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. OpenGeoDa is the cross-platform, open source version of Legacy GeoDa. While Legacy GeoDa only runs on Windows XP, OpenGeoDa 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). We are working towards eventually releasing OpenGeoDa as an open source program.

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 lattice data (points and polygons). 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 latest version contains several new features such as a cartogram, a refined map movie, parallel coordinate plot, 3D visualization, conditional plots (and maps) and spatial regression.

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."

OpenGeoDa can be downloaded at: http://geodacenter.asu.edu/software

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

We will use the same US 2000 Census data from New York City that was used in the previous GeoDa Exercise.

The shapefile nyc2000.shp is the map of New York City with Census 2000 data from summary file 3. These are socioeconomic attributes for 2219 Census tracts in five boroughs. It includes the following variables:

nyc2000.shp

Variable name	Label
POLYID	Polygon ID
STATE	State FIPS
COUNTY	County FIPS
TRACT	Census Tract ID
sctrct00	FIPSID
hvalue	Median housing value
t0_pop	Total population
t0_nhw_f	Total number of non-Hispanic white persons
t0_nhb_f	Total number of non-Hispanic black persons
t0_hsp_f	Total number of Hispanic persons
t0_asn_f	Total number of Asian persons
t0_min	Total number of minority persons
pctnhw	Percent non-Hispanic white persons
pctnhb	Percent non-Hispanic black persons
pcthsp	Percent Hispanic persons
pctasn	Percent Asian persons
pctmin	Percent minority persons
chn00	2000 Chinese
fil00	2000 Filipino
jap00	2000 Japanese
ain00	2000 Asian Indian
kor00	2000 Korean
m0_mex	2000 Mexican(Mumford estimates)
m0_prn	2000 Puerto Rican(Mumford estimates)
m0_cbn	2000 Cuban(Mumford estimates)
m0_dom	2000 Dominican(Mumford estimates)
t0_afa	2000 African American
t0_car	2000 Afro-Carribbean
t0p_own	Percent homeowners
t0p_vac	Percent vacant housing
t0p_coll	Percent college educated
t0p_prf	Percent of people employed in professional/managerial occupations
t0p_uemp	Percent of people unemployed
t0p_nat	Percent persons born in the United States
t0p_for	Percent foreign born persons
t0p_rec	Percent recent immigrants
t0p_old	Percent older immigrants
t0p_only	Percent persons who speak only English at home
t0p_oth	Percent persons who speak language other than English at home
t0_minc	Median household income
t0_pcinc	Per capita income
t0p_poor	Percent total population below poverty
m0p_poor	Percent minority population below poverty

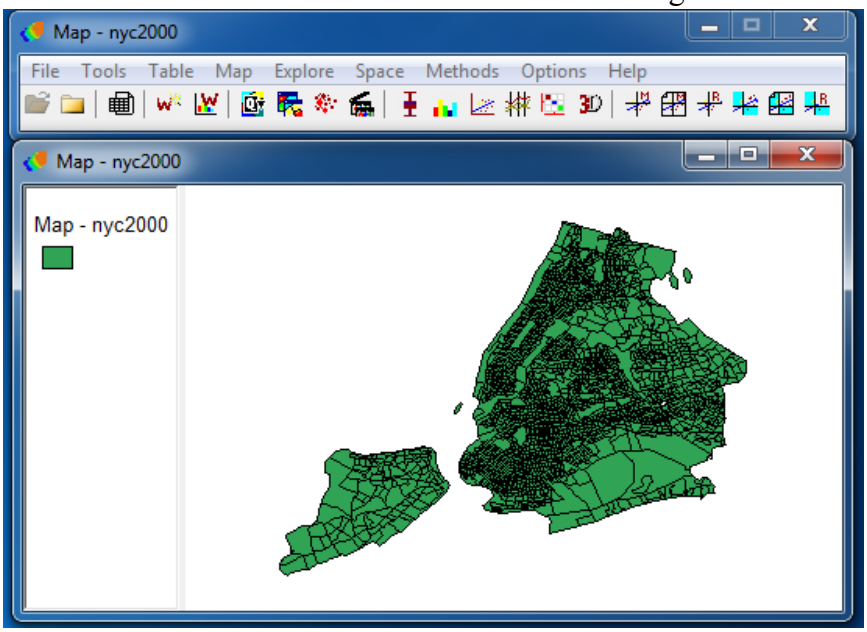
Manage tables in GeoDa

1. Start **GeoDa**

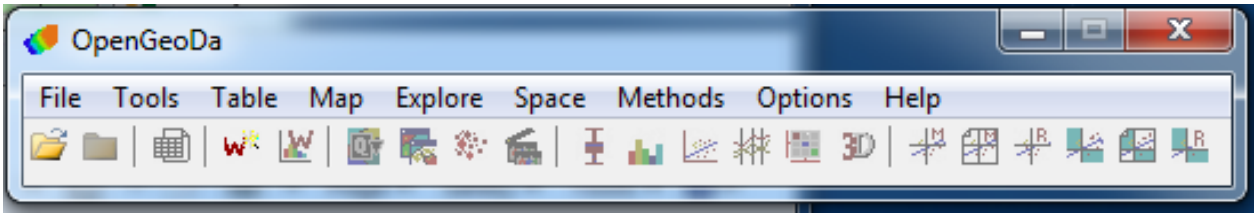
2. Go to **`File > Open Shapefile.`**

3. **`Browse`** through the folders to find and select the shape file, **`nyc2000.shp`**.

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



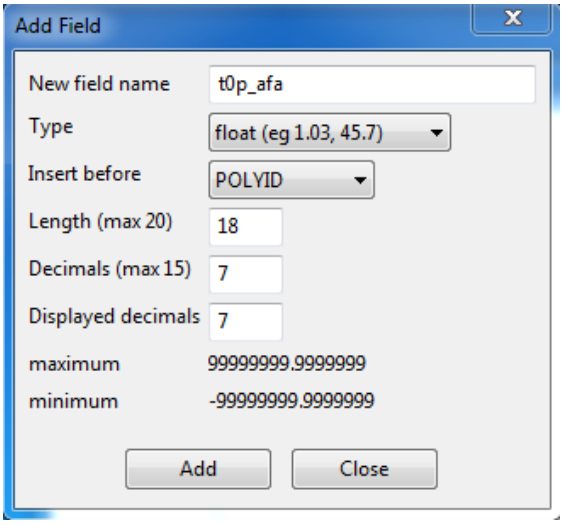
5. Click Table icon on the tool bar to open the attribute table.



6. To compute additional variables: First add variable columns

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

`type t0p_afa` (for percent African American) in the **`Input Column Name`** box, select **`real`** as the **`Type`**, select a location (logical locations would be near to_afa or near the beginning of the data (after HVALUE), and click **`Add (accept defaults for length, etc).`**



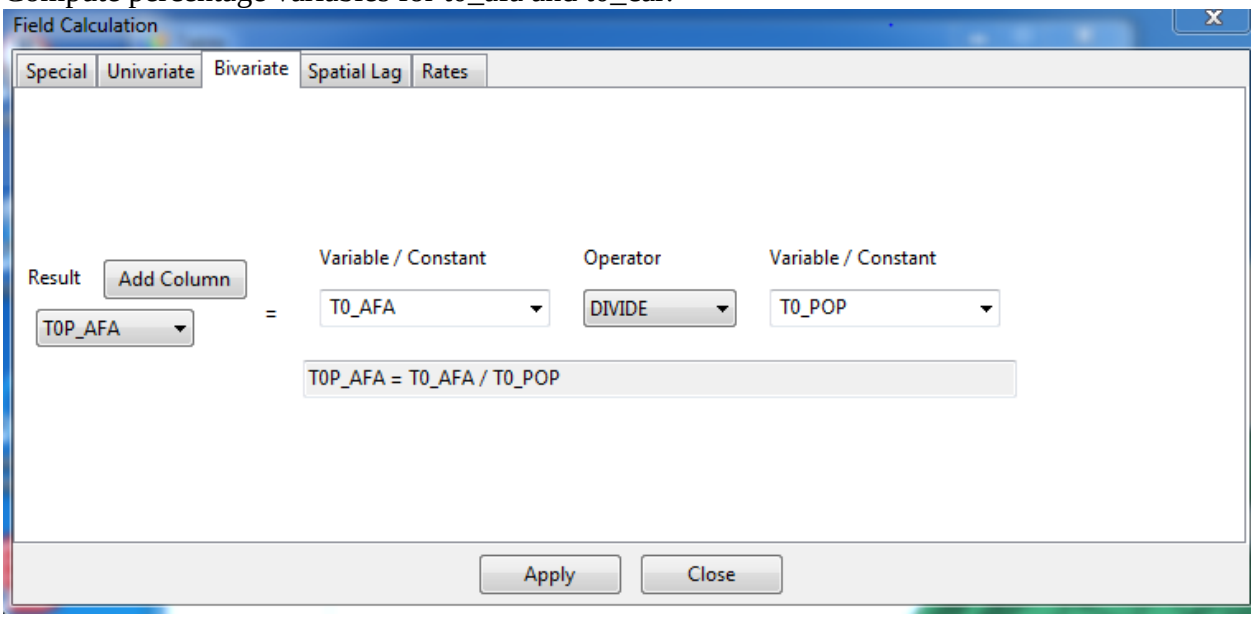
A blank column titled "**T0P_AFA**" will appear in the specified location.

Repeat the steps to add another column titled "**T0P_CAR**" (for percent West Indian).

Calculate values for the two added columns

Go to **`Table > Variable Calculation`** to open the dialogue box

Compute percentage variables for t0_afa and t0_car.



Choose **`Bivariate Operations > Select T0P_AFA`** in the **Results** listing, **`Select T0_AFA`** in the **Variables-1** listing. **`Select DIVIDE`** in the **Operators** listing. **`Select T0_POP`** in the **Variables-2** listing. Then Click **`OK`**.

For the T0P_CAR,

Choose **Binary Operations > Select T0P_CAR** in the **Results** listing, **Select T0_CAR** in the **Variables-1** listing, **Select DIVIDE** in the **Operators** listing, and **Select T0_POP** in the **Variables-2** listing, then Click **OK**.

Save the results
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 **Table > Save Copy of Shape file** and **Type NYC-new,** and click **Save.**

d) Close the current project,

Go to **File > Close All**.

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:

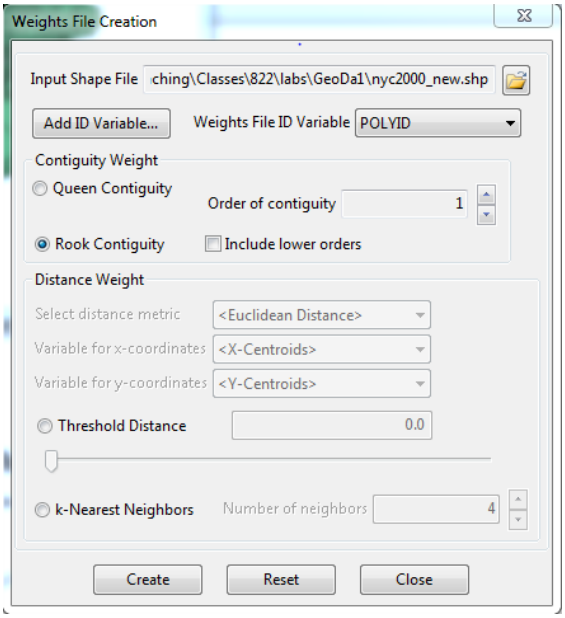
$w_{ij} = \frac{C_{ij}}{\sum_{j=1}^{N} C_{ij}}$ with C_{ij} =1 when i is linked to j, and C_{ij} =0 when otherwise.

1. Go to **File > Open Shapefile**.

2. **Browse** through the folder to find and select the shape file, **nyc2000-new.shp**.

3. Click **OK**.

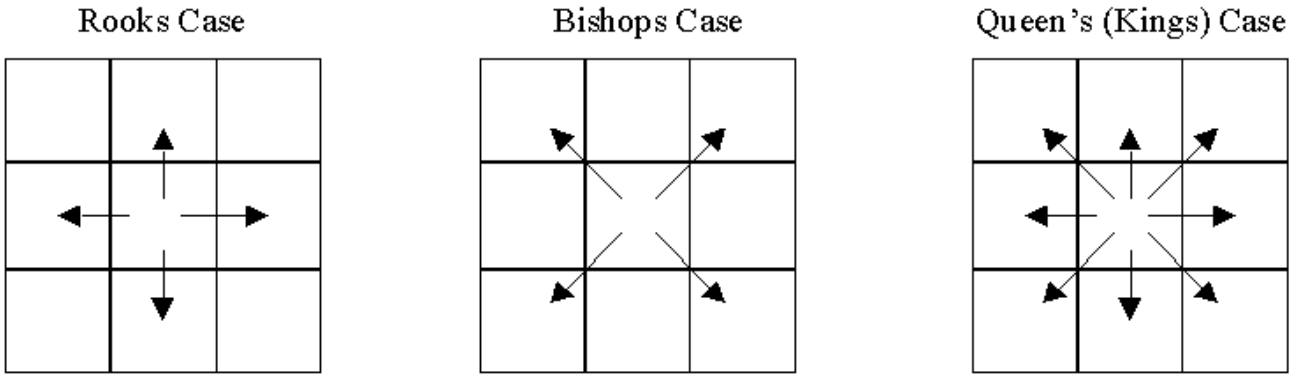
4. Go **to Tools > Weights > Create** to open the **Creating Weights** dialogue box.



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 neighbours 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 neighbours

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

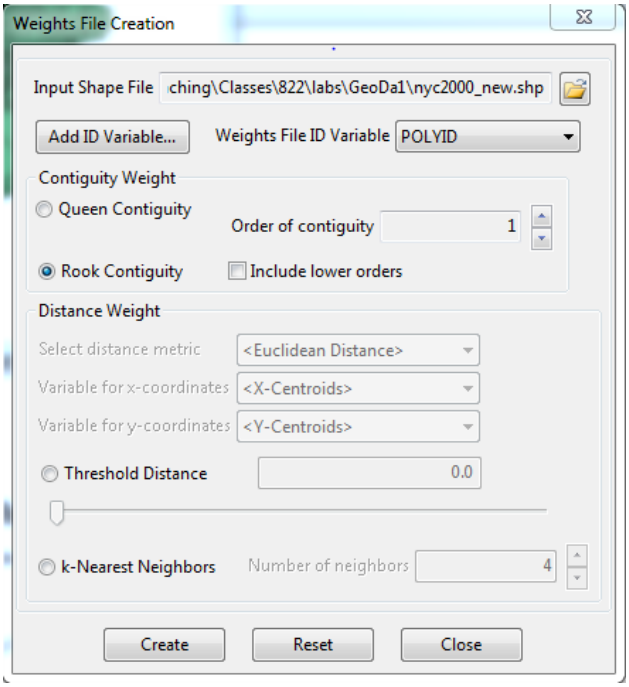
Here, we will calculate weights based on rook contiguity. In the **Creating weights dialogue** box:

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 **nyc2000-new.shp** as the input, **POLYID** as the **Weights File ID Variable (DON'T Add ID Variable).**

Select **Rook Contiguity**, click **Create**, then **Done**

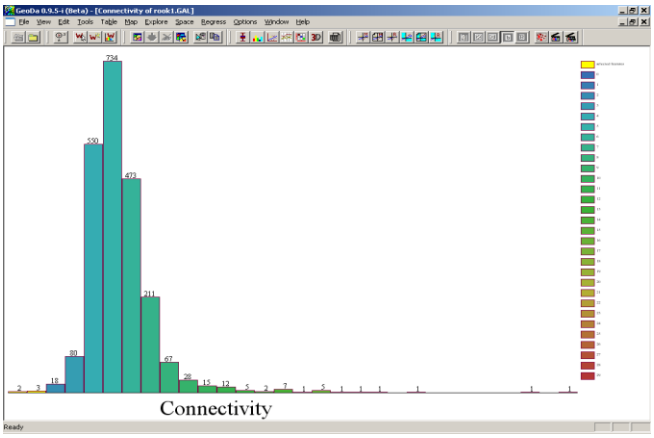
.



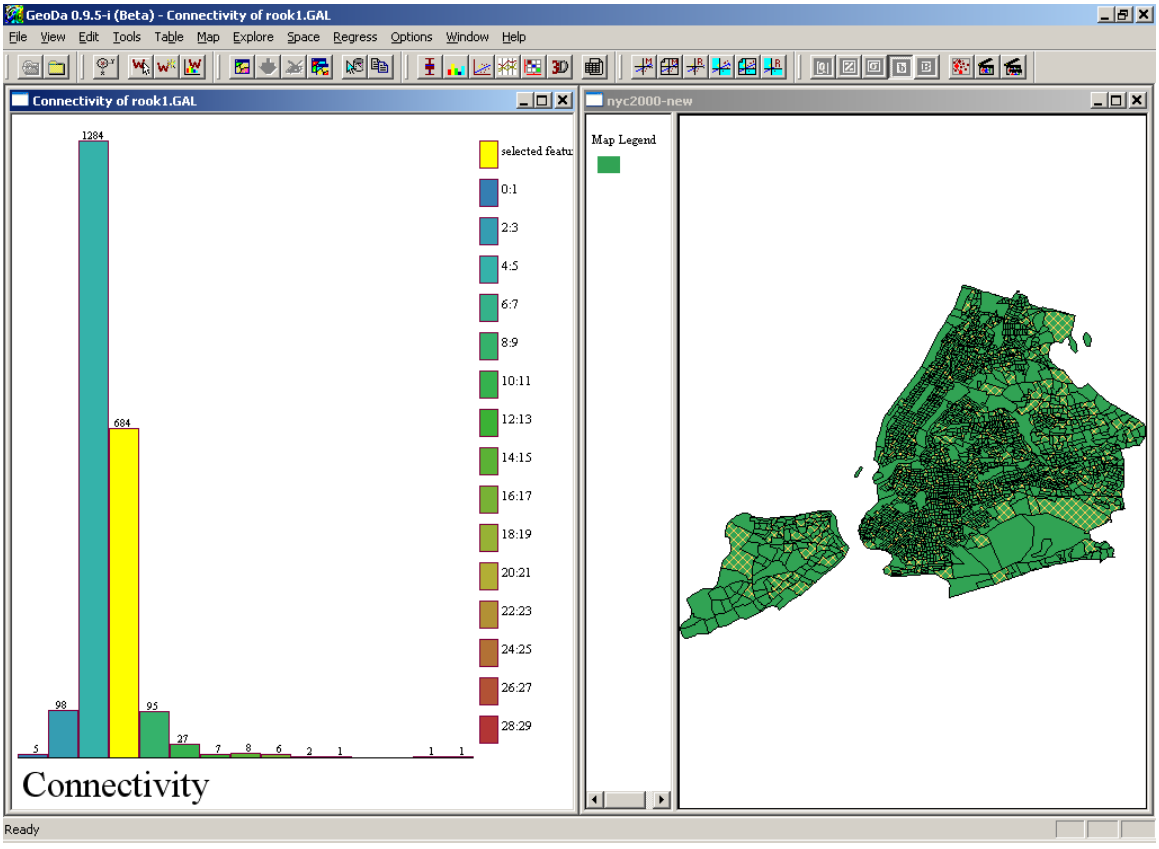
5. Characteristics of a weights matrix.

To explore the characteristics of the weights matrix, go to **Tools > Weights> Properties,** This opens up a **"Weights Characteristics"** dialogue box where you can select the relevant file**.** Select **Rook.GAL,** then click **OK.**

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.



7. 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% blacks 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.

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

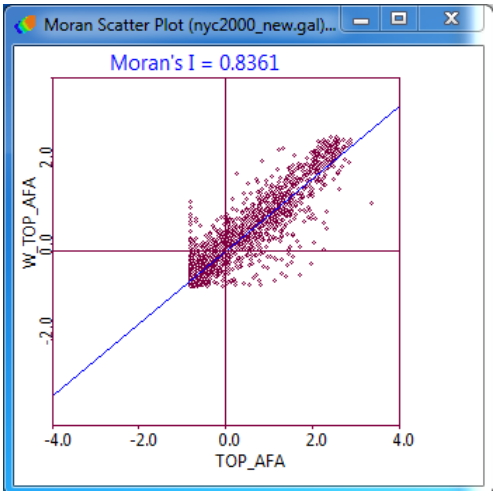
where there are N units, the attribute value for each unit *i* is y_i, and w_{ij} is the weight (or connectivity) for units *i* and *j*. Notice that the locational information for this formula is found in the weights. For non-neighboring tracts, the weight is zero, so these add nothing to the correlation.

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`**

This opens the variable selection box. Locate our newly created T0P-AFA -percent of African American- (at the end of the list), leave the **`Set the variables as default`** box unchecked, Click **`OK.`** (A Moran's I scatterplot will appear).

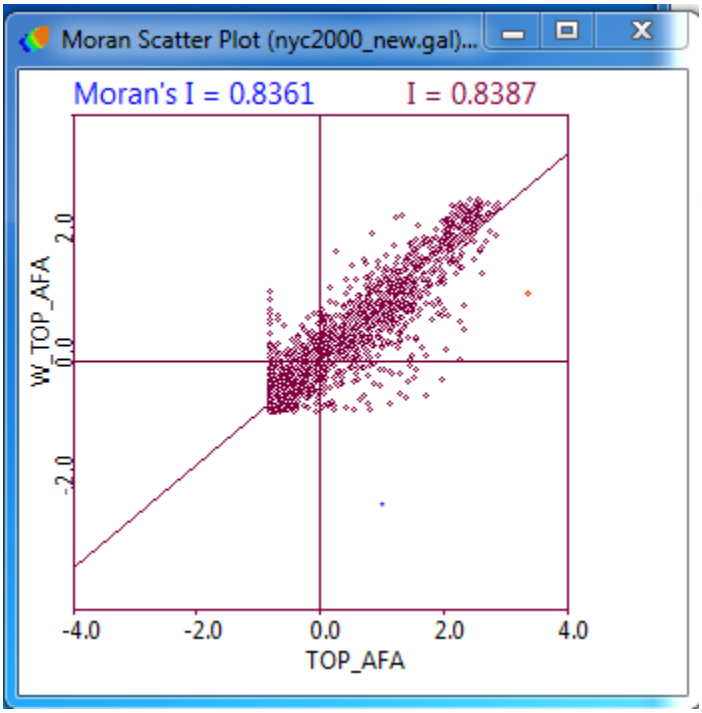


This scatterplot shows the value of original variable (% African American in the tract) on the horizontal axis and the spatial lag of the variable (average % African American in the tract's neighbors) on the vertical axis. Both variables are standardized and the graph is divided into four quadrants: 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

There are a number of options for the Moran scatterplot. First, you can exclude selected data points – such as the outliers in the scatterplot.

Click the lone case (or draw a box around it) on the far right side, then go to **`Options > Show Regression of Selected Excluded`**.



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

You can also select multiple cases by clicking cases while holding the **`Shift`** button. Alternatively, with the **`Ctrl`** button depressed, you can **`Left click and then drag and release`** to make a rectangle. Move the rectangle around, the observations covered by it will be excluded, and corresponding slope is calculated.

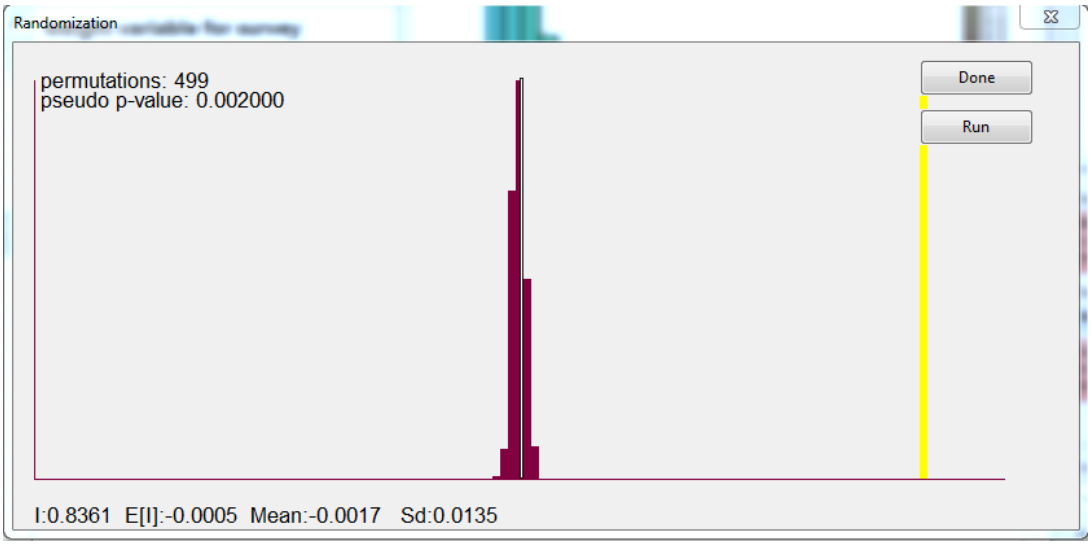
c) You can compute a reference distribution to assess the significance of a Moran's I spatial autocorrelation statistic.

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

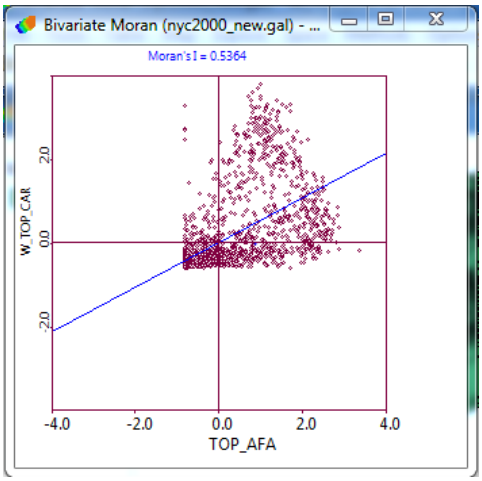
GeoDa allows you to save the calculated Moran results into a table so that you can keep them in your shapefile.

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.

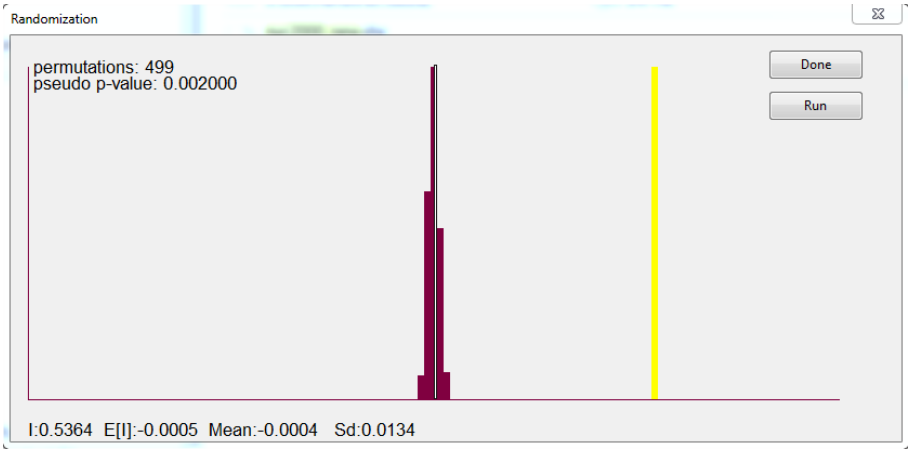
a) Go to **Space > Multivariate Moran**... In the **Variables settings** dialogue window, select **TOP_CAR** as the second variable.



This scatterplot shows the spatial lag of the first variable (T0P_AFA) on the vertical axis and the value of the second variable (T0P-CAR) 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.

b) Repeat the cases exclusion for outliers. Notice the slope changes.

c) 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 % Caribbean in a tract and the % African American 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.

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

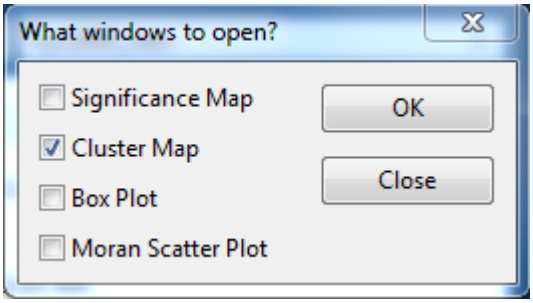
Where z_i and z_j are standardized scores of attribute values for unit i and j, and j is among the identified neighbors of i according to the weights matrix w_{ij}.

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 T0P-AFA -percent of African American, 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.



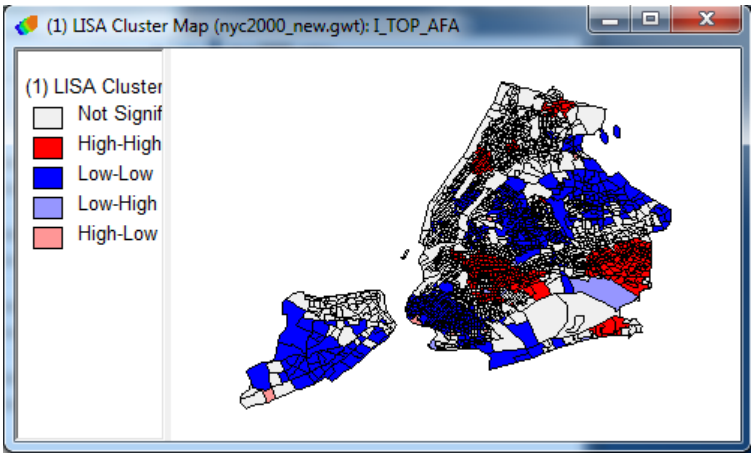
This gives user the option of creating three figures:

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

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

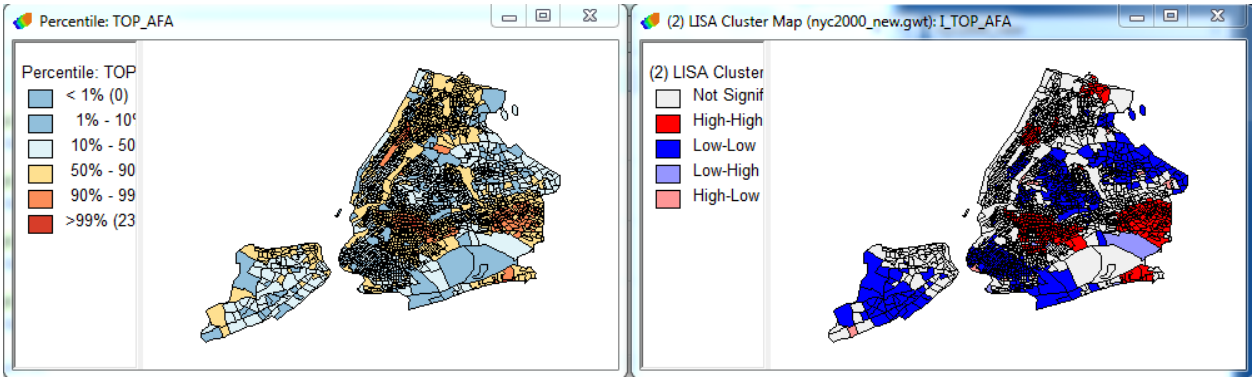
c) The Moran Scatter Plot shows the global Moran's I scatterplot.

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



b. Compare the LISA map to a Percentile map of T0P_AFA.

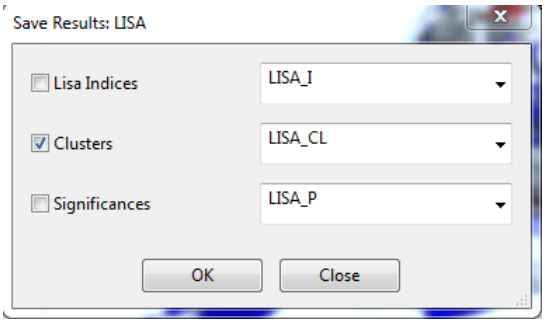
Go to **Map > New Map Window**



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 T0P_CAR.

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 **`Table > Save Copy of Shape File (nyc2000_cluster)`**, then click **`Save.`**

EXTENSION (complete this later so you have time to complete GeoDa3)

6. Making an ethnic black map in ArcGIS

This will demonstrate how to incorporate **`GeoDa`** results into other GIS software for analysis and presentation.

1) **`Start ArcMap > Open`** new project, Add layer **`nyc2000_cluster.shp,`**

2) Create a new variable- **ethblk**, with three categories: **African American cluster alone**, **Afro-Caribbean cluster alone**, and **clusters for both ethnic black groups**.

3) Create a thematic map based on **ethblk**. How do we interpret the map?

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.