ASSIGNMENT 1

Before beginning this assignment, look at Part IV (SOFTWARE) of the class NOTEBOOK. In particular, look at the three sections: "Opening JMP", "Opening ARCMAP", and "Opening MATLAB". These give you general instructions on how to access the software for the class and set up appropriate paths to the class directory inside the software. Also, read over the **Example Assignment** and **Example Answer** (especially the introductory paragraph). Your answer to each problem below should present a short but *self-contained* study of the data set presented.

(1) This first study is an extension of Exercise 1.5 of B&G, based on a set of *Burkitt's Lymphoma* data from the West Nile region in Uganda. The key source material for this study is the paper by Williams et al. on "Burkitt's Lymphoma" in the **Reference Materials**. You can find additional material on the web with key words such as "Burkitt's Lymphoma" and "West Nile".

(a) In ARCMAP open the file **Lymphoma.mxb** in the class directory (**Course Data Directories (T:)\ese502\arcview\ Projects\Lymphoma\Lymphoma.mxd**). Before doing anything else, save this file to your home directory (say **S:\home**). Now you will be able to save your work to your home directory. As you will see from the reference material, Burkitt's Lymphoma is most common among children. To study this population, you are going to construct [in a manner similar to the logical (indicator) variable used in Exercise 1.5] a selection of those cases with ages from 5 to 10, i.e., with $5 \leq \text{Age} \leq 10$. To do so:

1. Right click on the layer 'lymph_data' in the *Table of Contents* (window on the left) and open the *Attribute Table*.

2. In the upper left, click on **Table Options → Select by Attributes**.

3. In the text box type: **("AGE" >=5) AND ("AGE" <=10)**

   (You can also use the menu items to do this). Now click **Save** and save this calculation as a file **5_to_10.exp** in your home directory. Then click **Apply** And close both calculator window and the attribute table. You should now see the selected points on the map.

4. Next you will make a new layer containing only these data points. To do so, right click the layer 'lymph_data' and click: **Selection → Create Layer from Selected Features**. A new layer will now appear in the Table of Contents. If you remove the check on the layer 'lymph_data' you will see that only the selected points appear on the map. [You can rename this new layer as '5-to-10 Cases' by right clicking the layer, and then clicking **Properties → General**. Also you can change the map symbol by left-clicking on the symbol in the

Table of Contents.] Once you are satisfied with the looks of this new layer, be SURE to save **Lymphoma.mxb** again, so that your work is not lost.

5. Next you will make a layer containing only those points **not** in the '5-to-10' cases. To do so, right click on layer 'lymph_data' and again open the Attribute Table. As before, click **Options → Select by Attributes**. Now click **Load** and reload the file **5_to_10.exp** from your home directory. The calculation will appear in the window. Click **Apply** and **Close**. The selections will appear in the attribute table. Now click **Table Options → Switch Selection**. Those points not in '5-to-10' will now be selected. Repeat step 4 to create a new layer 'All other Cases'.

6. To compare these layers visually, it is more appropriate to use population density as the "reference" layer. To do so, remove the check for the 'W_NILE_BD' layer, so that the 'pop_density' layer underneath should now be visible. (You should also choose a color for the dots that *contrasts* with the pop_density map, such as 'bright red'.)

7. Finally, copy the maps for these layers into **WORD** [these notes are for **WORD 2016**] for a visual comparison.

   (1) To do so, display only the '5-to-10 Cases' layer on the map and in the main menu click: **File → Export Map**. Save the file to your home directory in Enhanced Metafile (**.emf**) format.

   (2) Now open WORD and click **Insert → Picture → From File**. The picture should now appear in the WORD document. (Depending on the state of the Network, this may take a while.)

   (3) When you right click on the picture that appears, the "Format" menu will automatically open in the Main Menu. Click the **Crop** option on the Format Menu, and crop the frame sides tightly around the map (by placing the mouse on the square for each side and dragging the side). Click outside the image to disengage the **Crop** tool.

   (4) Now click the picture again and drag the corners to achieve the image size you want on the page.

   (5) Finally, to position the picture, click **Wrap Text > Behind Text**and the picture can now be dragged into position.

8. Now repeat the procedure in step 6 for the layer 'All other Cases'. You can arrange them side by side and label them using **Insert > Text box > Draw Text Box** to label them as '5-to-10 Cases' and 'All other Cases'.

9. Finally, compare these two patterns visually.

    1. How do the patterns of points relate to the population density?

    2. Are there any clear differences between the spatial pattern of '5-to10 Cases' and 'All other Cases'?

(b) The remaining part of this study will use both MATLAB and JMP. You are going to look for possible 'spatial diffusion' effects in these cancer cases.

1. First, open MATLAB and be sure that the path (on the top of the window) is set to the class directory (**T:\ese502\matlab**) using the browser button. Then open the data file **lymphoma.mat** in the Workspace. (Check the Workspace window to be sure it was loaded.) There you will see that **L** is a 188x2 matrix containing the coordinates (x_i, y_i) of the $i = 1,..,188$ case locations in West Nile. You will now construct the nearest neighbors to each of these points using my program **neighbors.m** written in MATLAB. Type: **» edit neighbors**, and you will see the program document which explains the INPUTS and OUTPUTS of the program. (You can also simply click on the file in the MATLAB directory to open it.) To use this program for our purposes, type the command:

   » out = neighbors(L,1);

   (If you get an error message telling you that the program cannot be found, be sure that you have set the path to the **class directory** correctly.) This will compute the **first** nearest neighbor to each point in **L**. [Don't forget the semicolon at the end. Otherwise the entire matrix **out** will be dumped onto the screen!] The first column of the output matrix, **out**, contains these row numbers. So they can be put in a column vector, **neigh**, by typing

   » neigh = out(:,1);

   Now save this data to your home directory, **S:\home**, as an ascii (.txt) file by typing:

   » save S:\home\neigh.txt neigh –ascii

   (NOTE: Don't copy-paste this command into Matlab. The WORD version uses different fonts and may give you strange error messages.)

2. Next you will import this data vector into JMP and use it to check for diffusion effects. To do so, open JMP and open the data file, **lymphoma.jmp**. (You will see that the data is the same as the Attribute file for the layer 'lymph_data' in ARCMAP.) First save this file to your home directory so that you can edit it. You are going to add **neigh** as a new column in this data set. To do so, click: **File → Open → S:\home\neigh.txt**. Before opening this file in JMP (assumed to be **JMP Pro 14** for these notes) left click on the file and notice that there are a number of "Open as" options at the bottom of the screen. By far the best option (which in my view should be the "default" option) is "Data, using best guess". If you click on this option and then click **OK**, the file should now open in JMP. To copy-and-paste this column into **lymphoma.jmp**, left click on the column heading, so that the entire column is selected (should be dark gray in color). In the main menu click: **Edit → Copy**. Now bring up **lymphoma.jmp** (click: **Window → Lymphoma**), add a new column to the table (**Col → New Column** [Name = **Neigh**] → **OK**), and (with the new column selected) click **Edit → Paste** .

3. You are now ready to do some analysis in JMP. First create a second new column **Neigh_Time** which is to contain the onset time (time of occurrence) of the nearest neighbor of each case. To do so, right click the column heading and click: **Formula**. In the formula window click: **Time**. You will subscript this variable to pick out the nearest-neighbor time value by selected **Row** in the left column and clicking: **Subscript → Neigh**. The entry in the window should now have the form:

Click **Apply** and **OK**. The first row of this new column should have a value 3500, corresponding to the **Time** entry for row 88 (the nearest neighbor to row 1).

4. You are now ready to compare the onset times for each case with the onset time for its nearest neighbor. (If these are similar, then this is evidence of a possible spatial diffusion effect.) To so do, use nearest-neighbor onset time to predict onset time by a simple regression: First click **Analyze → Fit Y by X**. Then click **Time → Y,Response** and **Neigh_Time → X,Factor**, and **OK**.

5. You will now see a point scatter of the two variables, which shows little monotonic relation at all. To verify this, right click on the bar 'Bivariate Fit of

Time by Neigh_Time' and click **Fit Line**. This will display the desired regression output.

6. Given this regression output, evaluate the existence of a diffusion relation in terms of the **R-Square** value for this regression, and the **P-Value** for the slope coefficient (denoted by **Prob > |t|**). What do these values mean, and how might this particular regression result be explained? In particular, how does it relate to the findings of the "Burkitt's Lymphoma" paper?

(2) This second study is an extension of Exercise 1.7 of B&G (see also pp.256-257 for further discussion, and p.290 for the data source.) You can find additional information on the web, with key words such as "1992 Presidential Election".

(a) Open the file **Clinton_data.jmp** in JMP, and save it to your home directory. You are going to construct an 'Ark_dist' variable describing the distance of every state centroid to Arkansas. (*Centroids* are discussed in Section 2.2.1 in Part III of the class NOTEBOOK.) To do so:

1. Click: **Col** $\rightarrow$ **New Column**, and label the new column as **Ark_dist**.

2. Right click on the column heading and click **Formula**.

3. Observe that Arkansas is the third row in the table, and that the centroid coordinates are columns **X** and **Y**. So the centroid coordinates for Arkansas are (X_3, Y_3). Hence for any state, with centroid coordinates (X, Y), the centroid distance to Arkansas is given by the formula:

4. $$\sqrt{(X - X_3)^2 + (Y - Y_3)^2}$$

   (use **Row** $\rightarrow$ **Subscript** $\rightarrow$ **3** to make the subscript "3").

5. If you have done the formula correctly, the centroid distance to Alabama (first row of column **Ark_dist**) should be 67.36.

(b) Next you will compare the percent votes for Clinton with the distance of each state from Arkansas. To do so:

1. Run a simple regression of **Perc_Clint** on **Ark_dist**.

2. Using **R-Square** and the **P-value** on the beta coefficient for Ark_dist, determine whether there is any significant relation. Is your conclusion at all surprising?

(c) Finally, you will see how these National results compare with those in the immediate neighborhood of Arkansas. This is most easily done in ARCMAP:

1. In ARCMAP open the file **Clinton.mxd** in the class directory (**T:\sys502\ arcview\Projects\United States\Clinton.mxd**). Again, save it to your home directory.

2. Activate the data frame **1992 Presidential Election** (if not already activated).

3. On the layout toolbar in the Main Menu click on the **Select Features** tool, and using the mouse (with the Shift key held down) select Arkansas, together with its six neighbors (Texas, Louisiana, Mississippi, Tennessee, Missouri, Oklahoma).

4. Now right click on the layer 'Percent for Clinton' in the Table of Contents, and click: **Selection → Create Layer from Selected Features**. Label the new layer as 'Arkansas Neighborhood'.

5. Open the attribute table for Arkansas Neighborhood, and right click on the column **Perc_Clint**. Select **Statistics** and observe the mean percentage. Compare this with the overall mean (for the same column in the Layer 'National Voting Data'.) What can you conclude from this?

6. Finally, try the same comparison with Arkansas removed. To do so, open the attribute table for 'Arkansas Neighborhood' and select 'Arkansas'. On the bottom of the window, select: **Options → Switch Selection**. Now all the surrounding states should be selected. With this selection active, again click on the column **Perc_Clint**, and select **Statistics**. Determine whether this make a difference in the above comparison between means, and interpret your findings.