

Spatial Statistics: Spatial Autocorrelation

Workshop Exercise

1/24/2013

Introduction

You will conduct tests for spatial autocorrelation in both Geoda and ArcMap. You will use median housing values for each census tract in Middlesex County, MA from the 2006-2010 American Community Survey.

Transferring the Data



1. Create a folder on your desktop and name it your last name.
2. Drag and drop or copy and paste the data folder into your folder from the location specified by the instructor.

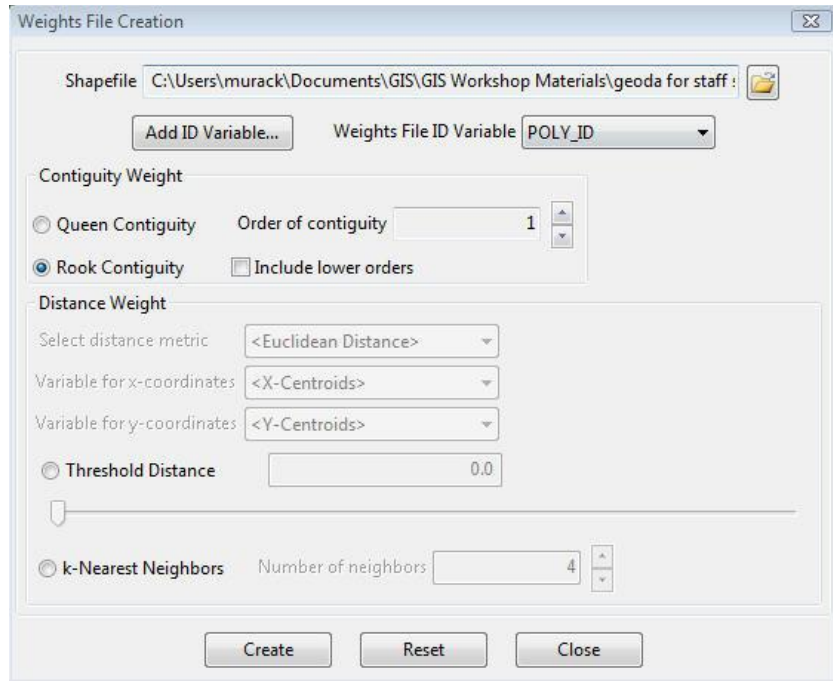
Geoda

You will first examine the data generally to determine if median household income is clustered or dispersed in Middlesex County, then you will run local statistical tests to see where clusters exist.

Creating a Spatial Weights Matrix

A spatial weights matrix is a way to numerically represent neighboring relationships and can be based on contiguity, distance, or the number of neighbors. You can create a spatial weights matrix while running a tool or before you begin an analysis. We'll create a weights matrix now to prepare for our spatial autocorrelation analyses.

1. Open Geoda.
2. Open the medianhousing.shp file by clicking File > Open Shapefile
3. Navigate to the folder on your desktop and open medianhousing.shp.
4. Open the table by clicking on the table button (). The column Median_val contains the median housing values and the column Median_v_1 contains the margin of error for each estimate. Close the table.
5. Click on the Weights File Creation button (.
6. Select POLY_ID as the ID variable. If you did not have an ID variable (or if Geoda did not recognize it), you could create one using the Add ID Variable... button.
7. Select Rook Contiguity and then click Create.




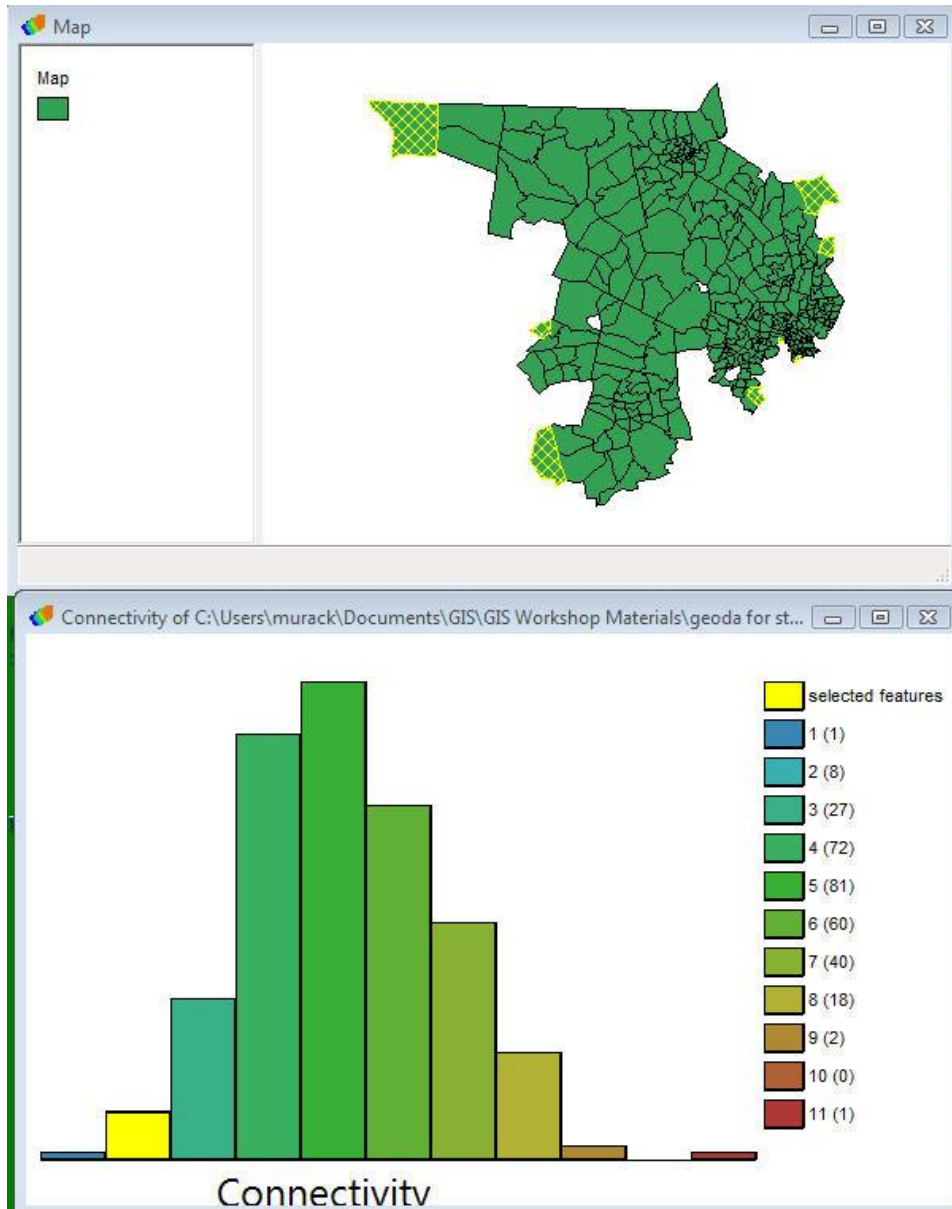
8. Save it to your folder on the desktop and call the file medianhousing_rook.
9. Close the weights creation box.

Q: What is rook contiguity? What neighbors are included in this matrix?

A: Any census tracts that share borders (but not vertices) with each other are neighbors.

To learn more about the neighbors, you can create a weights neighbor histogram.

1. Click on the Weights Neighbor Histogram button ().
2. Select the weights file you just created and click OK.
3. Select a census tract from the map and the corresponding bar on the histogram will be highlighted in yellow at the bottom.
4. Similarly, select a column or category from the histogram and the corresponding polygons will be highlighted on the map. This is called brushing. You can use this tool to identify polygons with large or small numbers of neighbors or to compare the number of neighbors using different spatial weights matrices.



Q: What is the most frequently occurring number of neighbors?

A: 5

Q: What is the smallest number of neighbors? How many tracts have that number of neighbors?

A: 1 is the lowest number of neighbors and 1 tract has only 1 neighbor

Q: Where do tracts with a small number of neighbors tend to be located?

A: Around the edges of the map. In future analyses you may want to analyze an area slightly larger than your area of interest so that your polygons are not cut off from their neighbors.

5. Close the histogram window.

Univariate Moran's I

Univariate Moran's I is a global statistic that tells you whether there is clustering or dispersion, but it does not inform you of the location of a cluster.

1. Click on Explore > Univariate Moran's I
2. Select Median_val as the variable and click Ok.

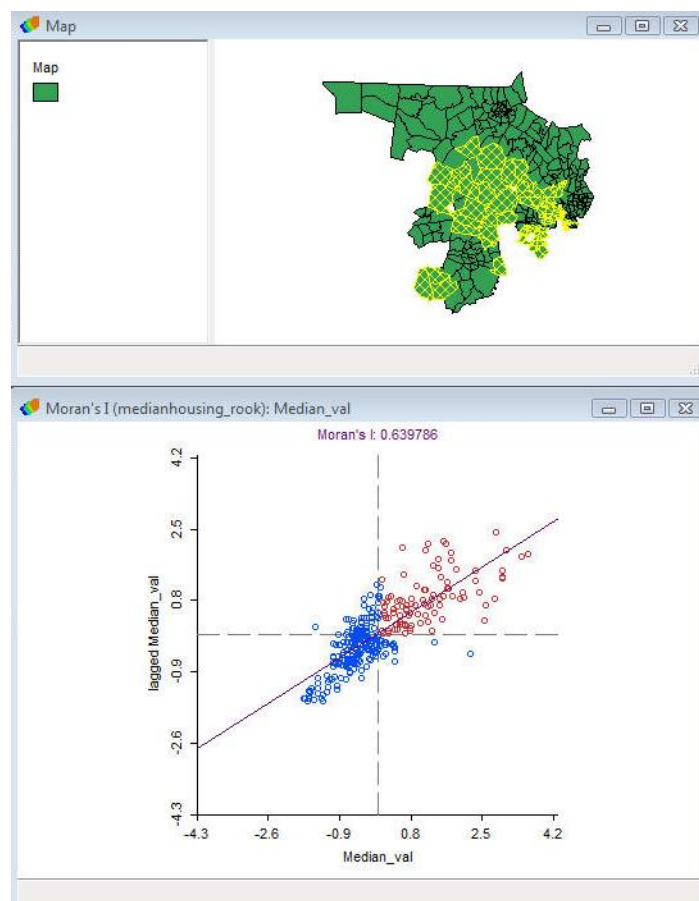
The result is a Moran's scatter plot with the I value displayed at the top. The X-axis is the value of I and the Y-axis is the spatial lag, which is the weighted average of neighboring values. The slope of the line is Moran's I.

Q: What is the I value? What does this mean?

A: $I = .639786$. Since it is positive, there is an overall pattern of clustering of median housing values. It is also quite close to 1, indicating the values are highly clustered.

The upper right and lower left quadrants represent positive spatial autocorrelation, which means clustering of like values while the lower right and upper left quadrants represent negative spatial autocorrelation or spatial outliers.

3. Select the points in the different quadrants by drawing a box around them to see the general areas where clustering and dispersion are occurring.



Q: Do there appear to be more clusters or outliers?

A: Clusters since most points fall into the upper right or lower left quadrants.

4. Right click on the scatter plot and select Randomization > 999 Permutations

This displays the I value (yellow line) in relation to the reference distribution. You can click Run over and over again to generate new reference distributions. If the yellow line overlapped the reference distribution, your distribution may be due to chance rather than autocorrelation.

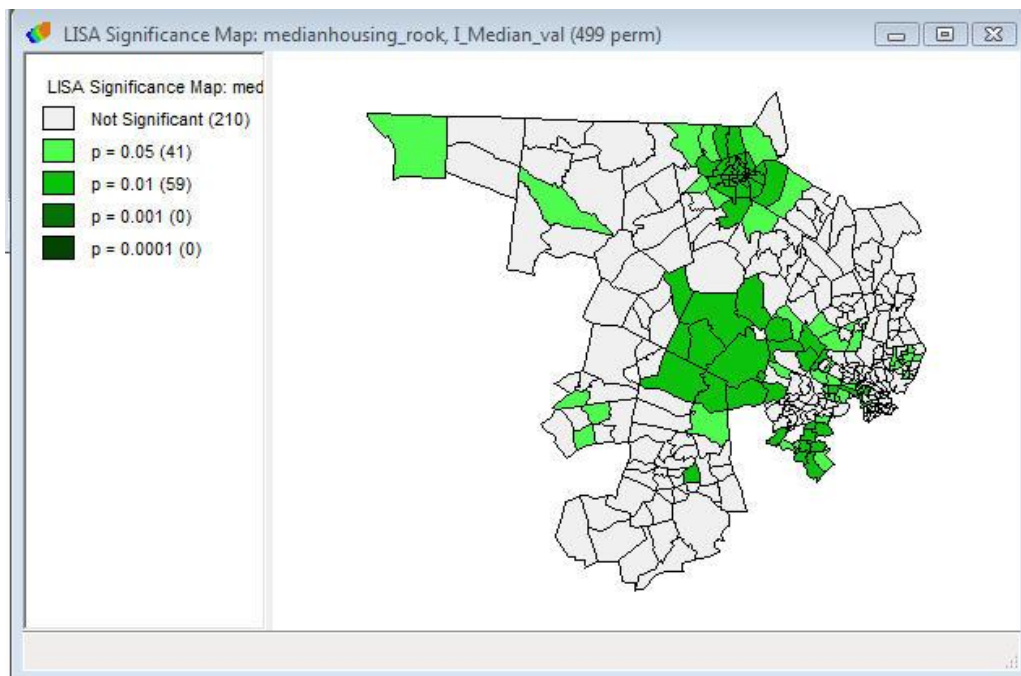
5. Click Done and close the Moran's scatter plot.

Univariate LISA (Local Moran's I)

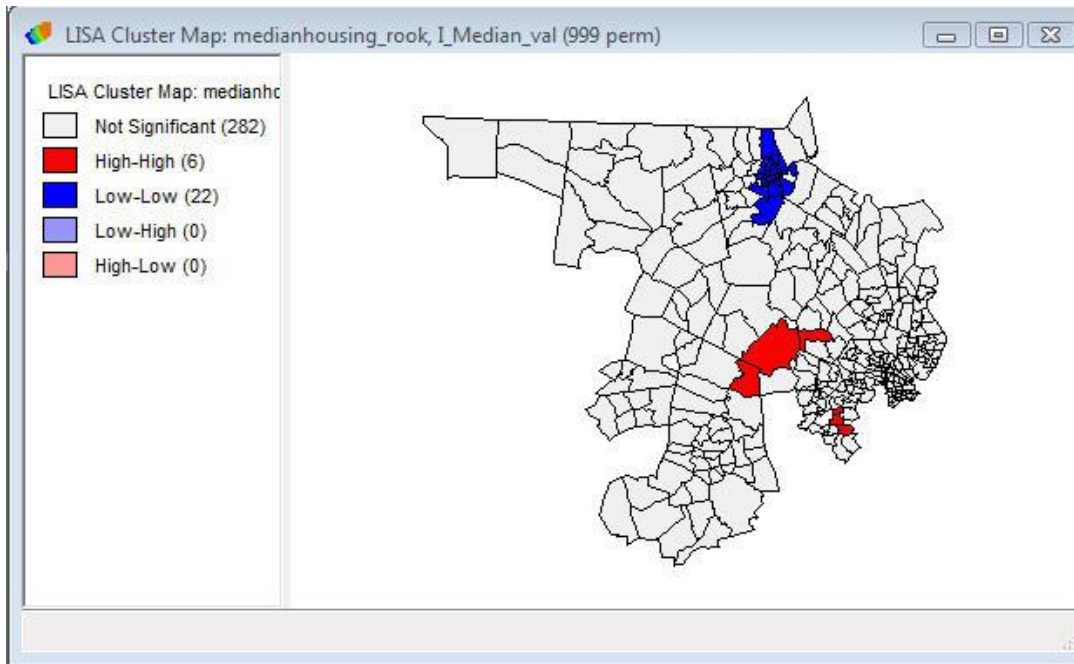
Univariate LISA calculates local Moran statistics to demonstrate local spatial autocorrelation.

1. Click Explore > Univariate LISA
2. Select Median_val as the variable and click Ok.
3. Check Significance Map and Cluster map Map and click OK.

The significance map shows the locations with significant local Moran statistics for various p values. When conducting spatial statistics, you'll often find significant results at $p=.05$, so it is good to look at the results for more restrictive p-values. You can adjust the significances displayed and the permutations by right clicking and selecting Randomization and/or Significance filter.



The LISA Cluster Map shows the significant locations color coded by type of spatial autocorrelation.



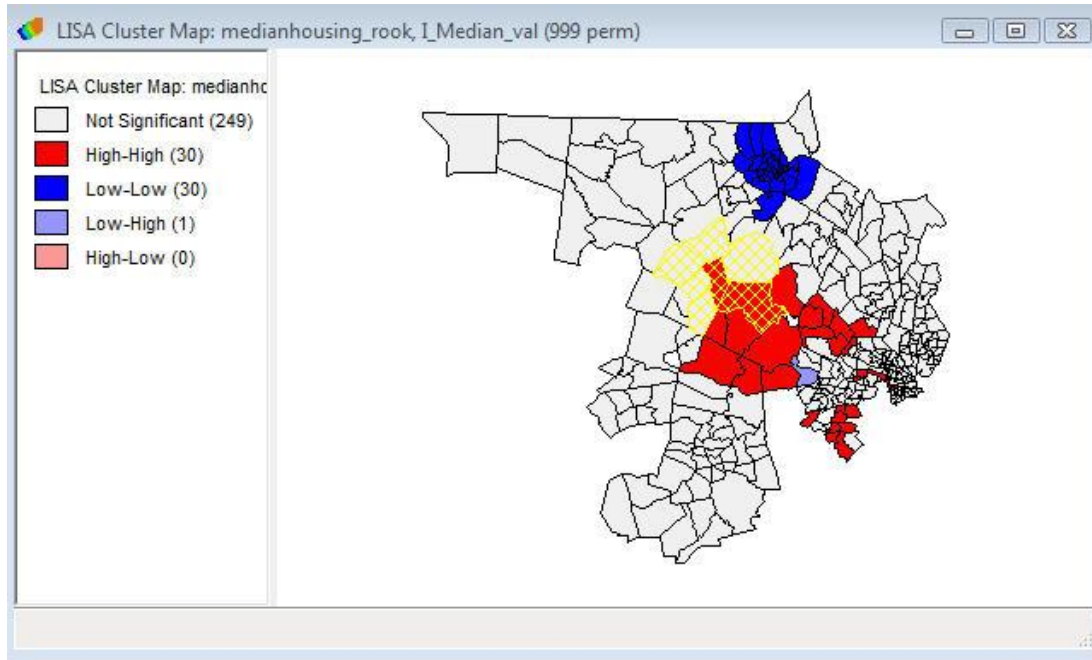
The above map shows $p < .001$ and 999 permutations.

- High-high and low-low = spatial clusters
- High-low and low-high = spatial outliers

The spatial clusters on the map refer to the core of the cluster. The cluster is classified as such when the value at a location (either high or low) is more similar to its neighbors (as summarized by the weighted average of the neighboring values, the spatial lag) than would be the case under spatial randomness. The cluster itself extends to its neighbors.

1. Select one of the census tracts that had significant clustering by clicking it on the map.
2. Right click in the white space on the map and select Add Neighbors to Selection

The map has now selected all the neighbors of the target census tract.



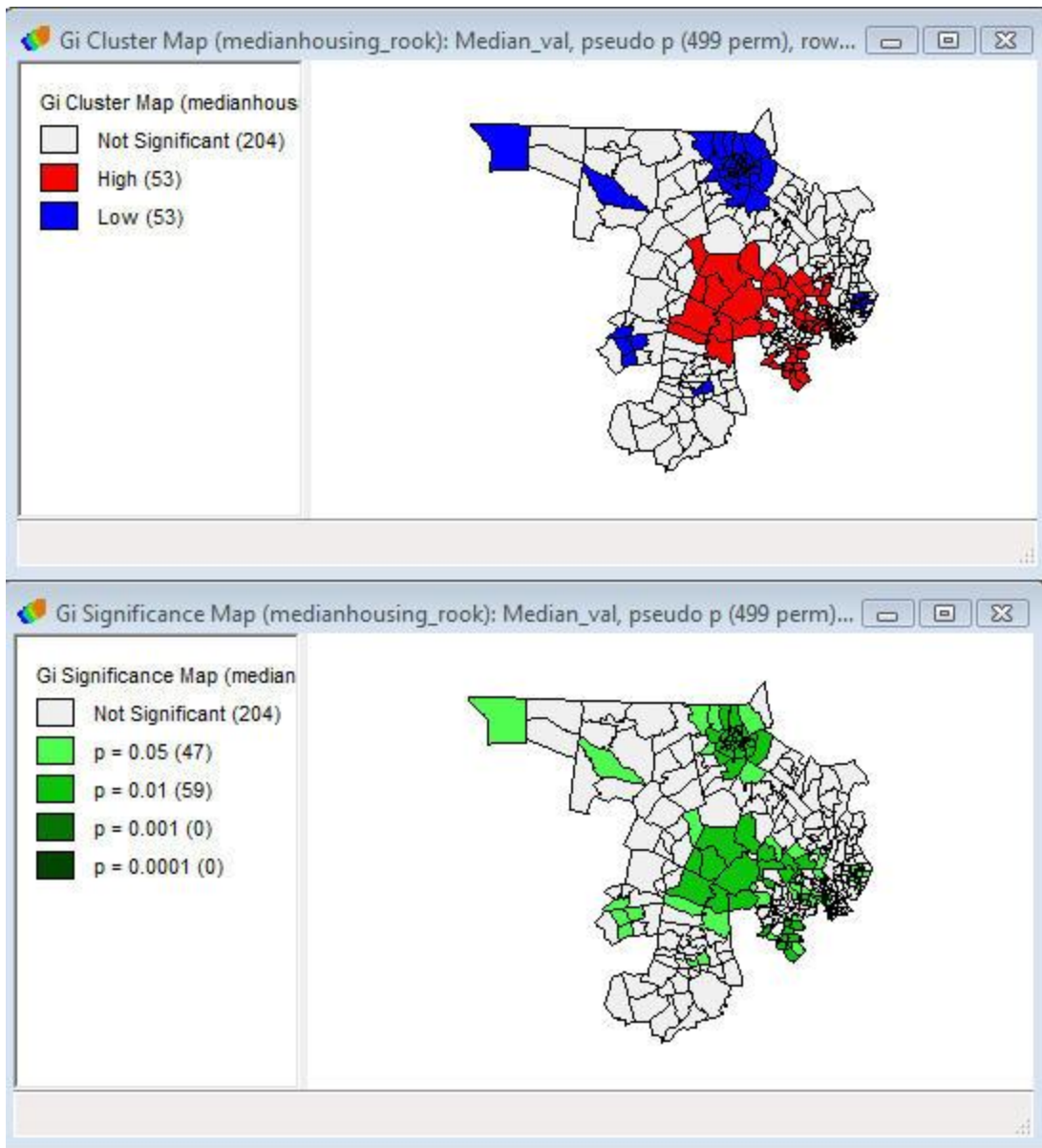
Q: Where are the clusters on your map? Where are the outliers?

A: There is a cluster of low values in the upper right corner and a cluster of high values in the center of Middlesex County. There is one outlier near the cluster of high values.

Local G Statistic

1. Minimize or close any maps you may have open.
2. Click on Space > Local G Statistics
3. Select Median_val as the variable and click OK.
4. Use the same spatial weights file you created earlier.
5. Check off all the options for Gi cluster map and show significance maps. Keep row-standardized selected. Click OK
6. Select your spatial weights file again.

The result will be a map showing clusters of significant high and low values and the significance map indicating the p-values for each polygon. You can change the number of permutations and the p-value as you did with the LISA maps.



ArcMap

You will first examine the data generally to determine if median household income is clustered or dispersed in Middlesex County, then you will run local statistical tests to see where clusters exist.

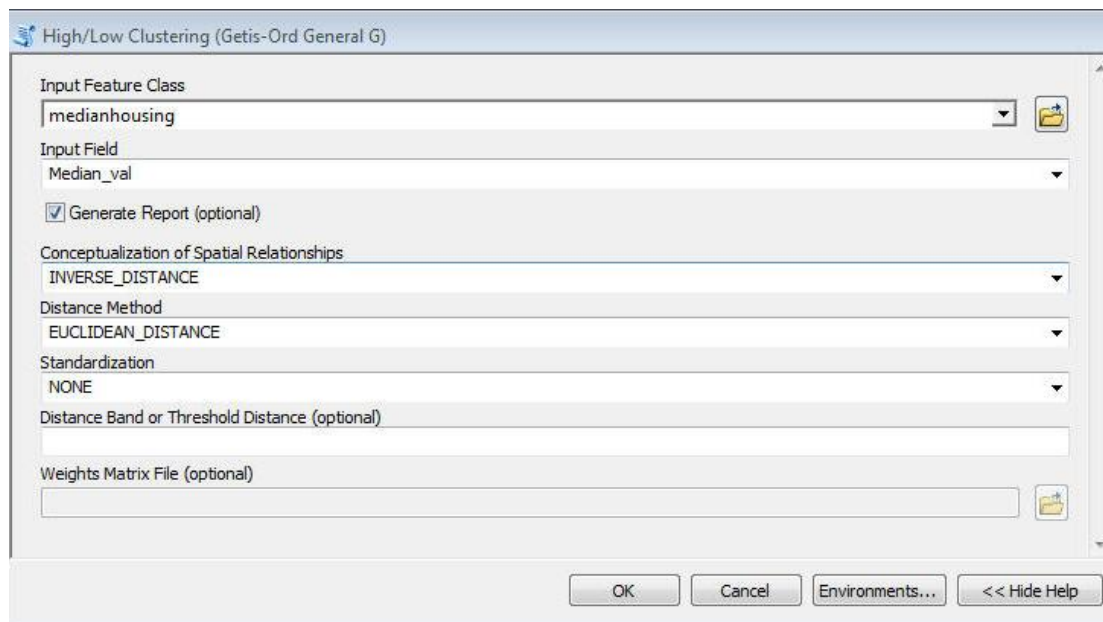
1. Open ArcMap and a blank map.
2. Add the data layer, “medianhousing” to the map from your folder on the desktop.

In the following exercises you will use the default setting for spatial weights in each tool, however you also have the option of creating a spatial weights matrix using this tool: Spatial Statistics > Modeling Spatial Relationships > Generate Spatial Weights Matrix. If you are familiar with network analysis you can construct a spatial weights matrix based on your road network using the Generate Network Spatial Weights tool.

Calculating General Clustering

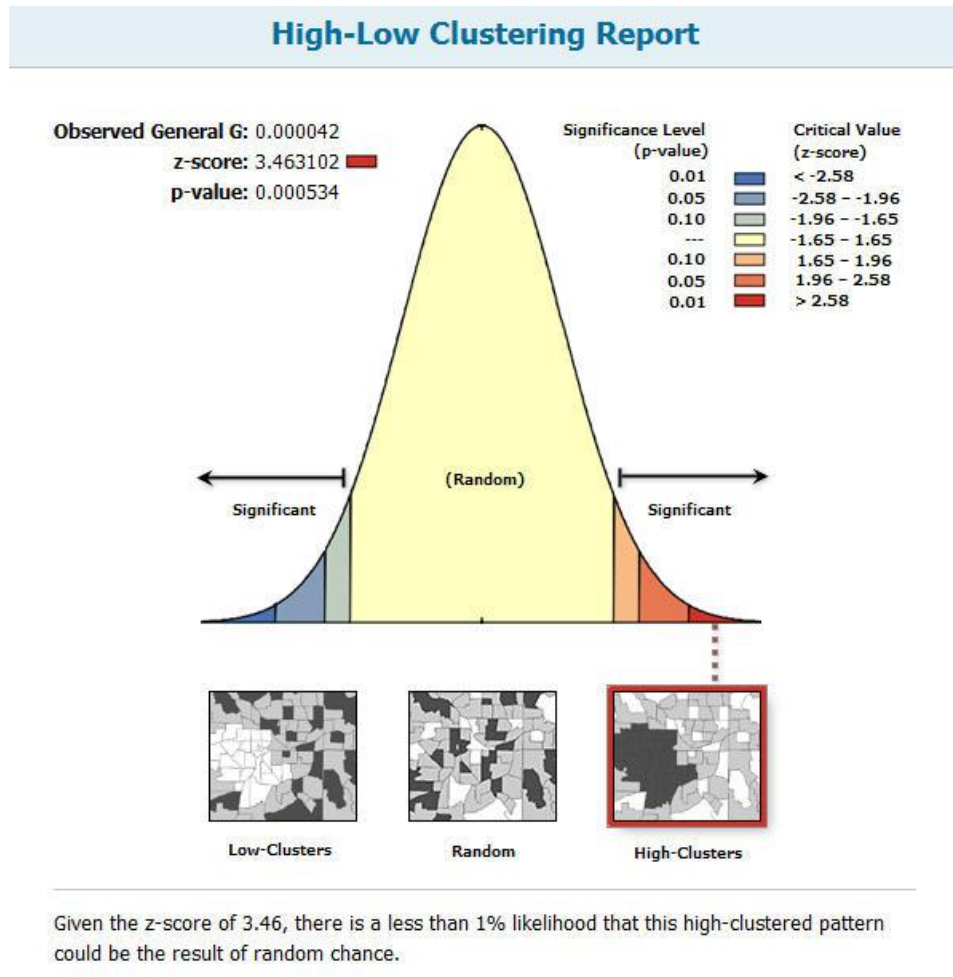
You will first run the **Getis-Ord General G** statistic to see if **areas with a high or low median income are clustered**.

1. Open the toolbox if it is not already open by clicking Geoprocessing > ArcToolbox.
2. Expand Spatial Statistics Tools.
3. From the *Analyzing Patterns* tools, select *High/Low Clustering (Getis Ord General G)*.
4. For the input feature class, enter “medianhousing”. For the input field, select “Median_val”.
5. Check the box next to *Generate Report*.
6. Keep the Conceptualization and Distance methods as the default. Feel free to experiment with other conceptualizations after you run this analysis.
7. Click *Ok*.



ArcMap warns you that it used the distance of 12597.9 meters for the distance of the search threshold. By default, ArcMap will use the minimum distance to insure all features have at least one neighbor. You can change this distance to make the analysis more appropriate for your project. The larger the value, the more neighbors that will be considered during the calculation.

8. Once the test has finished running, click Geoprocessing > Results (in the top menu bar)
9. Expand Current Session and double click Report File: GeneralG_Results.html.



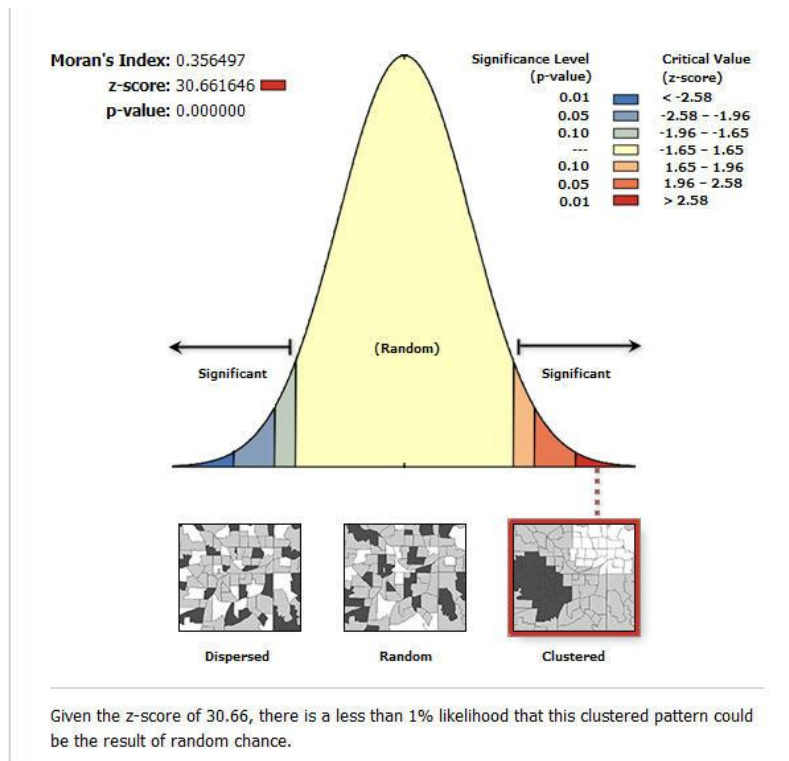
Q: What is the G score? What is the Z score? What does this tell you?

A: Because the Z score is highly significant, the median household income in each tract is significantly more clustered than a random pattern.

Next you will run a **spatial autocorrelation** test to see if the **general pattern of features is clustered or dispersed (as opposed to clustering specifically of high or low values)**.

10. Select *Spatial Autocorrelation* from the *Analyzing Patterns* menu and input the same information as you did for the General G test.
11. Check the box for "generate report"
12. Once the test has finished running, expand the Spatial Autocorrelation category under Current Session in the results window and double click on the report, Moransl_Results2.html.

13. Double click on the HTML report.



Q: What are the results? What does this tell you?

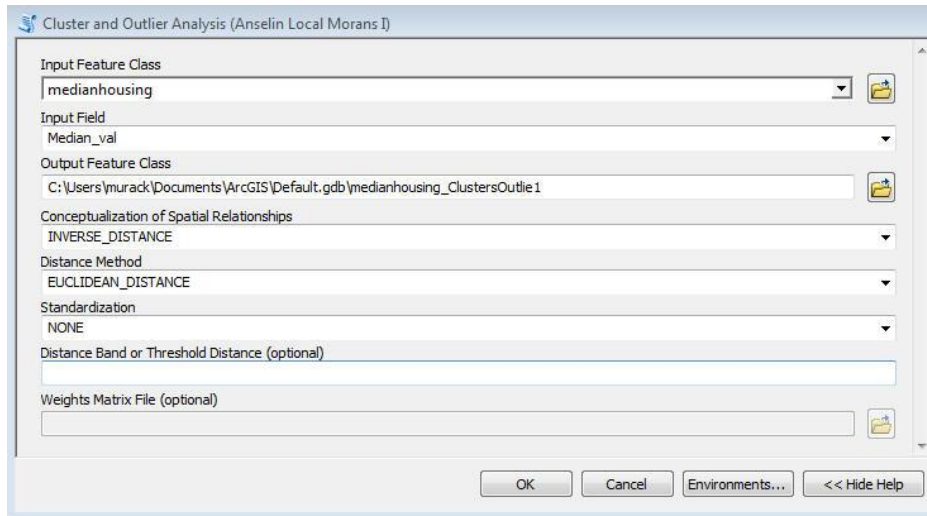
A: The median housing values are clustered at a significance level of $p < .001$. This means that there is clustering of similar values, but the test does not tell us which values are clustered (positive or negative).

Calculating Local Statistics

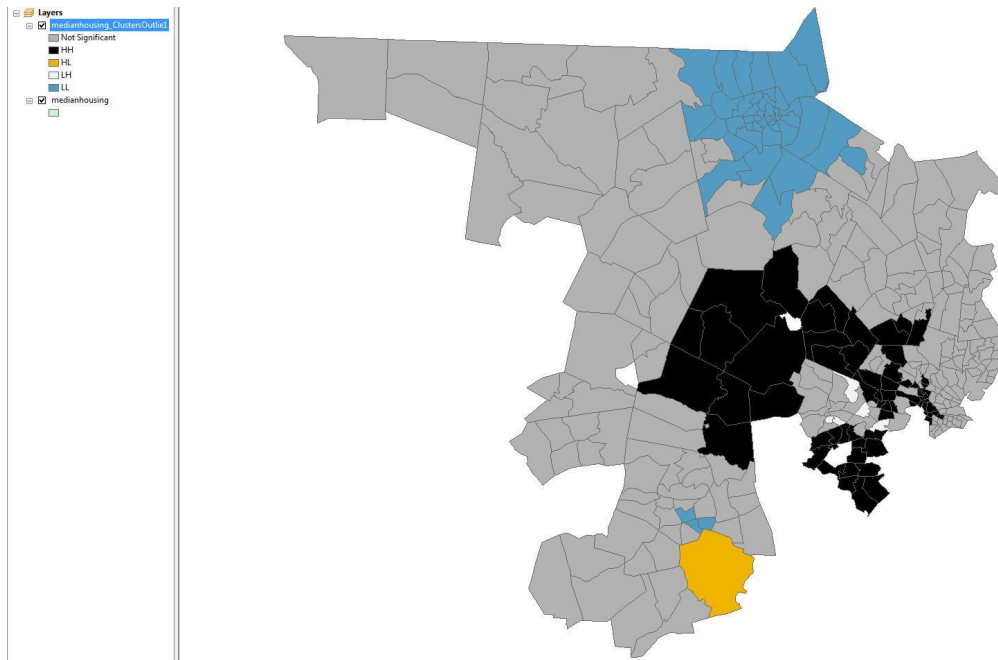
While the General G and Spatial Autocorrelation tests told you something about the overall pattern of median housing values in Middlesex County, you will now use local statistical tests to learn more about where clusters are located.

Local Moran's I

1. Expand the *Mapping Clusters* tool set, choose *Cluster and Outlier Analysis (Anselin Local Morans I)*.
2. Use the same Input Feature Class and Input Field as you did for the two previous tests.
3. For the Output Feature Class, click on the folder icon and save the file to your desktop folder, using a name that makes sense to you.
4. Click *Ok*.



A map showing the types of significant relationships is added to your map.



Q: Are there any clusters or outliers present?

A: You will notice a large cluster of low-low values, a large cluster of high-high values, and a polygon with a high-low relationship.

HH indicates that the relationship is High-High. These tracts are surrounded by tracts with high median housing values and their values are significantly higher than their neighbors. **This indicates that although there are high median housing values in this area in general, there are significantly higher values in these census tracts.** The inverse is true for the Low-Low relationships. For the tract that has a significant High-Low relationship with its neighbors, median housing values are high in this tract and it is surrounded by tracts with significantly lower housing values.

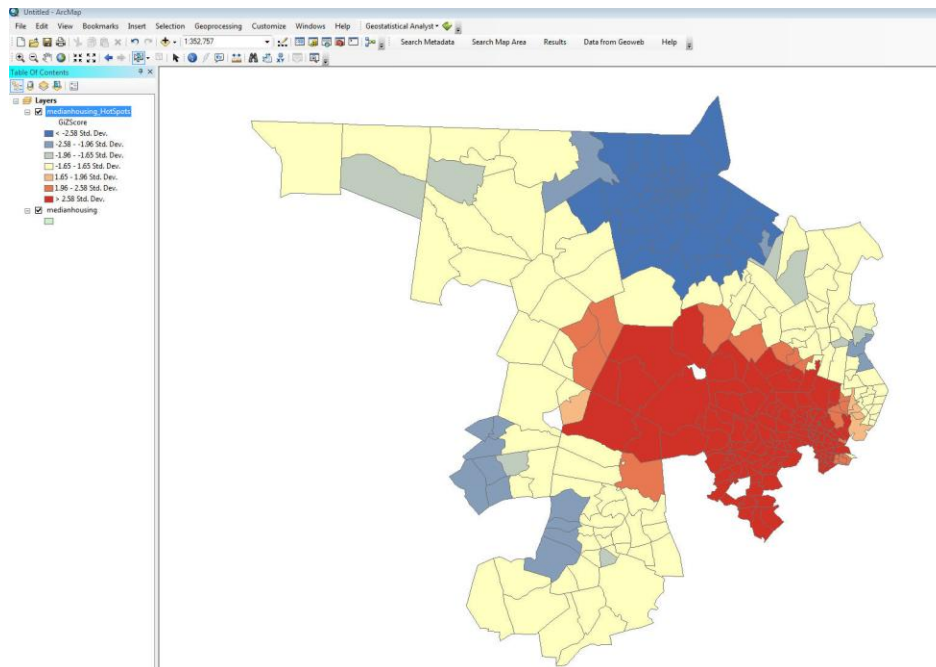
- Right click on this layer name in the table of contents and select *Open Attribute Table*. Scroll through the table and notice the fields for Z-score and P-value. If you were interested, you could map these values as well.

Hot Spot Analysis (Getis-Ord G_i^*)

To learn more about where median housing values are clustered, you will calculate a **local** version of the **G score** that you calculated earlier.

- Close the table if you haven't already.
- From the *Mapping Clusters* toolset, select *Hot Spot Analysis (Getis-Ord G_i^*)*.
- Use the same inputs as you did for the Local Moran's I test. Save the output feature class to your folder on the desktop. Click *Ok*.

A graphic display of the Z scores will be added as a layer to your map.



Q: What do you see?

A: You will notice a large statistically significant (Z score < 2.58 Std. Dev or > 2.58 Std. Dev.) “hot spot” (dark red) and a large statistically significant “cold spot.” There are also some sections with moderately high or low Z scores.

To be a statistically significant hot spot, the tract needs to be surrounded by tracts with high, positive values and have a much higher, positive value than its neighbors. The inverse is true for the cold spots. This is similar to the HH or LL relationships that the Local Moran's I test detects. However, Local Moran's I can also detect HL or LH relationships, whereas Hot Spot Analysis just looks for clusters of similar high or low values.

Take a look at the results of both of the local Moran's I tests (They are both copied on the next page). There are some differences due to defining spatial weights in different ways and using different p-values, however you can still clearly see clusters in certain areas.

Q: What did you learn about median housing values in Middlesex County? Feel free to add a basemap to ArcMap to give your data context. You can do this by clicking on the drop down arrow next to the Add Data button (plus sign in the menu bar).

A: Low housing values are clustered together in northern Middlesex County around Lowell while high values are clustered in the Western suburbs of Boston.

If you want to examine the median housing values on the map, you can use the Map menu in Geoda to map the Median_val column. In ArcMap, right click on the medianhousing layer, select Properties > Symbology and select Quantities and then Median_val from the drop down menu.

