

An Introduction to EDA with GeoDa

Luc Anselin
Spatial Analysis Laboratory
Department of Agricultural and Consumer Economics
University of Illinois, Urbana-Champaign
<http://sal.agecon.uiuc.edu/>
June 16, 2003

Introduction

This is a quick tour of GeoDa, illustrating its main features for exploratory data analysis (EDA). It assumes very little, and should be doable “out of the box” without having to refer to more extensive information. It assumes you are familiar with the technical concepts related to ESDA, but not with the way GeoDA implements them. It does not replace the User’s Guide. This note refers to GeoDa 0.9.3., June 4, 2003.

Starting a Project

Start GeoDa by double-clicking on its icon on the desktop, or run the GeoDa executable in Window’s Explorer (in the proper directory). A welcome screen will appear. In the File Menu, select New Project, or click on the New Project toolbar button, as shown in Figure 1.

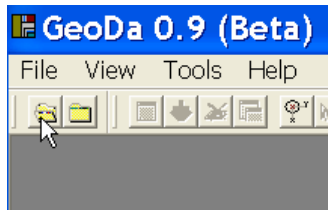


Figure 1. Start new Project

In the file selection dialog, select the SIDS sample data set as the Input Map File, with FIPSNO as the Key variable. You can either type in the full path name for the shape file, or navigate in the familiar Windows file structure, until the file name appears (only shape files are listed in the dialog). Then click on OK to launch the map, as in Figure 2.

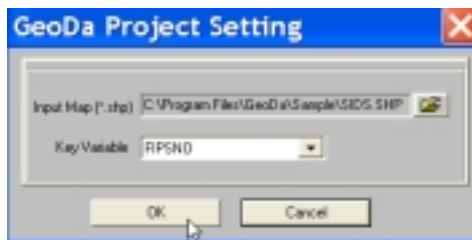


Figure 2. Select input file

A blank map will appear, showing the counties of North Carolina, as in Figure 3. You can change settings by right clicking in the map window and select characteristics such as color (background, shading, etc.) and the shape of the selection tool (see below). If you click on the thicker bar on the left hand side and move it to the right, space will appear for the legend, as in Figure 4.

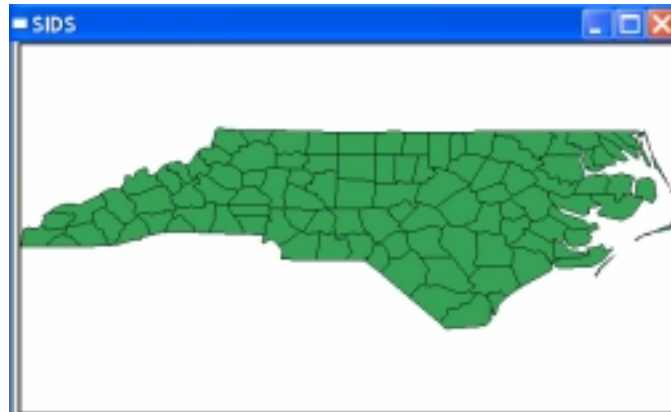


Figure 3. Initial map.

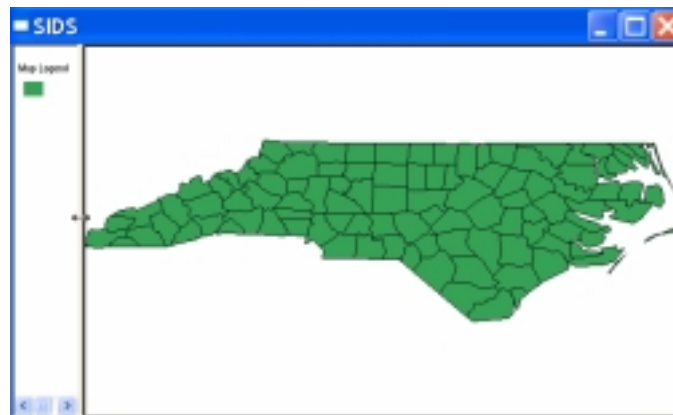


Figure 4. Initial map with legend bar moved.

Practice

Start a second instance of GeoDa with a project using the NCOVR St Louis county homicide data set (stl_hom.shp) and FIPSNO as the key variable.

Making a Simple Map

The SIDS data set is taken from Noel Cressie's (1993) *Statistics for Spatial Data* and contains variables for the count of SIDS deaths by county in two time periods, here labeled SID74 and SID79. In addition, there are the count of births in each county (BIR74, BIR79) and a subset of this, the count of non-white births (NWBIR74, NWBIR79).

Consider constructing a quantile map to compare the number of non-white births and total births in 74 (NWBIR74 and BIR74). Click on the map to make it “active” (the latest clicked window is active). In the Map Menu, select Quantile. A dialog will appear, allowing the selection of the variable to be mapped. In addition, a table will appear in the background. This can be ignored for now (the first time a variable has to be selected in any function, this table will appear). You may need to minimize the table to get it out of the way, but you will return to it later, so don’t remove it.

Select NWBIR74 as in Figure 5, and click OK. Note the check box in the dialog to set the selected variable as the “default”. If you do this, you will not be asked for the variable name the next time around. To undo the default, use Edit > Select Variable at any time. A second dialog will ask for the number of categories in the quantile map: select 4 for now and click OK. A quartile map (four categories) will appear, as in Figure 6. You can obtain the same result by right-clicking on the map, and selecting Choropleth Map > Quantile.

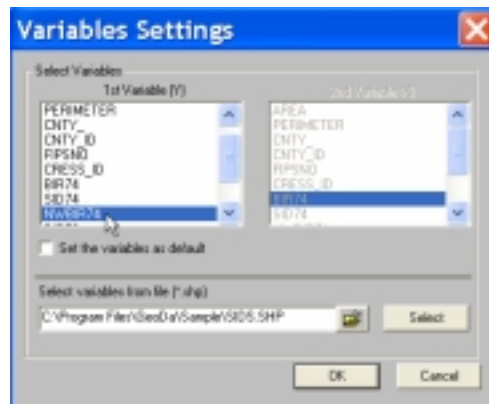


Figure 5. Variable selection.

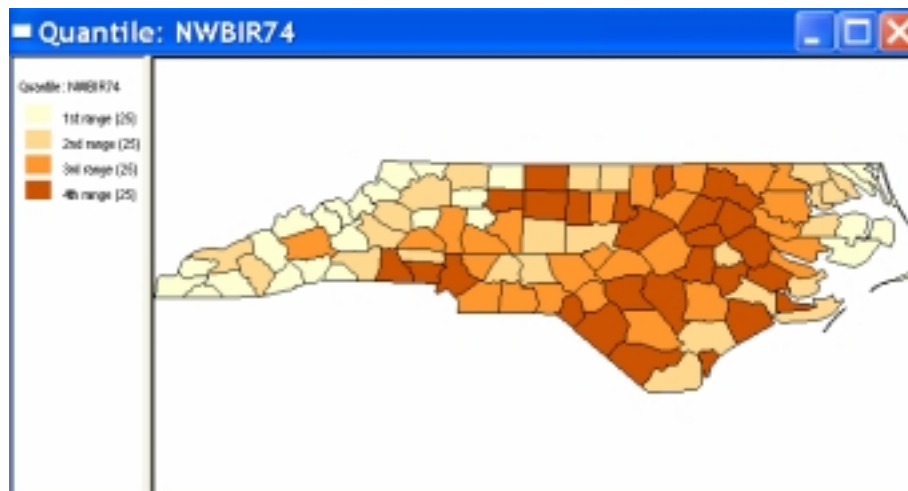


Figure 6. Quartile map for count of non-white births (NWBIR74).

Note how to the right of the legend the number of observations in each category is listed in parentheses. Since there are 100 counties in North Carolina, this should be 25. The legend also lists the variable name. Now, create another quartile map, this time for SID74. What do you notice?

There are two problems with this map. One, it is a choropleth map for a “count” or a so-called “extensive” variable. This illustrates “size” and is often inappropriate. A rate or density is a so-called “intensive” variable, which is more appropriate for choropleth maps. The second problem pertains to the computation of the break points. For a distribution such as the SIDS deaths, which more or less follows a Poisson distribution, there are many ties among the low values (0, 1, 2). The computation of breaks is not reliable in this case and quartile and quintile maps, in particular, are misleading.

Let’s stay with the NWBIR74 variable for now (remake the map if you tried out the SID74 variable). Create another map by clicking on the Duplicate Map toolbar button, shown in Figure 7. Alternatively, you can select Edit > Duplicate Map from the menu.

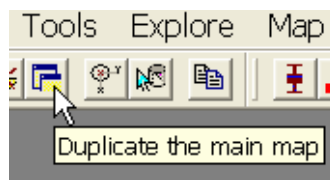


Figure 7. Duplicate map toolbar button.

This creates a blank (green background) map of the NC counties. Make sure this map is active and create a quartile map for the total births in 74 (BIR74). Your map should look like as in Figure 8. You can use Window > Tile Vertical or Tile Horizontal to rearrange the maps on your screen.

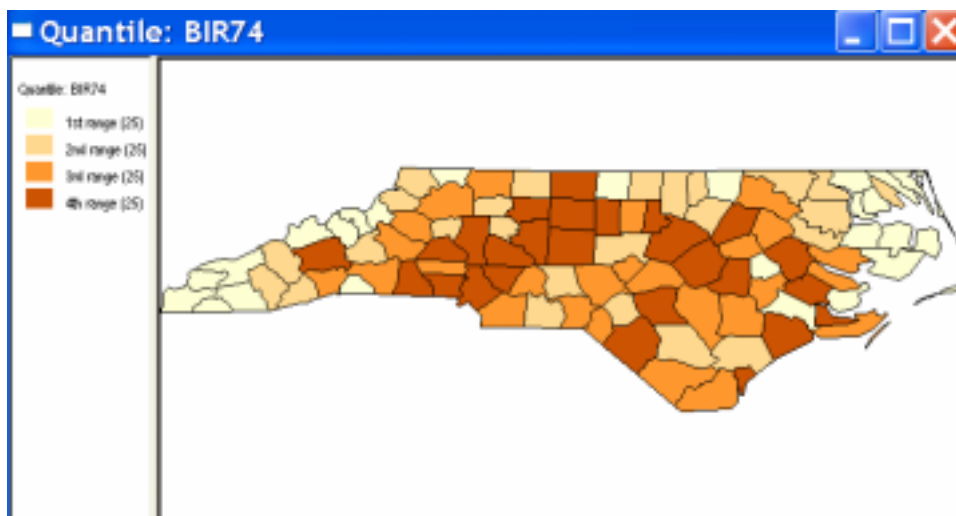


Figure 8. Quartile map for count of births (BIR74).

You can save the map to the clipboard by Edit > Copy to Clipboard. This only copies the map part. To also get a copy of the legend, right click on the legend pane and select Copy Legend to Clipboard. You can also save a bitmap of the map (but not the legend) to a .bmp formatted file by selecting File > Export > Capture to File and specifying a file name (and path, if necessary). Use a graphic converter software package to turn the bmp format into other formats.

Practice

Create a quintile map (5 categories) for the St Louis homicide rate in the period 84-88 (HR8488) and one for the period 88-93 (HR8893). Use both the menu as well as the right click approach to build the choropleth map. Save one of the maps as a bmp file and insert into a MS Word file. How would you get the legend into the Word file as well?

Outlier Maps

Quantile maps only give a very crude and simplified view of the spatial distribution of a variable. A better picture of what may be locations with “extreme” values, or so-called outliers is given by some specialized maps in GeoDa, specifically the Box Map and Percentile Map. The Box Map is essentially a quartile map, but where outliers are highlighted. Outliers are defined in the usual fashion, by computing a lower and upper “fence,” which is a value either 1.5 or 3.0 times the interquartile range (the difference between the 75% and 25% value) lower or higher than the 25% and 75% value. You create the Box Map from the menu (Map > Box Map) in the same fashion as a quantile map, except that you must also select the value of the hinge (1.5 is the default). Make sure to select the correct variable. The box maps for NWBIR74 and BIR74 are shown in Figure 9. Note how each has 7 outliers, but some are not the same. Try to identify them.

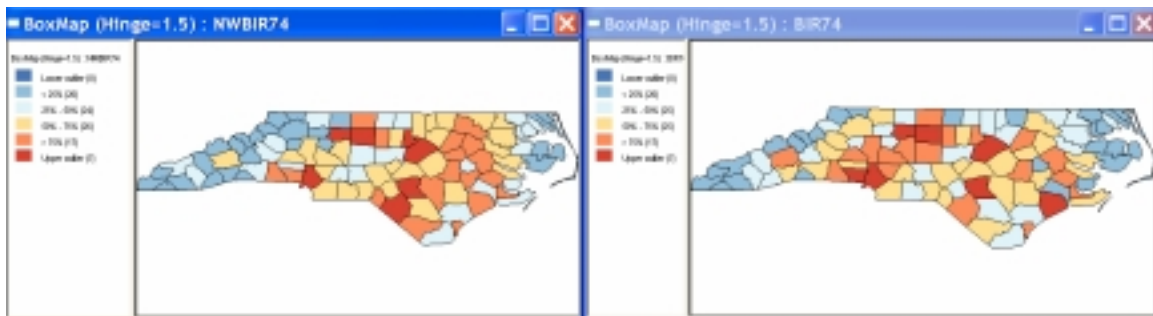


Figure 9. Box Maps for NWBIR74 and BIR74.

A percentile map uses special cut-off points to highlight extremes in the distribution. One can think of this as drawing attention to the tails of the distribution, rather than the center. The observations are sorted and shown as 0-1%, 1-10%, 10-50%, 50-90%, 90-99% and 99-100%, as in Figure 10 for the same two variables.

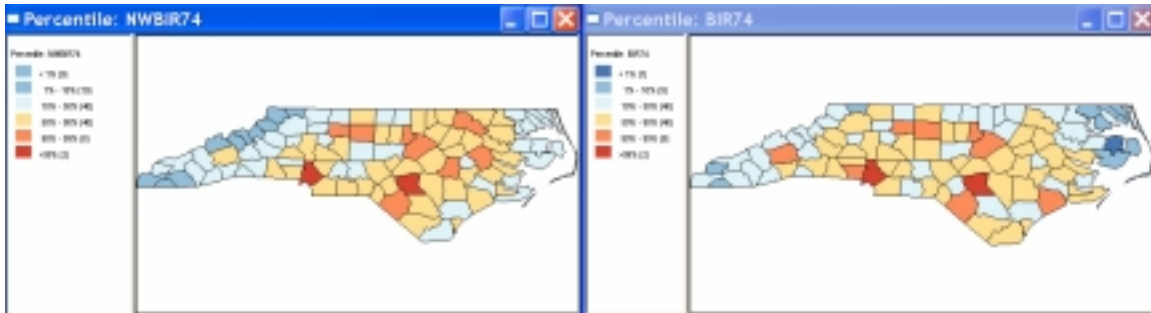


Figure 10. Percentile Maps for NWBIR74 and BIR74.

Note that the highest value(s) in the percentile map are not necessarily outliers, but if there are outliers then some or all of them will belong to the highest percentile (compare Figures 9 and 10).

Practice

Create a box map for the homicide rates in the St. Louis data set. How would you interpret the different number of outliers in each year (Note: box maps only pertain to the relative spatial distribution in a cross-section, but don't suggest a higher or lower level overall). What happens if you tighten the criterion for outlier (hinge = 3.0)? Would it make sense to construct a percentile map for St. Louis? Why or why not?

Selection and Linking

So far, the maps have been “static.” Dynamic maps include ways to select specific locations and to link the selection between maps. GeoDa includes several selection “shapes,” such as point, rectangle, polygon, circle and line. Point and rectangle shapes are the default, respectively for simple clicking (click on a county to select it) and for dragging (click on a point, drag the pointer to a different location to create a rectangle, and release). For point selection, you can add or remove locations from the selection by shift-click. To clear the selection, click anywhere outside the map. Other shapes can be selected by right clicking on the map and choosing one of the options in the Selection Shape drop down list, as in Figure 11. Each map has its own selection tool and they don't have to be the same across maps.

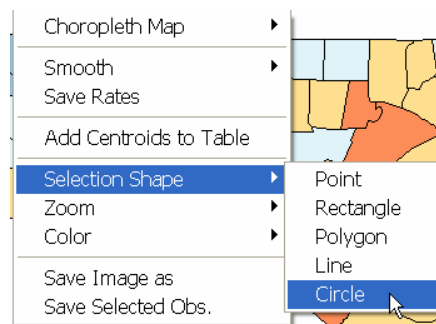


Figure 11. Selection shapes drop down list.

For example, choose circle selection (as in Figure 11), then click in the right-most outlier in the percentile map for NWBIR74 and select some counties by moving the circle out (see Figure 12).

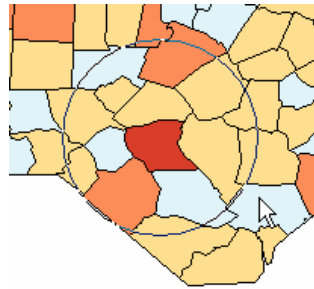


Figure 12. Circle selection.

As soon as you release the mouse, the counties with their centroids within the circle will be selected, shown as a cross-hatch (Figure 13). Note that the same counties are selected in both maps. This is referred to as *linking* and pertains not only to the maps, but also to the table and to all other statistical graphs active at the time. For example, in Figure 13, a subset of the two North Carolina percentile maps is shown with the selected counties. You can change the color of the cross-hatch as one of the map options (right click Color > Shading).

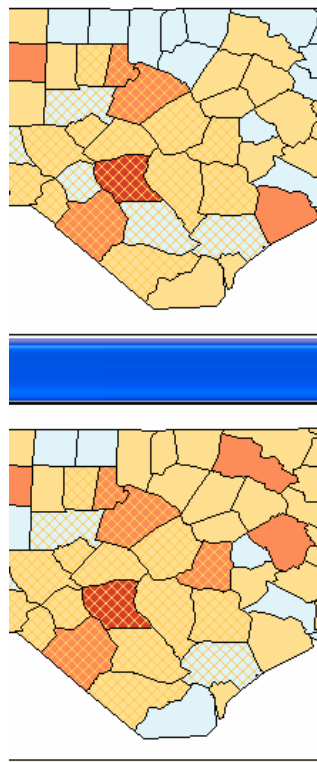


Figure 13. Selected counties in linked maps.

At this point, bring the table back up, scroll down to row 82, Cumberland County, and note how it is selected in the table as well. In other words, the maps and table are “linked.”

	AREA	PERIMETER	CNTY_	CNTY_ID	NAME	STATE_NAME	ST
78	0.131000	1.677000	2082	2082	Macon	North Carolina	37
79	0.241000	2.214000	2083	2083	Sampson	North Carolina	37
80	0.062000	1.389000	2085	2085	Pamlico	North Carolina	37
81	0.120000	1.686000	2088	2088	Cherokee	North Carolina	37
82	0.172000	1.835000	2090	2090	Cumberland	North Carolina	37
83	0.121000	1.978000	2091	2091	Jones	North Carolina	37
84	0.163000	1.716000	2095	2095	Union	North Carolina	37
85	0.138000	1.621000	2096	2096	Anson	North Carolina	37
86	0.096000	1.262000	2097	2097	Hoke	North Carolina	37

Figure 14. Selected counties in linked table.

You can promote (i.e., move to the top of the table) the selected counties to get a summary view of which ones are selected by invoking the Promote item from the table drop down menu (right click anywhere in the table), Figure 15. The same items are also available in the Menu, under Options. The promoted selection is as in Figure 16.

You clear the selection by clicking anywhere outside the map area in the map window (i.e., in the “white” part of the window), or by double clicking on the first column in the table (the columns with sequence numbers).

Practice

Use the selection tools in the box maps for St Louis to find the names of the counties that are outliers in 88-93 and not in 84-88 (point select in map and promote in table). Practice with line and circle tools to select counties in the map and check their properties in the table.

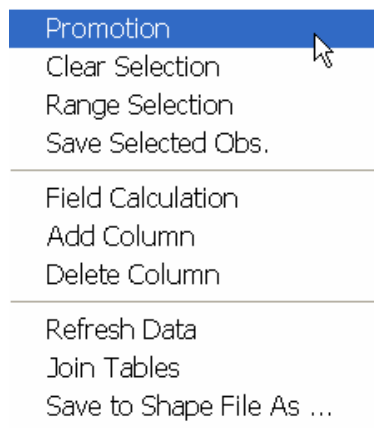


Figure 15. Table drop down menu.

	AREA	PERIMETER	CNTY_	CNTY_ID	NAME	STATE_NAME	STATE_FIPS
67	0.181000	1.980000	2040	2040	Moore	North Carolina	37
60	0.065000	1.093000	2026	2026	Lee	North Carolina	37
63	0.154000	1.680000	2030	2030	Hamett	North Carolina	37
54	0.207000	1.851000	1989	1989	Johnston	North Carolina	37
96	0.225000	2.107000	2162	2162	Bladen	North Carolina	37
94	0.240000	2.004000	2150	2150	Robeson	North Carolina	37
82	0.172000	1.835000	2090	2090	Cumberland	North Carolina	37
86	0.098000	1.262000	2097	2097	Hoke	North Carolina	37
79	0.241000	2.214000	2083	2083	Sampson	North Carolina	37
92	0.080000	1.188000	2123	2123	Scotland	North Carolina	37
98	0.240000	2.365000	2232	2232	Columbus	North Carolina	37
80	0.121000	1.855000	2107	2107	Richmond	North Carolina	37

Figure 16. Table with selected rows promoted.

	CNTY_FIPS	FIPS	FIPSNO	CRESS_ID	BIR.74	SID74	NWBIR.74
32	121	37121	37121	61	671.000000	0.000000	1.000000
90	043	37043	37043	22	284.000000	0.000000	1.000000
22	011	37011	37011	6	781.000000	0.000000	4.000000
38	115	37115	37115	58	765.000000	2.000000	5.000000
78	113	37113	37113	57	797.000000	0.000000	9.000000
1	009	37009	37009	5	1091.000000	1.000000	10.000000
2	005	37005	37005	3	487.000000	0.000000	10.000000

Figure 17. Table sorted on NWBIR74.

	CNTY_FIPS	FIPS	FIPSNO	CRESS_ID	BIR.74	SID74	NWBIR.74
68	119	37119	37119	60	21588.000000	44.000000	8027.000000
82	051	37051	37051	26	20366.000000	38.000000	7043.000000
94	155	37155	37155	78	7889.000000	31.000000	5904.000000
26	081	37081	37081	41	16184.000000	23.000000	5483.000000
37	183	37183	37183	92	14484.000000	16.000000	4397.000000
25	067	37067	37067	34	11858.000000	10.000000	3919.000000
30	063	37063	37063	32	7970.000000	16.000000	3732.000000

Figure 18. Table reverse sorted on NWBIR74.

Table Sorting and Selection

The new table feature in GeoDa also has a few useful sorting functions. Double-clicking on the column header for any column sorts the observations in ascending order for that variable (a small triangle pointing up appears next to the variable name). Double-clicking again reverses the order. Double clicking on the first column (with the sequence numbers) clears the sorting. For example, in Figure 17, the observations are sorted in ascending order for NWBIR74, while in Figure 18, they are sorted in descending order.

Individual rows can be selected by clicking on their sequence number in the left-most column of the table. Shift-click adds or removes observations from the selection. You can also drag the pointer down over the left-most column to select multiple records. The selection is immediately reflected in all the linked maps (and other graphs). You clear the selection by right click > Clear Selection (or Options > Clear Selection).

GeoDa also implements a limited number of queries, primarily geared to selecting observations that have a specific value or range of values. A logical statement can be constructed to select observations, depending on the range for a specific variable (but for one variable only at this point).

To build a query, right click in the table and select Range Selection from the drop down menu (or use Option > Range Selection in the menu). A dialog appears that allows you to construct a range (Figure 19). Note that the range is inclusive on the left hand side and exclusive on the right hand side (\leq and $<$). To find those counties with 500 or fewer live births in 74, enter 0 in the left text box, select BIR74 as the variable and enter 500.1 in the right hand side text box, next click Apply. This will activate the Recoding dialog, which allows you to create a new variable (default REGIME) with value = 1 for the selected observations and zero elsewhere. If you don't want this, click OK. If you do want the extra variable, first click Apply and then OK (Apply only won't do it).

The selected rows will show up in the table highlighted in blue. To collect them together, choose Promote from the drop down menu. The result should be as in Figure 20. Note the extra column for REGIME. However, the new variable is not permanent and can become so only after the table is saved (see below). Because of the linking, the matching counties are also selected in the map in Figure 21 (here highlighted in red).

Practice

Use the table sorting features to find in the map those counties that had no homicides in 84-88 ($HC8488 = 0$). Now use the range selection feature to find those counties with less than 5 homicides in 84-88 ($HC8488 < 5$). Create a dummy variable (type a different name instead of REGIME) for each of the selections. For practice, use different variables and/or different ranges.

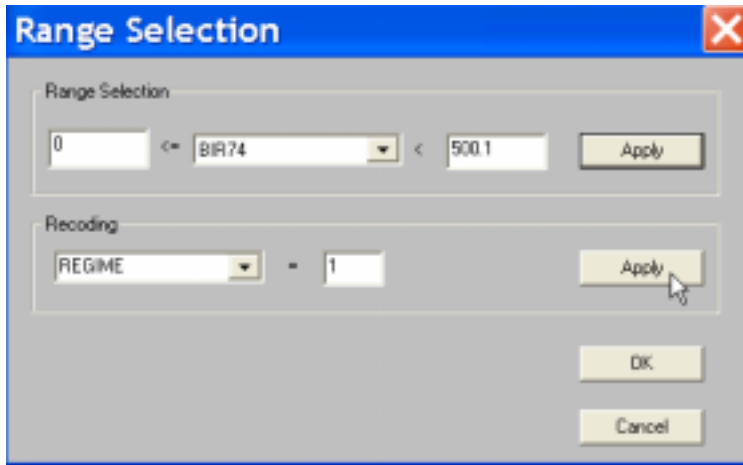


Figure 19. Range selection dialog.

	FIPS	FIPSNO	CRESS_ID	BIR74	SID74	NWBIR74	BIR79	SID79	NWBIR79	REGIME
87	37095	37095	48	338.000000	0.000000	134.000000	427.000000	0.000000	169.000000	1
2	37005	37005	3	487.000000	0.000000	10.000000	542.000000	3.000000	12.000000	1
7	37029	37029	15	286.000000	0.000000	115.000000	350.000000	2.000000	139.000000	1
73	37075	37075	38	415.000000	0.000000	40.000000	488.000000	1.000000	45.000000	1
20	37143	37143	72	484.000000	1.000000	230.000000	676.000000	0.000000	310.000000	1
90	37043	37043	22	284.000000	0.000000	1.000000	419.000000	0.000000	5.000000	1
8	37073	37073	37	420.000000	0.000000	254.000000	594.000000	2.000000	371.000000	1
45	37177	37177	89	248.000000	0.000000	116.000000	319.000000	0.000000	141.000000	1
9	37185	37185	93	968.000000	4.000000	748.000000	1190.000000	2.000000	844.000000	0

Figure 20. Counties with fewer than 500 births in 74, table view.

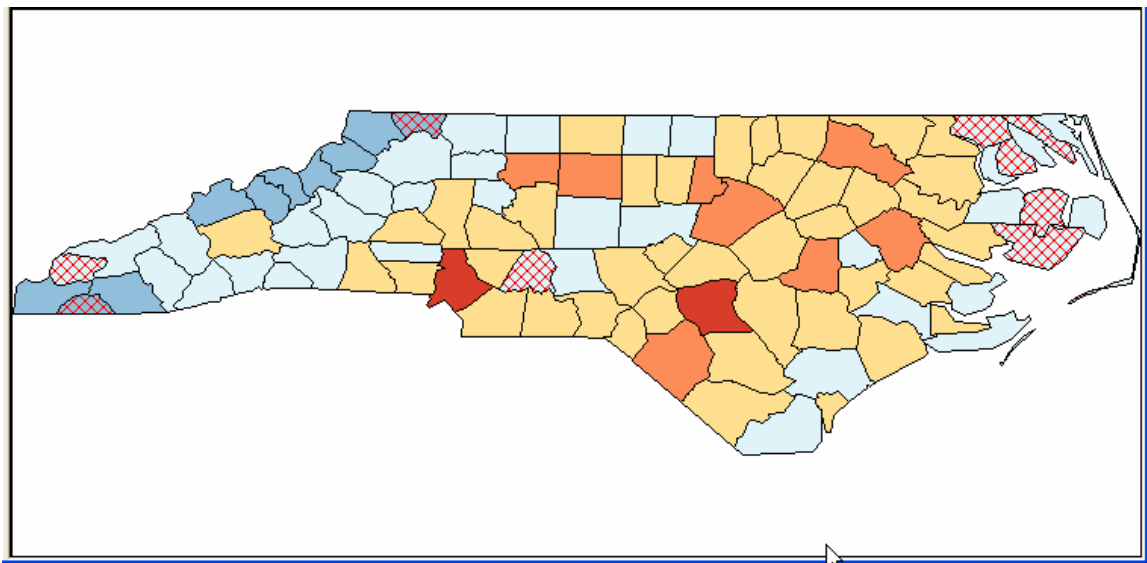


Figure 21. Counties with fewer than 500 total births in 74, in NWBIR74 map view.

Table Calculations

The table also has a limited “calculator” functionality, so that new variables can be added, transformations carried out on current variables, etc. You invoke the calculator from the drop down menu (right click on the table) or from the Options menu (when the table is active). The calculator dialog has tabs on the top to select the type of operation you want to carry out.

For example, you may have noticed that the SIDS data set contains only the counts of births and deaths, but no rates. To create a new variable for the SIDS death rate in 74, select Add Column from the drop down menu, and enter SIDR74 for the variable name, followed by a click on Add, as in Figure 22. A new empty column appears on the extreme right hand side of the table (Figure 23).

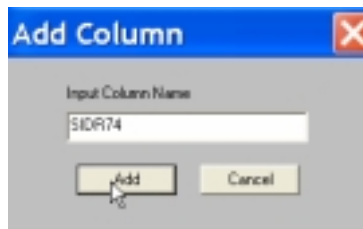
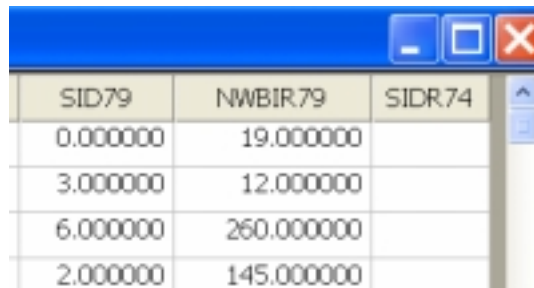


Figure 22. Adding a new variable to the table.



SID79	NWBIR.79	SIDR.74
0.000000	19.000000	
3.000000	12.000000	
6.000000	260.000000	
2.000000	145.000000	

Figure 23. Table with new empty column.

To calculate the rate, choose Field Calculation in the drop down menu and click on the right hand tab (Rate Operations) in the Field Calculation dialog. This invokes a dialog specific to the computation of rates (including rate smoothing). For now, select the raw rate option and make sure to have SID74 as the “event” and BIR74 as the “base,” as illustrated in Figure 24. Click OK to have the new value added to the table (Figure 25).

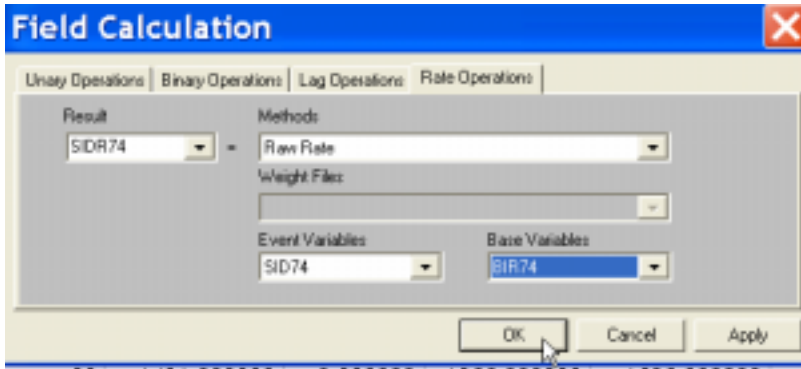


Figure 24. SIDS death rate calculation using raw rate option.

SID79	NWBIR79	SIDR74
0.000000	19.000000	0.000917
3.000000	12.000000	0.000000
6.000000	260.000000	0.001568
2.000000	145.000000	0.001969
3.000000	1197.000000	0.006334
5.000000	1237.000000	0.004821

Figure 25. Computed SIDS death rate added to table.

As shown in Figure 25, the rate may not be the most intuitive to interpret. You may want to rescale it as a number per 100,000 births, for example. Invoke the Field Calculation again, and this time, select the second tab for Binary Operations. Rescale the SIDR74 as SIDR74 MULTIPLY 100,000 (simply type the 100,000 over the variable name AREA), as in Figure 26. Click on OK to replace the SIDS death rate by its rescaled value, as in Figure 27.

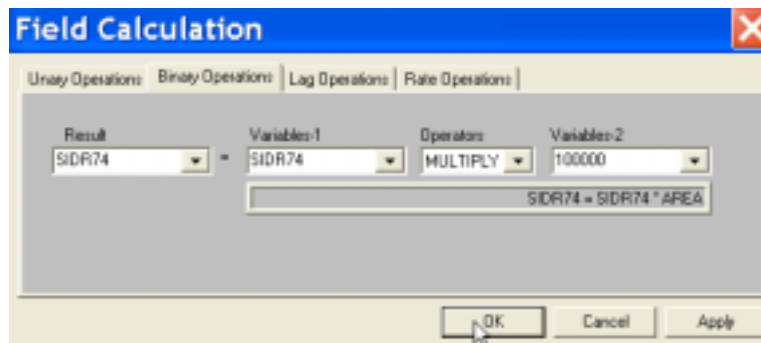


Figure 26. Rescaling the SIDS death rate.

SID79	NWBIR79	SIDR74
0.000000	19.000000	91.700000
3.000000	12.000000	0.000000
6.000000	260.000000	156.800000
2.000000	145.000000	196.900000
3.000000	1197.000000	633.400000

Figure 27. Rescaled SIDS death rate in table.

The newly computed values can be used in all the maps and statistical procedures. For example, in Figure 28, a Box Map is shown for the SIDR74 variable.

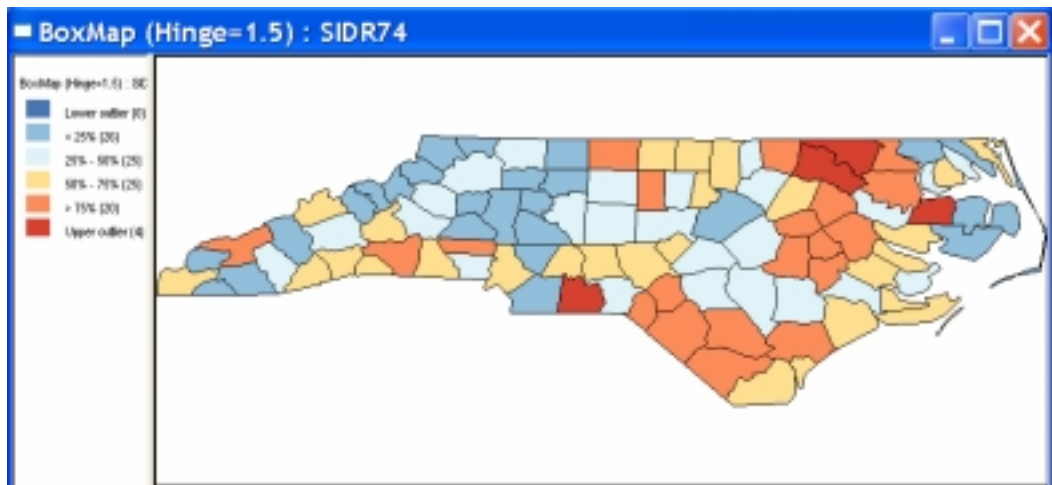


Figure 28. Box Map for SIDS death rates 74.

It's important to remember that the newly calculated variables are "temporary" and can be removed (in case you made a mistake) by selecting Refresh Data from the Table drop down menu. They become permanent only after you save them. You can save the new table to a new shape file using the Save to Shape File As ... option. The saved shape file will use the same "map" as the currently active shape file but with the newly constructed table as its "dbf" file. If you don't care about the shape files, simply remove the new .shp and .shx files and use the dbf file (e.g., in a statistics program). Use the procedure outlined above to create a SIDR74 and SIDR79 variable and add them permanently.

Practice

Construct homicide rate variables for the St Louis data (HC and PO are the event and base) and compare to the data already in the table. Rescale the rates to a different base and save the new table as a shape file under a different name. Clear GeoDa and load the new shape file. Check in the table that all the new variables are there. You may experiment with some of the other calculation options.

Linking Histograms

GeoDa contains histogram, box plot and scatter plot functions to carry out standard EDA. These statistical graphs are linked with each other as well as with the maps and table in such a way that any selection in any of the graphs yields the same selection in all other graphs and maps. This allows you to select on the map to find subsets of the data in the statistical graphs and vice versa. Selection in the statistical graphs is with point selection (click) or rectangle selection (click and drag).

The statistical graphs are invoked by clicking on the matching toolbar button or by means of the Explore menu (Figures 29-30). After the type of graph is selected, a dialog appears to choose the variable. Checking the default option will not ask you again for the variable. This is useful when you want to create multiple graphs for the same variable. You can turn it off and select a different variable with Edit > Select Variable.



Figure 29. Toolbar buttons for box map, histogram and scatter plot.

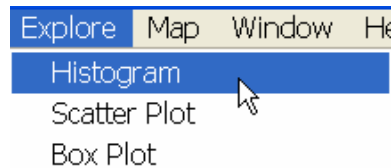


Figure 30. Menu items for statistical graphs.

Select Histogram (Figure 30) and choose the variable SIDR74 you created to make a histogram for the SIDS rates. The default number of categories is 7. This may not be appropriate in any given case. Change the number of categories with Options > Intervals and set to 10 (make sure the histogram window is active). Your histogram should look like Figure 31. Click on the right most bar to find out where in the map the county with the highest SIDS death rate is located. Alternatively, you can get a view of a “regional” distribution in a subset of the data by selecting the locations in the map and checking the histogram for the selected records. In Figure 32, this is shown for two bands of counties on the Western boundary of the state: the yellow part of the histogram corresponds to the selected counties. As before, you can go to the table, promote the selected counties, check their names, etc. In the map view, you can create a dummy variable for the selected locations by right clicking and choosing Save Selected Obs. A new variable will be added to the table.

Practice

Use the St. Louis data set to investigate regional distributions (e.g., east vs west, core vs. periphery) for the homicide rates. Create dummy variables for the regional selections.

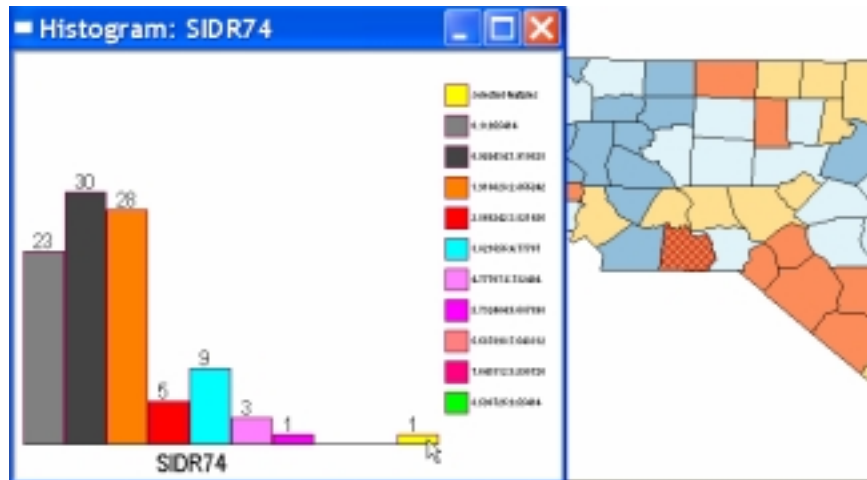


Figure 31. Histogram selection and matching county selection on the box map.

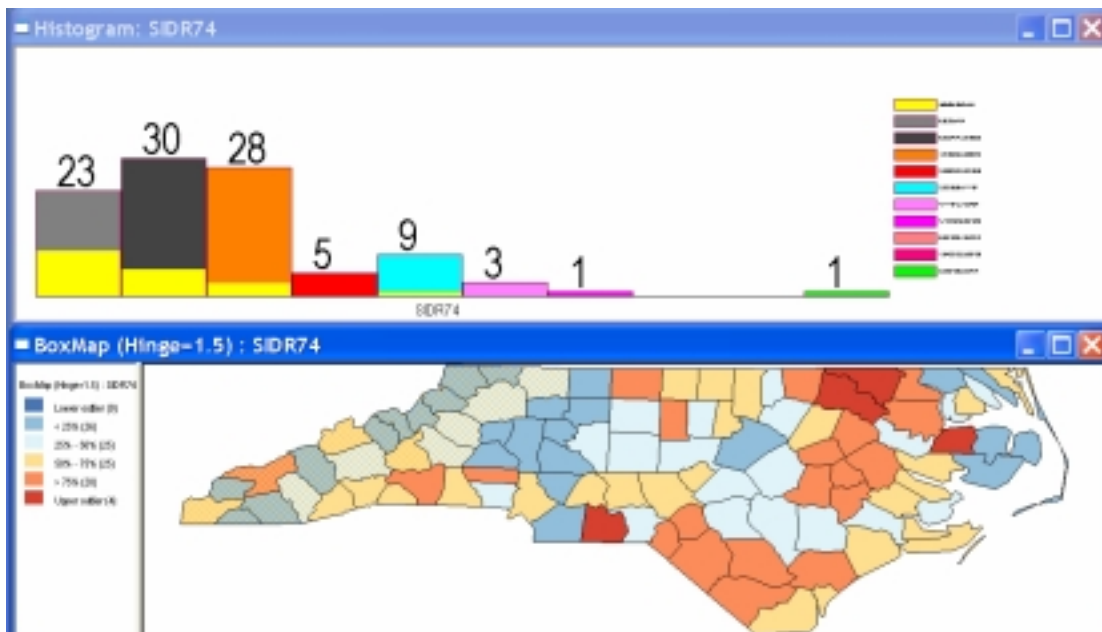


Figure 32. Map selection (box map) and corresponding histogram.

Linking Box Plots

Create a Box Plot for the SIDR74 variable by selecting Box Plot from the Explore Menu or by clicking on the corresponding toolbar button and choosing SIDR74 as the variable. Compare the outliers in the box plot to those in the box map: drag a rectangle around the points above and on the fence to select them and see where they are on the map (Figure 33). As you might expect, the “outliers” on the box “map” are the same as those in the box plot. A more interesting use of the link between the box plot and a map is to apply each to a different variable. For example, a box plot for SIDR79 and a box map for SIDR74 in Figure 34.

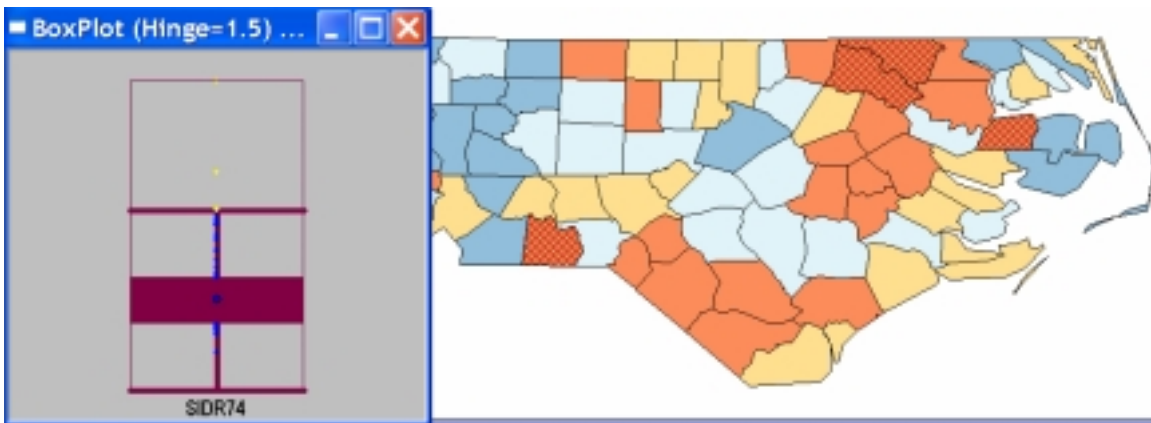


Figure 33. Linked box plot and box map for the same variable (SIDR74).

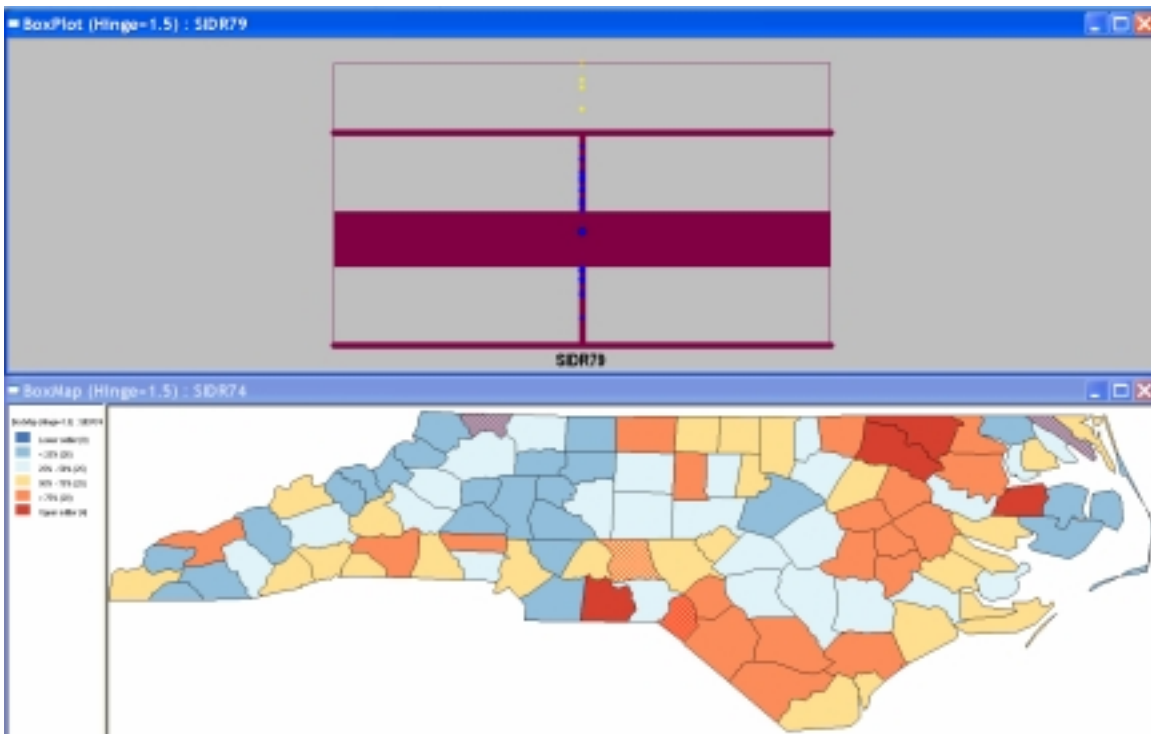


Figure 34. Linked box plot for SIDR79 to box map for SIDR74.

To make things look a little better, the customization tools were used to set the shading color to red and the background for the box plot to grey in Figure 34 (feel free to experiment with this as well). Note how none of the outliers in 79 were outliers in 74. In fact, two of them were below the median in 74. This is likely a problem due to the great variance instability of rates and requires adjustment, but a discussion of smoothing is beyond the current scope. As before, the selected locations are also linked to any other graph (such as the histogram just created), map and table.

Practice

Use the box plot, histogram and box map to compare outliers in the homicide rates (HR) for the three periods included in the data set. Use the table to identify the names of the counties and the actual number of homicides they had (HC).

Linking Scatter Plots

The scatter plot is a bivariate plot and thus two variables need to be specified. In all other respects, it works and starts as the other graphs. The Variable Select dialog shows two active columns, as in Figure 35. The first one corresponds to the Y-axis (dependent variable) the second one to the X-axis (explanatory variable). Select SIDR74 and NWR74 (the raw rate of non-white births over total births; needs to be constructed first if it is not in your data set) and click OK to create the scatter plot, as in Figure 36.

The scatter plot has two very useful options. One turns it into a correlation plot by standardizing both variables (Figure 37). With the scatter plot active, select Options > Scatterplot > Standardized data.

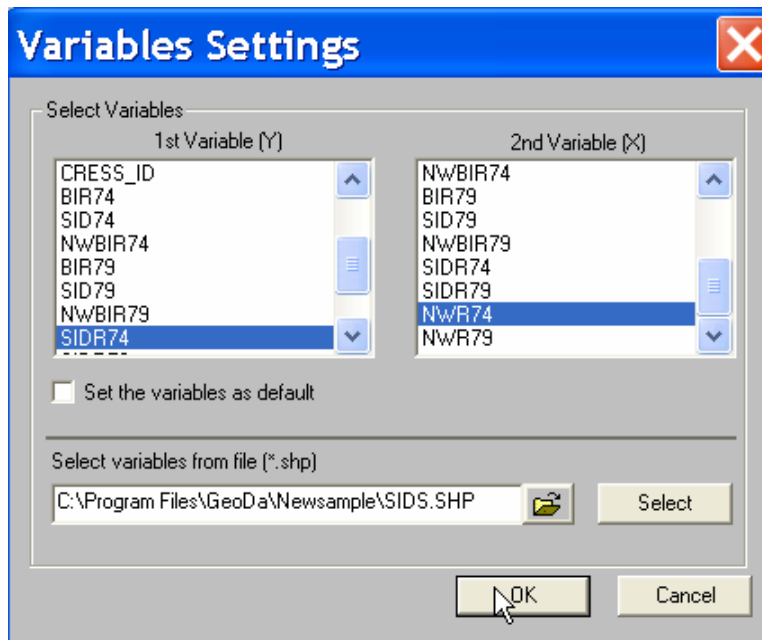


Figure 35. Bivariate variable selection.

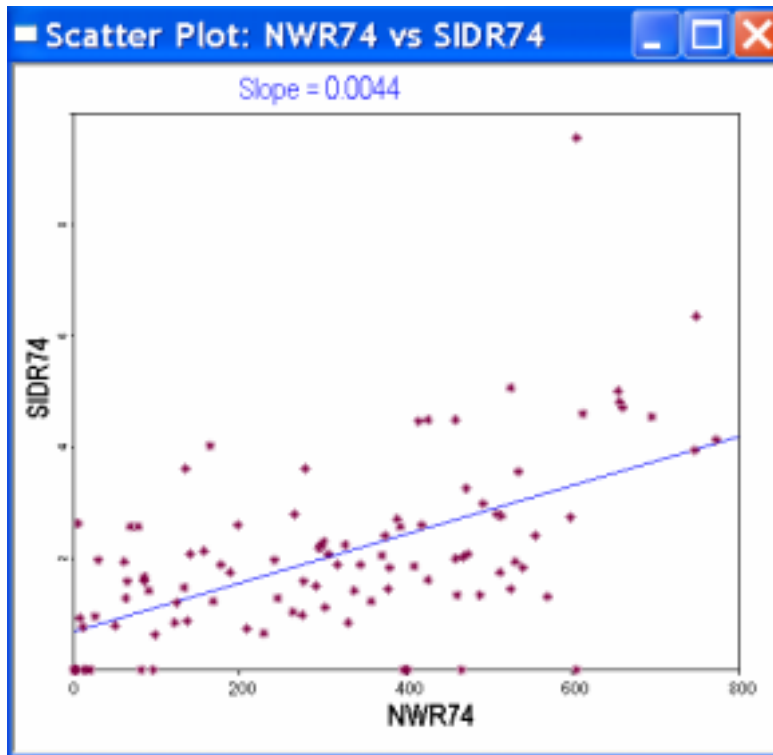


Figure 36. Scatterplot of NWR74 on SIDR74.

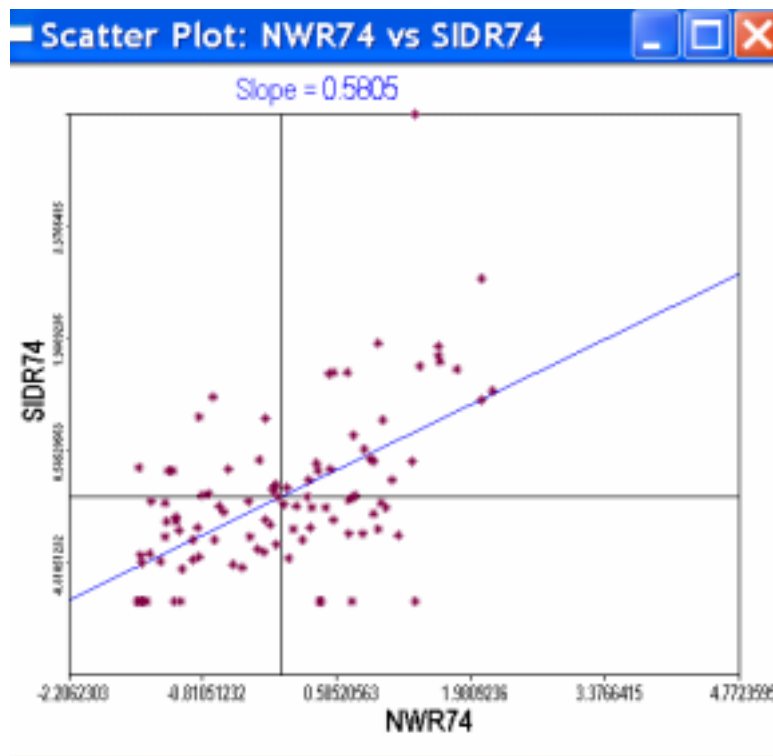


Figure 37. Standardized scatterplot, correlation between SIDR74 and NWR74.

The second option is particularly useful in the context of linking (and brushing, see below). When the Exclude Selected option is checked and a subset of points are selected in the scatter plot (using a point select or rectangle select), the slope is recomputed for the remaining points. So, the new slope (listed on top of the graph in brown) and the new regression line pertain to all observations *except* the selected ones (they are excluded from the sample). You can use this to assess the sensitivity of the regression slope to specific observations (such as outliers) or to compare the results in subsets of the data. For example, in Figure 38 the slope is recalculated with the selected counties excluded, the outliers for SIDR74 as identified in the Box Map. Note that double clicking in the scatter plot reverses the selection (and thus shows the slope with the previously excluded observations now included). Experiment with this (and follow the selected observations in the other graphs and table) for the selection of outliers, boundary counties, central counties, etc. Keep in mind that there must be a reasonable number of observations left for the scatter plot to be meaningful (a scatter plot with one observation is not meaningful).

Practice

Use the scatter plot to assess the degree of correlation between the homicide rates in St Louis and a deprivation index (RDAC). Assess the sensitivity of the regression slope to outliers (on any one of the variables in the data set) and to regional effects (e.g., select with and without St Louis city and East St Louis, Eastern counties vs Western counties, etc.). Use the different select tools and outlier maps, as needed.

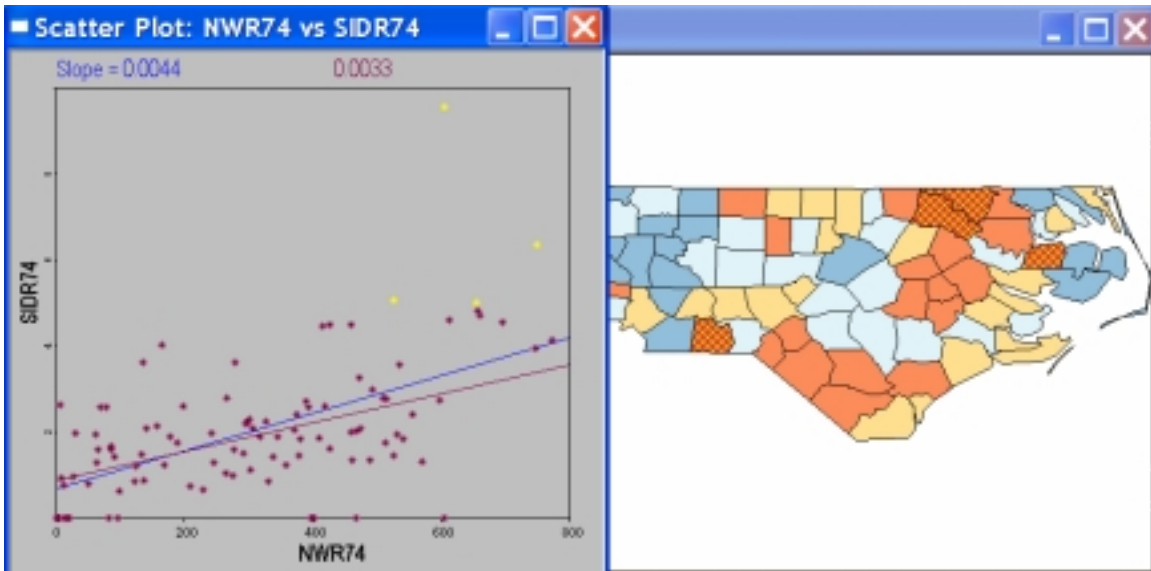


Figure 38. Linked scatter plot with selected counties excluded.

Brushing

An extension of linking to a fully dynamic framework is so-called brushing. It consists of constructing a rectangle in one of the maps or graphs and then moving it over the graph. As the rectangle moves, the selected observations change and this affects all the current graphs. The most useful graphs to brush on are the scatter plot (where the concept originated), box plot and map. It is especially powerful when the “exclude selected” option is active (as the brush moves, the slope of the scatter plot is recalculated on the fly).

The brush is created by dragging the pointer to a small rectangle, followed by CTRL. After a few moments, the rectangle will start to blink, which indicates it is active. This can be done in a map as well as in a scatter plot or box plot. Drag the rectangle over the graph with the mouse. You end the process by clearing the selection (click on white space in the map or on the left hand column in the table).

This is hard to illustrate on a static piece of paper, but try it out on the SIDR74 vs NWR74 scatter plot. Alternatively, start with the map and move the brush over different regions in the map.

Practice

Brush the scatter plots and maps you constructed for the analysis of the relation between homicide rates and the deprivation index. Try it both from the scatter plot and from the map end. Note how the brush moves across all maps.