

## ASSIGNMENT 4

(Due on Tuesday, March 31)

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 your 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, **e:\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 **Values** = “MN”. All else is the same.
  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 than the Cobalt data.) This non-normality

cannot be analyzed further in ARCMAP, so you will now export this data to JMPIN for further analysis.

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

1. Open EXCEL, and click: **File** → **Open**, and navigate to your home directory, **e:\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 exactly like the Attribute Table in ARCMAP (minus the first two columns).
4. Save this file to your home directory as **Manganese.txt** using the option “Text (Tab delimited)(\*.txt)”.
5. Now you can open this file in **JMPIN** using the “Text Import Preview” option, and then the “Delimited” option.
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 → Exclude**).
4. Now try the **Normal Quantile Plot** once again with these two point removed, and comment on the results.
- a. Click on these two row 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 bottom of the table clicking: **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 **Calculate Values...** 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 bottom of the table click: **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 JMPIN 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 JMPIN or EXCEL.)
    - a. In JMPIN, 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 set **Export Column Names** = “No”.

- b. Open MATLAB and set the path to the class MATLAB directory. (Also, use **File** → **Preferences** to change Numerical Format to **short g**.) Now click: **File** → **Import Data**, and load the file **Manganese.txt** from your home directory. You should now see this file in the workspace as a 284x4 matrix. [You need only load “data”. So you can uncheck “textdata” and “colheaders”, both of which contain the column labels.]
- (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 JMPIN.
  - (ii) Right click on **data** in the Workspace and rename it as **Manganese**.
  - (iii) 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 1<sup>st</sup>, 2<sup>nd</sup>, and 4<sup>th</sup> columns. **Mn** should appear in the Workspace as a 284x3 matrix).

3. Now estimate a spherical variogram using the default “Dmax/2” bandwidth:

» **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 BANDWIDTH**

**(Depending on data, you may wish to increase bandwidth)**

\*\*\*\*\*

Note also from the screen output that the **default bandwidth** used by the program was 37559 (= Dmax/2).

5. Now try a somewhat larger bandwidth 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 bandwidth.)

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**, 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 (**File** → **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 Problem 1 [Part (c).6] 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 (= “unknown units”) is just big enough to include the seven closest neighbors of point **L** (in a crescent shape around **L**).

3. Hence an appropriate Krige prediction in MATLAB is 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 = 7.321** 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?
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**.
  - a. How does this prediction interval compare with the Manganese values the seven predictors? To make a “fair” comparison here, it is appropriate to compare the *prediction interval* for a new sample based on this data, with the prediction interval calculated above.
    - (i) Recall that if the mean and standard deviation of a data set,  $(x_i : i = 1, \dots, n)$ , are denoted by  $\bar{x}_n$  and  $s_n$ , then a  $100(1 - \alpha)\%$  *prediction interval* for any new sample,  $X$ , from this estimated population is given by (see for example the explanation in *Wikipedia*, [http://en.wikipedia.org/wiki/Prediction\\_interval](http://en.wikipedia.org/wiki/Prediction_interval) )

$$\bar{x}_n \pm t_{\alpha/2, n-1} s_n \sqrt{1 + \frac{1}{n}}$$

where for comparison with the kriged prediction interval, we set

$\alpha = .05$  . [Remember that this is *wider* than the standard *confidence interval for the mean*, given by  $\bar{x}_n \pm t_{\alpha/2, n-1} s_n / \sqrt{n}$  ]

(ii) The above *t-value* can be obtained from MATLAB by writing

» **alpha = .05; n = 7; t = abs(tinv(alpha/2,n-1));**

(iii) In making your comparison, remember also that the normal approximation,  $t_{\alpha/2, \infty} = z_{\alpha/2} = 1.96$ , was implicitly used in the kriged prediction interval. So this value already yields a “tighter” interval than it really should.

b. What does this suggest to you about the possible shortcomings of using log transformations to achieve normality?

(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 **Input Data** = “Log-Manganese Data” and **Attribute** = “LnMN”. Be sure **Methods** = “Kriging”.
  - b. In the **Step 1** window select **Ordinary Kriging → Prediction Map**.
  - c. In the **Step 2** window, set **Lag Size** = 460 and **Lags** = 100 [so (460)x(100) = 46000 with about the same number of bins as in the MATLAB estimate of the variogram]. The estimation may take a few seconds to complete. Click **Next**.
  - d. In the “Searching Neighborhood” window, click on the first (hollow circle) icon next to **Shape Type** (so you will have a simple isotropic krige prediction). Next set **Neighbors to Include** = 7.
  - 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 “Ordinary Kriging”. 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 reset any of the kriging options. In particular, rather than choosing “Prediction Map” in “Geostatistical Methods”, now choose “Prediction Standard Error Map”, and proceed to the “Step 3” window. With “Test Location” set to **L** as before, you should now be able to read the standard error value as **Error**.

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 in JMPIN with only the columns (**X**, **Y**, **Res\_2**). Now open NOTEPAD as before, and make a text file **Res\_2.txt** with column headings (**X**, **Y**, **Res\_2**). Now make a copy of the file (copy and paste to the same directory) and rename this second file with a new extension, **Res\_2.tab**. [If you are in the Towne Lab, it may not be necessary to do this. Try adding it as in (2) below to see if it works. Open it to be sure. If you have difficulty adding the file to ARCMAP or renaming the file extension, you can find a more detailed discussion in Section 1.2.1 of Part IV of the NOTEBOOK on importing text files to ARCMAP.]

2. Next, open the map document **s\_amer.mxd** in ARCMAP and then (as in Problem 2 [Part (c).3] of Assignment 3) add **Res\_2.tab** and display it (using **Display XY Data...**). The new layer **Res\_2.tab 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, 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**.

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 (**X**, **Y**, **XY**, **Y<sup>2</sup>**) 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.txt**, 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, **data**, 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 variogram of these residuals, it is instructive to use a larger bandwidth than  $D_{max}/2$  (which is about 5). So use the command:

```
» opts.maxdist = 10; var_spher_plot(Res_2,opts);
```

3. You will see that the variogram has the classic shape up to about distance 4, and then declines. Comment on why you think this might be so.
  - a. To gain some insight here, use the “Measure” tool in ARCMAP to check distance ranges visually. Here you must change map units (the X and Y units in the data set turn out to be in Decimal Degrees):
    - (i) Open the **Properties** window for the active data frame containing **Res\_2** and in the click on the tab “General”. In the “Units” window, set **Display** = “Decimal Degrees”. Click **OK**.
    - (ii) If you use the **Identify** tool to click on any map data point, you will now see that the (X,Y) values agree with the Decimal Degree coordinates at the bottom of the screen. This means that the Measure tool will now yield distances in units agreeing with (X,Y).
  - b. In terms of these units, how do you think the *shape* of South America might be affecting the variogram results?
  - c. Now redo the variogram estimation above with a bandwidth of 4. By comparing the two variogram plots, discuss the effect of this reduction in bandwidth.
  - d. 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*. (This is always possible in least-squares estimation.) To see that a zero nugget is actually reasonable in this case, notice that you can kriging the data set **Res\_2.tab Events** using the Geostatistical Analyst in ARCMAP. Just proceed up to the variogram estimation (Step 2 of 4) and reset the lag size and number to .4 and 10, respectively, to yield a total bandwidth of 4. Observe that the parameter estimates are now roughly the same as in MATLAB, with a nugget of about 0.02. The smoothing procedure used there has not quite produced a zero nugget. But if you change the variogram model to “exponential” you will see that the nugget is now exactly zero (i.e., that the internal nonnegativity constraint is in force).

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

1. These observations suggest that a reasonable bandwidth for the variogram is about 4. So by using the **vnames** vector already constructed in **s\_amer.mat**, fit this data with the command:

```
» opts.maxdist = 4; 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 ( $A^{-1}y - A^{-1}X\hat{\beta}$ ) for the transformed problem.
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 variogram fits using **var\_spher\_plot** (where it is assumed that **opts.maxdist = 4**)

```
» 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 Geo\_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,  $\gamma$ , with unknown parameters  $(r, s, a)$  as follows:

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

How does this variogram 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 variograms in terms of the above model?
6. Finally, open the text file, **Res\_T.txt** in JMPIN 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 variogram 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?