

## ASSIGNMENT 6

(Due on Tuesday, April 28)

(1) In this study you will local  $G^*$ -statistics analysis to the Irish Blood Group data, which is displayed in the ArcMap file **F:\sys502\arcview\projects\eire\eire.mxd**. This study is discussed in Bailey and Gatrell (p.253) and in the Upton-Singleton section of the class BULKPACK (pp.267-276). Additional background material has been provided in the above directory, **F:\sys502\arcview\projects\eire**:

- First, the original study of Irish blood groups by Dawson (1964) is provided in the Reference Material on the class web page (**Irish\_Blood\_Groups.pdf**). Two more recent studies (**Blood\_Group\_paper\_1.pdf** and **Blood\_Group\_paper\_2.pdf**) are also included in the directory above, **F:\sys502\arcview\projects\eire**.
- Also, the key reference by Freeman (1969) in which the “Pale” counties were established is included in the file **Irish\_Pale.pdf** in the Reference Material. Additional material can be obtained by Googling “Irish Pale”, etc.
- Finally the original reference on  $G$ -statistics by Getis and Ord (1992) is included in the Reference Material as **G\_Statistics.pdf**. This should provide additional background material on the use of these statistics. (*NOTE: Our present application of the  $G^*$  statistic is more general than their original application to radial distance matrices.*)

The analysis of this data will be carried out in MATLAB and then exported to ARCMAP.

- (a) First, reproduce this map document in your home directory by creating a directory, say **S:\home\eire**, and copying the three files **Eire(.shp,.dbf,.shx)** from the class directory **F:\sys502\arcview\projects\eire**. You can then display the blood group data as is done in **eire.mxd** and make a map document file of your own. Name the data frame as “Eire Data” and the blood group layer as “Blood Group A”. Save this map document in your own directory as **eire.mxd**.
- (b) Open MATLAB and import the file **F:\sys502\matlab\eire.txt**. You should now have a  $26 \times 7$  matrix, **eire**, in the workspace.

1. Define the *centroid locations*, **L**, and *blood-group percentages*, **z**, by:

```
» L = eire(:,1:2);
» z = eire(:,3);
```

2. Now make an exponential weight matrix by writing:

```
» info.type = [4,10,1];
» info.norm = 1;
» W = dist_wts(L,info);
```

Here the argument **4** denotes the “exponential-weight-matrix” option, **10** denotes the exponent value, and **1** denotes the “include-diagonal-terms” option.

(c) Next, you will perform a **random-permutation test** using local  $G^*$ -statistics and export the results to ARCMAP:

1. To perform this test in MATLAB using spatial data, **z**, and weight matrix, **W**, with 999 random permutations, write:

```
» GP = g_perm_loc(z,W,999);
```

2. The (26 x 2) output matrix, **GP**, contains the  **$G^*$ -statistics** together with their associated **P-values** for each of the 26 counties in Eire.
3. You are going to join this data table to the attribute table for “Blood Group A”. This requires that the two tables have a common identifier. Since every attribute table in ARCMAP has an identifier “FID” that numbers all rows starting with the value “0”, the simplest procedure is to add a column to **GP** that is exactly of this form. To do so, simply write:

```
» GP(:,3) = [0:25]';
```

Here the prime ( ' ) transforms the row vector **[0:25]** to a column vector.

4. To export this data to your home directory as a file **G\_stats**, write:

```
» save S:\home\G_stats.txt GP -ascii
```

5. To import this data into ARCMAP, you must first convert **G\_stats.txt** into proper format using the procedure in (7) of Assignment 5. Here the three column titles of the new file, **G\_stats.tab**, should be **G\_star**, **P\_Val**, and **ID**, respectively.

(d) With your map document **eire.mxd** open in ARCMAP, use **File** → **Add Data** to bring **G\_stats.tab** into your document. It should now appear in the Table of Contents. To join this file to the attribute file for the layer “Blood Group A”:

1. Right click on “Blood Group A” and select **Joins and Relates** → **Join**.
2. In the **Join** window, set (1) = “FID”, (2) = “G\_stats.tab”, (3) = “ID”, and click **OK** to finish.

3. If you open the attribute table for “Blood Group A” you will now see that new data appears in the last three fields.
  4. To make this join permanent, right click on “Blood Group A” and select **Data** → **Export Data**. Now add this modified data to the map document as a new shape file, **eire\_g.shp**. Before proceeding, it is a good idea to remove the join on “Blood Group A”:
    - (i) Right click on “Blood Group A” and then select **Joins and Relates** → **Remove Join(s)**.
    - (j) Then select **Remove All Joins** and close. This will return the attribute table to its original state.
  5. Rename the new layer from “eire\_g.shp” to “P-Values”. Open the properties of “P-values” and navigate to the **Symbology** window and select **Quantities**.
  6. In the “Fields” box set **Value** = “P\_Val”, and in the “Classification” box, set **Classes** = 5 and click on **Classify**.
  7. In the Classification window which opens, set **Method** = “Manual” and on the right hand side set the **Break Values** to be (0.002, 0.01, 0.05, 0.10, 0.999). Click **OK**.
  8. Set an appropriate **Color Ramp**, and exit with **Apply** and **OK**.
- (e) Finally, using the procedure in part (1).(a).6 of Assignment 1, display the maps for layers “P-values” and “Blood Group A” side by side in WORD for comparison. (Include this display in your report.)
1. By comparing the distribution of Blood Group A percentages with the P-values obtained for each county, comment on whether these “concentration” results are what you would expect.
  2. Look at the specific county of Waterford (FID 22), which is part of the original *Pale* defined by Freeman. How does its *Blood Group A percentage* compare with the mean value for *Eire*? What is its ranking among all counties in *Eire*? Now look at its associated *P-value*. How might you explain this apparent discrepancy?
  3. Finally, take a look at the counties of Kildare (FID 8) and Meath (FID 16), both of which were also part of the *Pale*. First compare their *Blood Group A percentages*, and then compare their *P-values*. How might you account for this apparent discrepancy?

(2) In this study you will lay the groundwork for a spatial regression analysis of **median housing values** in Philadelphia. This analysis will be continued in the final assignment (Assignment 7). A great deal of interesting information on the Philadelphia housing market can be found by Googling topics like “Philadelphia Housing Market”. Just to get you started, I have included several items:

- The first is a set of summary profiles of trends in Philadelphia housing prices compiled by Kevin Gillen. This is included in the Reference Materials as **Gillen\_Housing\_Indices.pdf**.
- In addition, I have included in the directory, **F:\sys502\arcview\projects\Phil\_Housing**, an article, **Phila\_Trends.pdf**, containing a brief discussion of recent housing market and demographic trends in Philadelphia.
- Also, there is a longer paper, **Philadelphia\_Housing\_Submarkets.pdf**, on the relation between traditional Philadelphia neighborhoods and housing submarkets.
- Finally, for those who may be interested, I have included a more detailed discussion, **Phila\_Nbhd\_Initiative.pdf**, of the *Neighborhood Transformation Initiative* (NTI) which is mentioned in the references above.

The relevant census tract data for Philadelphia is displayed in the ARCMAP file **F:\sys502\arcview\projects\Phil\_Housing\Phil\_Housing.mxd**. To analyze this data, first copy-and-paste all “**tracts90**” files into a directory of your own, say **S:\home\Phil\_Housing**. Now open a new project in ARCMAP and use **Add Data** to add the shape file **tracts90** to the data frame. You will observe that the default projection for Philadelphia looks wrong. Also, if you open the Attribute Table and scroll to the last two columns, you will see that the coordinates for this map are in decimal degrees. Hence to obtain a more natural projection in terms of Euclidean distance (in feet), your first task is to *project* this shapefile into State Plane Feet. This two-step procedure requires you to *define* and *construct* the projection. To do so:

(a) First open **ArcToolbox** on the main menu, and click:

**Data Management Tools → Projections and Transformations  
→ Raster  
→ Define Projection**

(Note that even though we do not have a raster file, projection definitions are located under “Raster” for some reason!)

(b) In the window that opens, set **Input Data Set** = “tracts90”. For the **Coordinate System** click the Browse icon to the right, and in the “Spatial Reference Properties” window that opens, click **Select...** and then proceed to

**Geographic Coordinate Systems  
→ North America  
→ North American Datum 1983.prj**

Click **Add** and **OK**, and the desired definition should now appear in the **Coordinate System** window. Click **OK**. When the process is completed, click **Close**.

(c) Next you will *construct* the projection. To do so return to **ArcToolbox** and click:

**Data Management Tools** → **Projections and Transformations**  
→ **Feature**  
→ **Project**

(d) In the **Project** window that opens, again set **Input Data Set** = “tracts90”, and name the new file, **tracts90\_feet**. For the **Output Coordinate System** click the Browse icon, and in the “Spatial Reference Properties” now click **Select...** and proceed to

**Projected Coordinate Systems**  
→ **State Plane** → **NAD 1983 (Feet)**  
→ **NAD 1983 StatePlane Pennsylvania South FIPS 3702 (Feet).prj**

As before, click **Add** and **OK**, and the desired definition should now appear in the **Output Coordinate System** window. Click **OK**. When the process is completed, click **Close**.

(e) To see the new projection, first close this map file (no need to save). Now open a new map file and add your file, **tracts90\_feet**. The projection should now look correct. Notice also that the default coordinates on the bottom of the map are now in *feet*.

(f) Your final task here is to recalculate the coordinates in terms of feet. To do so, open the Attribute Table for **tracts90\_feet** and scroll to the last two columns, which list the coordinates in the old decimal degree units. You will now recalculate these using the small programs in the utility sub-directory of the class directory as follows:

- First right click on the **X\_coord** column and select **Calculate Values...**
- In the “Field Calculator” select **Load**, and in the class sub-directory, **F:\sys502\arcview\projects\utilities**, open **X-centroid.cal**.
- Click **OK**, and when the calculation is complete, you should now see that the coordinate values are now in feet rather than decimal degrees.
- Now repeat this procedure for the **Y\_coord** column, using the file **Y-centroid.cal**.

Finally, rename the data frame as **Philadelphia Tract Data (1990)**, and save this project in your home directory as **Phil\_Housing.mxd**. You are now ready to begin!

(a) But before proceeding to the analysis, there is some additional “data cleaning” that needs to be done. Open the attribute table for the layer “tracts90” and sort the column MEDIANVAL by right clicking on the column label and choosing the “Sort Ascending” option. You will see that there are 14 census tracts with MEDIANVAL = 0, which either involve missing data or zero housing units. These tracts should be eliminated from the analysis. [You will also see a large number of tracts with MEDIANVAL = 14999. The reason for this is that the U.S. Census convention is to report all lower housing values as \$14999. So be aware that this will necessarily create some bias in the later analysis. But since this Census reporting convention is standard, we shall keep these tracts in the analysis.]

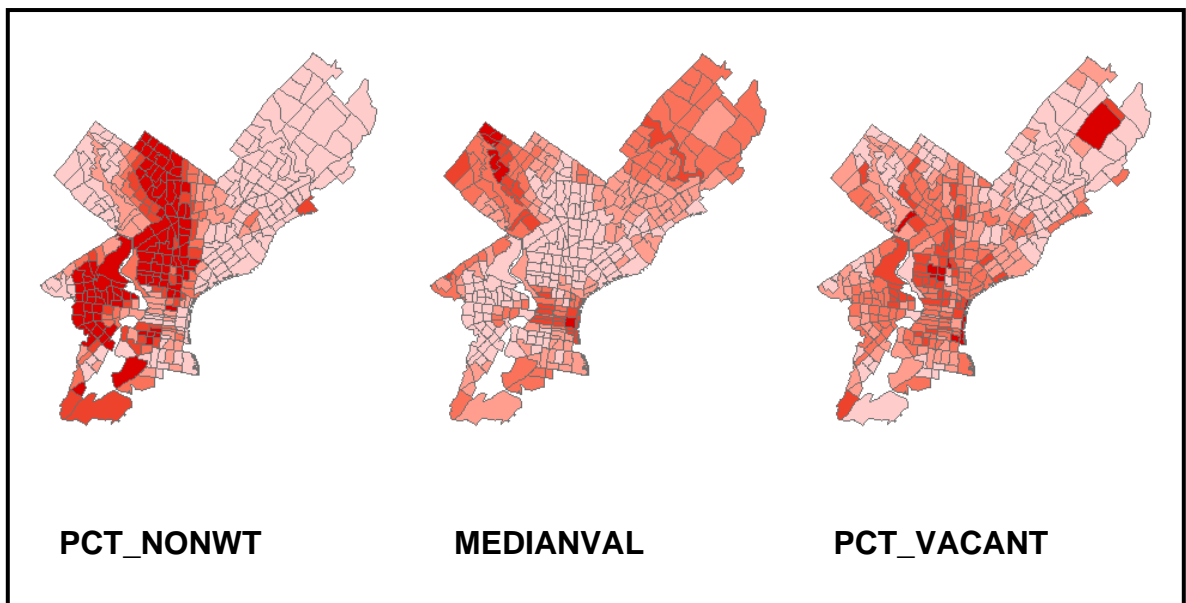
1. To eliminate these 14 tracts, use the “Select by Attributes” option [as in part (1).(a).5 of Assignment 1] to create a new layer containing only those census tracts actually containing houses (“MEDIANVAL” > 0).
2. Make a new shape file **Final\_Tracts.shp** and add it to the data frame. Before proceeding, open the attribute table of this new layer and check on the bottom to be sure that you now have 353 records. (You can also remove the original layer since it will no longer be used.)
3. Name this layer as “Median Housing Value” layer, and in the **Symbology** window for this layer, display the MEDIANVAL data on the map, with an appropriate color ramp. This will be the key dependent variable for the regression analysis.

(b) Next you will construct two possible explanatory variables, namely, the **percent of vacant housing** and the **percent of nonwhite population** in each census tract.

1. To do so, open the attribute table for “Median Housing Value” layer and observe that there are two columns “HOUSUNITS” and “VACANT”. The first desired quantity “PCT\_VACANT” is constructible in terms of these two quantities as  $100 \times (\text{VACANT}/\text{HOUSUNITS})$ .
  - (a) To construct this quantity use **Options** → **Add Field** to create a new column (**Name** = “PCT\_VACANT”, **Type** = “double”, **Precision** = “6”, **Scale** = “3”). [Here **Precision** denotes the maximum number of digits to be used, and **Scale** denotes the number of decimal places.]
  - (b) When the new field appears, right click on the column head and select **Calculate Values**.
  - (c) Finally, in the **Field Calculator** window for “PCT\_VACANT = ” you should write:  $100 * [\text{VACANT}] / [\text{HOUSUNITS}]$ . (The first value in the column should be 6.732.)
2. To construct the second explanatory variable, observe that in the attribute table there are also columns “PERSONS” and “WHITE”. So to make a new column

“PCT\_NONWT”, proceed as part 1 above by writing  $100 * ([\text{PERSONS}] - [\text{WHITE}]) / [\text{PERSONS}]$  in the “PCT\_NONWT = ” window. (The first value in the column should be 3.792.)

3. Now you are ready to display these new variables as maps.
  - (a) First, use **Copy** and **Paste Layer** [as in part (c).3.b of Problem 2 in Assignment 3] to make two copies of “Median Housing Prices” layer in the data frame.
  - (b) Rename the first new layer as “Percent of Vacant Housing” and the second as “Percent NonWhite”.
  - (c) In the Symbology window for “Percent of Vacant Housing”, select the PCT\_VACANT variable for display, and choose the same color ramp as for the “Median Housing Value” layer.
  - (d) Repeat this process for the “Percent NonWhite” layer, and display the PCT\_NONWT variable.
4. Next, construct a visual comparison of these three layers by successively copying each map to WORD, and resizing them so that they can be displayed side by side. Finally, label the variable displayed in each figure using **text boxes**. If you have done all steps correctly, you obtain a final display similar to the one shown below (which uses a red color ramp for all three maps):



(c) By comparing these three figures visually:

1. What can you say about the expected signs of correlations between **MEDIANVAL** and each of the explanatory variables?
2. What can you say about possible spatial autocorrelation between the values of each variable across census tracts?

(d) Next you will export the data to JMPIN for analysis.

1. To do so, first open the data base file **Final\_Tracts.dbf** in EXCEL, and remove all columns except “**MEDIANVAL**”, “**X-COORD**”, “**Y-COORD**”, “**PCT\_VACANT**”, and “**PCT\_NONWT**”. Save the file as a tab delimited text file, **Phil\_Housing.txt**.
2. Now open this file in JMPIN (using **Text Import Preview**), and save the file in your own directory as **Phil\_Housing.jmp**.
  - (i) For convenience, rename the columns as **MV**, **X**, **Y**, **%Vac**, and **%NW**.
  - (j) For later use, reorder these columns by moving the two centroid coordinate columns, **X** and **Y**, to the left of **MV** (using **Cols → Reorder Columns → Move Selected Columns**).
  - (k) Finally, check to be sure that you have exactly 353 rows of data. If you have extra rows (such as repeats of row 353 or zero-value rows) delete these from the table. [Sometimes data transfers create such errors.]
3. Display the frequency distribution of median values, **MV** (using **Analyze → Distribution → MV**), and observe that these values are highly skewed. To remove this effect, take logs:
  - (i) First make a new column **lnMV** just to the right of **MV** (using **Cols → Add Multiple Columns**).
  - (j) Then right click on **lnMV** and use **Formula → Transcendental → Log** to construct the natural log of **MV**.
  - (k) Comment on your result by comparing the Normal Quantile Plots of these two distributions.
4. Next examine the frequency distributions of the two explanatory variables, **%Vac** and **%NW**.

- (i) Notice first that **%Vac** is also somewhat skewed. However, we cannot take logs without excluding rows (95,164,353), which contain (informative!) zero values. So we shall leave this variable as is. [Note that one could also use the transformation,  $\log\{1 + \%Vac\}$ , which works quite well.]
- (j) Next observe the dramatic *bimodal* nature of the **%NW** distribution. How might you interpret this finding? [To see how this distribution might be “flattened” a bit, try a **logit** transformation,  $\log\{\%NW/(100-\%NW)\}$ .]

[**NOTE:** It is *not* essential that explanatory variables be normally distributed. For example, dummy variables can never be “normal”. What *is* important, however, is to have a good spread of values (good “leverage”) in order to help identify slope parameters. So for the dummy-variable case, one would ideally like to have a reasonable balance between one’s and zero’s.]

5. You are now ready to do a multiple regression of these variables.

- (i) Regress **lnMV** on **%Vac** and **%NW** (using **Analyze → Fit Model**).
- (j) Comment on the results in terms of **Adjusted R-Square** for the whole model, as well as the **P-Values** for each variable coefficient.
- (k) Notice in the **Residual Plot** (at the bottom of the regression output window) that there is a concentration of **lnMV Predicted** values at both ends.
  - a. To gain further insight, save these predicted values in JMPIN (right click on the top bar, **Response lnMV**, and then use **Save Columns → Predicted Values**).
  - b. Now plot these predicted values against **%NW** and discuss how this relation helps to explain the above pattern of **lnMV Predicted** values.
- (l) Before proceeding, right click on the top bar (**Response lnMV**) in the regression window, and use **Save Columns → Residuals** to save the regression residuals as a new column, which you should label as **Res**. Also delete the column, **Predicted lnMV**, which will not be used further.

(e) Next you will export the data to MATLAB and construct a **spatial weight matrix**:

1. Save the JMPIN file as a Text Export File, **Phil\_matrix.txt**. To do so, open the **Options** window and:
  - (i) Set **Export Column Names** = “No”.

- (j) Also set **End of Field** = “Tab”.
2. Now open MATLAB and load this text file into the workspace. Check the workspace to be sure that **Phil\_matrix** now appears as a  $353 \times 7$  matrix. (If it is not there, open **Phil\_matrix.txt** in NOTEPAD and check the first line to be sure the labels are gone, and the last line to be sure that no extra spaces have been added at the end.) Save the workspace to your home directory as **Phil\_Housing.mat**.
3. Recall from the column labels in **Phil\_Housing.jmp** that the first two columns (**X,Y**) are the coordinates of the centroids of each census tract. To construct a symmetric nearest-neighbor weight matrix from these centroids:

(i) Make a matrix of centroid locations, **L**, by writing:

```
» L = Phil_matrix(:,1:2);
```

(j) Next make the desired weight matrix, **W**, using option 2 in **dist\_wts.m** :

```
» info.type = 2;
» info.norm = 1;
» W = dist_wts(L,info);
```

(k) Recalling that the regression residuals, **Res**, were in the last column of **Phil\_matrix**, construct the vector of symmetric nearest-neighbor residuals, **nn\_res**, by writing:

```
» res = Phil_matrix(:,7);
» nn_res = W*res;
```

Before leaving MATLAB, test for spatial autocorrelation in **res** by using **sac\_perm**:

```
» sac_perm(res,W,999);
```

You will compare these results with a JMPIN analysis below.

(l) Finally, export these weighted residuals back to JMPIN by writing:

```
» save 'e:\home\nn_res.txt' nn_res -ascii
```

Before analyzing these residuals, note that it is **not** possible to obtain this same result by using the procedure in part (a).3 of Problem (2) in Assignment 3. Nearest neighbors are **not** the same as *symmetric* nearest neighbors.

- a. To see this, examine the weight matrix **W** in MATLAB by writing: » **find(W(275,:))** . This will produce the column numbers of the **nonzero** components in row 275 of **W**. You should see **three** column numbers, which by definition identify the *symmetric* nearest neighbors of the tract in row 275.
  - b. Recall that that FID numbers in ARCMAP attribute tables start counting from “0” rather than “1”, so that row 275 corresponds to tract FID number 274 in ARCMAP. Similarly by subtracting “1” from each column number above, you will obtain the FID numbers of each symmetric nearest neighbor tracts.
  - c. Open ARCMAP and examine these tracts. Do the results of MATLAB make sense in terms of these tract locations? Comment.
4. With **Phil\_Housing.jmp** open in JMPIN, import **nn\_res.txt** to JMPIN.
- (i) Copy-and-paste this data into **Phil\_Housing.jmp** as a new column, **nn\_Res**.
  - (j) Notice that weighted residuals in rows 77 and 78 are *identical*. How is this possible? (Look at the relative locations of these tracts in ARCMAP.)
5. Finally, use **Analyze → Fit Y by X** to regress **Res** on **nn\_Res**.
- (i) What does the **P-value** on **nn\_Res** tell you about the multiple regression above?
  - (j) How do these results relate to the permutation test done in part 3.(k) above? Do they support one another?
  - (k) How does the value of **rho** in the output of **sac\_perm** relate to the present slope estimate for **nn\_Res**? Why? (You might also try using **Analyze → Fit Model**, and repeat the regression of **Res** on **nn\_Res** using the **No Intercept** option at the bottom of the Fit Model window.)
  - (l) What additional information about this regression result is added by the output of **sac\_perm**? (Compare the value of **rho** with the simulated *range* of values for **rho** reported by **sac\_perm**).
  - (m) What do you expect to happen if you redo this analysis using **spatial autoregression (SAR)**? [You will do so in the final assignment for this course.]