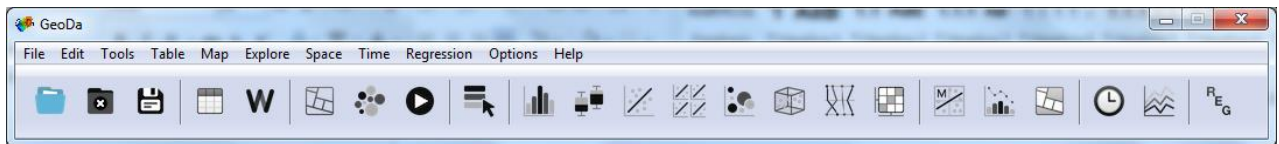


Session 2 Exercise: Spatial weights, spatial clustering analysis

A. Spatial Weights

In a computational environment, data management and the theoretical background should be relevant all the time, because they ensure the quality and robustness of the analysis (and save you from trouble afterwards, actually). Beyond those requirements, however, the first practical issue is the generation of spatial weights. Spatial weights formalize neighborhood relationships in the observations. The entity to-be-generated is a text file (but it does not have familiar extensions, such as *.txt or *.csv). This file is a requirement to work with spatial statistics.

In our class we are using a free program called GeoDa. The program looks like the following image, with additional floating windows opening depending on the operations:



GeoDa is able to generate contiguity and distance-based weights:

- Contiguity spatial weights files have the extension **.gal**
- Distance-based spatial weights files have the extension **.gwt**

Keep in mind that:

- These are connectivity files, later converted to weights, but we will name them weights,
- Contiguity weights use common polygon edges,
- Distance-based weights use Euclidean or other measures.

After spatial weights are computed, you have everything in place to conduct statistical tests for spatial autocorrelation and compute relevant metrics, and then move on to estimating spatial regression specifications.

0 – Locate software and data, change decimals notation if needed

Software: You will find GeoDa at the lab's computer.

Data and Working folder: You'll find the data for each exercise in the corresponding section in Moodle. Please use your workstation's hard drive or USB drive to create a working folder for the exercise and download and unzip the data in that working folder.

Decimals mark: Go to Control Panel > Clock, Language, and Region > Change the date, time, or number format. Or type region and language in the start menu's search box and press enter. In the window that opens, click the button Additional Settings and in the new window make sure that the decimal symbol is the "point" (.) and not "comma" (,).

1 – Inspect the data, Generate contiguity weights

Contiguity-based weights work with polygons only, because they are computed based on common edges, or edges and vertices. Two common types are:

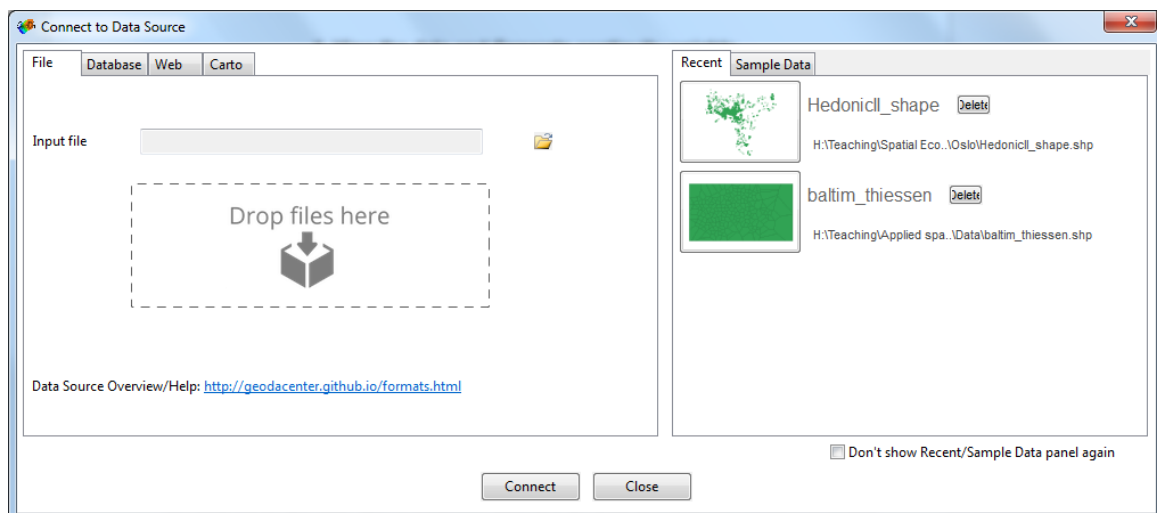
- “Rook” contiguity: polygons with **common edges** are considered neighbors
- “Queen” contiguity: polygons with **common edges and vertices** are neighbors

Regardless of rook or queen contiguity, you can further distinguish:

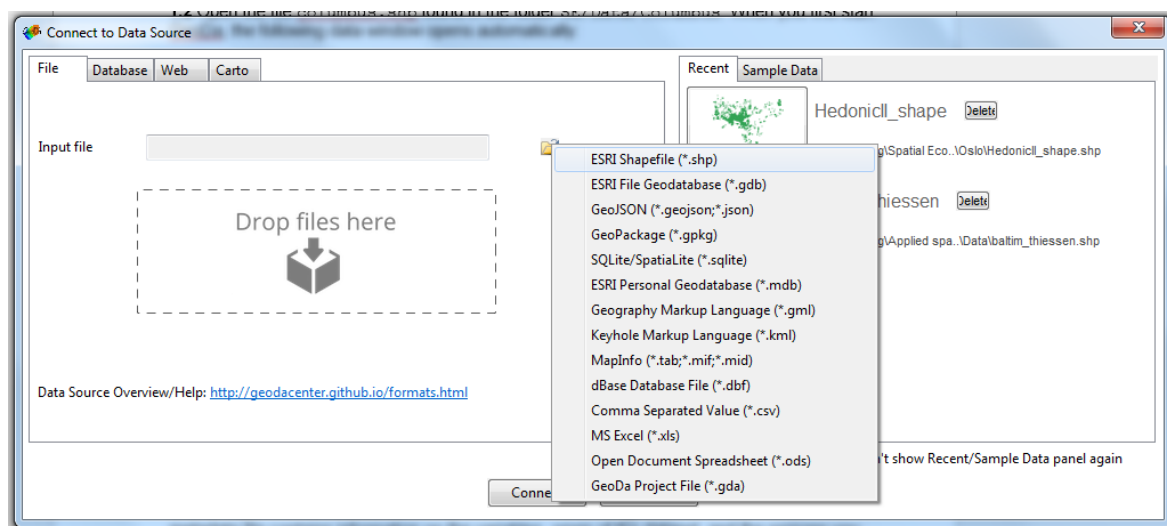
- First-order contiguity: the immediate “ring” of neighbors
- Higher-order contiguities: additional rings of neighbors.

1.1 Start GeoDa.

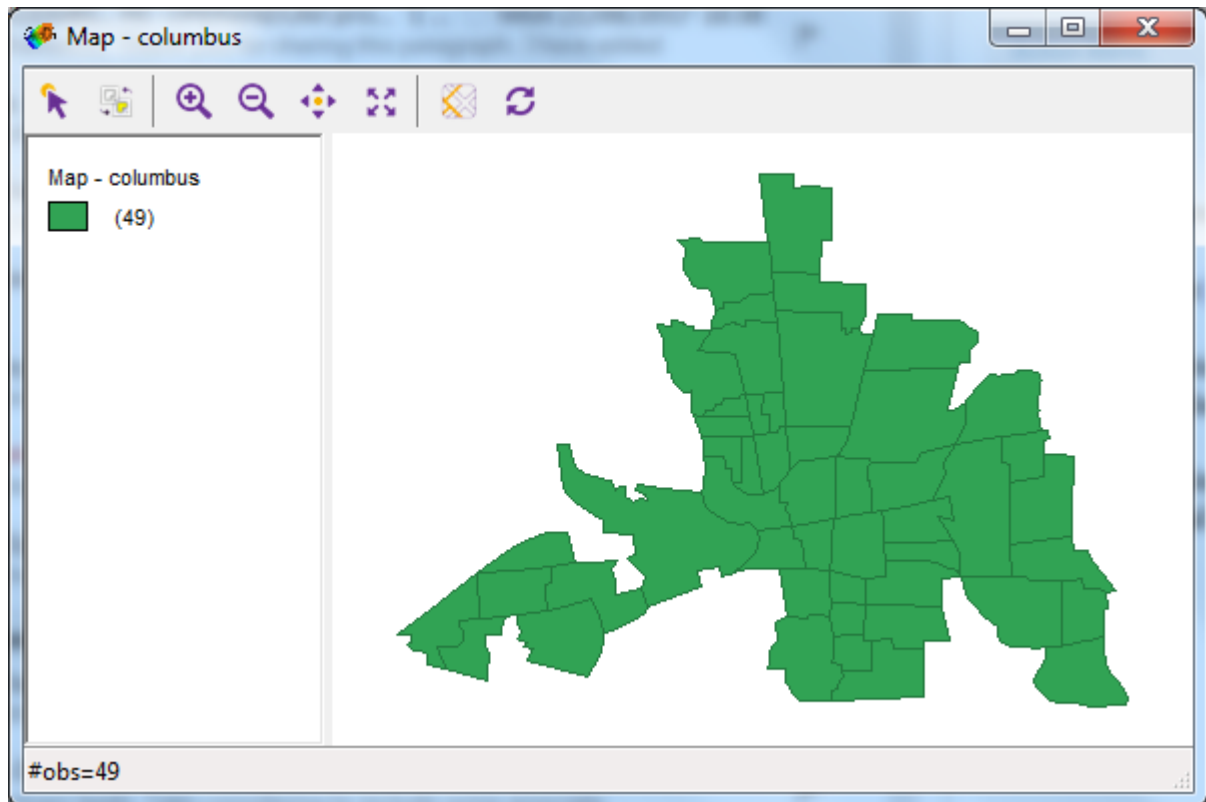
1.2 Open the file `columbus.shp`. When you first start GeoDa, the following data window opens automatically:



Press the folder icon that is next to the Input file field, then click on ESRI Shapefile (*.shp) and navigate to the file `columbus.shp`. Alternatively, just drag and drop the file in the area named Drop files here.

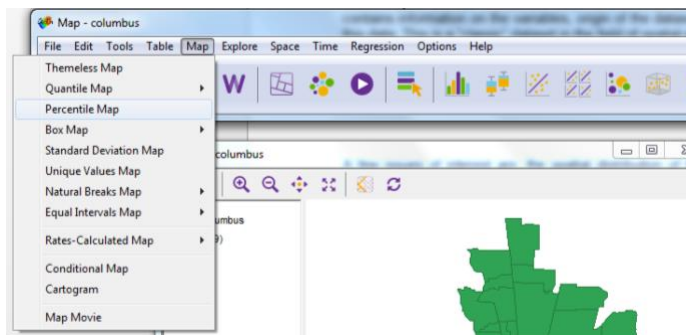


A new window will open, showing the polygons of the various districts of the city:

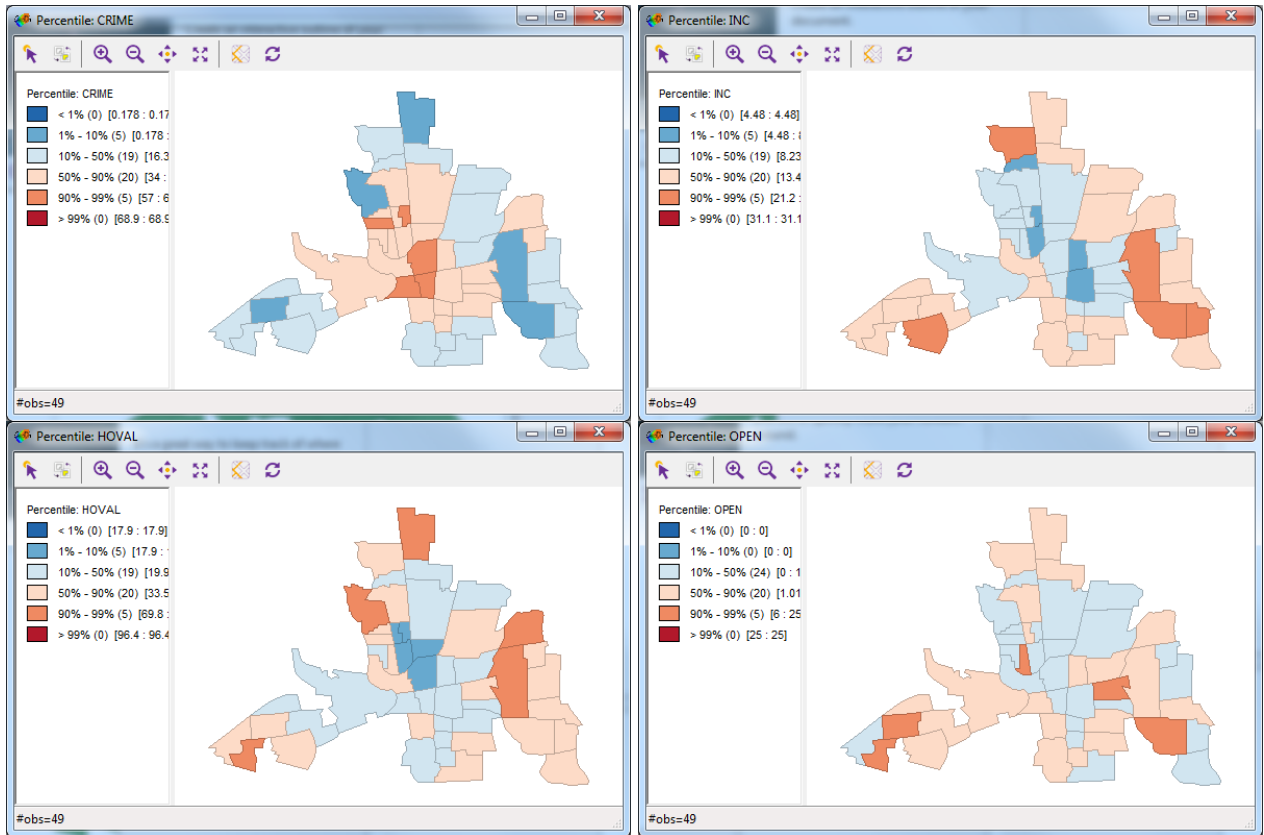


1.3 The shapefile contains the districts of Columbus, which is the capital city of Ohio State in the United States. Most importantly, the file contains a number of socio-economic variables per district. Open the file `columbus.html`. This metadata file contains information on the variables, origin of the dataset, and the persons you should thank for this data. This is a “classic” dataset in the field of spatial econometrics.

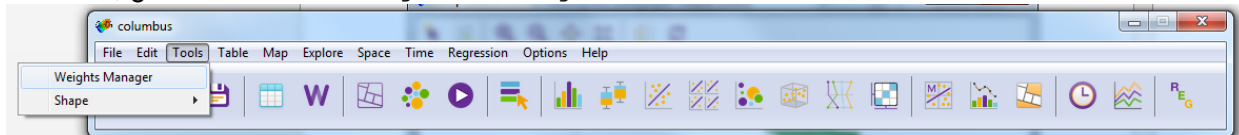
1.4 We will make a few thematic maps of Columbus for the crime variable `CRIME`, income variable `INC`, and housing value variable `HOVAL`. This can be done through the **Map** tab of GeoDa’s menu bar, selecting, among others, **Quantile Map**, **Percentile Map**, or **Natural Breaks Map**:



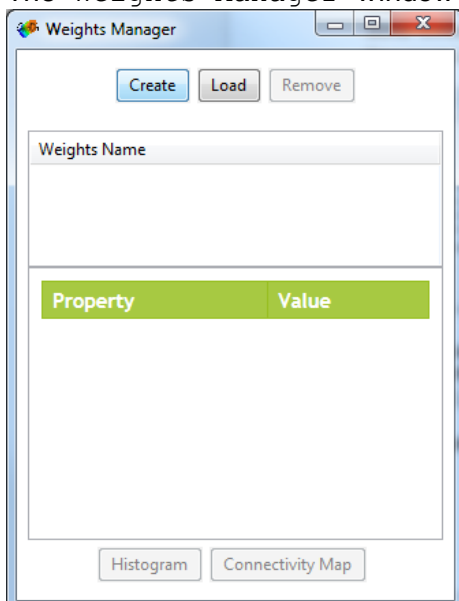
Percentile maps for crime (top left), income (top right), property value (bottom left), and percent open space (bottom right):

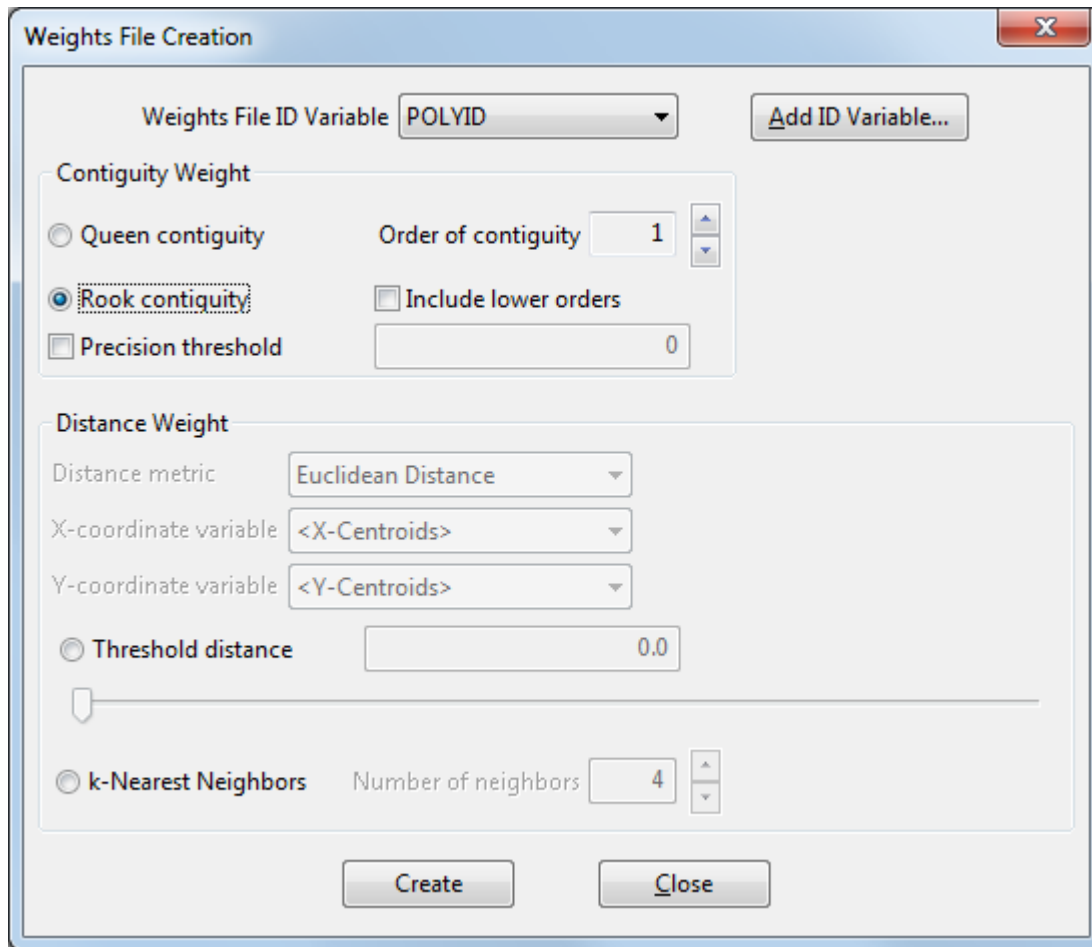


1.5 Now, go to Tools > Weights Manager



The "Weights Manager" window opens. Press Create.

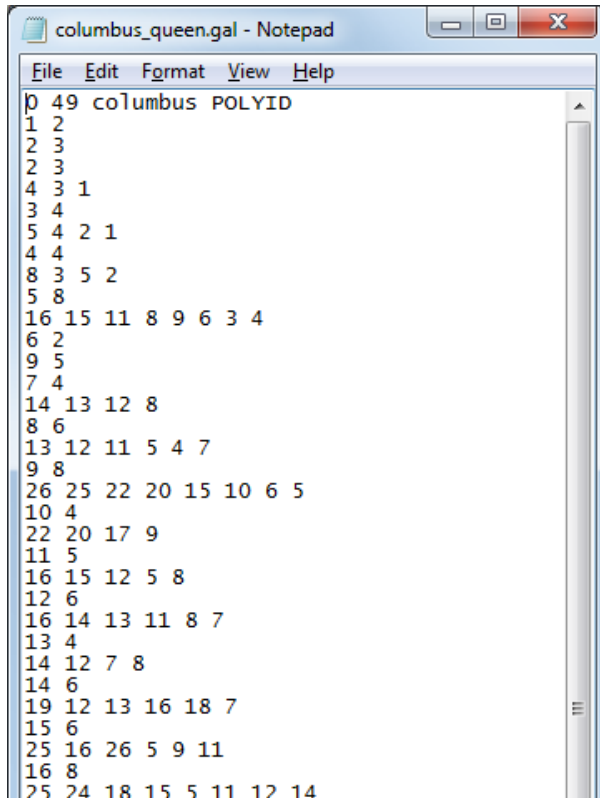




The “Weights File Creation” window contains several options for weights creation. **Most importantly, the shapefile must contain a field with unique numeric (integer) ID’s for each polygon, other than the FID or other internal fields of ArcGIS.** In the case of the Columbus data POLYID is such a field so you do not have to create a new one. When you work with files without such a field, you can create one on the fly with GeoDa by clicking the Add ID Variable button, which will walk you through the creation of such a field.

1.6 Create two weights files, one by Rook contiguity (i.e. von Neumann neighborhood) and one by Queen contiguity (i.e. Moore neighborhood):

- From the drop-down list `Weights File ID Variable` select the `POLYID` variable as your unique ID.
- Select the `Rook Contiguity` radio button.
- Set the `Order of Contiguity` = 1, as shown in the picture above.
- Click the `Create` button. In the new window that opens, save the file as `columbus_rook`. You do not need to enter the extension `.gal` manually – GeoDa takes care of this.
- Do the same for Queen contiguity and save the file as `columbus_queen`.



```
columbus_queen.gal - Notepad
File Edit Format View Help
0 49 columbus POLYID
1 2
2 3
2 3
4 3 1
3 4
5 4 2 1
4 4
8 3 5 2
5 8
16 15 11 8 9 6 3 4
6 2
9 5
7 4
14 13 12 8
8 6
13 12 11 5 4 7
9 8
26 25 22 20 15 10 6 5
10 4
22 20 17 9
11 5
16 15 12 5 8
12 6
16 14 13 11 8 7
13 4
14 12 7 8
14 6
19 12 13 16 18 7
15 6
25 16 26 5 9 11
16 8
25 24 18 15 5 11 12 14
```

As mentioned earlier, the weights files can be viewed with any text viewer such as Notepad. For example, the `columbus_queen.gal` file looks like the image on the left.

Notice that the first row is a header, that is, it displays information on the file.

The rest of the rows display groups of polygons that are considered as neighbors. The polygons are marked by the unique identifier (`POLYID`) that you set earlier while creating the weights.

Technically speaking, this is not a weights matrix, but a sort of connectivity index. Nevertheless, this file is used by the software to calculate one.

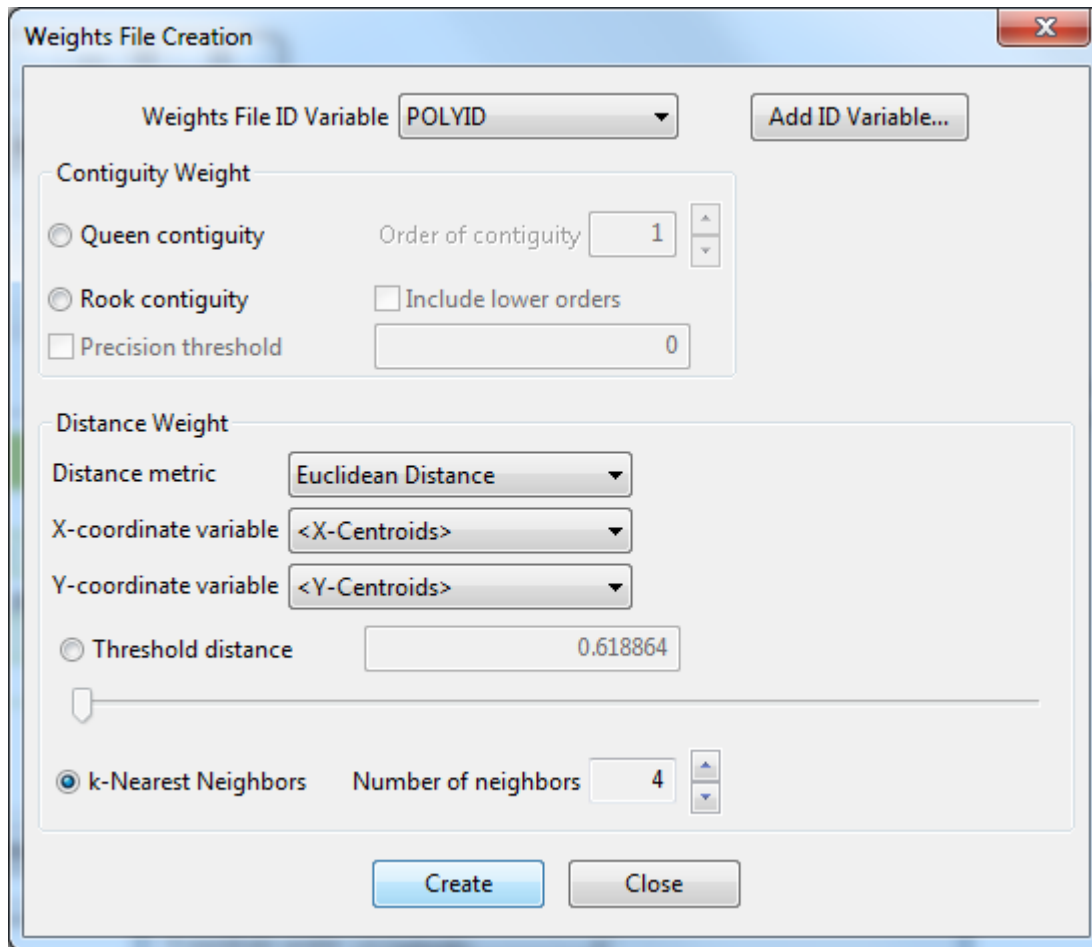
2 – Generate *k*-nearest-neighbors weights

Distance-based weights, such as the commonly used *k* nearest neighbors (*knn*, with $k = 1, 2 \dots n$) is another way to model the neighborhood structure of your data. Contrary to rook and queen contiguities (which look at common edges/vertices), distance-based weights search either for a specified number of nearest entities, or for any entity within a specified distance. Thus:

- A *knn* weights procedure identifies the *k* closest entities of each item and assigns them to its neighbor list – usually by nearest centroid;
- A pure distance weight finds anything within a given distance and assigns it as a neighbor. Distance can be Euclidean distance or more complex distances such as cost distance, socioeconomic distance, Manhattan distance, and so on.

While very complex neighbor rules can be created, it is preferable to not use complex expressions because you might end up imposing too much pre-conceived structure on your data. Of course, there might be theoretical reasons to do that, in which case more complex modeling will be understandable.

2.1 Use the `Weights Manager` (as in steps 1.5-6 above to) create a 4 nearest neighbors weights file. Use the options as shown in the image below, and save the file as `columbus_4nn`. Again, there is no need to manually type the extension `.gwt` (denoting this type of weights, while `.gal` denoted contiguity weights in the previous cases):



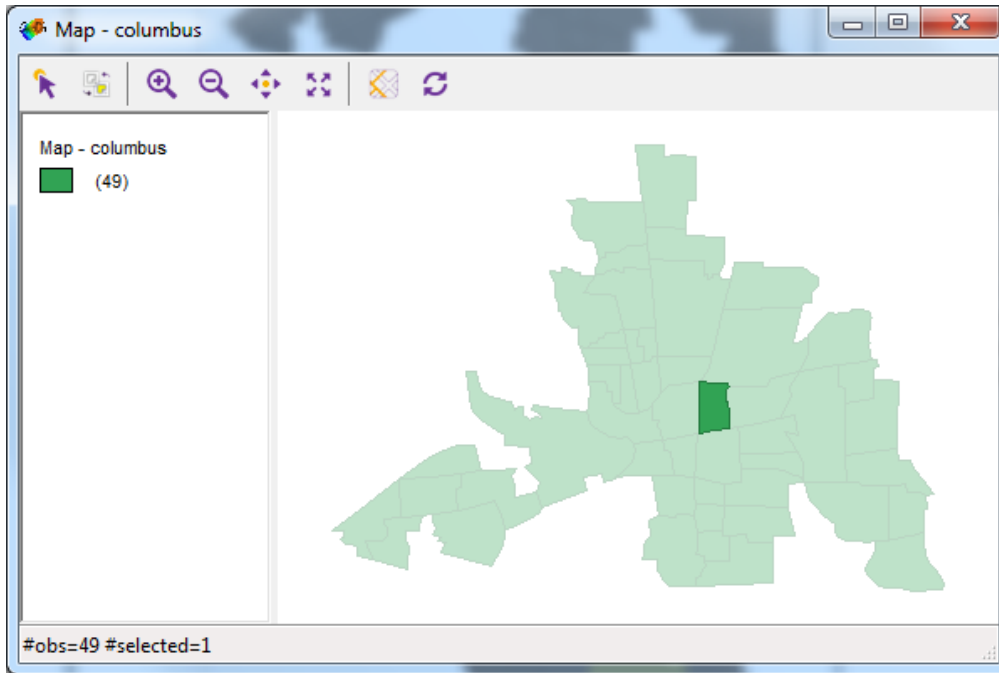
3 – Visualize weights with GeoDa

GeoDa has useful capabilities for exploring and visualizing weights. Visualizing and comparing the various weights can help you understanding whether alternative neighborhood conceptualizations alter significantly the imposed connectivity structure of your observation (remember: with spatial statistics we make explicit assumptions about space). For instance, it helps understand why spatial regressions may yield different coefficients when using different weights.

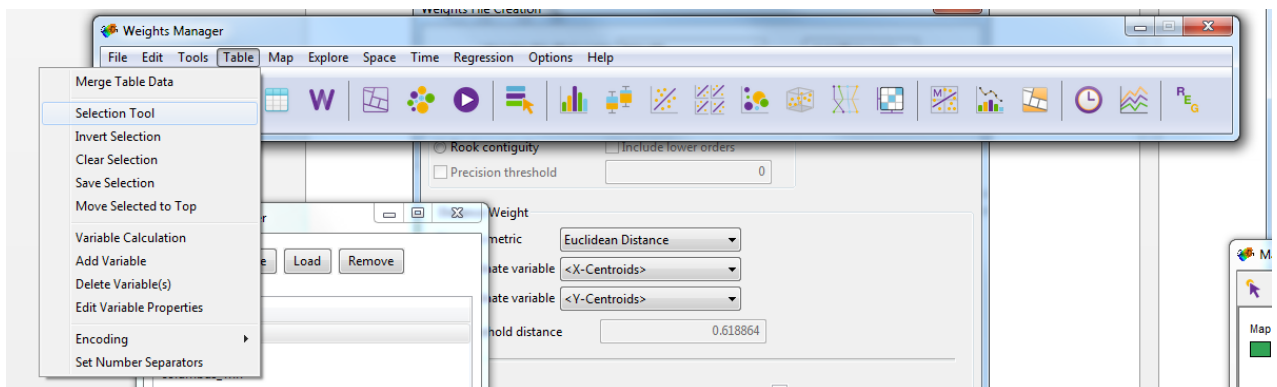
3.1 In the `Weights Manager` window (`Tools > Weights Manager`), you can click the `Load` button to navigate through the different weights files you might have created.

3.2 Now, make sure you have a map window open displaying the Columbus shapefile. **If not**, you can open a new one by clicking `Map > Themeless Map`.

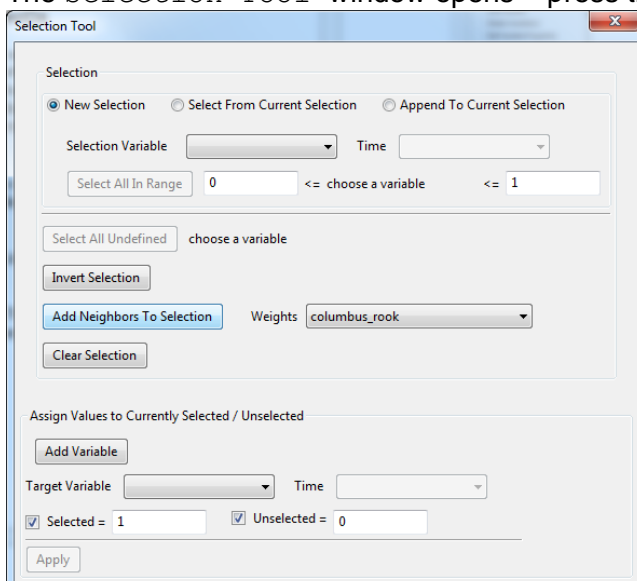
Inside the map, select one district of the city by clicking on it. The selected polygon will remain vivid, while the rest of the (unselected) polygons will be dimmed:



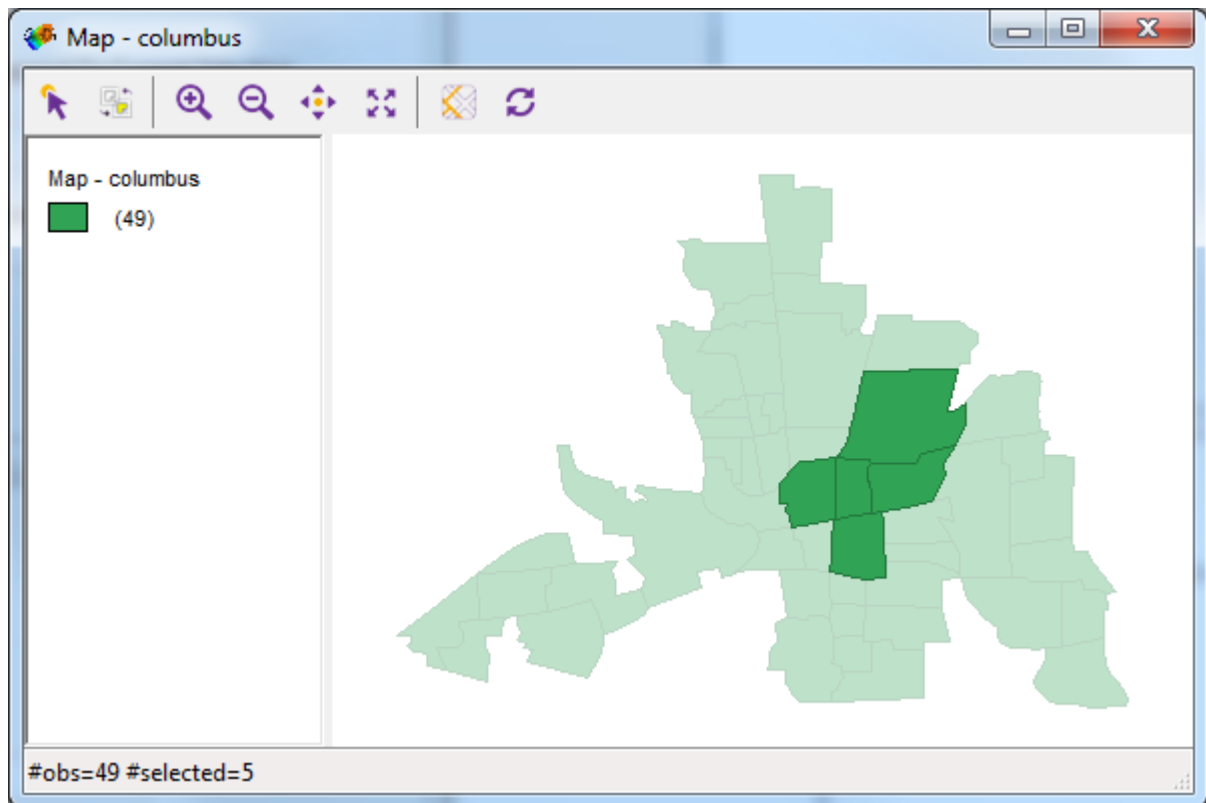
3.3 Next, making sure that you indeed have selected **one single** polygon, click Table > Selection Tool from the menu bar:



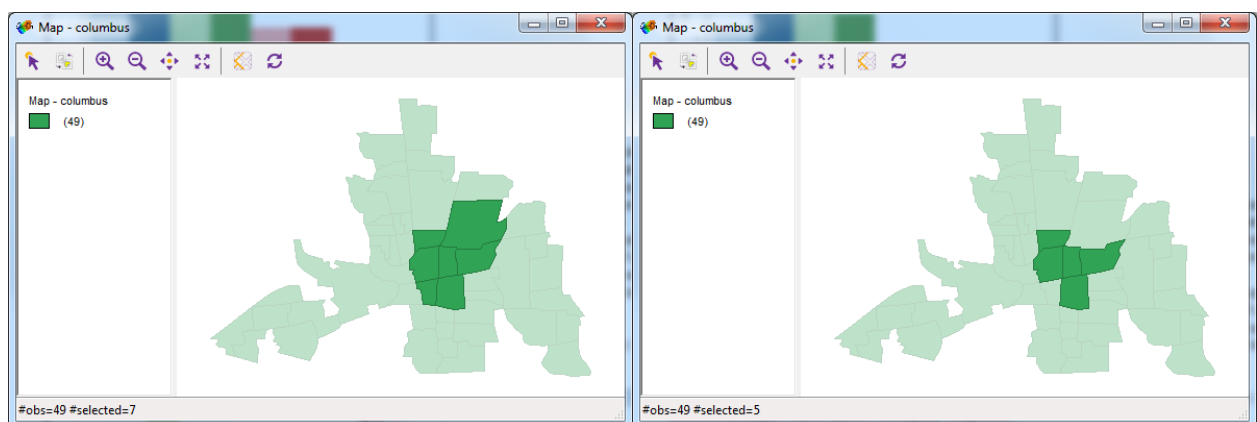
The Selection Tool window opens – press the button Add Neighbors To Selection:



This command cooperates with the **specific** weights file that is currently active and adds the neighbors of the previously selected polygon, given the conceptualization of neighbors that the weights file represents (rook, queen, knn, etc.). The map will now display vividly the neighbors, in addition to the previously selected polygon. (**Note:** to clear the selections, left-click somewhere on the white area of the map window. If you select another polygon and perform the same procedure, the new “neighborhood” will be shown). The image below shows the rook neighbors of one district:

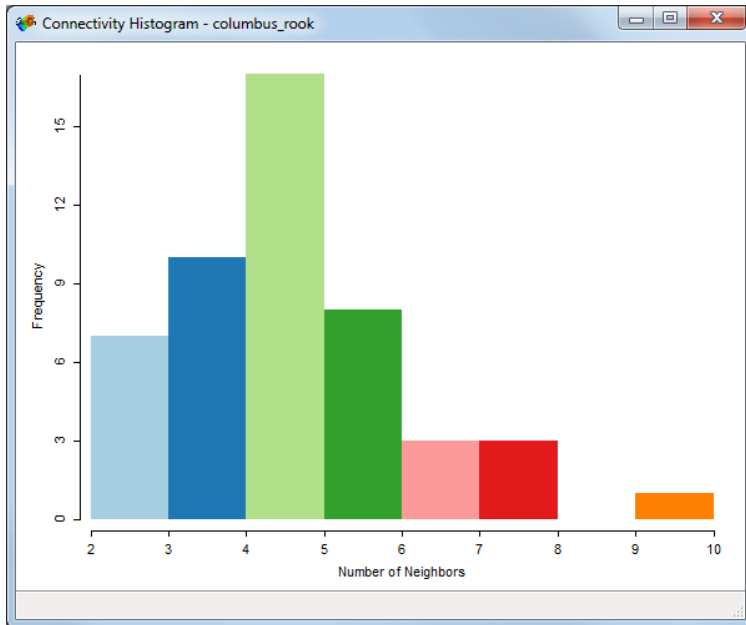


3.4 Notice that the Selection Tool has a Weights drop-down list on the right of the Add Neighbors To Selection button. You can use this list to repeat steps 3.2-3.3 for the remaining two weights files that you created earlier in section 2, namely the files `columbus_queen.gal` and `columbus_4nn.gwt`:



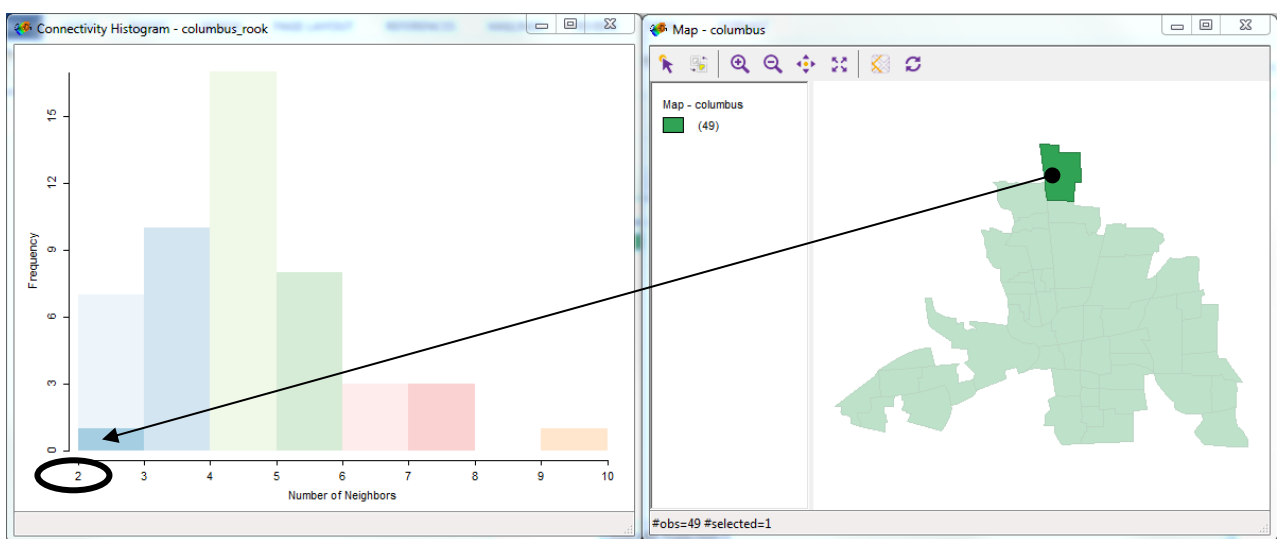
3.5 Another useful function of GeoDa is its capacity to visualize the connectivity of your data in an alternative manner: a histogram. To achieve this, go back to the Weights Manager window (Tools > Weights Manager). Load the weights file `columbus_rook.gal`. Then, still in the same Weights Manager window, click the button Histogram. A new window opens that gives

you a synoptic overview of the connectivity distribution in your data, **if** you model connectivity in the sense of amount of neighbors and through rook contiguity:

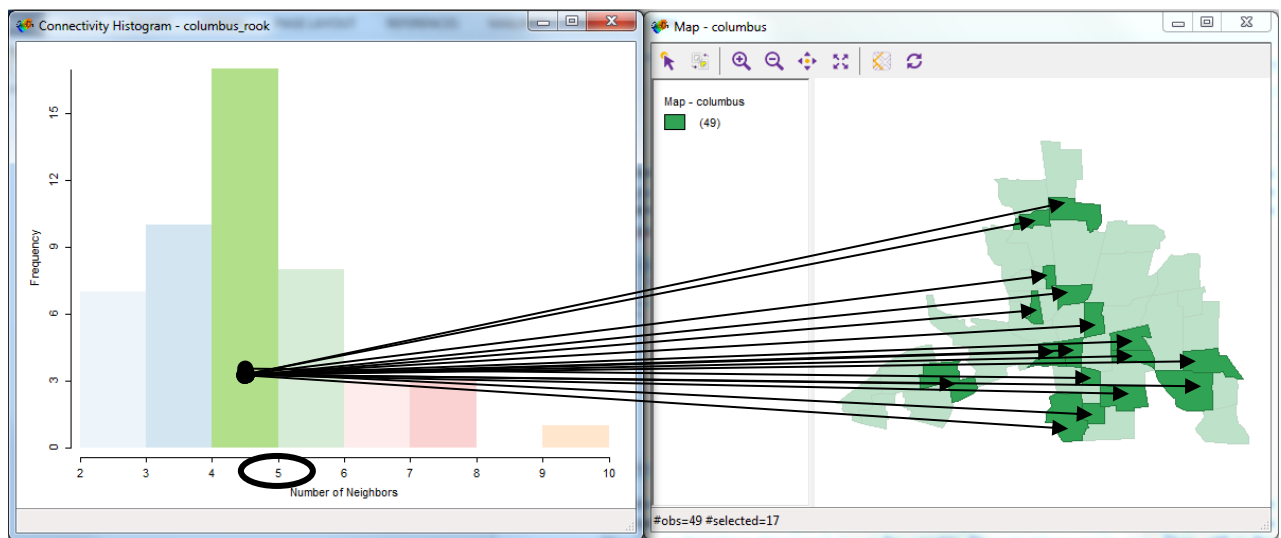


Remember that GeoDa is an interactive program. Having both a map window and a connectivity histogram window open, you can interactively click either on the different histogram bars or on the different map polygons in order to: see which are the polygons that have a certain amount of neighbors (by clicking on the histogram's bars); or how many neighbors does a polygon have (by clicking on a polygon on the map window). This is a small part of what is called ESDA (exploratory spatial data analysis). It helps you understand what the dynamics of your data are, and what your modeling of neighbors might suggest for the phenomenon you are trying to study. For instance:

Selecting a polygon on the map (right hand) shows its place in the left-hand histogram: the selected polygon has 2 rook contiguity neighbors:

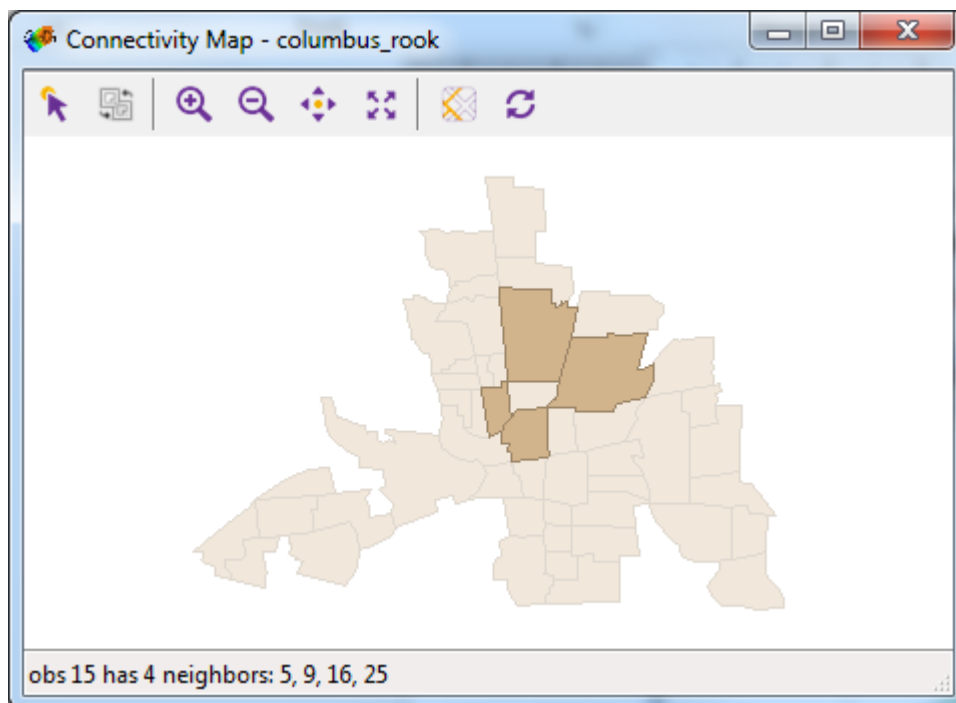


Selecting the histogram bar for 5 neighbors on the right hand shows all the polygons that have 5 neighbors on the right-hand map.



The situation in the two images above would be different if you used a different weights file, of course, which demonstrates that there is a great deal of assumptions imposed on the data. But these assumptions are made explicit in spatial statistics.

Lastly, the `Weights Manager` window has a button called `Connectivity Map`. This tool is similar in functionality to the more sophisticated `Selection Tool` that we used in steps 3.2-3.4. Its difference is that it is interactive, displaying the immediate neighbors of the polygon upon which you hover:



B. Spatial Autocorrelation Tests

You might have noticed that the heading above is spatial autocorrelation *tests*; not spatial autocorrelation. The human eye is great at distinguishing patterns, including spatial autocorrelation. So, simple visualization usually gives a sense of the spatial behavior of a phenomenon. Still, it is not possible for most people to say e.g. where the statistically significant clusters are, or to compute 9999 iterations on the fly in order to verify the validity of visual hints.

This (among other things) is why statistical tests are needed. In the scientific context, it is required to give non-visual evidence of clustering, thus a map will not suffice. The statistical tests for spatial autocorrelation provide a first solid evidence of the presence of identifiable spatial behavior in your data.

Spatial autocorrelation tests have the same mentality with other hypothesis tests and measures of statistical significance, such as the *t*-test, *z*-score, or *p*-value. They tell you how likely it is that something could have occurred randomly. A typical procedure in spatial statistics is:

1. State the null hypothesis H_0 and also have ready an alternative hypothesis H_1
2. Compute the appropriate statistic (a numeric value)
3. Check the significance of that value (a probability)
4. Reject or fail to reject the null hypothesis based on the probability in step 3.

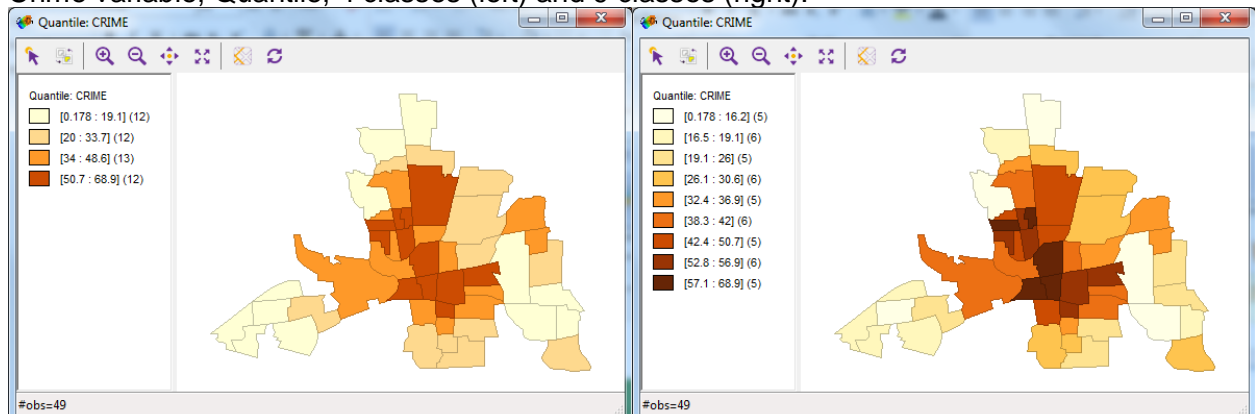
NB! There exists some controversy on hypothesis terminology, e.g. rejecting or accepting, failing to reject or accept. The most important is to be able to say which case is unlikely given a certain threshold, and based on that to be able to say in free language what does that mean.

4 – Clustering: Visual evidence

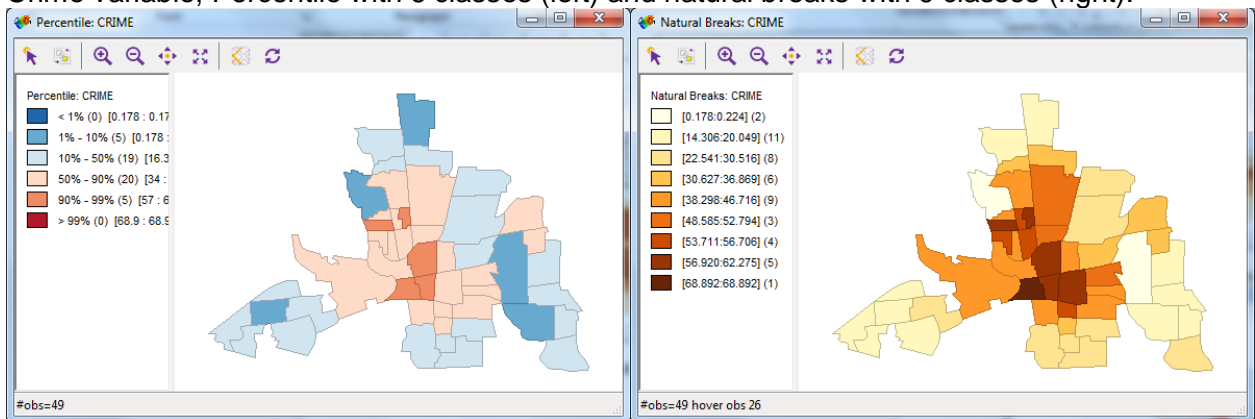
4.1 We continue with the file `columbus.shp`. For now, we are interested in the variable `CRIME`, which records the residential burglaries and vehicle thefts per 1000 households, for each district of Columbus.

4.2 As we saw earlier in step 1.4, we can generate nice thematic maps in GeoDa. Its built-in color-maps are curated to not bias the viewer. You can use the maps to examine **evidence of clustering**, i.e., what sort of geographical pattern does a variable follow. However, can you be certain about it? As we saw in step 1.4, there are alternative algorithms for generating thematic maps, and these are the kinds of maps that have special appeal to policy makers, stakeholders, journalists, and the public. Below are four alternative thematic maps of the crime variable:

Crime variable, Quantile, 4 classes (left) and 9 classes (right):



Crime variable, Percentile with 6 classes (left) and natural breaks with 9 classes (right):



5 Clustering – statistical evidence

While the previous evidence provides a good starting point for getting familiar with the spatial behavior of a variable, the proper practice is to utilize statistical tests. One such tests for global spatial autocorrelation is Moran's I (Moran 1950).

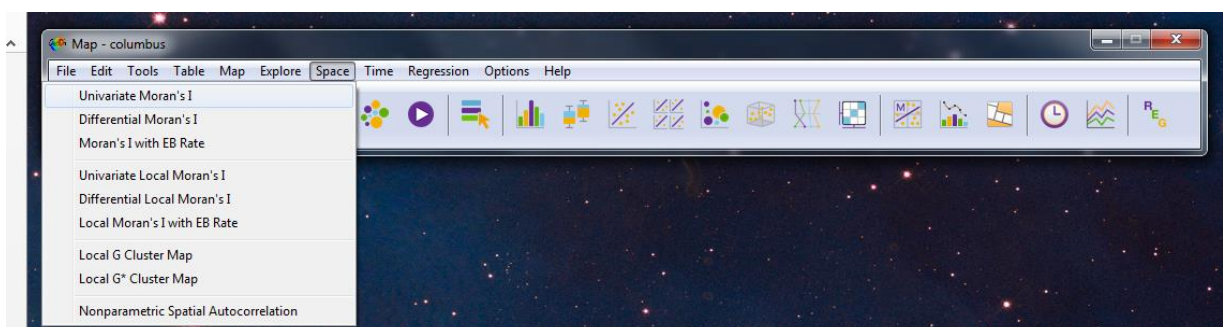
You can look up a bit of [information](#) for Pat Moran himself. In Moodle you can also find Moran's 1950 article in which he discusses his statistic for spatial autocorrelation.

The testing flow for global spatial autocorrelation, using Moran's I and GeoDa, is as follows:

1. Null hypothesis H_0 : The variable does not exhibit spatial autocorrelation.
2. Alternative hypothesis H_1 : The variable exhibits spatial autocorrelation.
3. Compute Moran's I selecting a spatial weights matrix.
4. Check its pseudo p -value. Conventionally, if $p < 0.05$ (95% certainty), you reject H_0 , meaning that there is enough evidence to state that the examined variable exhibits spatial autocorrelation. Other thresholds are possible for the p -value as well (0.1, 0.01), but most people stick to 0.05.
5. Check the actual value of Moran's I statistic (not its p -value). Negative values indicate dispersion, while positive values indicate clustering. The closer the value to -1 or +1, the stronger the effect.

Implement the above workflow in GeoDa (using the `columbus_rook.gal` weights file):

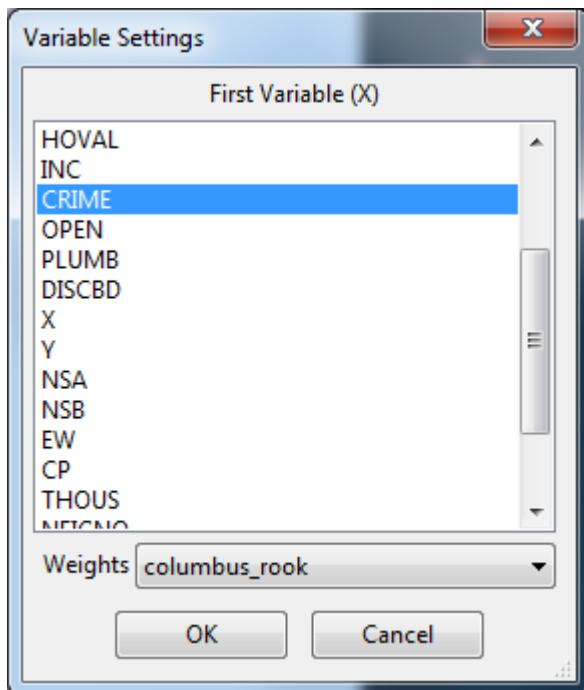
5.1 With the `columbus.shp` loaded, click Space > Univariate Moran's I:



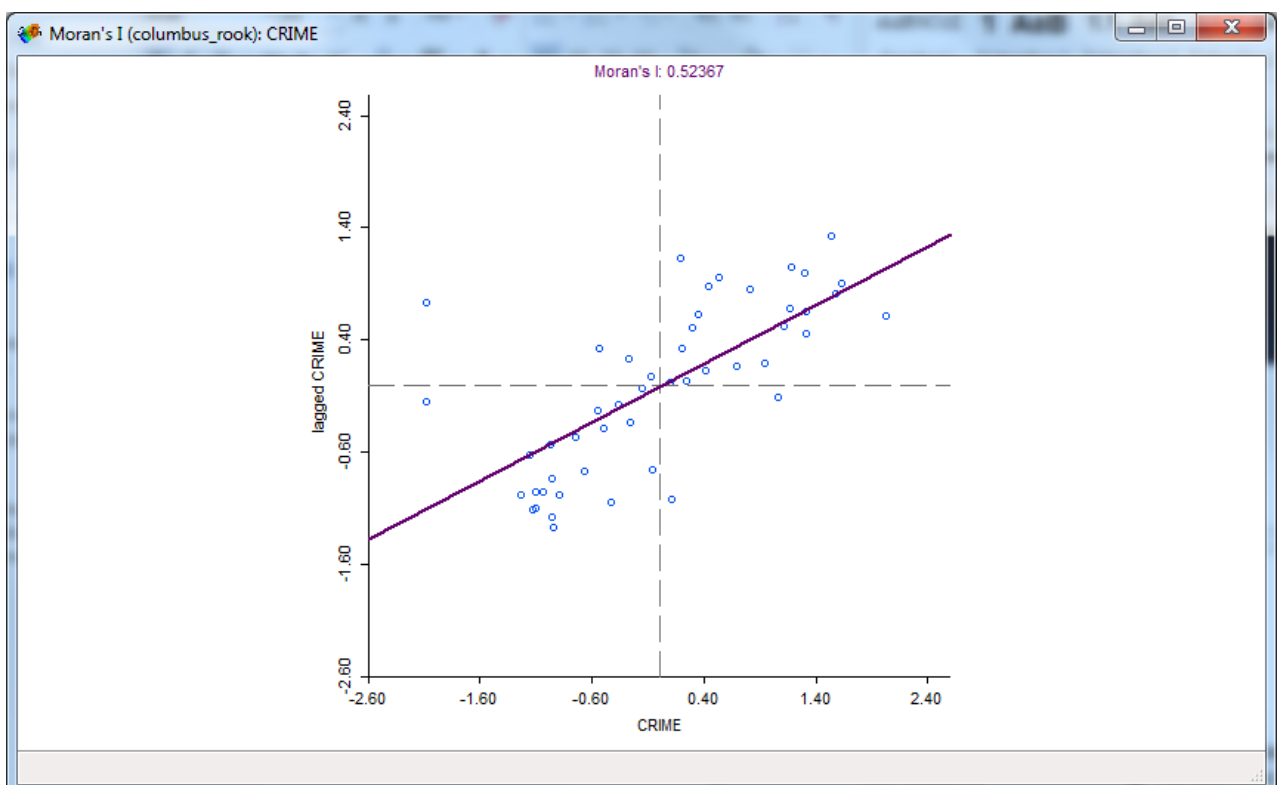
5.2 A window will come up, asking you which variable you want to test for spatial autocorrelation.

The window also provides a drop-down list to select a weights file – use `columbus_rook`.

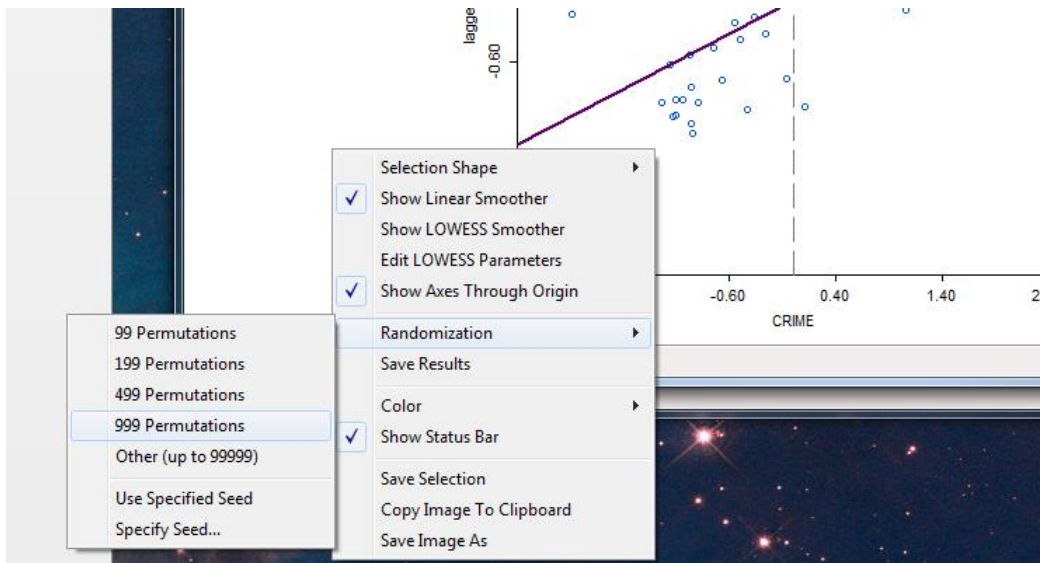
Select `CRIME` and hit OK:



5.3 A Moran's I scatterplot for `CRIME` opens. It is a plot of `CRIME` against its spatially lagged version lagged `CRIME`. At the top of the plot you can find the computed I -value of 0.5. Also notice that the graph is divided in four quarters of pairwise combinations of high and low Y and Y_L :

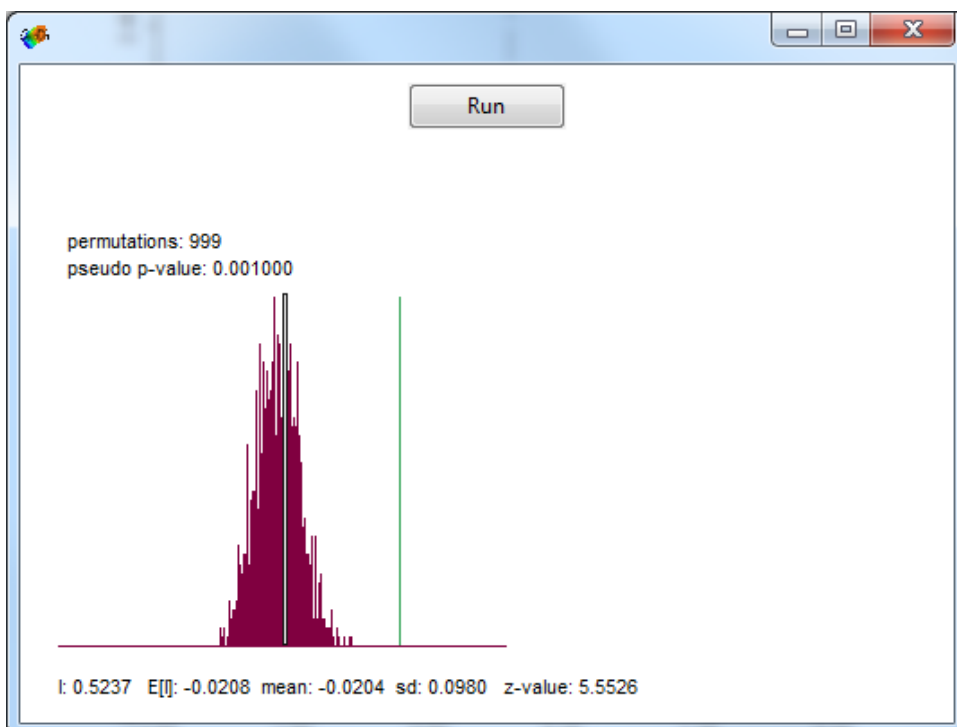


5.4 Test for the significance of this value. Right-click inside the scatterplot; in the pop-up menu click **Randomization** > 999 permutations:



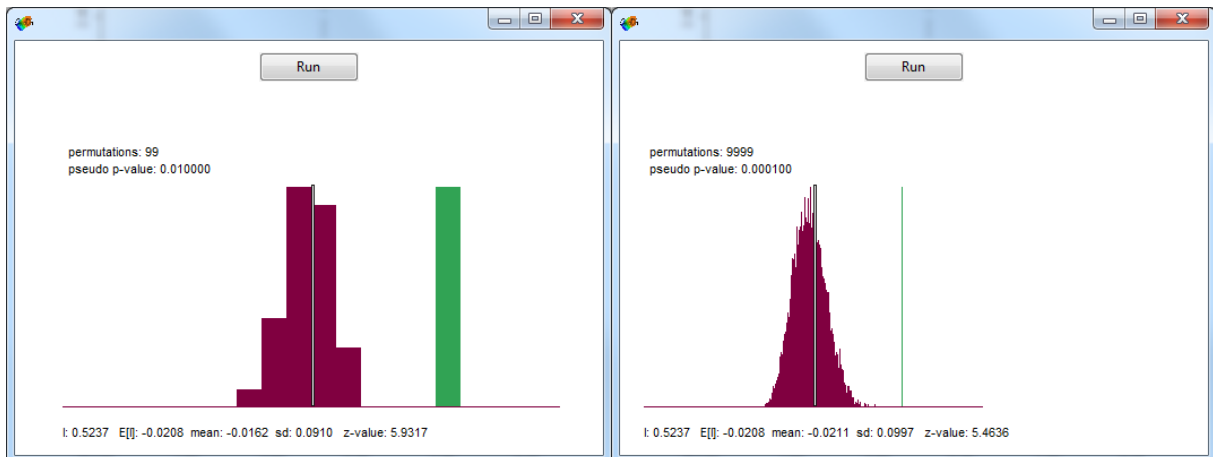
The following window opens, showing you important information, namely:

- number of permutations
- pseudo p -value
- Moran's I value (I)
- Expected Moran's I value according to the 999 random reshufflings ($E[I]$)
- The mean of the 999 permutations (mean)
- The standard deviation of the 999 permutations (sd)
- The z -value based on the 999 permutations ($z\text{-value}$)
- The histogram, which displays the random distribution of the value of Moran's I (red piles) and the Moran's I value of the actual data (vertical green line).



5.5 The pseudo p -value is permutation-based. This has two main effects:

- The p -value may vary slightly per run.
- The p -value's order of magnitude depends on the number of permutations. That is, 99 permutations will yield a different p -value from 9999 permutations. **However, if the spatial autocorrelation is statistically significant (say at the 95% margin), the p -value will always be smaller than 0.05** (i.e. it may be 0.001 or 0.000001; the number of permutations influences the fraction). E.g., the same histogram with 99 (left) and 9999 (right) permutations is:



5.6 Let's conduct a global spatial autocorrelation test, using GeoDa and Moran's I statistic, for the variable `HOVAL` (average housing value in \$1000):

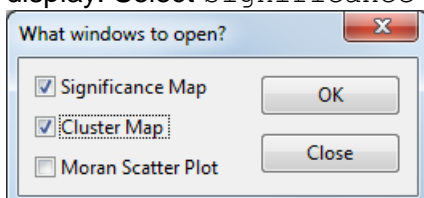
- Null hypothesis: Housing value does not exhibit spatial behavior.
- We generate a Moran's I scatterplot.
- We perform a significance evaluation (99999 permutations): p -value < 0.05.
- We fail to accept the null hypothesis of spatial randomness. Housing value exhibits statistically significant, weak positive spatial autocorrelation ($I = 0.2$)

6 – Clusters: statistical evidence and saving to shapefile

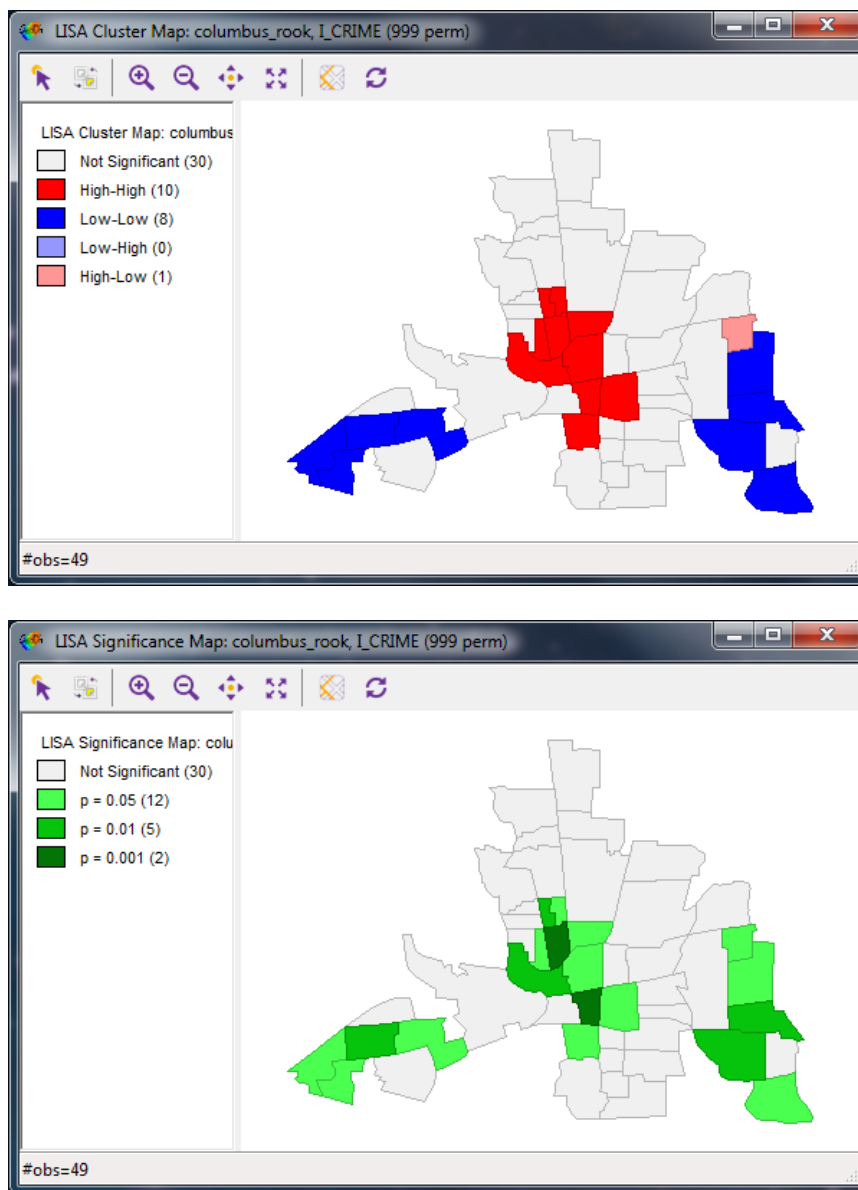
Global autocorrelation testing tells you whether the values of a variable cluster or not together in space, and if that is statistically significant. If you would like to further investigate where in space clusters occur, then you can employ the LISA approach (Anselin 1995, in your `Literature` folder). The idea behind LISA is that it decomposes the Global Moran's I to local instances, guided by the Global Moran's I scatterplot and its four quarters of pairwise high and low values of the examined variable.

6.1 In GeoDa click `Space > Univariate Local Moran's I`. The familiar `Variables Settings` window will appear, asking you which variable you would like to investigate for local clusters. Select `CRIME` and hit `OK`.

6.2 Lastly, the final dialog that pops up asks you about what kind of results you would like to display. Select `Significance Map` and `Cluster Map`, and hit `OK`.



6.3 Two maps appear, following your selections in the `What windows to open?` dialog. If we compare them to the quantile maps we produced before and also to the ArcGIS maps at the very beginning, we will understand many of the insights that spatial statistics can offer.

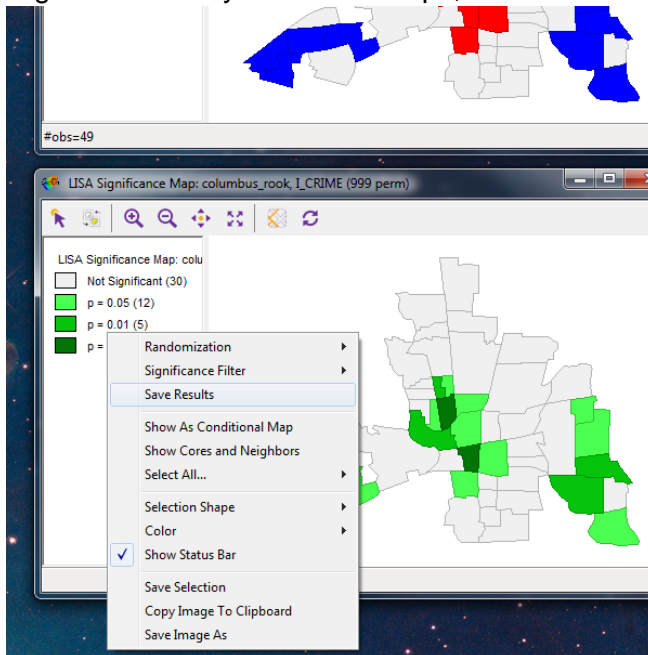


Similar to how the p -value depended on the number of permutations in Moran's I case above, the significance of the displayed clusters of the LISA test (the second. Green-themed map above) will be different according to the permutations you used. **However, significant clusters (regardless of their varying but always significant p -value) will “always” come up and insignificant clusters will “always” be shown as such.** You can check this by right-clicking on any of the two maps and setting different amounts of permutations. By the same right-click menu, you can also tell GeoDa what is your acceptable cut-off significance value.

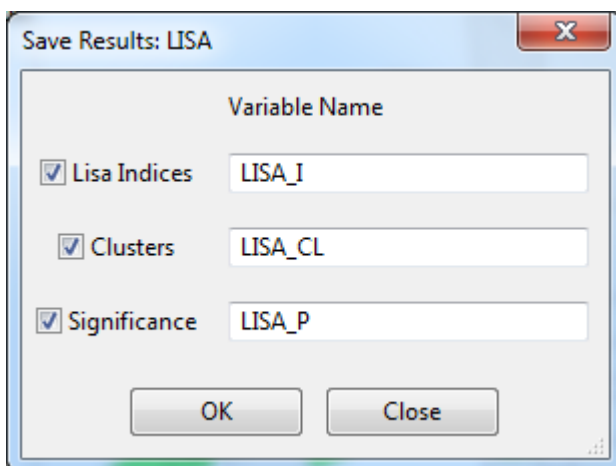
6.4 As this is exploratory spatial data analysis, such meddling with the permutations number can help us assess the stability of clusters in the dataset. Typically, most clusters remain unchanged regardless of the permutations. But non-significant clusters sometimes appear as significant, and sometimes as insignificant – this indicates uncertain borderline cases.

6.5 You can save an instance of the local clusters and their significances to the shapefile's database, so you can carry on the analysis in other GIS software.

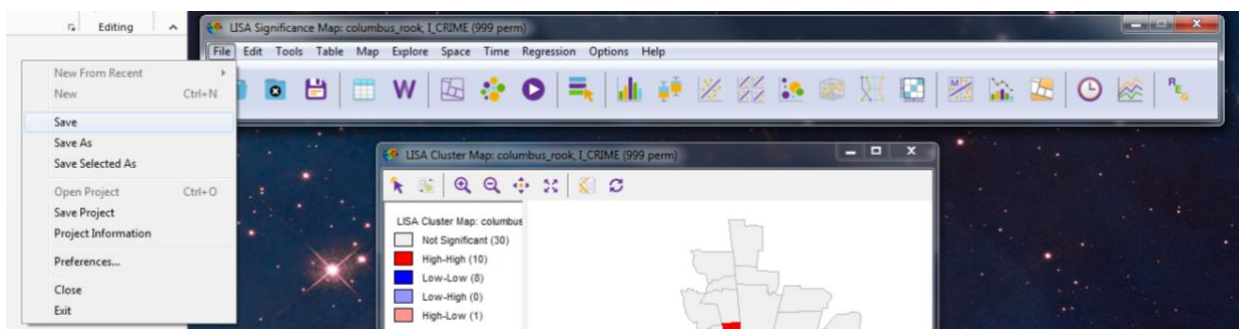
Right-click on any of the two maps, and select **Save Results**:



In the dialog that opens, select **Lisa Indices**, **Clusters**, and **Significance** and hit **OK**:



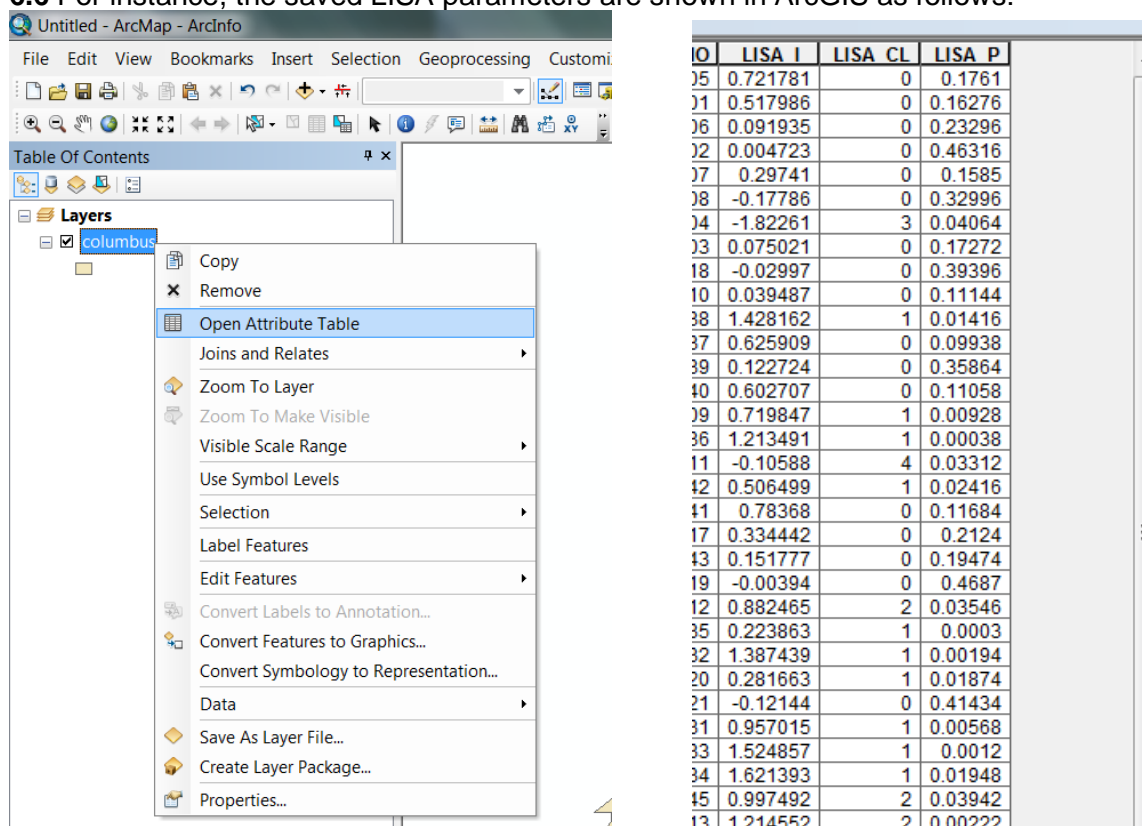
However, this is temporary and lasts only till you close GeoDa. To make the change permanent for the shapefile, click **File > Save**:



The LISA information that is being stored is the following:

- Lisa Indices is a local decomposition of Moran's global index for location, so to speak.
- Significances give the pseudo p-value of LISA and clusters
 - Clusters give a cluster code for each polygon:
 - 0: not significant
 - 1: High-High cluster (cluster of high values)
 - 2: Low-Low cluster (cluster of low values)
 - 3: Low-High cluster
 - 4: High-Low cluster
 - 5: Neighborless polygons (islands, if your data contains them)

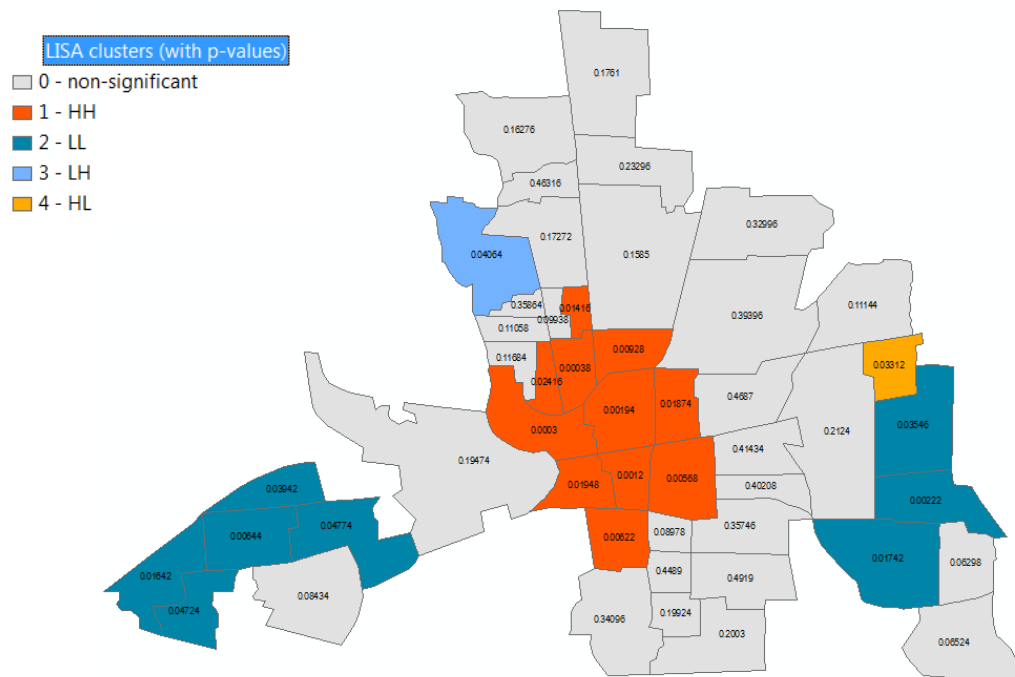
6.6 For instance, the saved LISA parameters are shown in ArcGIS as follows:



The screenshot shows the ArcGIS interface with the 'columbus' layer selected in the Layers panel. A context menu is open over the layer, with 'Open Attribute Table' selected. The attribute table is displayed on the right, showing the following data:

ID	LISA I	LISA CL	LISA P
05	0.721781	0	0.1761
01	0.517986	0	0.16276
06	0.091935	0	0.23296
02	0.004723	0	0.46316
07	0.29741	0	0.1585
08	-0.17786	0	0.32996
04	-1.82261	3	0.04064
03	0.075021	0	0.17272
18	-0.02997	0	0.39396
10	0.039487	0	0.11144
38	1.428162	1	0.01416
37	0.625909	0	0.09938
39	0.122724	0	0.35864
40	0.602707	0	0.11058
09	0.719847	1	0.00928
36	1.213491	1	0.00038
11	-0.10588	4	0.03312
42	0.506499	1	0.02416
41	0.78368	0	0.11684
17	0.334442	0	0.2124
43	0.151777	0	0.19474
19	-0.00394	0	0.4687
12	0.882465	2	0.03546
35	0.223863	1	0.0003
32	1.387439	1	0.00194
20	0.281663	1	0.01874
21	-0.12144	0	0.41434
31	0.957015	1	0.00568
33	1.524857	1	0.0012
34	1.621393	1	0.01948
45	0.997492	2	0.03942
13	1.214552	2	0.00222

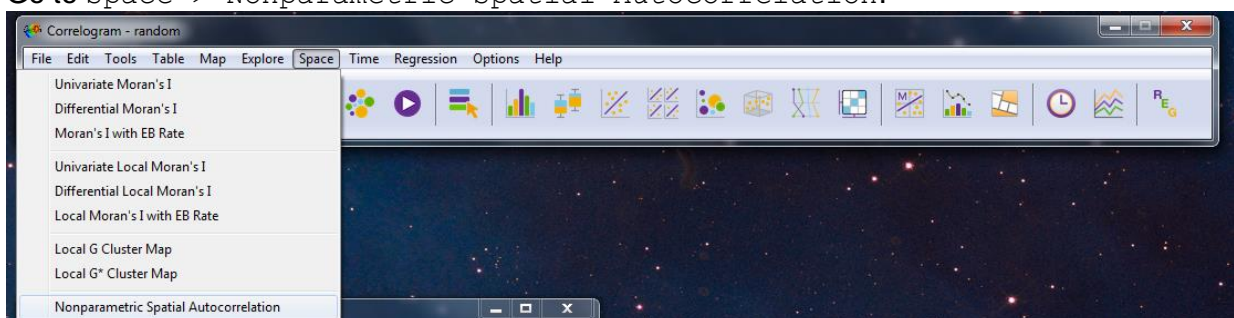
You can therefore carry on the analysis elsewhere and produce more sophisticated visualizations:



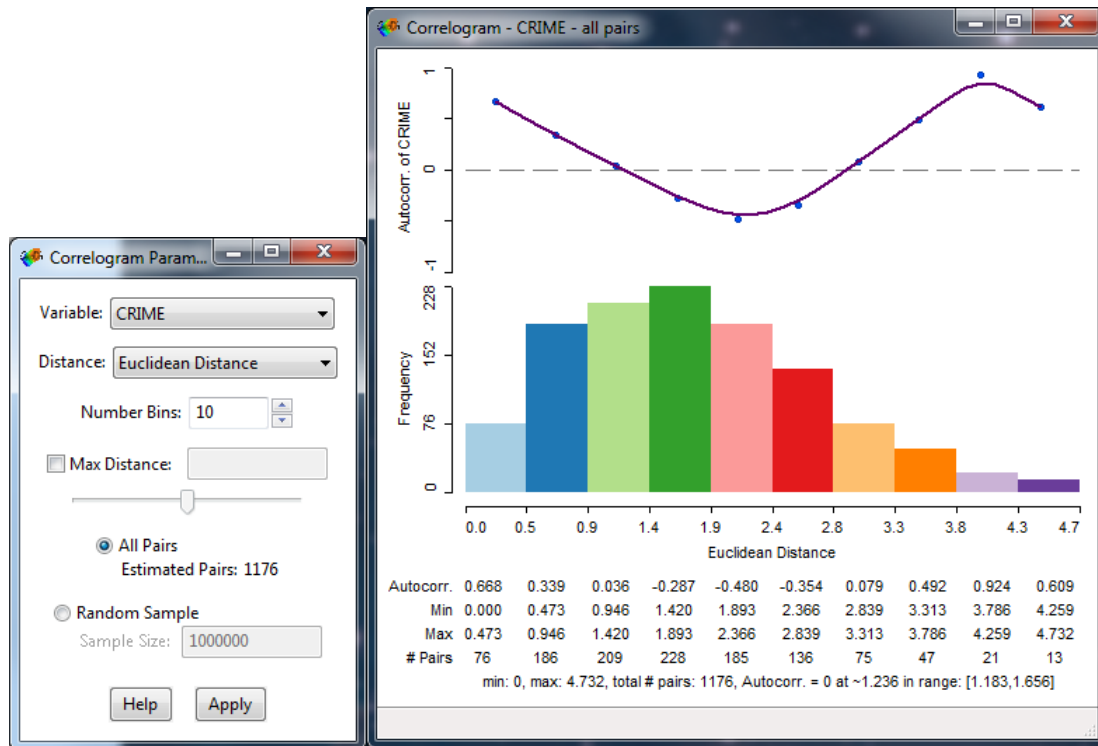
6.6 Notice also an inherent implication of the local clustering approach of LISA (but also of other similar methods such as the Getis-Ord hot spot analysis: they sacrifice a big part of mid-range values and focus on high-valued and low-valued outliers, so that they can delineate statistically significant clusters.

6.7 Another functionality of GeoDa is its non-parametric exploration of spatial autocorrelation. This function provides an intuition of how the magnitude of autocorrelation proceeds when increasing the distance between variables.

Go to Space > Nonparametric Spatial Autocorrelation:



The following windows appear, one for selecting on-the-fly the variable of interest and other settings (left) and one for displaying the results (right). For the crime variable, they look as follows:



The right-hand image above contains a univariate spatial correlogram, revealing the closeness of spatial econometrics to time series econometrics. This is typically used when deciding how many lags will be included in a model and, by extension, it reveals the dilemmas present in constructing model specifications in both fields. The approaches for selecting and validating the number of lags (orders) present in time series econometrics are not, however, present in spatial econometrics.

6.8 Lastly, the global and local autocorrelation explorations can be extended to two-variable relationships. In that case, the phenomenon is called bivariate spatial correlation. GeoDa does not provide such functionality any longer (older versions did).