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

GIS AND PUBLIC HEALTH EXERCISE 6 – CLUSTER ANALYSIS (ArcGIS 10)

PREPARATION

Download the **exer6** 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 **exer6**. 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 ArcGIS10 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 10** to start ArcCatalog.

Find the button labeled **Connect to Folder** and click the button. Navigate to **c:\exer6** 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.

EXPLORING HEALTH OUTCOME DATA

Go to **Start ⇒ Programs ⇒ ArcGIS ⇒ ArcMap 10** to start ArcMap.

In the “ArcMap – Getting Started” window, close the window you would use to open an existing map document or make a new map using a template.

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 Cluster. Click OK. The name of the Data Frame in the Table of Contents window should now appear as Mapping.

Add a Database of Breast Cancer Data for ZIP Codes

To begin, add a breast cancer database for ZIP codes in the Chicago, Illinois, area. These data were prepared from a public data set available for download from the Illinois Department of Public Health site (Illinois Department of Public Health, Illinois State Cancer Registry, public dataset, 1986-2008, data as of November 2010). The data show the percent of breast cancer patients living in the ZIP code whose cancer was diagnosed late, after the cancer spread beyond the initial site. Patients whose cancer is diagnosed late have a much higher risk of mortality and morbidity

than patients diagnosed at an early stage. A high rate of patients diagnosed late represents a poor health outcome or high need for health care.

Find the button labeled **Add Data** and click the button. You should find the **c:\exer6** folder in your catalog. If not, please connect to the folder using the **Connect to Folder** button.

Navigate to **c:\exer6\data\agency\shapes** and add the **latebreast.shp** shapefile.

This database is stored in the agency subfolder because it has been modified by the user and is no longer the same data downloaded from the Illinois Department of Public Health site.

Use the Save button or go to **File** ⇒ **Save** the map document. Navigate to **c:\exer6\mapdocs** and save the file as **exer6.mxd**.

Map Late Diagnosis Rates by ZIP Code

Create a map showing the percent of breast cancer cases diagnosed late.

Right click the latebreast.shp data layer and select **Properties** from the menu. Then click the **Symbology** tab.

In the “Show:” window to the left, click on **Quantities** and Graduated colors to make a graduated color choropleth map of the percentage of late stage cases by ZIP code.

Under “Fields:”, select LATEBRST as the “Value Field” from the pull-down menu.

Next, click the Classify button under “Classification” to open the “Classification” window. Choose Standard Deviation as the classification “Method:” from the pull-down list. Then click Apply and OK.

The resulting map should have five class intervals based on standard deviations. A group of ZIP codes along the western border of the study area in the north should have percent late stage breast cancer cases more than 1.5 standard deviations above the mean.

Save the map document.

GLOBAL CLUSTERING ANALYSIS USING MORAN’S I

To investigate whether or not this apparent pattern arises from a statistically significant pattern of global clustering, you can calculate Moran’s I. This calculation requires two steps. First, you will generate a spatial weight matrix for use in the analysis. Second, you will use the spatial weights matrix to calculate Moran’s I.

If you do not have a version of ArcGIS 10 which allows you to generate a spatial weights matrix using contiguity of ZIP code boundaries and corner points (like queen’s case in chess), read through to the section below for **Calculating Moran’s I** and use the spatial weights matrix already generated for you.

Creating a Spatial Weights Matrix

Click on the **ArcToolbox window** button to open the ArcToolbox window.

Click on the **Spatial Statistics Tools** ⇒ **Modeling Spatial Relationships** ⇒ **Generate Spatial Weights Matrix** to create the spatial weights matrix used in the analysis.

In the “Generate Spatial Weights Matrix” window, click on the button next to “Input Feature Class” and navigate to **c:\exer6\data\agency\shapes** to add the **latebrst.shp** database as the input data for which you will generate spatial weights.

From the pull-down list, select the ZIPID field as the unique integer field identifying each ZIP code area.

Click on the button next to “Output Spatial Weights Matrix File” and navigate to **c:\exer6\data\agency\shapes** and save the file as **exer6.swm**.

From the pull-down list under “Conceptualization of Spatial Relationships”, select CONTIGUITY_EDGES_CORNERS. This means that you will be using all of the ZIP codes that share a boundary or a corner point with a ZIP code to develop the spatial weights matrix.

For the “Distance Method”, select EUCLIDEAN from the pull-down menu.

Then click OK.

After the spatial weights matrix has been generated successfully, close the “Generate Spatial Weights Matrix” window.

Save the map document.

Calculating Moran’s I

Click on the **ArcToolbox window** button to open the ArcToolbox window.

Click on the **Spatial Statistics Tools ⇒ Analyzing Patterns ⇒ Spatial Autocorrelation (Moran’s I)** to calculate Moran’s I as a measure of global spatial autocorrelation or overall clustering in the data.

In the “Spatial Autocorrelation (Moran’s I)” window, select LATEBRST as the “Input Feature Class”.

Select LATEBRST as the “Input Field”. This is the field containing percent late stage breast cancer values.

Check the box to “Generate Report”.

Then, under “Conceptualization of Spatial Relationships”, select Get Spatial Weights from File from the pull-down list.

In the “Weights Matrix File (optional)” space below, click on the button and navigate to **c:\exer6\data\agency\shapes** and add the **exer6.swm** spatial weights matrix file.

If you were not able to create the spatial weights file, navigate to **c:\exer6\data\agency\swm** and add the **exer6.swm** spatial weights matrix file already created for you.

Then click OK.

To view the results, click the **Geoprocessing ⇒ Results** menu in the main interface. The result is displayed in the “Results” window. Double click the “HTML Report File: MoransI_Result.html. The Moran’s I index is a positive value and equals 0.54. The Z Score is 5.81 standard deviations which is significant at the 0.01 level. This means there is less than 1% likelihood that the observed pattern of percent late stage breast cancer values could have occurred by chance. The positive value means that like areas are clustered together.

Close the two “Report” windows.

Save the map document.

The result of the global clustering analysis indicates that there is a significant pattern of positive spatial autocorrelation present in the data but it does not identify particular clusters. To identify groups of ZIP codes representing significant clusters of late stage breast cancer cases, you will need to perform a local clustering analysis.

LOCAL CLUSTERING ANALYSIS USING LISA

Click on the **Spatial Statistics Tools ⇒ Mapping Clusters ⇒ Cluster and Outlier Analysis (Anselin Local Morans I)** to calculate local indicators of spatial autocorrelation statistics.

In the “Cluster and Outlier Analysis (Anselin Local Morans I)” window, select LATEBRST as the “Input Feature Class”.

Then select LATEBRST as the” Input Field”.

Set the “Output Feature Class” to:

c:\exer6\data\agency\shapes\latebreast_ClustersOutliers.shp

Then, under “Conceptualization of Spatial Relationships”, select Get Spatial Weights from File from the pull-down list.

In the “Weights Matrix File (optional)” space below, click on the button and navigate to **c:\exer6\data\agency\swm** and add the **exer6.swm** spatial weights matrix file.

If you were not able to create the spatial weights file, navigate to **c:\exer6\data\agency\swm** and add the **exer6.swm** spatial weights matrix file already created for you.

Then click OK.

When the analysis has been completed successfully, close the “Cluster and Outlier Analysis (Local Moran’s I)” window.

You will see that the **latebreast_ClustersOutliers.shp** layer has been added to the map document automatically and that the ZIP code areas have been symbolized base on Local Moran’s I Z Scores.

There are three areas in the study region where there are significant clusters of ZIP codes with based on percentage late stage breast cancer.

Right click on the latebreast_ClustersOutliers.shp layer and select **Open Attribute Table** from the menu. Explore the table. The LMiIndex field displays the LISA statistic for the ZIP code. The LMiZScore field displays the associated Z Score. The LMiPValue gives the P value for the Z Score. The COType field indicates whether the ZIP code is part of a cluster where neighbors have high percentage late stage breast cancer values (HH), part of a cluster where neighbors have low percentage late stage breast cancer values (LL), or whether a ZIP code with a high or low value is surrounded by ZIP codes with very different values.

Close the attribute table.

One cluster is HH, meaning that these neighboring ZIP codes all have high percentages of late stage breast cancer cases. The other two clusters are LL, meaning that these neighboring ZIP codes all have low percentages of late stage breast cancer cases.

If you wish, right click the latebreast.shp layer and select **Properties** from the menu. Click the **Labels** tab and select LATEBRST as the label field. Click Apply and OK. Then, right click the latebreast_ClustersOutliers.shp layer and select **Label Features** from the menu. This will display the late stage breast cancer rates.

Save and **Close** the map document.