Online Supplement to *GIS and Public Health, 2nd Edition*
Ellen K. Cromley and Sara L. McLafferty

GIS AND PUBLIC HEALTH EXERCISE 10 – GEOGRAPHICALLY WEIGHTED REGRESSION (ArcGIS 9.3.1)

PREPARATION

Download the **exer10** folder you will need for this exercise from the online supplement.

All of the databases and files used in the exercise will be stored in various subfolders within the folder called **exer10**. The following instructions are written for this folder to be located on the **c:** drive. If the folder is located on another drive, the path names shown below should be modified accordingly. Some of the folders are empty. They have been included because you may need to save the results of an operation to one of these folders.

The map documents created using ArcGIS 9.3.1 reference the spatial databases and tables in the application based on the directories and paths where the data are stored. Changing the locations of databases in the system can prevent a GIS application from working properly.

Connecting to the Exercise Folder

Go to **Start** ⇒ **Programs** ⇒ **ArcGIS** ⇒ **ArcCatalog** to start ArcCatalog.

Find the button labeled **Connect to Folder** and click the button. Navigate to **c:\exer10** then click OK and look at the Catalog tree in the left window to see that the folder has been added.

Within the data folder, data can be organized in folders identifying the agency that produced the data and then by the format of the data. For these exercises, you will consider yourself to be working for the organization called “agency” that is creating the GIS.

As you work through the exercises, you will be retrieving data from and saving data to specific folders. Please make sure you understand the System Design for the exercises.

Use the **File** ⇒ **Exit** menu to close ArcCatalog.

GEOGRAPHICALLY WEIGHTED REGRESSION

Go to **Start** ⇒ **Programs** ⇒ **ArcGIS** ⇒ **ArcMap** to start ArcMap.

In the “ArcMap Start using ArcMap with” window, click the radio button labeled “A new empty map” and then click OK.

Rename the Layers data frame by right clicking the word Layers and selecting the **Properties** item in the menu. Then select the **General** tab and enter the name GWR. Click OK. The name of the Data Frame in the Table of Contents window should now appear as GWR.

In this exercise, you will be exploring patterns of very low weight births for counties in Illinois. Data on the 1998-2002 rate of very low birth weights (weights of less than 1,500 grams) per 10,000 births by county of residence were obtained from a report by the Illinois Department of Public Health (Birth Defects and Other Adverse Pregnancy Outcomes in Illinois 1998-2002: A Report on County-Specific Incidence, Illinois Department of Public Health, May, 2004). Data from the 2000 Census Summary File 3 on each county have been downloaded from the Census web site for analysis as variables that might explain very low birth weight. You will be using the GIS software to perform geographically weighted regression using these data.

Add the Database of Very Low Birth Weight Rates

To begin, add a database of very low birth weight rates for counties of Illinois.

Find the button labeled **Add Data** and click the button. You should find the **c:\exer10** folder in your catalog. If not, please connect to the folder using the **Connect to Folder** button. Navigate to **c:\exer10\data\agency\shapes** and add the **vlbw.shp** shapefile. The coordinate system is NAD_1983_StatePlane_Illinois_West_FIPS_1201_Feet. Map units are feet.

Right-click the vlbw data layer and select **Open Attribute Table** from the menu. There are 102 counties in Illinois. The VLBW9802 field provides the 5-year (1998-2002) very low birth weight rate per 10,000 births. The FIPS field provides the 5-digit state and county FIPS codes for each county. The state FIPS code for Illinois is 17. The NAME field shows the county name. The PERBLACK field shows the percent of the population identifying as Black or African American only, the PBELPOV field shows the percent of individuals for whom poverty status is determined who are below the poverty level, the FPTLSHS shows the percent of females 18 to 24 who have less than a high school education, and PFORBORN field shows the percent of the total population who were foreign born.

Close the table.

Use the **Save** button or go to **File** ⇒ **Save** to save your map document. Navigate to **c:\exer10\mapdocs** and save the file as **exer10.mxd**.

Symbolize the Data Layer

Right click the vlbw layer and select **Properties** from the pull-down menu. Then click the **Symbology** tab. Under “Show:” select **Quantities** ⇒ **Graduated colors**. Under “Fields:” in the “Value:” section, select VLBW9802 from the pull-down menu as the field containing the values to be mapped. Under “Classification” use the Natural Breaks (Jenks) classification and set the number of classes to 5. Select a color ramp and then click OK.

Examine the pattern of very low birth weight rates by county.

Repeat this for each of the independent variables to visualize the spatial patterns in the data.

Save the map document.

Develop the GWR Model

Click the **Show/Hide ArcToolbox Window** button to open the “ArcToolbox Window.”

Double click **Spatial Statistics Tools** ⇒ **Modeling Spatial Relationships** ⇒ **Geographically Weighted Regression** to open the “Geographically Weighted Regression” window.

Select vlbw as the “Input feature class” from the pull-down.

Select VLBW9802 as the “Dependent variable” from the pull-down.

Select PERBLACK, PBELPOV, FPTLSHS, and PFORBORN from the pull-down and add them as “Explanatory variable(s)” to the model.

Set the “Output feature class” to:

c:\exer10\data\agency\shapes\GeographicallyWeightedRegression1.shp

Under “Kernel weight” select Adaptive from the pull-down menu. This allows the kernel to vary in extent as a function of county density.

Under “Bandwidth method” select BANDWIDTH PARAMETER to use a fixed number of neighboring counties as the bandwidth parameter.

Under “Number of neighbors (optional)” enter 20.

Click OK.

The “Geographically Weighted Regression” window will be updated to show the execution and results of the analysis. This information is also written to a supplementary table along with summary information about model variables and parameters which will automatically be added to the Data Frame when you close the window. When the regression has executed successfully, close the window.

Save the map document.

Explore the Results

You will see that a new data layer has been added to the Data Frame. This layer GeographicallyWeightedRegression1 is automatically symbolized based on the StdResid field of Standardized Residuals using Standard Deviation as the classification method.

You will also see that the supplementary table, GeographicallyWeightedRegression1_supp has also been added to the Data Frame. You can see the table if you click the **Source** tab at the bottom of the Table of Contents.

First, explore the table.

Right click the GeographicallyWeightedRegression1_supp table and select **Open** from the menu.

The value for the number of neighbors is 20.

The ResidualSquares is the sum of the squared residuals of the model. The smaller this measure is, the closer the fit of the GWR model to the observed data.

The EffectiveNumber is related to the choice of bandwidth. As the bandwidth increases, the coefficient estimates will become closer to those for a global Ordinary Least Squares regression model.

Sigma is the square root of the normalized residual sum of squares. Smaller values of this statistic are preferable.

AICc is a measure of model performance. Models with lower values are preferred.

R2 is a measure of goodness of fit. Values range from 0.0 to 1.0 and higher values are better.

R2Adjusted compensates for the number of variables in a model by taking into account degrees of freedom in the numerator and denominator. In GWR, the effective number of degrees of freedom is related to bandwidth. For this reason, AICc is preferred as a measure for comparing different models.

Close the table.

Next, explore the GeographicallyWeightedRegression1 data layer by right-clicking the layer and selecting **Open Attribute Table** from the menu.

In addition to regression residuals seen in the Data Frame, the table includes fields for observed and predicted values of the dependent variable, condition number, local R2, residuals, and explanatory variable coefficients and standard errors.

The condition number in the Cond field is a diagnostic measure for evaluating local collinearity. Results associated with condition numbers larger than 30 may be unreliable. Right click the Cond field header and select **Sort Descending** from the menu. The condition value is larger than 30 for only 4 of the 102 counties. Use the cursor to select these 4 records in the table. The counties will be highlighted on the map. They are all adjacent and in the northwest border area of the state. Then, click the **Options** button and select **Clear Selection** to clear the selections.

Local R2 values range between 0.0 and 1.0 and indicate how well the local regression model fits observed dependent variable values. Very low values indicate that the local model is not performing well.

The Predicted field gives the estimated very low birth weight rate values computed by GWR.

The Intercept and independent variable parameter estimates are given for each record followed by the Residual and StdError fields. The residuals are obtained by subtracting the estimated independent variable values from the observed values of very low birth weight rates. The standardized residuals have a mean of zero and a standard deviation of 1. The map of these values is automatically added to the Data Frame when GWR executes successfully.

The remaining fields include fields showing the standard errors for each coefficient. Confidence in the parameter estimates is higher when the standard errors are small in relation to the coefficient values. Large standard errors may indicate local collinearity.

Close the table.

Mapping GWR Local R2 and Parameter Estimates

First, map the Local R2 values to see where the GWR predicts well and where it predicts poorly. Right click the GeographicallyWeightedRegression1 layer, select **Properties** from the menu and click the **Symbology** tab. Under “Show:” select **Quantities** ⇒ **Graduated colors**. Under “Fields:” in the “Value:” section, select LocalR2 from the pull-down menu as the field containing the values to be mapped. Under “Classification” use the Natural Breaks (Jenks) classification and set the number of classes to 5. Select a color ramp and then click OK.

Examine the pattern of Local R2 values across the state. The model yields higher Local R2 values for counties in the northwestern areas of the state with values declining to the southeast.

Next, map the parameter coefficient values and parameter standard errors for the PERBLACK variable. Right click the GeographicallyWeightedRegression1 layer, select **Properties** from the menu and click the **Symbology** tab. Under “Show:” select **Quantities** ⇒ **Graduated colors**. Under “Fields:” in the “Value:” section, select C1_PERBLAC from the pull-down menu as the field containing the values to be mapped. This field gives the parameter values for the first variable PERBLACK. Under “Classification” click the Classify button and select Natural Breaks (Jenks) and set the number of classes to 5. Then click OK. Select a color ramp and then click OK.

You should see that the coefficients for the PERBLACK variable are positive and higher in the northwestern counties of the state and in some southeastern counties. Remember that some coefficients might not be significant.

Next, right click the GeographicallyWeightedRegression1 layer, select **Properties** from the menu and click the **Symbology** tab. Under “Show:” select **Quantities** ⇒ **Graduated colors**. Under “Fields:” in the “Value:” section, select StdErrC1_P from the pull-down menu as the field containing the values to be mapped. This field gives the standard error for the first variable PERBLACK. Under “Classification” click the Classify button and select Natural Breaks (Jenks) and set the number of classes to 5. Then click OK. Select a color ramp and then click OK.

You should see that the standard error values for the PERBLACK variable are lowest in the counties in southern tip of the state and higher in groups of counties in other parts of the state.

If the standard error value is high in relation to coefficient value, the coefficient may not be significant. To compare these values, you might wish to add a field to the table and calculate the field value by dividing C1_PERBLAC by StdERRC1_P. This would divide the coefficient value by the standard error value. Then map the values to see which counties have high values in relation to the standard error values. Counties in the eastern and southern parts of the state have high coefficient values in relation to the standard errors values for the PERBLACK coefficient.

Explore the other coefficient standard error and parameter values.

GWR analyses can yield important information for public policy formulation. Significant variables showing little regional variation point to problems which might best be addressed at a statewide level. Significant variables showing strong local variation point to problems which might best be addressed at the local level. For each independent variable in this analysis, the local parameters are significant in most counties, yet the variable is positively related to very low birth weight rates in some counties and negatively related in others. This suggests that these variables are locally significant but may not be globally significant and that policies to address the problem of very low birth weight rates need to be developed based on the situations in groups of counties.

Use the **Save** button to save the map document and then use the **File** ⇒ **Exit** menu to close ArcMap.