GeoDa™ 0.95i Exercise

Stuart Sweeney and Rob Farrell
University of California – Santa Barbara

July 12, 2004

Outline

1. Background

2. Data set introduced ("Santa Barbara" home sales data)

3. GeoDa introduced

4. Bivariate and Multivariate exploratory analysis

5. Spatial autocorrelation and spatial EDA

6. Spatial regression and diagnostics

In this handout all "commands" are printed in **`Courier New (Bold)`**

1. Background

GeoDa™ is a trademark of Luc Anselin.

GeoDa is a collection of software tools designed for exploratory spatial data analysis (ESDA) based on dynamically linked windows, and this software replaces the DynESDA Extension for Arcview 3.x. GeoDa is freestanding and does not require a specific GIS system.

GeoDa has evolved from efforts to couple SpaceStat and DynESDA with ESRI products (e.g., Arcview 3.x) via extensions. GeoDa adheres to ESRI's shapefiles as the standard for storing the information, using MapObjects LT2 technology for spatial data access, mapping and querying.

Luc Anselin suggests http://www.csiss.org/learning_resources/content/syllabi#gis for extended course notes and examples dealing with an introduction to spatial data analysis, and requests that users ***please*** **report "anything that seems like a bug"** to anselin@uiuc.edu (or post to the Openspace mailing list: mailto:openspace@sal.agecon.uiuc.edu).

2. Data set introduced

GeoDa comes with several data files as samples (e.g., Crime in Columbus [tracts], SIDS in North Carolina [counties]).

In this exercise we will use home sales data from Santa Barbara. There are individual home attributes, including sale price for 418 locations, as well as fixed effects for school district and sales region. Also included are fixed neighborhood effects at three different geographies, allowing the user to assess the effect of aggregation on the spatial analysis and regression. A complete listing of the data elements can be found in the Data Dictionary.

The exercises that follow are illustrative only and are not intended to cover the fully functionality of GeoDa.

3. GeoDa introduced

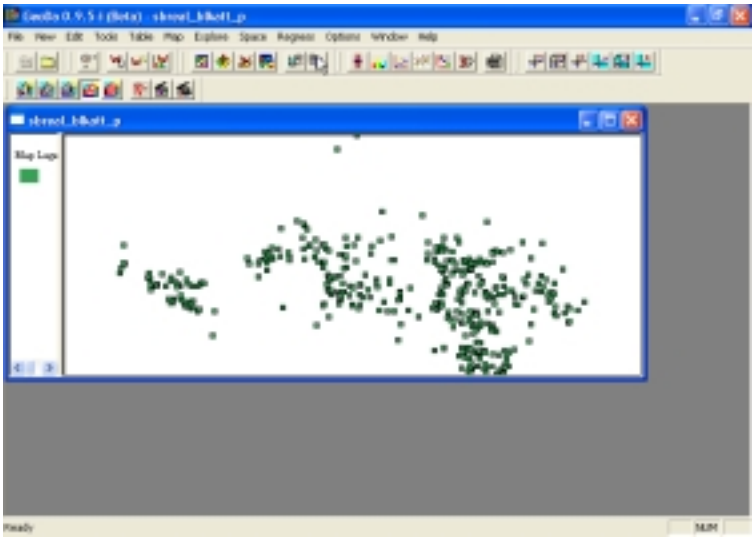
Start GeoDa

a) **File > Open Project** (This opens the GeoDa Project Setting dialog box).

b) **Browse** through the folder to find and select the shape file, **sbreal_blkatt_p.shp** (should be in the SB housing lab folder). Please note for this exercise to run the data must be on the local machine (i.e., not accessed over a network).

c) Select **ID** as the "Key Variable" (by scrolling and clicking on variable name). The Key Variable must have a unique value for each observation (i.e., in this case district). The unique value is used to implement the link between maps and statistical graphs. *For later exercises (optional), the ID field is at the end of the scroll list.*

d) Click **OK**. Your screen should now look something like … (you will have to **click** on the left side of the map window and then **drag** to the center of the map to view the legend area).



e) It is a good idea to maximize the GeoDa window. **Click on the full screen button for GeoDa,**  **.** Note, all windows in GeoDa can be resized and positioned anywhere within the main program window.

The GeoDa menu bar contains twelve menu items

1. File (Project Toolbar)
2. View
3. Edit (Edit Toolbar)
4. Tools (Weights Toolbar)
5. Table
6. Map
7. Explore
8. Space
9. Regress
10. Options
11. Window
12. Help

with the important menu items matched by a "button" on the toolbar. Toolbar components [Project, Tools, Edit/Map Window, Explore, Space, and Map] can be moved and docked anywhere within the main program window. For example,

The ***File*** Menu ***(Project Toolbar)*** contains the standard project management commands (e.g., open a new project, close project windows).

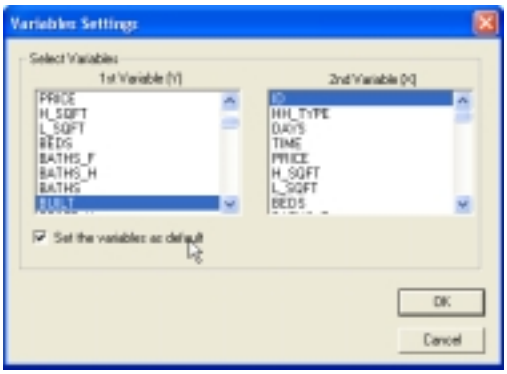
The ***Edit*** Menu ***(Map Window Toolbar)*** allows the user to (a) manipulate maps and layers (e.g., new map, duplicate map, add layer, remove layer), (b) select variables, and (c) make use of the Windows clipboard.

Basic data exploration

a) First, we will improve the background color for more clarity. **Right-click** on the map, and choose **Color > Background**, and choose a light gray.

b) Create histograms. For example, we could explore the distribution of "year built".

c) To select a variable, choose **Edit > Select Variable**. This opens a variable settings dialog box. (Note: for univiate operations, the choice of second variable is ignored).



d) In the dialog box **select** the variable of interest (**BUILT**) by scrolling down the 1st variable Y listing.

e) Notice that the box "select variables as default" is checked. If the variable is set as a default (i.e., checked) then all mapping or statistical graph options assume this is the focal variable. If this check box is not marked, the variable selection dialog box will open for each mapping or statistical operation. In the example above, BUILT (year built) is selected.

f) Click **OK**.

> Note: Tables may appear following certain operations. The user may close, move, or minimize the table. Instructions will assume that the table is closed.

g) To produce a basic (non-spatial) histogram of BUILT, select **Explore > Histogram** (if BUILT was not checked as the default than select BUILT if necessary). With so many classes, the histogram is a bit difficult to interpret, but the lowest class is '1885', and values proceed left to right.

h) The spatial component of this distribution can be viewed, however, using the Map Movie functionality. Choose **Map > Map Movie > Cumulative**. Make sure the new window and the histogram are visible and set the speed of play (using the continuous toggle) to about 200. Choose **Play**. Note the build-out of Santa Barbara and the later suburbanization. Also note, the linked display of the histogram with the map movie. When finished, close the

map movie and histogram windows. Also, maximize the map and click anywhere on it (in order to "de-select" those points selected by the map movie function.

i) We will now map individually the variables for year built, house price per square foot, and price.

j) Choose **Edit > Select Variable**, and choose BUILT (1st variable)—don't worry about choosing a second variable, only applies in multivariate situations (that comes later!).

k) Choose **Map > Quantile**, and select 9 categories. How much clustering is present in this variable?

l) Repeat steps j & k for the other two variables, PPSF_H and PRICE.

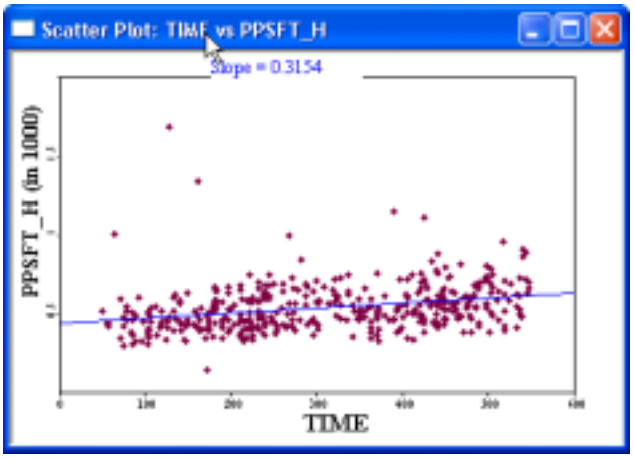
4. Bivariate and multivariate exploratory analysis

Scatterplots

a) Create scatterplots for price per square foot (house) versus time, and price per square foot (lot) versus time. To do this, first you must set the variables using the Edit menu.

b) Choose **`Edit > Select Variables`**, and select the **`PPSF_H`** and **`TIME`** as the variables.

c) Choose **`Explore > Scatterplot`**. The plot should look something like this…



What does this mean for price per square foot of housing in Santa Barbara?

d) Repeat b & c for price per square foot of the lot.

e) As you noticed in the map movie, GeoDa has the ability to link selections across multiple windows. Let's examine some of the points in the scatter plot simultaneously on the map!

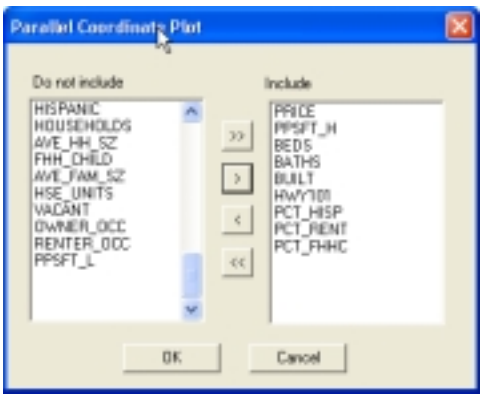
f) Tile vertically your quantile map and scatter plot. Click and drag to select a group of points in the scatter plot and notice that they become "highlighted" in the map window.

g) Now, with the **`Ctrl`** button depressed, click, drag, and release to create a small box in the scatter plot window. It will flash for a couple of seconds, and then become continuously active. You can move it around the scatter plot and dynamically highlight portions of the plot ("brushing"), all the while viewing the active selections in the map ("linking"). Simply click the mouse to end the brushing.

h) Close the scatterplot window.

Parallel Coordinate Plots

a) Choose **`Explore > Parallel Coordinate Plot`**. A window should appear allowing you to chose the variables you wish to examine…

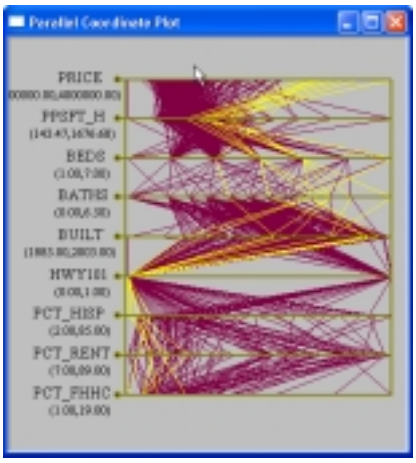


b) Add these variables in the following order – PRICE, PPSFT_H, BEDS, BATHS, BUILT, HWY101, PCT_HISP, PCT_RENT, PCT_FHHC, and click **`OK`**.

c) Maximize the plot window and make the background color gray, by **`right-click > Background Color`**.

d) By right-clicking, the user can toggle between original data and standardized scores.

e) Examine the data by selecting groups of data points based on some variable and assessing their simultaneous values on all the others. For example, none of the highest priced properties are within the Highway 101 buffer.



f) Just like in previous exercises, the selected items are linked to the map window, so that the currently selected are "highlighted" in the map window. Close the plot window.

5. Spatial autocorrelation and spatial EDA

Spatial autocorrelation measures such as Moran's I require the user to define a weights matrix that essentially defines the local neighborhood around each point. The value at each location is compared to a weighted average of the values of its neighbors. The weights file defines this neighborhood. For a point file, the only option is a neighborhood based on distance.

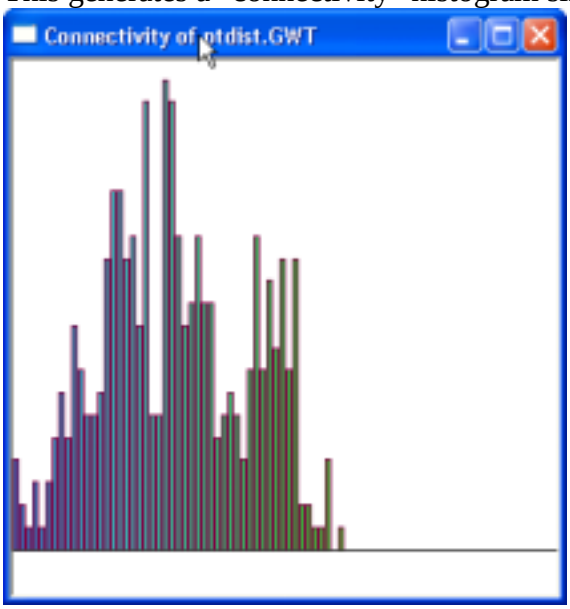
Creating a weights matrix

a) To construct a spatial weights file select **`Tools > Weights > Create`**. This opens up the "Creating Weights" dialog box. Select the **`sbreal_blkatt_p.shp`** file as the input. The general weights (output) file is saved as a .gal file—in the save output as box, choose a name for your weights file. The weights file used **`ID`** as the ID variable and the distance weight is based on a threshold distance. Set the threshold distance to 2000.

b) Click on **`Create`**, then **`Done`**.

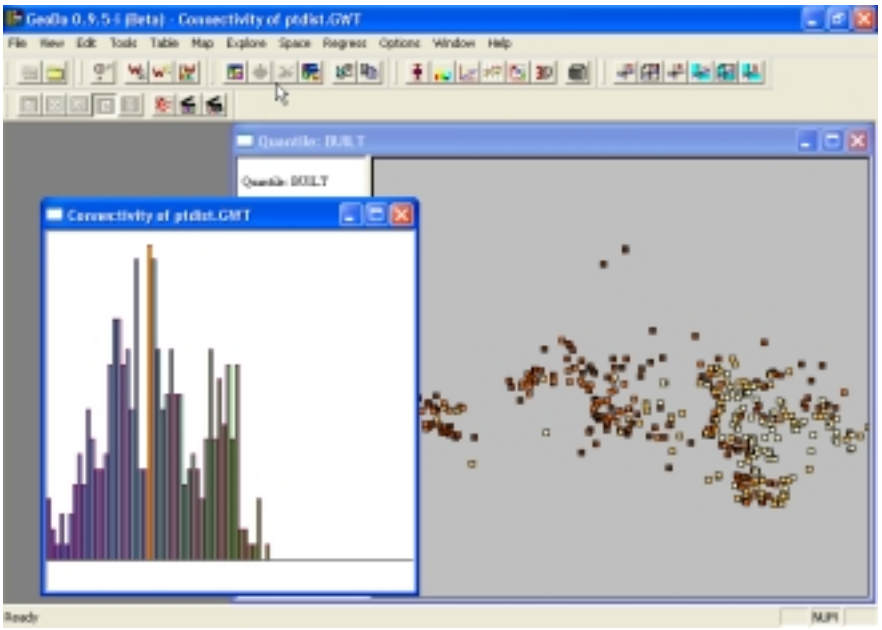
c) To explore characteristics of the weights file select **`Tools > Weights > Properties`**. This opens up a "Weights Characteristics" dialog box where one can browse for the relevant file. After selecting the file click on **OK**.

This generates a “connectivity” histogram similar to the one below.



With so many values, the histogram classes are difficult to discern, but we can tell that some locations only have a few neighbors, while others have 30 or more.

d) The histogram can be queried by **clicking** on a vertical bar. The corresponding points (that have that number of neighbors are highlighted in the map window.



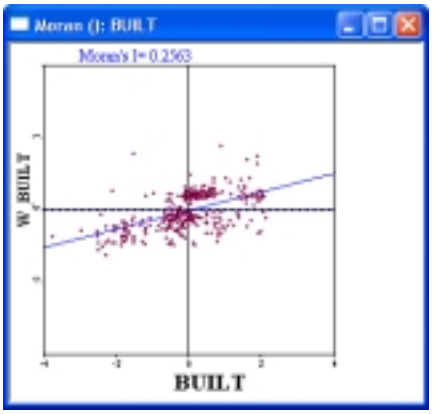
e) Close the histogram window by **clicking** on the  and deselect the “selected” districts by **clicking** anywhere on the map.

Univariate Moran

a) First, select a variable using the **Edit > Select Variables > BUILT** (as the 1st variable – for this exercise the 2nd variable won't play a role).

b) Choose **Space > Univariate Moran** (then select appropriate weights file – the one saved earlier.

c) Click on **OK**. (A Moran scatterplot will appear and will need to be resized).



This shows the spatial lag of the variable on the vertical axis and the original variable on the horizontal axis (based on threshold distance). The variables are standardized and the graph is divided in to four quadrants: high-high (upper right) and low-low (lower left) for positive spatial autocorrelation; and high-low (lower right) and low-high (upper left) for negative spatial autocorrelation. The slope of the regression line is Moran's I.

d) There are a number of options for Moran scatterplots. First, there is the option to exclude selected data points. To select some districts in the scatterplot (e.g., outliers), **Left click** and then **drag and release** to make a rectangle.

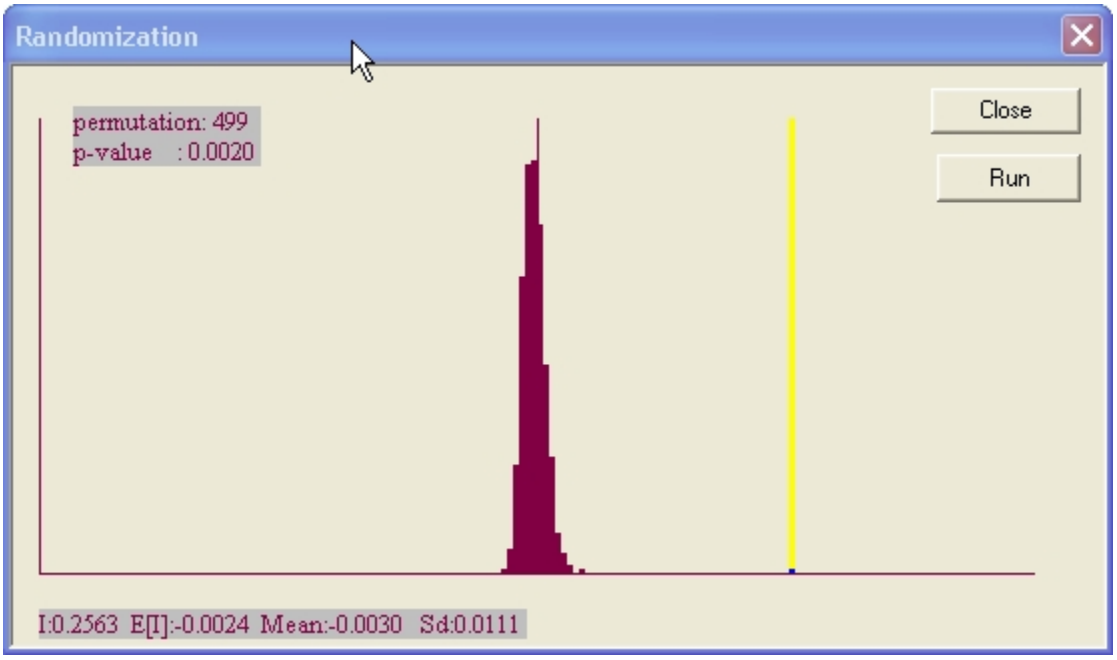
e) Use **Options > Exclude Selected**

When the excluded selection is active, the selection of observations in the graph(s) will result in the recalculation of Moran's I for a layout without the selected observations. The new regression line is shown in brown. Note, you may need to redraw the scatterplot using **Options > Exclude** once or twice to view the plot above.

f) To 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 be used in the computation of a reference distribution to assess the significance of a Moran's I spatial autocorrelation statistic and then generates a "randomization histogram" for a reference distribution, with the observed Moran's I shown as a yellow bar and a pseudo-significance level or p-value. Also listed on the graph are the Moran's I, the theoretical mean for Moran's I, and both the mean and standard deviation for the reference distribution.



g) In the randomization graph one can re-**Run** to generate another set of simulated values.

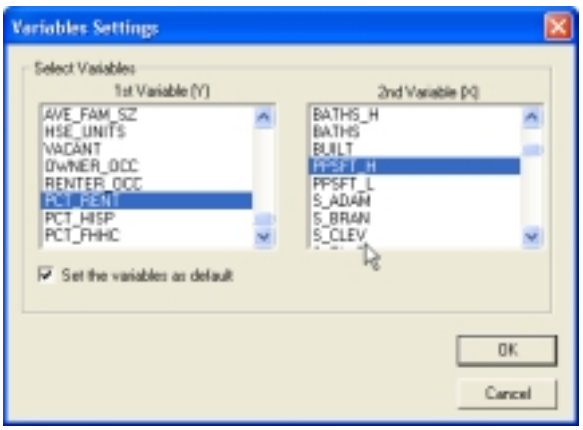
h) Just like with the Parallel Coordinate Plots, one can "brush" and link to the map. Do this for the high-high and low-low sections of the Moran scatter plot.

In this example we focused on the year built (BUILT). Perform steps a through h for the variables PPSFT_H, PCT_HISP, and PCT_RENT.

i) Close the windows, and de-select any selected locations in the map window.

Multivariate Moran

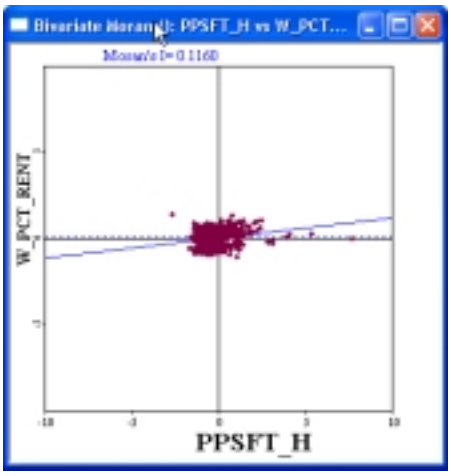
a) Select two variables using **`Edit > Select Variable`**.



b) In the dialog box **`select`** the variables of interest, **`PCT_RENT`** by scrolling up the 1st variable Y listing, and **`PPSFT_H`** by scrolling through the 2nd variable X listing. Notice that the box "select variables as default" is checked. Then click **`OK`**.

c) **`Space > Multivariate Moran`**. This opens up a dialog box prompting for the name of the weights file.

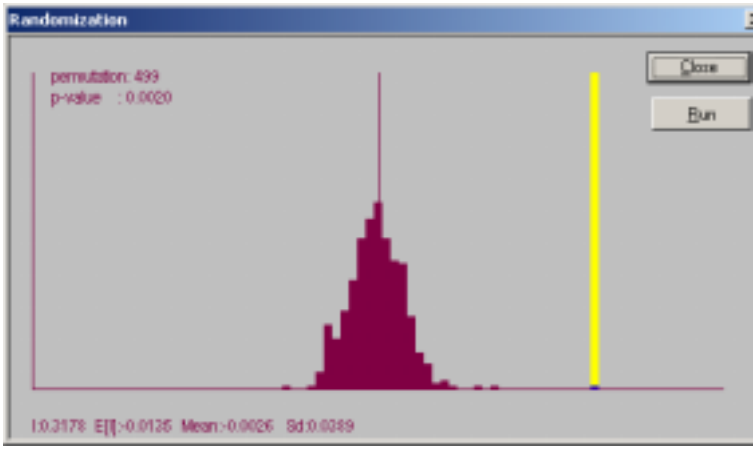
d) Browse for the weights file, as necessary and then click on **OK**.



This scatterplot shows the spatial lag of the first variable (PCT_RENT) on the vertical axis and the second variable (PPSFT_H) on the horizontal axis. Both variables are standardized. The slope of the regression line shows the degree of linear association between the variables at neighboring locations (as defined by the spatial weights file).

e) Inference is based on randomization.

`Options > Randomization > 499 permutations`



f) **`Close`** the Randomization histogram.

g) Close the Moran Scatterplot graph by clicking on **`File > Close`**.

h) Perform a multivariate Moran for the following pairs of variables: PCT_RENT – PRICE, PCT_HISP – PPSFT_H, PCT_HISP – PRICE, and any others you might want to explore. Can you explain the relationships described by the Moran values??

i) Close all scatter plot windows and make sure all locations have been de-selected.

Univariate LISA

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

The LISA statistics requires that both a variable(s) be specified (Edit > Select Variable) as well as a spatial weights file (see Tools > Weights > Create if this file does not exist).

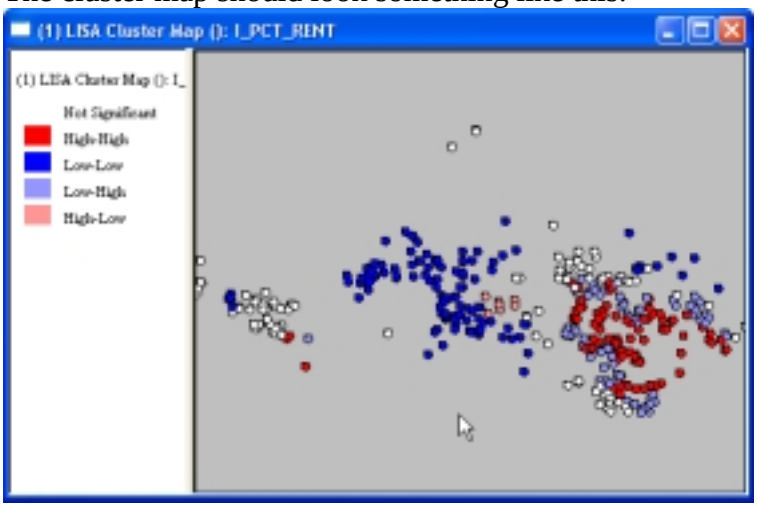
a) Change the default variable to PPSFT_H using **`Edit > Select Variable`**.

b) Univariate LISA statistics are invoked by **`Space > Univariate LISA`**.

c) Browse for the weights file, as necessary and then click on **OK**.

This gives the user the option to generate 4 figures **`(choose only "cluster map")`**

The cluster map should look something like this:



d) Compare the LISA map to a quantile map of PPSFT_H. What is the benefit of the LISA?

e) Create a histogram for the S_WASH variable.

f) Click on the "1" value in the histogram and look at the LISA cluster map for just those with children at that school. Do the same with a histogram of the HWY101 variable.

g) Perform steps a through f for the PRICE variable.

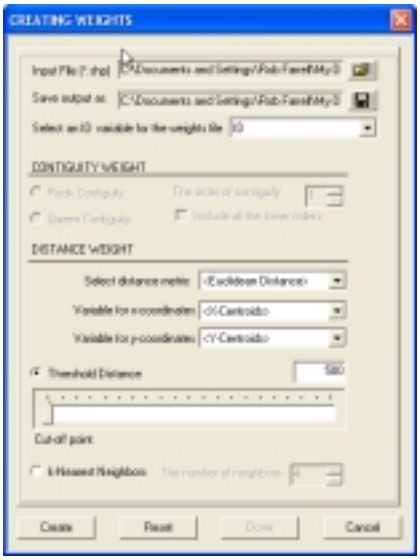
Close the LISA windows and de-select all locations.

6. Spatial Regression and diagnostics

This form of regression analysis incorporates the spatial autocorrelation present in the data. This correlation can take the form of spatially related error terms or spatially related dependent variables (leading in turn also to related error terms). Ordinary Least Squares regression that ignores this correlation runs the risk of inefficient estimators (in the spatially related error term case) or both biased and inefficient estimators (in the spatially related dependent variable case). This exercise examines the potential benefits of incorporating the spatial component into the regression analysis.

Creating a weights matrix

a) Create a weights matrix as you did in section 5 of this lab. Use the same **`sbreal_blkatt_p.shp`** file, and call the weights file "dist500.gwt". Choose a threshold distance of 500.



b) Now create a weights matrix that uses the 4 nearest neighbors, and call is "NN4.gwt".

c) Examine the two weights matrices by **`Tools > Weights > Properties`**. Examine the distributions—how are they different?

d) Create univariate Moran's I scatter plots for price using each of the weights matrices: **`Space > Univariate Moran`**. How are they different? Why might this be? Notice that 11 of the lagged price values for the Moran scatter plot "dist500.gwt" are exactly equal to 0—why do you think this is?

Non-normality of the dependent variable

Let's create a new variable, L_PRICE, which is the natural logarithm of the original PRICE variable.

a) Open the data table by clicking on the table icon in the toolbar - -  .

b) **Right-click** on the table and choose **Add Column**.

c) Name this new column L_PRICE.

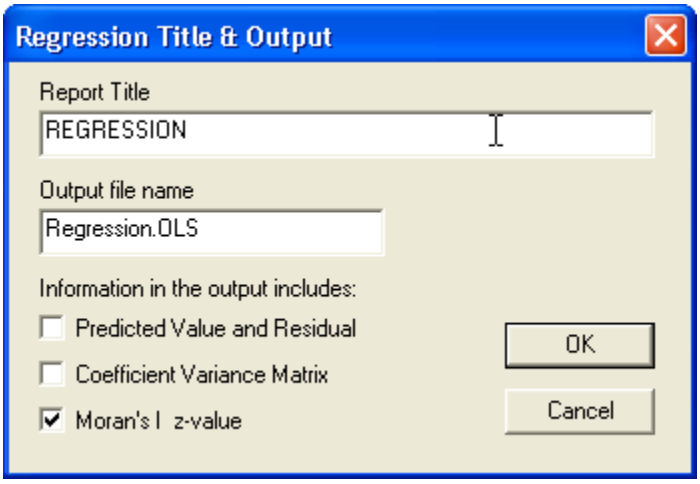
d) **Right-click** again on the table and choose **Field Calculation**. Define L_PRICE as the log(PRICE).

e) Create histograms, Explore > Histogram, of L_PRICE and PRICE. Note the varying level of distributional symmetry in the two variables.

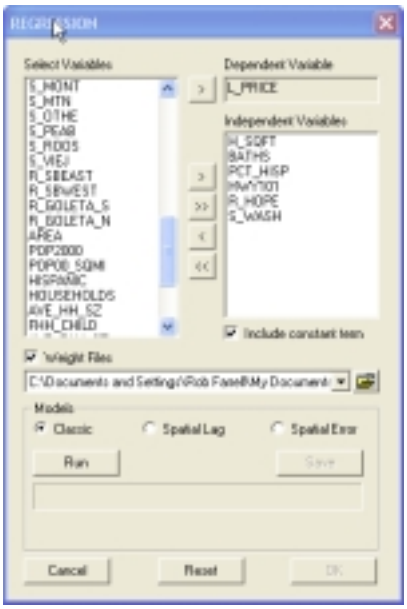
Spatial Regression and Diagnostics

Now, we will perform spatial regression and assess the benefit of incorporating the spatial component into the regression!

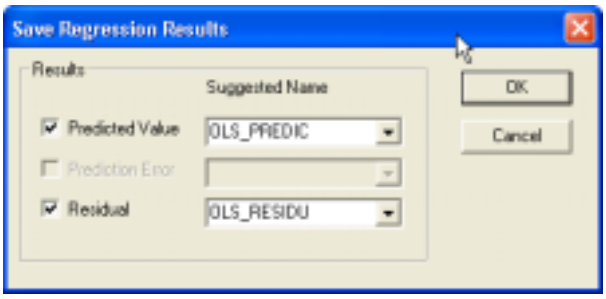
a) Choose **Regress** from the Menu bar. Check the box for **Moran's I z-value** as indicated below and choose **OK**.



b) A box will appear requesting you to define the regression. Identify L_PRICE as the dependent variable, {H_sqft,baths,pct_hisp,hwy101,r_hope,s_washington} all as the independent variables, "dist500.gwt" as the weights file, and Classic as the regression type. Click **Run**.



c) After the model executes, click **Save**, and then click the two check boxes that appear, as shown below:



Click **OK**.

d) When you are returned to the regression window, change the model type to **Spatial Lag**, and **Run** the model again. Click **Save**, and check all 3 boxes and click **OK**.

e) When you are returned to the regression window, change the model type to **Spatial Error**, and **Run** the model again. Click **Save**, and check all 3 boxes and click **OK**.

f) These values have been "saved" as new variables in the table. Click **OK** in the REGRESSION window. The results open in a new window.

Note that the diagnostics at the bottom of the first page (the OLS results) indicate the presence of spatial autocorrelation in the errors. Below that there are diagnostics for the Spatial Lag, Spatial Error, and SARMA models.

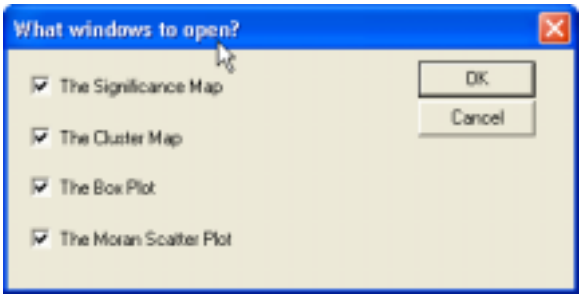
Assessing Residuals for Autocorrelation

Now, let's check visually (and locally!) the autocorrelation of the residuals from the regression models, using LISA (Local Indicators of Spatial Association).

a) From the menu bar, choose **`Explore > Univariate LISA`**. As you scroll down, you should see all of the saved variables from the regression model executions.

b) Choose **`OLS_ RESIDU`** from the list, and click **`OK`**. Choose the **`"dist500.gwt"`** weights file, and click **`OK`**.

c) Choose to view all of the options, by checking each box, and then clicking OK.



d) What do these local indicators reveal? The Moran's I value should match that from the regression execution. Now, do the same thing for the Spatial Lag Model Residuals (LAG_RESIDU) and for the Spatial Error Model Residuals (ERR_RESIDU). You should find that the error residuals are quite a bit better behaved (having less local spatial autocorrelation) than the spatial lag residuals.

Now, consider re-running part or all of this exercise with one or both of the aggregated data sets, and note the differences in the results!!

I encourage you to check the Illinois Spatial Analysis Lab site, administered by Luc Anselin, where you can find the latest versions and materials for the software -- http://sal.agecon.uiuc.edu/geoda_main.php.