

1. Check for normality by right clicking on the "MN" title bar in the "Distribution" window and selecting **Normal Quantile Plot**. Comment on the normality of this data.

2. Next try taking logs to remove skewness. (Make a new column "lnMN", right click on the column heading and click **Formula > Transcendental > Log**)

3. Again examine the **Normal Quantile Plot** of this data and comment. On the histogram for MN you will see two (rather suspicious) outliers in this data, which suggest some type of local anomalies at these sites. Try removing them.

a. Click on these outliers in the **Box Plot** to the right of the frequency plot, and you will see their row numbers.

b. Select these two rows in the data table (**ctrl-click**) and remove them from the analysis (**Rows > Hide and Exclude**).

4. Now try the **Normal Quantile Plot** once again with these two points removed, and comment on the results.

   a. Click on these two rows again, and remove the outliers completely (**Rows > Delete Rows**). [There should now be 284 rows.]

   b. This **reduced data set** will be used for the rest of the analysis.

(d) Before proceeding, it is necessary to make the corresponding changes in ARCMAP, so that this **reduced data set** can be viewed properly.

1. To do so, you must first construct the appropriate **log transformation** of the Manganese data in ARCMAP. Start by opening the attribute table for "Manganese Data" in **Manganese.mxd**, and at the top of the table clicking: **Table Options > Add Field**.

   a. Set **Name** = "lnMN", **Type** = "double", **Precision** = 6 and **Scale** = 3. Click **OK**.

   b. Now right click on the "lnMN" heading and select **Field Calculator.** Ignore the warning about "calculating outside of an edit session" and click **Yes**. (If you make any mistakes, simply right click on "lnMN", select **Delete Field**, and start over.)

   c. In the "Field Calculator" window first select "Log()" from **Functions**, and then select "MN" from **Fields**. You should now have the expression "Log([MN])" in the calculator. Click **OK**, and you should now see values for the log of Manganese in the column.

2. Next you will eliminate the two outliers by creating a subset of the data, as in done in Assignment 1 for the Lymphoma data.

   a. First reopen the attribute table for "Manganese Data", right click on the "MN" column and select **Sort Ascending**.

   b. You will now see the two low-level outliers with values ("2" and "5") which are dramatically smaller than the rest.

c. To eliminate these values, first select them (by pressing **ctrl** and then clicking on the tabs at the left end of their rows).

d. Now at the top of the table click: **Table Options > Switch Selection**. You will now see that all points except these two are selected. (This is a VERY useful tool!)

e. Finally, proceed as in Problem 1 [Part (a).4] of Assignment 1 to create a layer containing **only** these points.

f. Before proceeding it is useful to create a new **shape file** containing only these points (for certain procedures can only be carried out on shapefiles, and not selections from shapefiles.)

(i) To do so, right click on the new layer "Manganese Data Selection" and click: **Data > Export Data** . Save the new file as **ed_data.shp**. [Note again that for later purposes of Kriging this data, it is advisable that **all path names** (including your directory names) contain **at most eight characters**. Otherwise you may get error messages during Kriging telling you that the file cannot be opened.]

(ii) Now add this shape file to the data frame, and name the new layer as "Log-Manganese Data". You can adjust the color ramp and symbol choices as before. (Also, you can remove the layer "Manganese Data Selection", since it will no longer be used.)

(e) Given this new data representation in ARCMAP, you will now Krige this data in MATLAB.

1. First you will export the data from JMP to MATLAB. (NOTE: exporting data from ARCMAP directly to MATLAB is not yet possible. You must first create an appropriate text version of the data, and that is most easily done in JMP or EXCEL.)

a. In JMP, save the file **Manganese.jmp** to your home directory as a "Text Export File", **Manganese.txt**.

(i) Click "Options" and set **End of Field** = "Tab".

(ii) Also *unselect* **Export Table Headers**.

b. Open MATLAB and set the path to the class MATLAB directory. (Also on the Main Menu, open **Preferences > Command Window** to be sure that **Numeric Format** is set to "short g"). Now on the Main Menu, click **Import Data**, and load the file **Manganese.txt** from your home directory.

Set **Output Type** = 'Numerical Matrix", click **Import Selection**, and you should now see this file in the workspace as a 284x4 matrix, **Manganese**.

(i) If the file is not there, open **Manganese.txt** in NOTEPAD and look to see that it has a clean matrix format. If not, either clean it up in NOTEPAD, or try re-exporting it from JMP.

(ii) Save this workspace to your home directory as **Manganese.mat**. You will use this workspace for all MATLAB calculations.

2. The next task is to estimate a **spherical variogram** (using **var_spher_plot**) that will serve as an input to the Kriging procedure. To apply **var_spher_plot** to the Manganese data you must first delete the third column containing MN-data. To do so in MATLAB, make a submatrix with the command:

   » Mn = Manganese(:,[1:2,4]);

   (Here "**[1:2,4]"** denotes a vector consisting of the 1st, 2nd, and 4th columns. **Mn** should appear in the Workspace as a 284x3 matrix).

3. Now estimate a spherical variogram using the default "$h_{\max}/2$" value for max distance:

   » var_spher_plot(Mn);

4. You should receive an error message similar to the following:

   Exiting: Maximum number of function evaluations has been exceeded
   - increase MaxFunEvals option.
   Current function value: 0.021881

   VAR_FIT: convergence not obtained in 324 iterations

   ****
   New estimation done with RANGE truncated to MAX DISTANCE

   (Depending on data, you may wish to reset max distance)

   ****

   Be sure to hit **Enter** again to display the resulting (derived) **Covariogram Plot**, so that you can complete the program. [Note also from the screen output that the **default max distance** used by the program was 37559 (= $h_{\max}/2$).]

5. Now try a somewhat larger max distance with the following command:

   » opts.maxdist = 46000; OUT = var_spher_plot(Mn,opts);

The above error message should no longer be present. What do you think has caused this problem? (HINT: Think about why the range was truncated, and why the error message suggested that you increase the max distance.)

6. The above output, **OUT**, is a MATLAB **cell structure**, in which the first cell contains the parameters of the estimated spherical variogram. To use this data type:

   » p = OUT{1}

   By leaving off the semicolon, **p** is shown on the screen to be a 3x1 parameter vector consisting of the estimated **range**, **sill**, and **nugget** values.

(f) Using these parameters, you will now **Krige** the (log) Manganese values in MATLAB as follows:

1. First open the program **o_krige** (Ordinary Kriging) and examine the required inputs for this program. You can construct the **lnMN** values, **y**, and data locations, **L0**, as follows.

   » y = Mn(:,3); L0 = Mn(:,1:2);

   For illustrative purposes, you will only determine the Krige prediction at a single location, **L**, defined by writing:

   » L = [612300, 579700];

   Be sure to save your workspace (in the Main Menu use **Save Workspace As**) in order to avoid having to reconstruct this data.

2. To determine a reasonable bandwidth for this prediction, use the procedure of Study 1 [Part (d).10] in Assignment 3 to locate the point **L** in the Manganese Region (in ARCMAP). [Be sure that the layer "Manganese Data" is turned off, so that only the new layer "Log-Manganese Data" is displayed on the map.] By using the "Measure" tool, you will see that a bandwidth of 4900 meters is just big enough to include the seven closest neighbors of point **L** (in a crescent shape around **L**). These will be taken to define the *relevant prediction* set for point **L**.

3. An appropriate Krige prediction in MATLAB is thus given by the command:

   » OUT = o_krige(y,L0,L,4900,p)

   The predicted **lnMN** level at **L** and the standard error, **std_err**, for this prediction can be displayed as follows:

» lnMN = OUT{3}

» std_err = OUT{4}

If you have run the program correctly, you should obtain **lnMN = 7.186** and **std_err = 0.4103**.

4. In order to check the reasonableness of this prediction, you can use the "Identify" tool in ARCMAP to determine the **lnMN** levels at each of the seven neighbors of **L** used in the Krige. To see these seven values in a better way:

   a. Select all seven using the **Select Features** tool (holding down **Shift**).

   b. Now open the attribute table for "Log-Manganese Data", then right click on the "LnMN" column heading, and select **Statistics**.

   c. You will then see the distribution of just these **seven points**, along with their **mean** and **standard deviation**.

   d. Does the Kriged value at **L** look reasonable compared to the mean of interpolating data set?

   e. Compare the **standard deviation** of the seven prediction points with the estimated **std_err** at **L**. How might you account for the big difference?

   (**NOTE:** Be sure to *clear* this seven-point selection before continuing.)

5. Finally, to translate this prediction into practical terms, use the results above to determine a **95% prediction interval** for the value of **MN** (not **lnMN**) at the point **L** [recall expressions (6.2.82) through (6.2.85) in the NOTEBOOK].

   a. How does this prediction interval compare with the range of Manganese values of the seven predictors?

   b. How might you account for the difference?

(g) For purposes of comparison, it is instructive to perform the same Krige in ARCMAP.

1. To Krige this data using **Geostatistical Analyst**, use the *same* procedure as for the Cobalt data in Problem 1 [Part (c)] of Assignment 3, with the following modifications

a. In the Kriging window set **Source Data** = "Log-Manganese Data" and **Data Field** = "LnMN". Be sure **Methods** = "Kriging/CoKriging".

b. In the **Step 2** window select **Kriging Type** = "Ordinary" and **Output Type** = "Prediction"

c. In the **Step 3** window, set **Lag Size** = 4600 and **Lags** = 10 [so (4600)x(10) = 46000 yields the same max distance used in the MATLAB estimate of the variogram]. The estimation may take a few seconds to complete. Click **Next**.

d. In the **Step 4** window, click on the "1 Sector" (hollow circle) icon under **Sector Type** (so you will have a simple isotropic krige prediction). Next set **Maximum Neighbors** = 7 and again use 5000 meters for Major and Minor semiaxes.

e. Before exiting this window, you can obtain the prediction at the point **L =** [612300, 579700] in the Test Location window. This will give you an exact comparison with the MATLAB krige prediction for **L** obtained above.

f. Click **Finish** and **OK**, and the krige will now appear in the Menu as "Kriging". Rename as "Ordinary Kriging", and drag the Mask above this output to improve the appearance. If you like, you can also change the color scheme by opening "Properties" for this display, and double clicking on the selected item, "Filled Contours".

g. It is important to note that this file is "live" and is linked directly to **Geostatistical Analyst**. (For example, if you click on **Properties** → **Extent**, you can change the extent of the display, and it will be instantly recalculated.) If you want to save this as a permanent file independent of **Geostatistical Analyst**, there are two options (neither of which is completely satisfactory).

   (i) Probably the best option is to save as a raster file by right clicking on the file name "Ordinary Kriging", and then clicking **Data > Export to raster**. You can save the file as "geo_grid", and include it in your map document. To display the map click "Properties" (ignore any warning about the large number of unique values by clicking **OK**) and in the "Layer Properties" window set **Show** = "Classified". You can then choose a number of classes (usually 10 is about right) and an appropriate color ramp. Notice that the map display is much less smooth than the original display. This is because the colors displayed are closer to the *actual kriged values* plotted in each pixel. Since the set of points used for the interpolation changes discontinuously over

space, the values appear as a "patchwork". The smoother version in the "live" display is actually somewhat less accurate in terms of the color ranges shown.

(ii) An alternative is to save as a vector file by clicking **Data > Export to vector**. You can save the file as "geo_vector", and include it in your map document. Here the map display shows the smoothed contours of the original kriged display. You can assign Max or Min Values to each contour band under **Properties > Symbology**, and can assign a color ramp to these values. However, while this smoothed version gives a nicer visual representation of the general kriged surface, all the detailed pixel information is *lost*. So while **Geostatistical Analyst** gives all this information in the same display, it *cannot* be saved as a single independent file.

2. To compare this result with MATLAB,

   a. First use the "Identify" tool to verify that Krige predictions very close the point **L** on the display of "Ordinary Kriging" are essentially the same as the kriged value calculated in (g).1.d above.

   b. Next to compare the standard errors, again right click on the "Ordinary Kriging" layer and this time select "Method Properties". You will see that this allows you to scroll back and reset any of the kriging options. In particular, rather than choosing "Prediction" in Step 1, now choose "Prediction Standard Error", and proceed to the **Step 3** window. With **Predicted Value** set to **L** as before, you should now be able to read the standard error value as **Value**.

   c. How do these values compare with those obtained in MATLAB?

(2) In this study you will apply the Geostatistical Regression procedure in the MATLAB program **geo_regr** to obtain better temperature estimates for the "South American Climate" data analyzed in Problem 2 of Assignment 3 [A3].

(a) First recall from Problem 2 [Part (b)] in Assignment 3 that even though the fit of the quadratic regression is superior to that of the linear regression, there remains a significant degree of spatial autocorrelation in the residuals. To analyze these residuals in more detail, it is convenient to examine them **spatially**.

1. First open **S_amer.jmp** and by using the procedures in Problem 2 [Part (c).2] of Assignment 3, make a new file by deleting all columns except (**X**, **Y**, **Res_2**), and save as an EXCEL file, **Res_2.xls** (or **Res_2.cvs** – do *not* use **.xlsx**).

2. Next, open the map document **s_amer.mxd** in ARCMAP and then (as in Problem 2 [Part d.(iii)] of Assignment 3) add **Res_2.xls** and display it (using **Display XY Data…**). The new layer **Res_2.xls Events** will now appear in the Data Frame. [There is no need to make a shape file here since the display is all that is needed.]

3. Finally, color the residual data using, say "red-to-blue" for "positive-to-negative" residuals (with "white" for the interval containing zero) and examine the results. Do you see any pattern to these residuals? In particular, how might you explain the coastal values?

(b) Next you will analyze these same residuals in MATLAB by constructing the associated **variogram** (and derived **covariogram**).

1. Start by opening the MATLAB workspace **s_amer.mat**. Use the workspace browser to observe that this workspace contains the "temperature" data in the vector **y** and "explanatory variable" data (Y,XY,YY) in the matrix **X**. There is also a matrix **L** containing only the coordinate data (X, Y). You will now add the residual data, **Res_2.xls**, to this workspace:

   a. In MATLAB, load this file (**File > Import Data**) using the same procedure as for Manganese.txt above. Rename the 76x3 matrix to **Res_2**.

   b. Now save the workspace to your home directory, in order to keep this data. [Each time you add data that you want to keep, you must resave the workspace in your home directory.]

2. The matrix **Res_2** provides the necessary data inputs to **var_spher_plot**. To gain better perspective on the structure of these residuals, it is instructive to use a larger max distance than $h_{\max}/2$ (which is about 4185 km). So use the command:

   » opts.maxdist = 8000; var_spher_plot(Res_2,opts);

3. Now focusing on the derived **covariogram** for these residuals, observe that it decreases (in classical form) up to about distance 2500 km, and then rises again. Comment on why you think this might be so.

   a. To relate this to the map of South America, use the **Measure** tool in ARCMAP to check distance ranges visually. In particular, if you open the **Measure** tool on the main menu, and draw a line across South America at roughly its widest latitude, this should yield a value of about 5000 km. In terms of these observations, how do you think the *shape* of South America might be affecting the covariogram results?

b. Now redo the derived covariogram estimation above with a max distance of **4000** (roughly $h_{\max}/2$). By comparing the two covariogram plots, discuss the effect of this reduction in max distance.

c. Notice that there continues to be an error message:

****
New estimation done with NUGGET truncated to ZERO
****

This is telling you that the (unconstrained) nonlinear least-squares procedure produced a nugget estimate that was *negative*. [Note that such violations of nonnegativity are always possible in least-squares estimation.]

d. Here it is of interest to compare this result with **Geostatistical Analyst** by kriging the data set **Res_2 Events** (above) using the "ordinary krige" option. [This item is for your information only, and should **not** be included in your report.] Just proceed up to the (semi)variogram estimation (**Step 3**) and reset the lag size and number to 400 and 10, respectively, to yield a total max distance of 4000. [As usual, we focus on the (semi) variogram and *not* the covariogram in ArcMap.] Observe that the parameter estimates here yield a *positive* nugget. But now look at the lower end of the variogram plot, say from $h = 0$ to $h = 1000\,km$ (note that the graph units here are $h \cdot 10^{-3}$). Does this look like a good fit to you? If you mentally force the nugget to zero (i.e., imagine pulling the end of the curve down to zero), does this look like a better fit? What is happening here is that the empirical variogram data shown (by red dots) in the figure is *not* the data used to fit this curve. In fact the program fits a *different* empirical variogram (based on what is called "geometric binning"). This is intended to give "better performance" for small distances, but in my view only serves to confuse the user. [More generally, you should be wary of hidden computations and smoothing adjustments used in commercial software like **Geostatistical Analyst.**]

(c) Finally, you will apply **geo_regr** to this data, utilizing the observations above.

1. These observations suggest that a reasonable max distance for the covariogram is about 4000 km. So, by using the **vnames** vector already constructed in **s_amer.mat**, fit this data with the command:

   » opts.maxdist = 4000; OUT = geo_regr(y,X,L,vnames,opts);

2. How do the results of this regression (in particular, the p-values) compare with the original OLS regression in Assignment 3? Discuss possible reasons for these differences.

3. Next, open the program **geo_regr**, and observe that the output is a 5x1 cell structure, with the third cell, **OUT{3}**, containing the final regression residuals ($y - X\hat{\beta}$) and the fourth cell, **OUT{4}**, containing the corresponding residuals ($T^{-1}y - T^{-1}X\hat{\beta}$) for the transformed problem (where T is the Cholesky matrix for the estimated covariance matrix).

4. Save these residual vectors to the workspace as follows:

   » Res = OUT{3};
   » Res_T = OUT{4};

   and also save **Res_T** as a text file in your home directory:

   » save S:\home\Res_T.txt Res_T -ascii

5. Given these residual vectors, next examine their estimated covariogram using **var_spher_plot** (where it is assumed that **opts.maxdist = 4000**)

   » M1 = [L,Res];
   » var_spher_plot(M1,opts);
   » M2 = [L,Res_T];
   » var_spher_plot(M2,opts);

   a. The parameters for the **Res** variogram should correspond to those in the screen output from **geo_regr** above. Interpret these results in terms of the original statistical model used for Geostatistical Regression, i.e. the linear model

   $$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \Sigma)$$

   with covariance matrix, $\Sigma = (\sigma_{ij})$, generated by a spherical variogram, γ, with unknown parameters (r, s, a) as follows:

   $$\sigma_{ij} = C(h_{ij}) = C(\|s_i - s_j\|) \quad \text{where} \quad C(h) = s - \gamma(h; r, s, a), \; h \geq 0$$

   How does this covariogram compare with that of the OLS residuals from Assignment 3?

   b. Next consider the variogram for **Res_T**. How can you account for the difference between these covariograms in terms of the above model?

6. Finally, open the text file, **Res_T.txt** in JMP and copy-and-paste it as a new column "Res_T" in **S_amer.jmp** (which should already be in your home directory).

   a. Make a new column "nn_Res_T" using the same procedure as before, and use **Fit Y by X** to regress **Res_T** on **nn_Res_T**.

   b. How does this new result compare with that for the regression of **Res_2** on **nn_Res_2** ? How does it relate to the covariogram for **Res_T** just constructed in MATLAB?

   c. What do all these residual analyses tell you about **spatial autocorrelation** in the above statistical model for South American temperatures?