Geographical Information Systems Institute

Center for Geographic Analysis, Harvard University

GeoDa: Spatial Autocorrelation

A. Background

From geodacenter.asu.edu:

"GeoDa is a free software program that serves as an introduction to spatial data analysis. GeoDa is the cross-platform, open source version of Legacy GeoDa. GeoDa runs on different versions of Windows (including XP, Vista and 7), Mac OS, and Linux. It is written in C++ and no longer relies on ESRI's MapObjects library (it uses wxwidgets instead).

GeoDa is the flagship program of the GeoDa Center, following a long line of software tools developed by Dr. Luc Anselin. It is designed to implement techniques for exploratory spatial data analysis (ESDA) on points and polygon). The free program provides a user friendly and graphical interface to methods of descriptive spatial data analysis, such as spatial autocorrelation statistics, as well as basic spatial regression functionality.

The development of GeoDa and related materials has been primarily supported by the U.S. National Science Foundation/ the Center for Spatially Integrated Social Science (CSISS) (Grant BCS-9978058).

Reference: Anselin, L., I. Syabri and Y Kho. (2005). GeoDa : An Introduction to Spatial Data Analysis. Geographical Analysis 38(1), 5-22."

GeoDa can be downloaded at: spatial.uchicago.edu/software

GeoDa does run on Mac OS. Scott runs it on a Mac. However, while stable, there are some quirks that can suggest instability. These instructions are for Windows, but the functionality (but maybe not the interface) is the same on Mac. Since we are all working on our own computers this has the potential to affect your workflow; before you download and start running GeoDa on Mac OS consider other places you might need the data or output (in a format that can be used in ArcGIS).

2. Major tasks in this lab:

1. Learn the table functions of Open**GeoDa**.
    i) add/delete fields,
    ii) variable calculation.

2. Spatial weights matrix.
    i) Introduction to weights matrix: contiguity, distance, and K-nearest neighbors.
    ii) Create Rook weights matrix for New York City.
    iii) Connectivity Histogram.

3. Analyze spatial lag using measures of spatial autocorrelation.
    i) Global Moran's I: univariate and multivariate.
    ii) LISA (Local Indicator of Spatial Association): univariate.
    iii) Export Moran's I results, and create thematic maps.

3. Data set

In this exercise we will use American Community Survey (ACS) 3 year data, ending in 2016. There is also a dataset for the same geography (New York City) from the 2000 decennial census, if you are interested. (Note: GeoDa comes with several data files as samples (e.g., Crime in Columbus [tracts], SIDS in North Carolina [counties])).

The data is in "GeoDa2018" in the GIS Data folder.

The shapefile NYC_Data.shp is the map of New York City with ACS 2016 data. These are socioeconomic attributes for census tracts in the five boroughs. It includes the following variables:

NYC_Data.shp

PCT_WHT	Percent of Population that is white
STATEFP	State FIPS* code
COUNTYFP	County FIPS* code
TRACTCE	Extended tract id
AFFGEOID	GeoID for joining
GEOID	
ALAND	area of tract
GEOID1	Copy of GeoID
GEOD2	Copy of GeoID
CT	Census Tract #
TOT-POP	Estimate of total population
TOT_WHT_POP	Estimate of total White population
TOT_BLK_POP	Estimate of total Black population
TOT_ASN_POP	Estimate of total Asian population
HSHLDS_TOT	Estimate of total households
FAM-TOT	Estimate of total families
Highschoo	Estimate count of population with a high school diploma
Postsec	Estimate count of population with some post-secondary education
Bachelo_o	Estimate count of population with a Bachelor's degree or higher
MED_House	Median House Value
MED_income	Median Household income (derived by compiling three other median income variables

*Federal Information Processing Standard

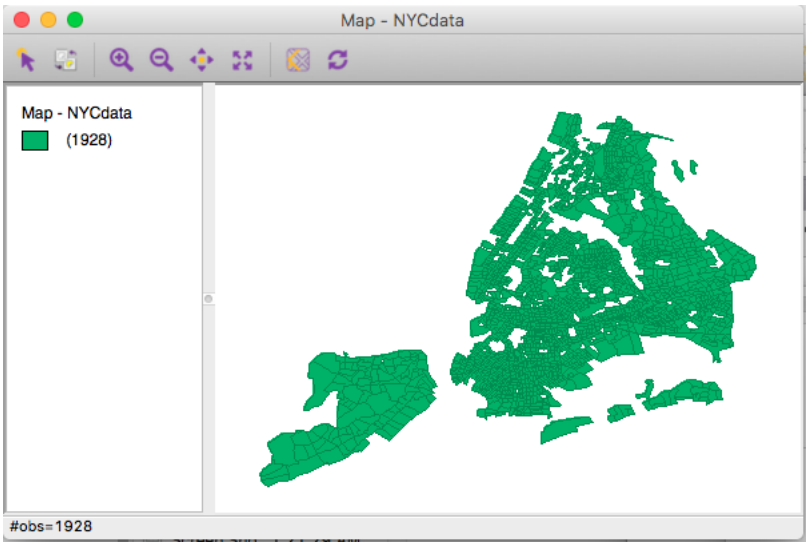
Manage tables in GeoDa

1. Start **GeoDa**

2. Go to **`File > New,`** click on the **`folder`** icon choose **`ESRI shapefile,`** **`navigate`** to nyc2000 and click **`Connect.`** If you are moving directly from GeoDa1, you are probably already already "connected" to this data.

3. **`Browse`** through the folders to find and select the shape file, **`NYC_data.shp`**.

4. Click **`OK`**. Your screen should now look something like …

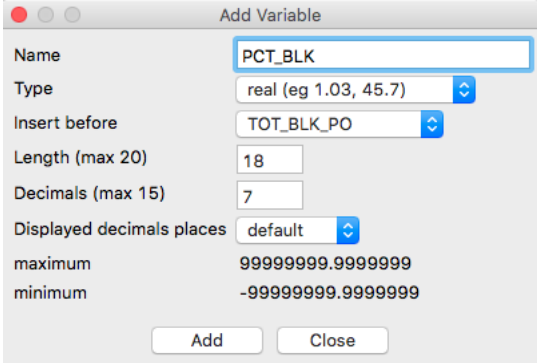


5. Click Table icon on the tool bar to open the attribute table.

6. To compute additional variables: First add variable columns

Go to **`Table > Add Variable,`**

`type PCT_BLK` (for percent Black) in the **`Name`** box, select **`real`** as the **`Type`**, select a location (logical locations would be near **`TOT_BLK_PO`** or near the beginning of the data (after `PCT_WHT`), and click **`Add (accept defaults for length, etc).`**



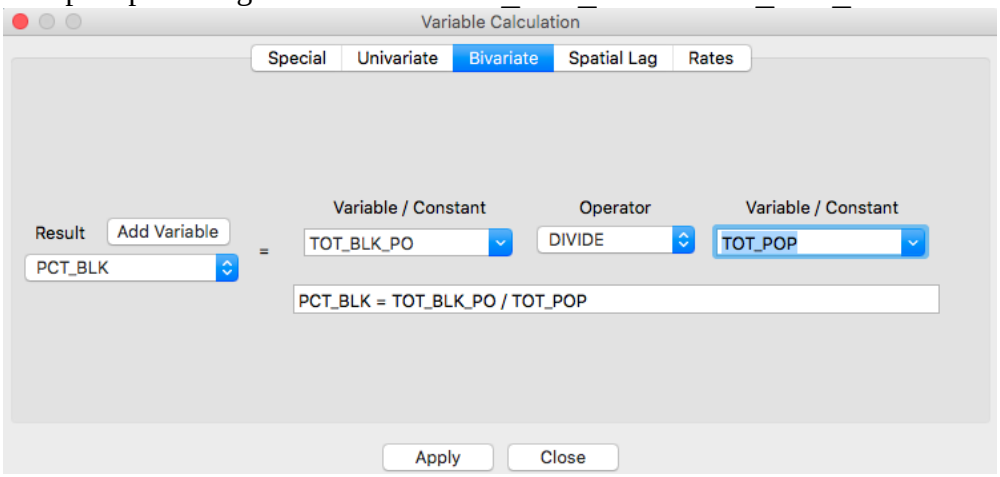
A blank column titled "**PCT_BLK**" will appear in the specified location.

Repeat the steps to add another column titled "**PCT_ASN**" (for percent Asian).

Calculate values for the two added columns

Go to **`Table > Calculator`** to open the dialogue box

Compute percentage variables for `TOT_BLK_PO` and `TOT_ASN_PO`.



Choose **`Bivariate Operations > Select PCT_BLK`** in the **Results** listing, **`Select TOT_BLK_PO`** in the **Variables-1** listing. **`Select DIVIDE`** in the **Operators** listing. **`Select T0T_POP`** in the **Variables-2** listing. Then Click **`Apply`**.

Repeat for PCT_ASN.

Save the results. In previous versions users had to save a NEW file to save new variables, so you can skip saving a new version (new name).

*(In order to keep computed variables, you need to save a new shapefile. (Note: **`GeoDa`** doesn't allow you to override the currently opened table.))*

*(Go to **File > Save as,** select **New Datasource** only, and **Type NYC-new,** and click **OK.**)*

d) Close the current project,

Go to **File > Close Project**. (You don't HAVE TO do this step, we can continue with the same file).

Create a weights matrix.

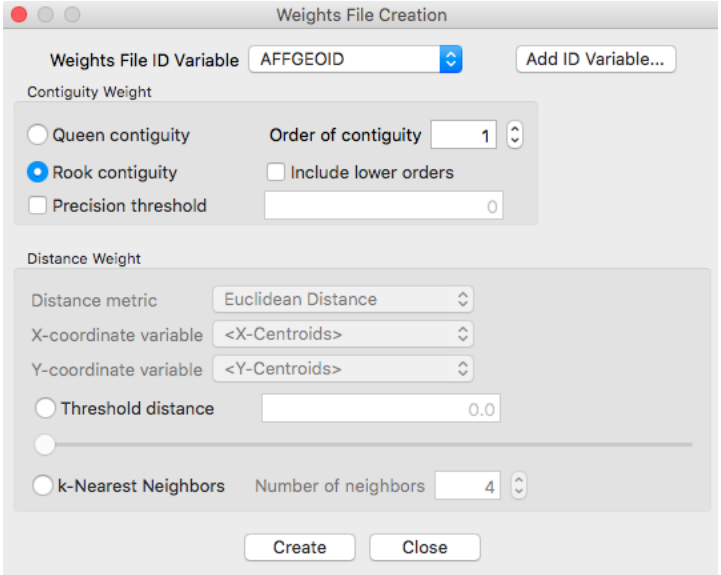
Spatial autocorrelation measures such as Moran's I require a weights matrix that defines a local neighborhood around each geographic unit. The value at each unit is compared with the weighted average of the values of its neighbors. A weights file identifies the neighbors. Weights can be constructed based on contiguity to the polygon boundary (shape) files, or calculated from the distance between points (points in a point shape file or centroids of polygons).

For most analyses, the spatial weights in **GeoDa** are in row-standardized form. This means that the row elements for each observation sum to 1, with zero on the diagonal and some non-zero off-diagonal elements. The formula for each weight is:

$$w_{ij} = \frac{C_{ij}}{\sum_{j=1}^{N} C_{ij}}$$ with C_{ij} =1 when i is linked to j, and C_{ij} =0 when otherwise.

*1. Go to **File > New,** click on the **folder** icon choose **ESRI shapefile,** **navigate** to NYC-new and click **Connect.***

2. Go **to Tools > Weights Manager,** click **Create** to open the **Create Weights** dialogue box.



There are three kinds of weight matrices: Contiguity, distance, and k-nearest neighbor.

a) Contiguity Weights File.

Most analyses of spatial autocorrelation adhere to a common definition of contiguity. Namely, either rook contiguity or queen contiguity. Contiguity refers to what polygons are selected as neighbors for a single target polygon. **GeoDa** also allows you to specify order of contiguity, for example, you may decide that the value of a unit is not only affected by the immediately contiguous units, but also the second order contiguous units. Below you can see the different cases of contiguity for spatial autocorrelation (bishops is rarely used).

b) Distance weights

GeoDa uses XY-coordinates to automatically calculate distance between points or centroids of polygons. You can specify the cut-off point (threshold distance) to determine the minimum distance for two units be considered neighbors.

c) k-Nearest neighbors

You can specify the exact number of neighbors that a unit should, and GeoDa will find those that are the nearest.

Here, we will calculate weights based on rook contiguity. In the **Creating weights dialogue** box:

First we have to specify the **Weights File ID variable** for which each polygon has a different value, in this case we can use polyid (usually you will have an ID variable available for this purpose).

Select **NYC-new.shp** as the input, **AFFGEOID** as the **Weights File ID Variable (DON'T Add ID Variable).**

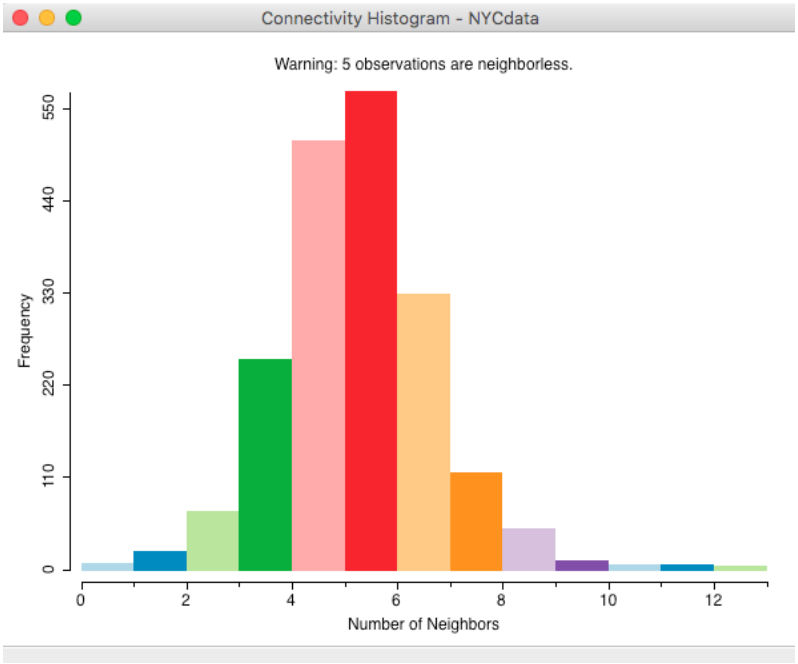
Select **Rook Contiguity**, click **Create**, name the file, then **Save,** then **Close.** Make sure you pay attention to where you save the file and what you called it.

You might get a "neighborless Observation" message, this could be real (legitimate). We know we are using a dataset with gaps in the geographic units of analysis that WE created, but there also actual islands. We're going to accept this constraint.Q

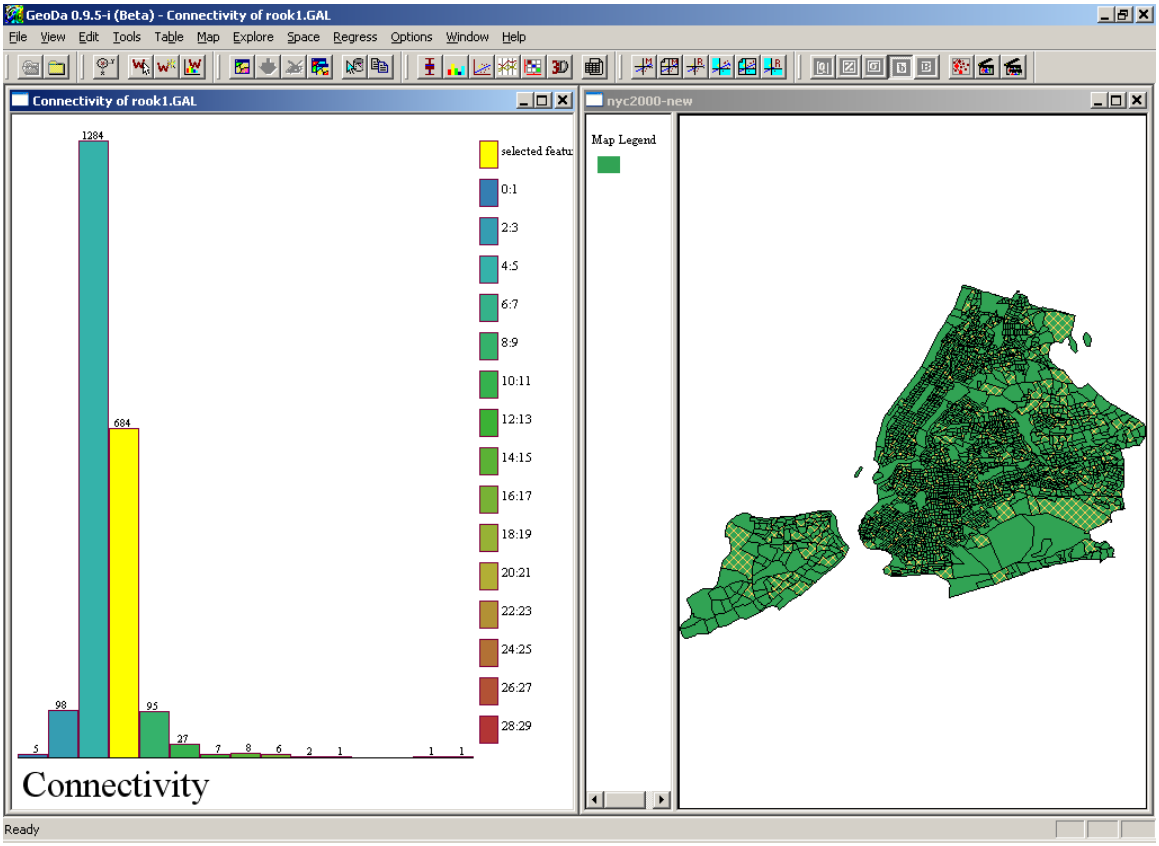
3. Characteristics of a weights matrix.

To explore the characteristics of the weights matrix, the **`Weights Manager`** opens and give you option that allow to you explore the weights distribution, click the **`histogram`** button. This opens up a **`"Histogram."`**

This generates a "connectivity" histogram similar to this one:



It shows the frequency distribution of neighbors according to this weights matrix. This means the histogram is displaying the frequency (number) of tracts by number of neighbors; as you move from left to right the total number of neighbors increases. The histogram can be queried by clicking on the vertical bar(s). The corresponding census tracts are highlighted on the map. And you can change the intervals by right clicking on the histogram, and specifying your desired number.



4. What does a weights matrix look like?

Use Microsoft Word to open the weights file: "rook.gal." The first line is the header that contains the name of the data set with the ID variable. Then the neighborhood information follows. The Polyid variable is used to identify the neighborhoods of each of the 2219 tracts in New York City. Can you see how they are organized?

Analyze spatial lag

A spatial lag of a specified variable is computed by taking the weighted average of neighboring polygons, as specified in the weights matrix. For example: a census tract with three neighboring tracts that had 10%, 15%, and 20% black population would have a spatial lag of 15%; that is, [(10%+15%+20%)/3]. The row-standardized spatial weights matrix is used in the calculation of the spatial lag.

Measuring spatial autocorrelation

1. Spatial autocorrelation (spatial association) vs. spatial randomness

What is spatial randomness?

a) Values observed at a location do not depend on values observed at neighboring locations;

b) the observed spatial pattern of values is equally likely as any other spatial pattern;

c) the location of values may be altered without affecting the information content of the data.

When spatial randomness is violated then there is spatial autocorrelation. There are two kinds of spatial autocorrelations: positive, when the relationship between the value at a location and the values of its neighbors is positive; otherwise, the spatial autocorrelation is negative.

Moran statistics are one class of measures of spatial autocorrelation.

2. Global Moran's I

The literal meaning of spatial autocorrelation is self-correlation (autocorrelation) of observed values of a single attribute, according to the geographical (spatial) ordering of the values. Global autocorrelation statistics provide a single measure of spatial autocorrelation for an attribute in a region as a whole.

$$I = \frac{N}{\sum_i \sum_j W_{ij}} \times \frac{\sum_i \sum_j W_{ij}\left(y_i - \bar{y}\right)\left(y_j - \bar{y}\right)}{\sum_i \left(y_i - \bar{y}\right)^2} \text{ or } I = \frac{N}{\sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} w_{ij}} \times \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} z_i w_{ij} z_j}{\sum_{i=1}^{N} z_i^2}$$

where there are N units, the attribute value for each unit *i* is y_i, and w_{ij} is the weight (or connectivity) for units *i* and *j*. Notice that the locational information for this formula is found in the weights. For non-neighboring tracts, the weight is zero, so these add nothing to the correlation.

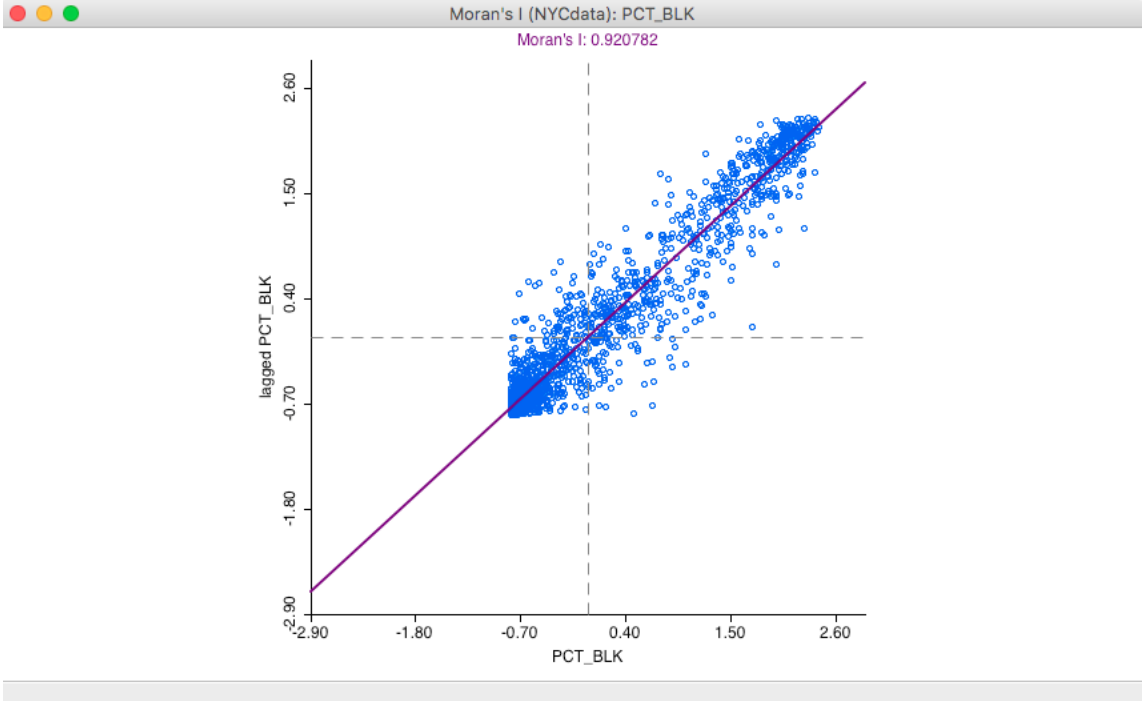
The expected value of Moran's *I* is -1/(*n* - 1), and the interpretation is similar to that of the product moment correlation coefficient. Informally, +1 indicates strong positive spatial autocorrelation (i.e., clustering of similar values), 0 indicates random spatial ordering, and -1 indicates strong negative spatial autocorrelation (i.e., a checkerboard pattern).

3. Univariate Moran's I

a) Click the map window, deselect any selected tracts (by clicking in the blank area), go to **Space > Univariate Moran's I**

This opens the variable selection box. Locate our newly created PCT_BLK -percent Black- (at the end of the list), leave the **Set the variables as default** box unchecked, Click **OK.** (A Moran's I scatterplot will appear).

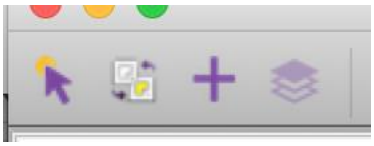
"Ignore" the message about isolates (a noun). The removal of "islands" is done automatically.



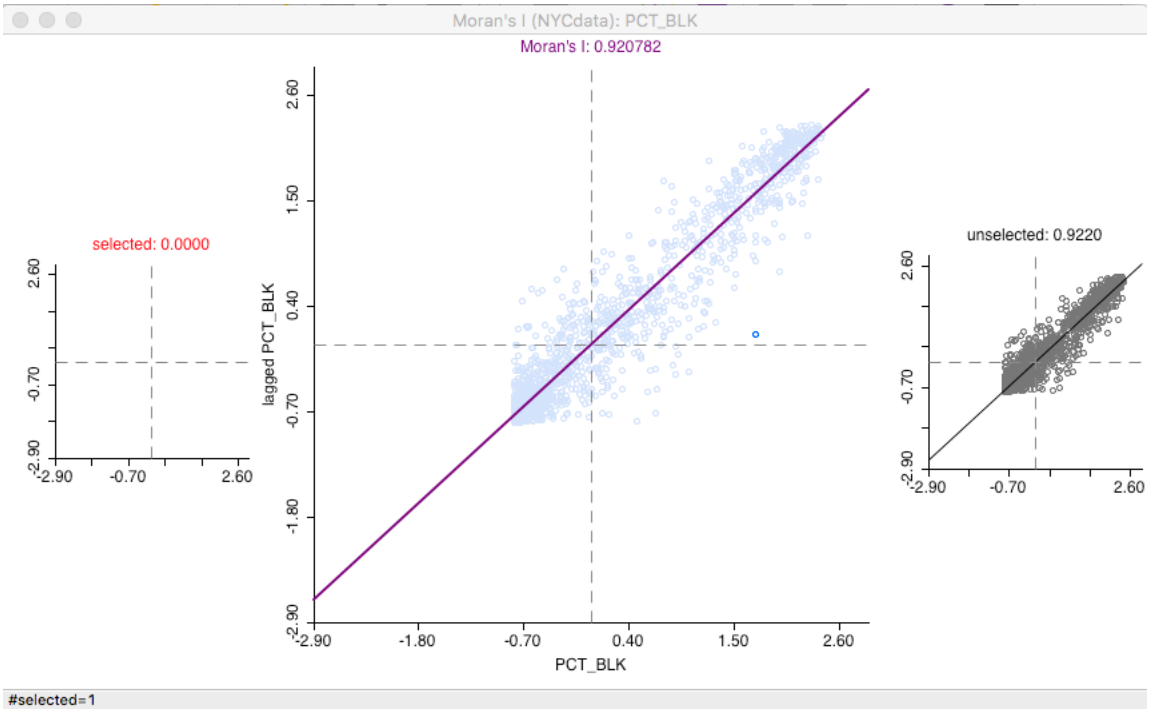
This scatterplot shows the value of original variable (% Black in the tract) on the horizontal axis and the spatial lag of the variable (average % Black in the tract's neighbors) on the vertical axis. Both variables are standardized and the graph is divided into four quadrants: high-high (upper right) and low-low (lower left) indicating positive spatial autocorrelation; and high-low (lower right) and low-high (upper left) indicating negative spatial autocorrelation. The slope of the regression line is Moran's I.

b) Observation exclusion (this tool does not appear to be working as it once did, sorry)

There are a number of options for the Moran scatterplot. First, you can exclude selected data points – such as the outliers in the scatterplot. Once selected you can inverse the selection by open a map of the same data and using "invert select" tool on the map window menu.



Click a lone case (or draw a box around it) on the far right side, then go to **`Options > Selection Shape, rectangle;`** this allows you to calculate the Moran's statistic for a subset of the distribution.



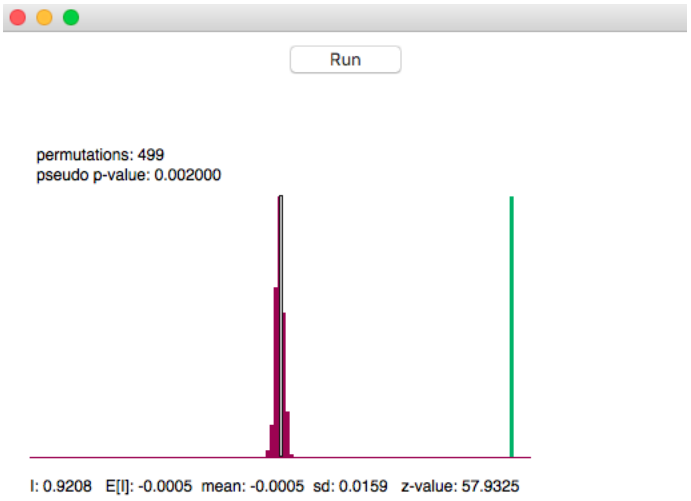
This results in the recalculation of Moran's I for a layout without the selected observations. Notice: a new regression line is drawn.

c) You can compute a reference distribution to assess the significance of a Moran's I spatial autocorrelation statistic.

With the Moran scatterplot window active, choose

`Options > Randomization > 499 permutations`

This sets the number of permutations to compute a reference distribution to 499, and then generates a **"randomization histogram"** for the reference distribution; the observed Moran's I is shown as a yellow bar and a pseudo-significance level is displayed on the top left under "permutations." Additionally, the graph also lists the Moran's I, the mean for Moran's I, and both the mean and standard deviation for the reference distribution.



In the Randomization graph, you can re-**Run** to generate another set of simulated values.

d) Save results

GeoDa allows you to save the calculated Moran results into a table so that you can keep them in your shapefile.

Close the **Randomization** window, then go to **Options > Save Results**... This will open the **Save Moran-Plot Results** window**, check** both **Standardized Data** and **Spatial Lag, GeoDa** assigns names to these variables automatically, but you can choose the name you like.

4. Multivariate Moran

Go back to the map window, deselect any selected observations.

a) Go to **Space > Bivariate Moran**... In the **Variables settings** dialogue window**,** select **PCT_ASN** as the second variable**.**

This scatterplot shows the spatial lag of the first variable (PCT_BLK) on the vertical axis and the value of the second variable (PCT_ASN) on the horizontal axis. Both variables are standardized. Again, the distribution is shown in four quadrants to indicate positive and negative spatial autocorrelation. The slope of the regression line shows the degree of spatial association between the variables at neighboring locations. This is somewhat troubling, as it compares the value for

the first variable in the target to the value of the second variable in the neighbors, without including in this calculation the value for the second variable from the target tract.

b) Inference based on randomization. Repeat the procedure to create a randomization histogram. This gives you a measure of the significance of the association between the % Asian in a tract and the % Black in neighboring tracts.

5. Local Moran's i

Local spatial autocorrelation statistics provide a measure, for each unit in the region, of the unit's tendency to have an attribute value that is correlated with values in nearby areas.

$$I_i = z_i \sum_j w_{ij} z_j$$

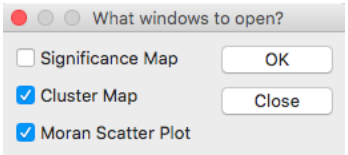
Where z_i and z_j are standardized scores of attribute values for unit i and j, and j is among the identified neighbors of i according to the weights matrix w_{ij}.

The local spatial autocorrelation analysis is based on LISA (Local Indicators of Spatial Autocorrelation) statistics. This computes a measure of spatial association for each individual location.

Go back to the map window, **deselect** any selected observations.

a. Go to **Space > Univariate Local Moran's I**

This opens the variable selection box. Locate our newly created PCT_BLK -percent Black, Click **OK.** You will have to select a weights file as well, if your initial weights file (.gal) doesn't work you can create a new one using k neighbors with 6 neighbors.



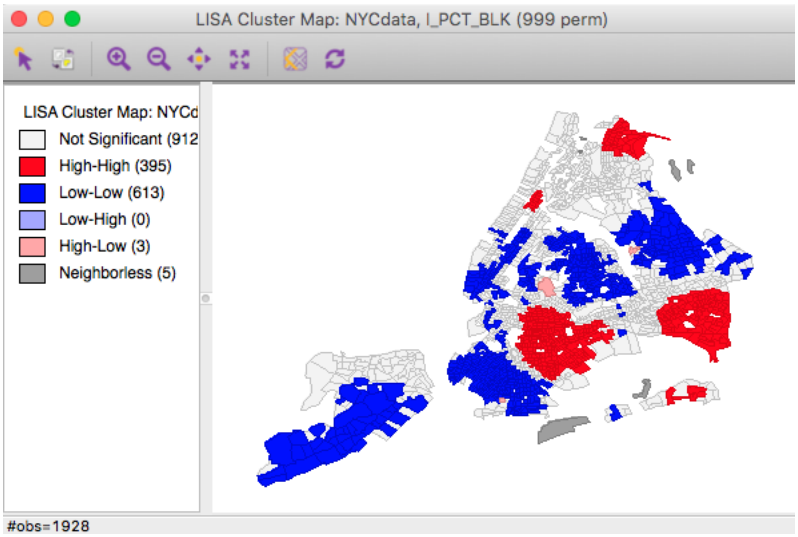
This gives user the option of creating three figures:

a) The Significance Map- indicates observations with significant Moran statistics

b) The Cluster Map- indicates significant cases and type of spatial association;

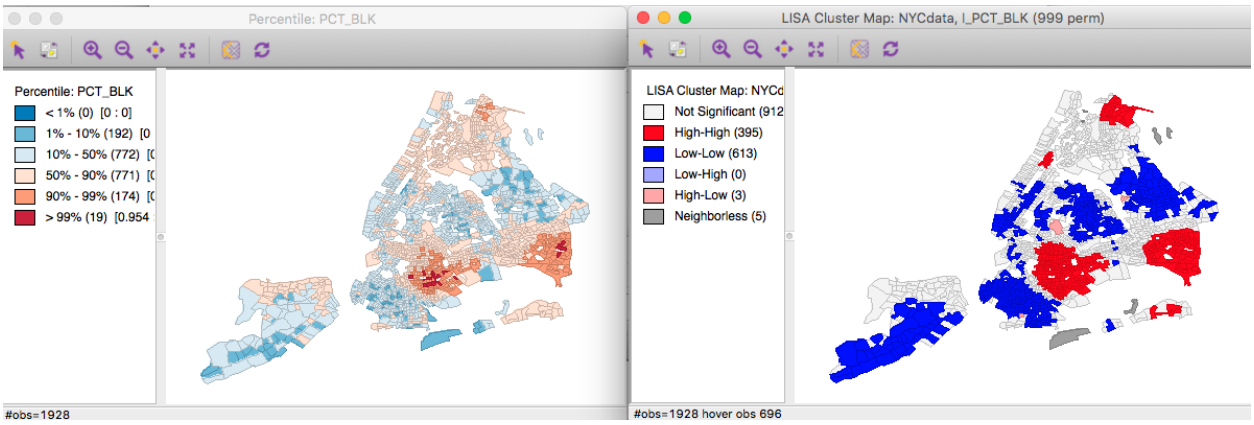
c) The Moran Scatter Plot shows the global Moran's I scatterplot, that we already created.

Choose only "**cluster map**," Click **OK.**



b. Compare the LISA map to a Percentile map of PCT_BLK.

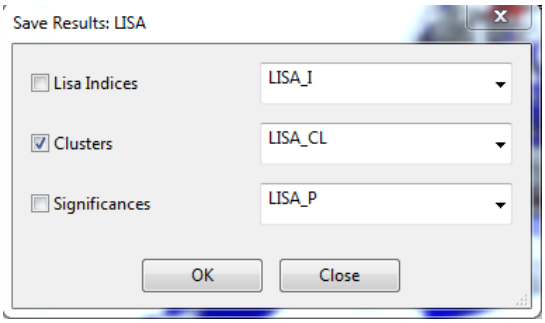
Go to **`Map > Percentile map`**



What is different between the maps? Look closely on the map to see if you can find what is identified in LISA map (significant spatial clusters) but not in the percentile map, and vice versa.

c. Save results

Click the Moran cluster map window, and go to **`Options > Save Results`**, in the **`Save LISA Results`** window, check Clusters, Click **`OK`**. The names on the right (LISA_I, LISA_CL, and LISA_P) are the field names that will be added to the table (in our case only LISA_CL will be added.



d. Repeat steps a through c for PCT_ASN.

e. Locate the **Table** window, slide the bottom bar to the right end, you should find the saved variables – these are values of Local Moran's i for each tract for the African American and Caribbean maps. **Important**: these added variables will be lost unless you save them into a new shapefile.

Go to **`File > Save as…, New datasource (nyc2016_cluster)`**, then click **`OK.`**

Summary

In this session, we learned the basic table functions of **`GeoDa`**. We had a brief introduction of the basic concepts of spatial autocorrelation, and we learned to calculate a measure of spatial autocorrelation: Moran statistics. We also did one exercise to learn how the results of **`GeoDa`** can be exported for analysis and map making in ArcGIS.