CSISS Center for Spatially Integrated Social Science

Exploring Spatial Data with GeoDa™ : A Workbook

Luc Anselin

Spatial Analysis Laboratory
Department of Geography
University of Illinois, Urbana-Champaign
Urbana, IL 61801

http://sal.agecon.uiuc.edu/

Center for Spatially Integrated Social Science

http://www.csiss.org/

Revised Version, March 6, 2005

Contents

List of Figures

Preface

This workbook contains a set of laboratory exercises initally developed for the ICPSR Summer Program courses on spatial analysis: *Introduction to Spatial Data Analysis* and *Spatial Regression Analysis*. It consists of a series of brief tutorials and worked examples that accompany the *GeoDa*$^{\mathrm{TM}}$ *User's Guide* and *GeoDa*$^{\mathrm{TM}}$ *0.95i Release Notes* (Anselin 2003a, 2004).[1] They pertain to release 0.9.5-i of *GeoDa*, which can be downloaded for free from http://sal.agecon.uiuc.edu/geoda_main.php. The "official" reference to *GeoDa* is Anselin et al. (2004c).

GeoDa$^{\mathrm{TM}}$ is a trade mark of Luc Anselin.

Some of these materials were included in earlier tutorials (such as Anselin 2003b) available on the SAL web site. In addition, the workbook incorporates laboratory materials prepared for the courses ACE 492SA, *Spatial Analysis* and ACE 492SE, *Spatial Econometrics*, offered during the Fall 2003 semester in the Department of Agricultural and Consumer Economics at the University of Illinois, Urbana Champaign. There may be slight discrepancies due to changes in the version of *GeoDa*. In case of doubt, the most recent document should always be referred to as it supersedes all previous tutorial materials.

The examples and practice exercises use the sample data sets that are available from the SAL "stuff" web site. They are listed on and can be downloaded from http://sal.agecon.uiuc.edu/data_main.php. The main purpose of these sample data is to illustrate the features of the software. Readers are strongly encouraged to use their own data sets for the practice exercises.

Acknowledgments

The development of this workbook has been facilitated by the continued research support through the U.S. National Science Foundation grant BCS-

[1] In the remainder of this workbook these documents will be referred to as *User's Guide* and *Release Notes*

9978058 to the Center for Spatially Integrated Social Science (CSISS). More recently, support has also been provided through a Cooperative Agreement between the Center for Disease Control and Prevention (CDC) and the Association of Teachers of Preventive Medicine (ATPM), award # TS-1125. The contents of this workbook are the sole responsibility of the author and do not necessarily reflect the official views of the CDC or ATPM.

Finally, many participants in various *GeoDa* workshops have offered useful suggestions and comments, which is greatly appreciated. Special thanks go to Julia Koschinsky, who went through earlier versions in great detail, which resulted in several clarifications of the material.

Exercise 1

Getting Started with GeoDa

1.1 Objectives

This exercise illustrates how to get started with *GeoDa*, and the basic structure of its user interface. At the end of the exercise, you should know how to:

- open and close a project
- load a shape file with the proper indicator (`Key`)
- select functions from the menu or toolbar

More detailed information on these operations can be found in the *User's Guide*, pp. 3–18, and in the *Release Notes*, pp. 7–8.

1.2 Starting a Project

Start *GeoDa* by double-clicking on its icon on the desktop, or run the *GeoDa* executable in Windows Explorer (in the proper directory). A welcome screen will appear. In the `File` Menu, select `Open Project`, or click on the `Open Project` toolbar button, as shown in Figure 1.1 on p. 2. Only two items on the toolbar are active, the first of which is used to launch a project, as illustrated in the figure. The other item is to close a project (see Figure 1.5 on p. 4).

After opening the project, the familiar Windows dialog requests the file name of a shape file and the `Key` variable. The `Key` variable uniquely identifies each observation. It is typically an integer value like a FIPS code for counties, or a census tract number.

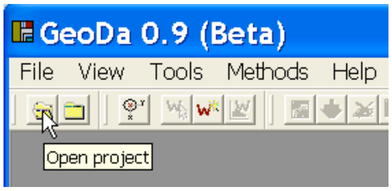
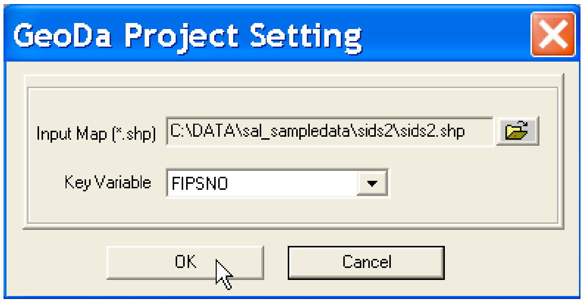


Figure 1.1: The initial menu and toolbar.

In *GeoDa*, only shape files can be read into a project at this point. However, even if you don't have your data in the form of a shape file, you may be able to use the included spatial data manipulation tools to create one (see also Exercises 4 and 5).

To get started, select the SIDS2 sample data set as the `Input Map` in the file dialog that appears, and leave the `Key` variable to its default `FIPSNO`. 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).[1]

Finally, click on `OK` to launch the map, as in Figure 1.2.



Input Map (*.shp) C:\DATA\sal_sampledata\sids2\sids2.shp

Key Variable FIPSNO

Figure 1.2: Select input shape file.

Next, a map window is opened, showing the base map for the analyses,

[1] When using your own data, you may get an error at this point (such as "out of memory"). This is likely due to the fact that the chosen `Key` variable is either not unique or is a character value. Note that many county data shape files available on the web have the FIPS code as a character, and *not* as a numeric variable. To fix this, you need to convert the character variable to numeric. This is easy to do in most GIS, database or spreadsheet software packages. For example, in ArcView this can be done using the `Table` edit functionality: create a new `Field` and calculate it by applying the `AsNumeric` operator to the original character variable.

depicting the 100 counties of North Carolina, as in Figure 1.3. The window shows (part of) the legend pane on the left hand size. This can be resized by dragging the separator between the two panes (the legend pane and the map pane) to the right or left.

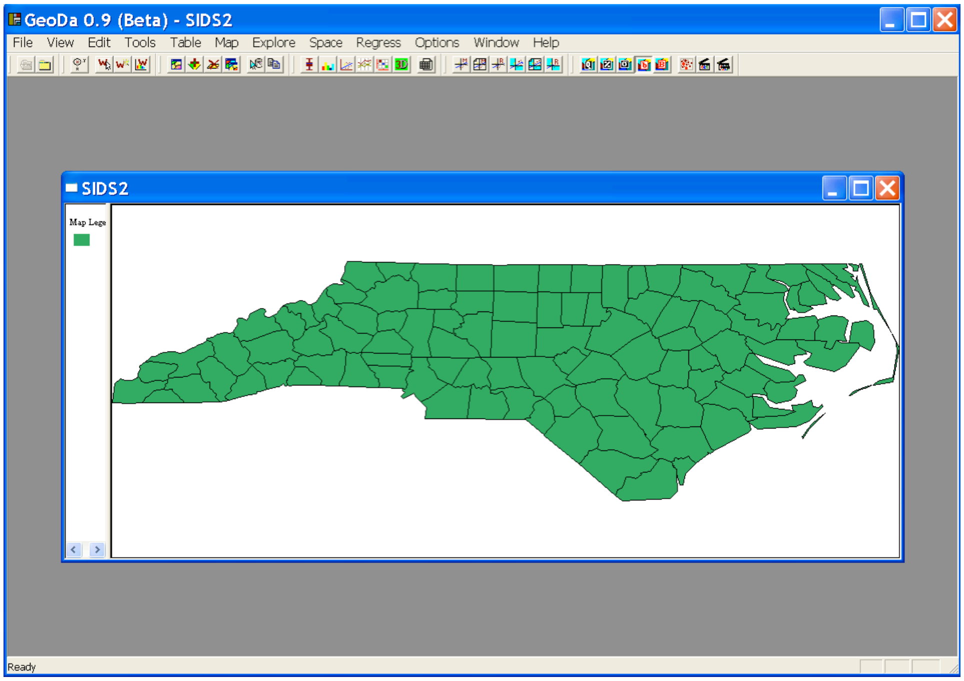


Figure 1.3: Opening window after loading the SIDS2 sample data set.

You can change basic map settings by right clicking in the map window and selecting characteristics such as color (background, shading, etc.) and the shape of the selection tool. Right clicking opens up a menu, as shown in Figure 1.4 (p. 4). For example, to change the color for the base map from the default green to another color, click `Color` > `Map` and select a new color from the standard Windows color palette.

To clear all open windows, click on the `Close all windows` toolbar button (Figure 1.5 on p. 4), or select `Close All` in the `File` menu.

1.3 User Interface

With a shape file loaded, the complete menu and all toolbars become active, as shown in detail in Figure 1.6 on p. 4.

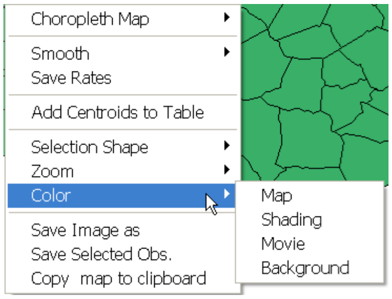


Figure 1.4: Options in the map (right click).

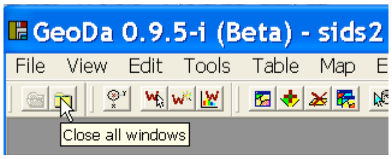

GeoDa 0.9.5-i (Beta) - sids2
File View Edit Tools Table Map E
Close all windows

Figure 1.5: Close all windows.

The menu bar contains eleven items. Four are standard Windows menus: `File` (open and close files), `View` (select which toolbars to show), `Windows` (select or rearrange windows) and `Help` (not yet implemented). Specific to *GeoDa* are `Edit` (manipulate map windows and layers), `Tools` (spatial data manipulation), `Table` (data table manipulation), `Map` (choropleth mapping and map smoothing), `Explore` (statistical graphics), `Space` (spatial autocorrelation analysis), `Regress` (spatial regression) and `Options` (application-specific options). You can explore the functionality of *GeoDa* by clicking on various menu items.

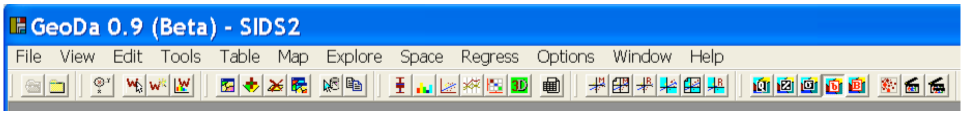

GeoDa 0.9 (Beta) - SIDS2
File View Edit Tools Table Map Explore Space Regress Options Window Help

Figure 1.6: The complete menu and toolbar buttons.

The toolbar consists of six groups of icons, from left to right: project open and close; spatial weights construction; edit functions; exploratory data analysis; spatial autocorrelation; and rate smoothing and mapping. As an example, the `Explore` toolbar is shown separately in Figure 1.7 on p. 5.

Clicking on one of the toolbar buttons is equivalent to selecting the matching item in the menu. The toolbars are dockable, which means that you can move them to a different position. Experiment with this and select a toolbar by clicking on the elevated separator bar on the left and dragging it to a different position.

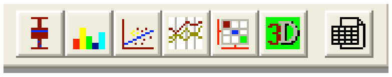


Figure 1.7: Explore toolbar.

1.4 Practice

Make sure you first close all windows with the North Carolina data. Start a new project using the St. Louis homicide sample data set for 78 counties surrounding the St. Louis metropolitan area (`stl_hom.shp`), with `FIPSNO` as the key variable. Experiment with some of the map options, such as the base map color (`Color` > `Map`) or the window background color (`Color` > `Background`). Make sure to close all windows before proceeding.

Exercise 2

Creating a Choropleth Map

2.1 Objectives

This exercise illustrates some basic operations needed to make maps and select observations in the map.

At the end of the exercise, you should know how to:

- make a simple choropleth map
- select items in the map
- change the selection tool

More detailed information on these operations can be found in the *User's Guide*, pp. 35–38, 42.

2.2 Quantile Map

The SIDS data set in the sample collection is taken from Noel Cressie's (1993) *Statistics for Spatial Data* (Cressie 1993, pp. 386–389). It contains variables for the count of SIDS deaths for 100 North Carolina counties 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`).

Make sure to load the `sids.shp` shape file using the procedures reviewed in Exercise 1. As before, select `FIPSNO` as the `Key` variable. You should now have the green base map of the North Carolina counties in front of you, as in Figure 1.3 on p. 3. The only difference is that the window caption will be `sids` instead of `SIDS2`.

Consider constructing two quantile maps to compare the spatial distribution of non-white births and SIDS deaths in 74 (`NWBIR74` and `SID74`). Click on the base map to make it active (in *GeoDa*, the last 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 data table will appear as well. This can be ignored for now.[1] You should minimize the table to get it out of the way, but you will return to it later, so don't remove it.[2]

In the `Variables Settings` dialog, select `NWBIR74`, as in Figure 2.1, and click `OK`. Note the check box in the dialog to set the selected variable as the default. If you should do this, you will not be asked for a variable name the next time around. This may be handy when you want to do several different types of analyses for the same variable. However, in our case, we want to do the same analysis for different variables, so setting a default is *not* a good idea. If you inadvertently check the default box, you can always undo it by invoking `Edit > Select Variable` from the menu.

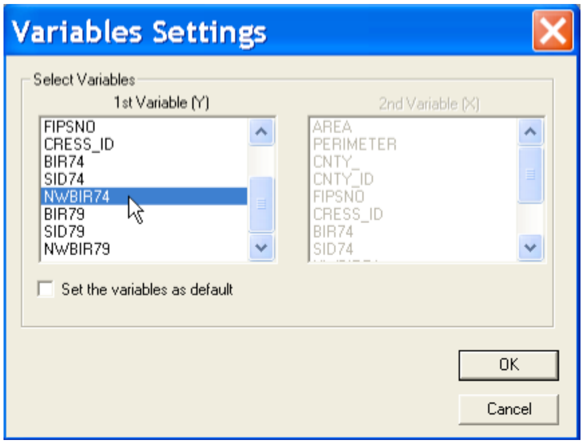


Figure 2.1: Variable selection.

After you choose the variable, a second dialog will ask for the number of categories in the quantile map: for now, keep the default value of 4 (quartile map) and click `OK`. A quartile map (four categories) will appear, as in Figure 2.2 on p. 8.

[1] The first time a specific variable is needed in a function, this table will appear.

[2] Minimize the window by clicking on the left-most button in the upper-right corner of the window.

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 in each of the four categories of the quartile map. The legend also lists the variable name.

You can obtain identical result by right-clicking on the map, which brings up the same menu as shown in Figure 1.4 on p. 4. Select `Choropleth Map` > `Quantile`, and the same two dialogs will appear to choose the variable and number of categories.

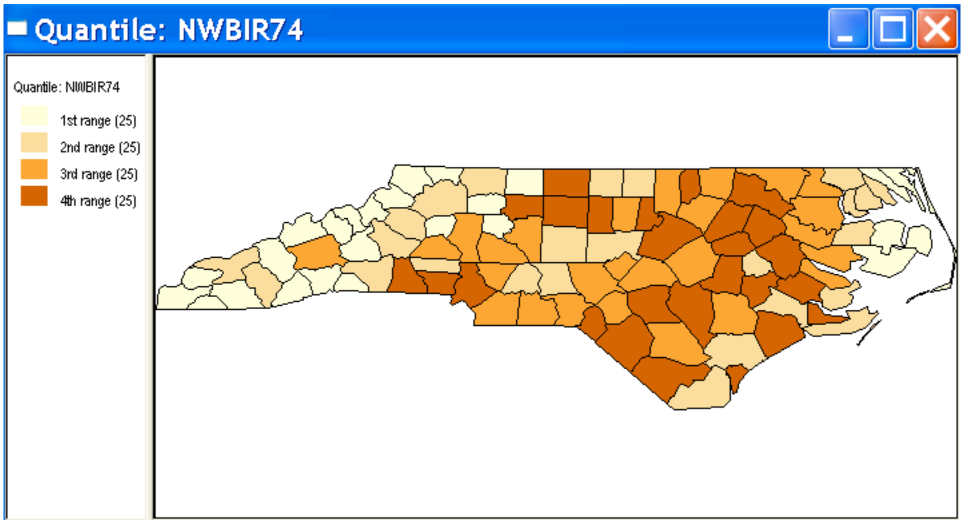


Figure 2.2: Quartile map for count of non-white births (`NWBIR74`).

Create a second choropleth map using the same geography. First, open a second window with the base map by clicking on the `Duplicate map` toolbar button, shown in Figure 2.3. Alternatively, you can select `Edit` > `Duplicate Map` from the menu.

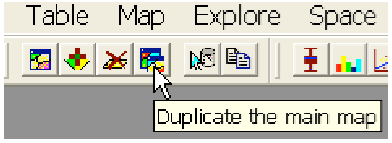


Figure 2.3: Duplicate map toolbar button.

Next, create a quartile map (4 categories) for the variable `SID74`, as shown in Figure 2.4 on p. 9. What do you notice about the number of

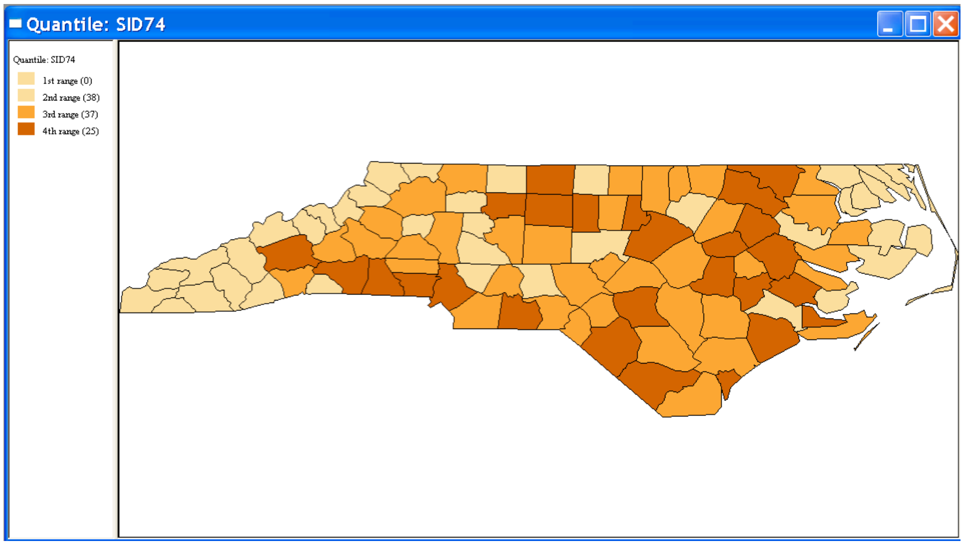


Figure 2.4: Quartile map for count of SIDS deaths (`SID74`).

observations in each quartile?

There are two problems with this map. One, it is a choropleth map for a "count," or a so-called *extensive* variable. This tends to be correlated with size (such as area or total population) and is often inappropriate. Instead, a rate or density is more suitable for a choropleth map, and is referred to as a *intensive* variable.

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. Note how the lowest category shows 0 observations, and the next 38.

You can save the map to the clipboard by selecting `Edit` > `Copy to Clipboard` from the menu. This only copies the map part. If you also want to get a copy of the legend, right click on the legend pane and select `Copy Legend to Clipboard`. Alternatively, you can save a bitmap of the map (but not the legend) to a .bmp formatted file by selecting `File` > `Export` > `Capture to File` from the menu. You will need to specify a file name (and path, if necessary). You can then use a graphic converter software package to turn the bmp format into other formats, as needed.

2.3 Selecting and Linking Observations in the Map

So far, the maps have been "static." The concept of *dynamic* maps implies that there are 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 for polygon shape files, whereas the circle is the default for point shape files. You select an observation by clicking on its location (click on a county to select it), or select multiple observations by dragging (click on a point, drag the pointer to a different location to create a rectangle, and release). You can add or remove locations from the selection by `shift-click`. To clear the selection, click anywhere outside the map. Other selection shapes can be used by right clicking on the map and choosing one of the options in the `Selection Shape` drop down list, as in Figure 2.5. Note that each individual map has its own selection tool and they don't have to be the same across maps.

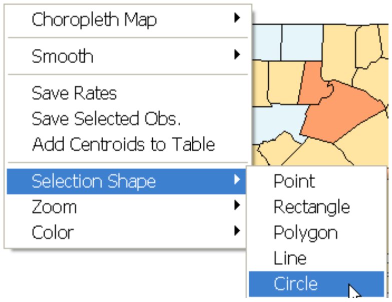


Figure 2.5: Selection shape drop down list.

As an example, choose circle selection (as in Figure 2.5), then click in the map for NWBIR74 and select some counties by moving the edge of the circle out (see Figure 2.6 on p. 11).

As soon as you release the mouse, the counties with their centroids within the circle will be selected, shown as a cross-hatch (Figure 2.7 on p. 12). Note that when multiple maps are in use, the same counties are selected in all maps, as evidenced by the cross-hatched patterns on the two maps in Figure 2.7. This is referred to as *linking* and pertains not only to the maps, but also to the table and to all other statistical graphs that may be active at the time. You can change the color of the cross-hatch as one of the map options (right click `Color` > `Shading`).

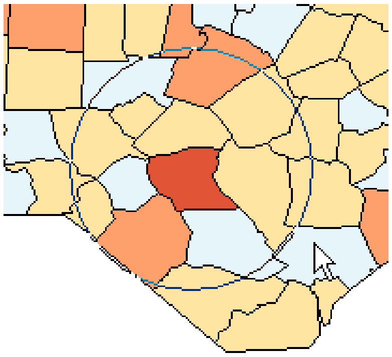


Figure 2.6: Circle selection.

2.4 Practice

Clear all windows, then start a new project with the St. Louis homicide sample data (`stl_hom.shp` with `FIPSNO` as the `Key`). Create two quintile maps (5 categories), one for the homicide rate in the 78 counties for the period 84-88 (`HR8488`), and one for the period 88-93 (`HR8893`). Experiment with both the `Map` menu as well as the right click approach to build the choropleth map. Use the different selection shapes to select counties in one of the maps. Check that the same are selected in the other map. If you wish, you can save one of the maps as a bmp file and insert into a MS Word file. Experiment with a second type of map, the standard deviational map, which sorts the values in standard deviational units.

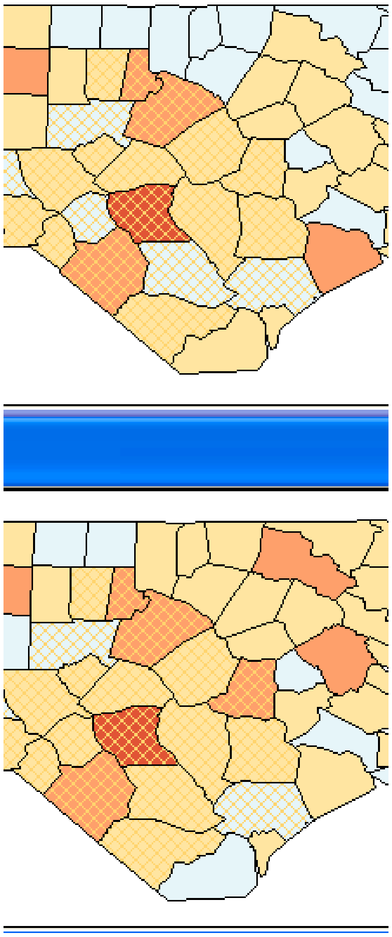


Figure 2.7: Selected counties in linked maps.

Exercise 3

Basic Table Operations

3.1 Objectives

This exercise illustrates some basic operations needed to use the functionality in the `Table`, including creating and transforming variables.

At the end of the exercise, you should know how to:

- open and navigate the data table
- select and sort items in the table
- create new variables in the table

More detailed information on these operations can be found in the *User's Guide*, pp. 54–64.

3.2 Navigating the Data Table

Begin again by clearing all windows and loading the `sids.shp` sample data (with `FIPSNO` as the `Key`). Construct a choropleth map for one of the variables (e.g., `NWBIR74`) and use the select tools to select some counties. Bring the `Table` back to the foreground if it had been minimized earlier. Scroll down the table and note how the selected counties are highlighted in blue, as in Figure 3.1 on p. 14.

To make it easier to identify the locations that were selected (e.g., to see the names of all the selected counties), use the `Promotion` feature of the `Table` menu. This can also be invoked from the table drop down menu (right click anywhere in the table), as shown in Figure 3.2 on p. 14. The

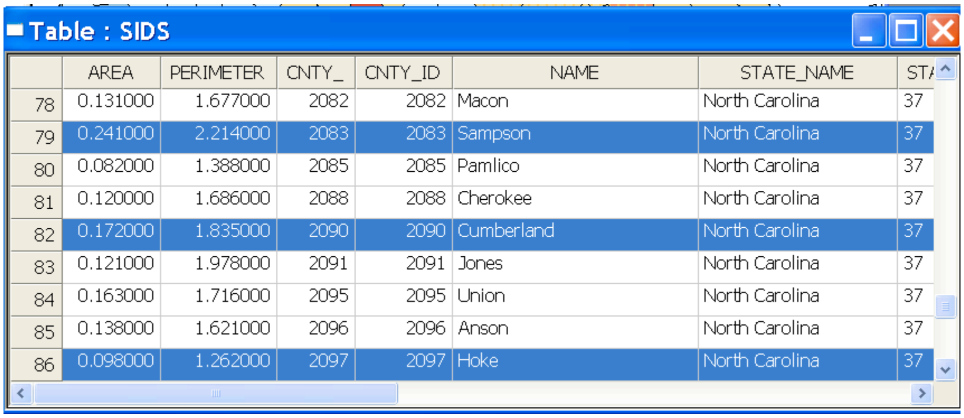


Table : SIDS

	AREA	PERIMETER	CNTY_	CNTY_ID	NAME	STATE_NAME	STA
78	0.131000	1.677000	2082	2082	Macon	North Carolina	37
79	0.241000	2.214000	2083	2083	Sampson	North Carolina	37
80	0.082000	1.388000	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	0.098000	1.262000	2097	2097	Hoke	North Carolina	37

Figure 3.1: Selected counties in linked table.

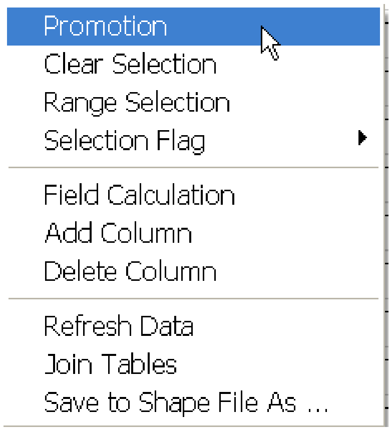
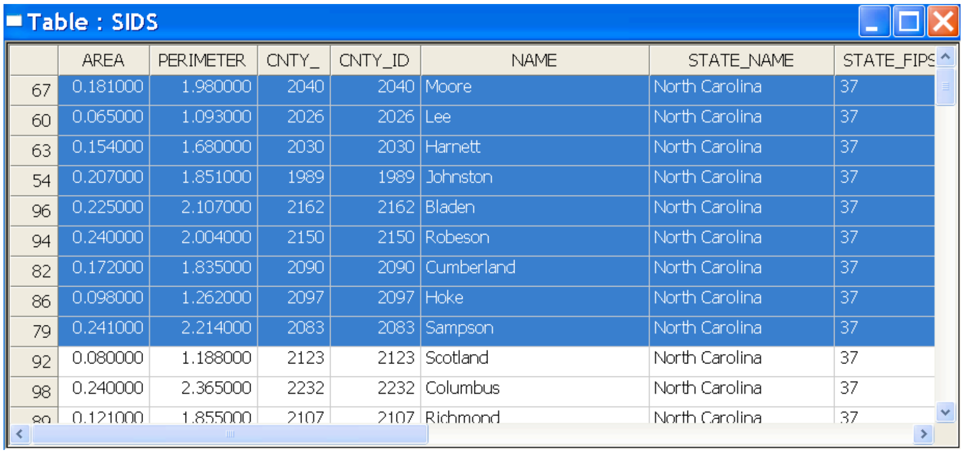


Figure 3.2: Table drop down menu.

selected items are shown at the top of the table, as in Figure 3.3 on p. 15. 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 selecting `Clear Selection` from the menu in Figure 3.2.

3.3 Table Sorting and Selecting

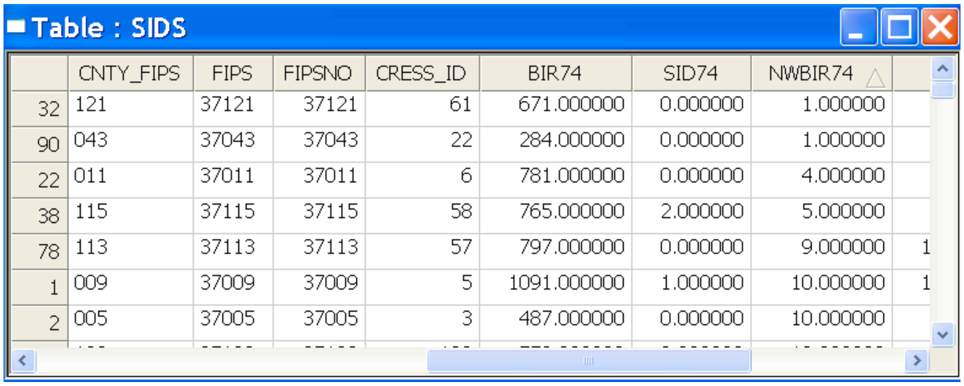
The way the table is presented at first simply reflects the order of the observations in the shape file. To sort the observations according to the value of



	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	Harnett	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
89	0.121000	1.855000	2107	2107	Richmond	North Carolina	37

Figure 3.3: Table with selected rows promoted.

a given variable, double click on the column header corresponding to that variable. This is a toggle switch: the sorting order alternates between ascending order and descending order. A small triangle appears next to the variable name, pointing up for ascending order and down for descending order. The sorting can be cleared by "sorting" on the observation numbers contained in the first column. For example, double clicking on the column header for NWBIR74 results in the (ascending) order shown in Figure 3.4.



	CNTY_FIPS	FIPS	FIPSNO	CRESS_ID	BIR74	SID74	NWBIR74 △	
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
1	009	37009	37009	5	1091.000000	1.000000	10.000000	1
2	005	37005	37005	3	487.000000	0.000000	10.000000	

Figure 3.4: Table sorted on NWBIR74.

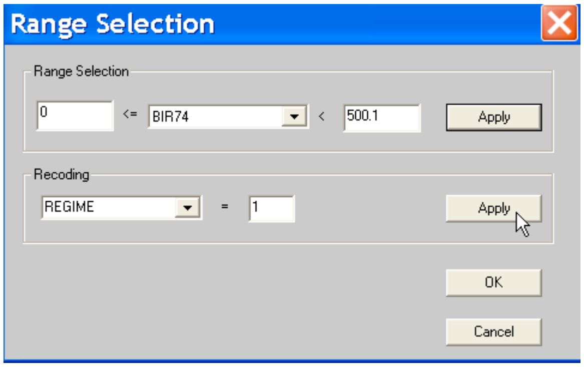
Individual rows can be selected by clicking on their sequence number in the left-most column of the table. Shift-click adds observations to or removes them from the selection. You can also drag the pointer down over the left-most column to select multiple records. The selection is immediately reflected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

3.3.1 Queries

GeoDa implements a limited number of queries, primarily geared to selecting observations that have a specific value or fall into a 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 `Table` > `Range Selection` in the menu). A dialog appears that allows you to construct a range (Figure 3.5). Note that the range is inclusive on the left hand side and exclusive on the right hand side (<= and <). To find those counties with 500 or fewer live births in 1974, enter `0` in the left text box, select `BIR74` as the variable and enter `500.1` in the right hand side text box. Next, activate the selection by clicking the top right `Apply` button.[1]

The first `Apply` will activate the `Recoding` dialog, which allows you to create a new variable with value set to `1` for the selected observations and zero elsewhere. The default variable name is `REGIME`, but that can be changed by overwriting the text box. If you do not want to create this variable, click `OK`. On the other hand, if you do want the extra variable, first click `Apply` (as in Figure 3.5) and then `OK`.



0 <= BIR74 < 500.1

REGIME = 1

Figure 3.5: Range selection dialog.

[1] This is the button above the one with the arrow on Figure 3.5.

The selected rows will show up in the table highlighted in blue. To collect them together, choose `Promotion` from the drop down menu. The result should be as in Figure 3.6. Note the extra column in the table for the variable `REGIME`. However, the new variable is not permanent and can become so only after the table is saved (see Section 3.4).

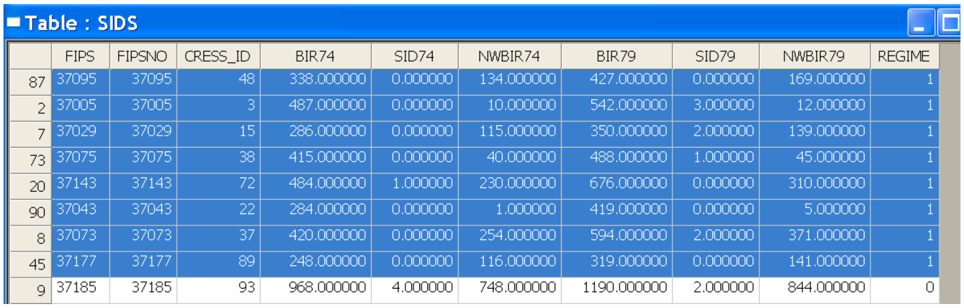


Table : SIDS

	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 3.6: Counties with fewer than 500 births in 74, table view.

3.4 Table Calculations

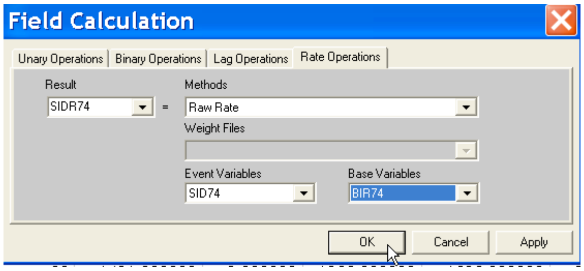
The table in *GeoDa* includes some limited "calculator" functionality, so that new variables can be added, current variables deleted, transformations carried out on current variables, etc. You invoke the calculator from the drop down menu (right click on the table) by selecting `Field Calculation` (see Figure 3.2 on p. 14). Alternatively, select `Field Calculation` from the `Table` item on the main menu.

The calculator dialog has tabs on the top to select the type of operation you want to carry out. For example, in Figure 3.7 on p. 18, the right-most tab is selected to carry out rate operations.

Before proceeding with the calculations, you typically want to create a new variable. This is invoked from the `Table` menu with the `Add Column` command (or, alternatively, by right clicking on the table). Note that this is not a requirement, and you may type in a new variable name directly in the left most text box of the `Field Calculation` dialog (see Figure 3.7). The new field will be added to the table.

You may have noticed that the `sids.shp` file contains only the counts of births and deaths, but no rates.[2] To create a new variable for the SIDS death rate in 74, select `Add Column` from the drop down menu, and enter `SIDR74`

[2] In contrast, the `sids2.shp` sample data set contains both counts and rates.



Field Calculation

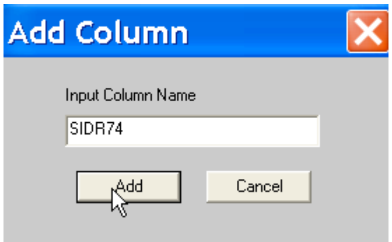
Result: SIDR74 = Methods: Raw Rate

Weight Files

Event Variables: SID74

Base Variables: BIR74

Figure 3.7: Rate calculation tab.



Add Column

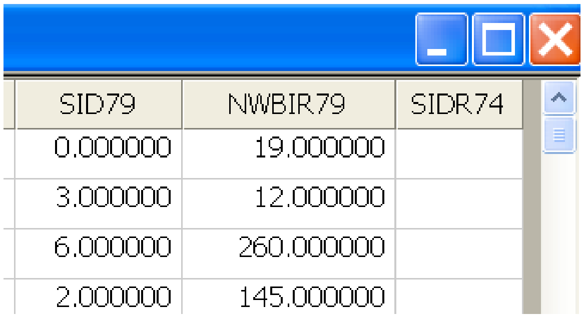
Input Column Name: SIDR74

Figure 3.8: Adding a new variable to a table.

for the new variable name, followed by a click on `Add`, as in Figure 3.8. A new empty column appears on the extreme right hand side of the table (Figure 3.9, p. 19).

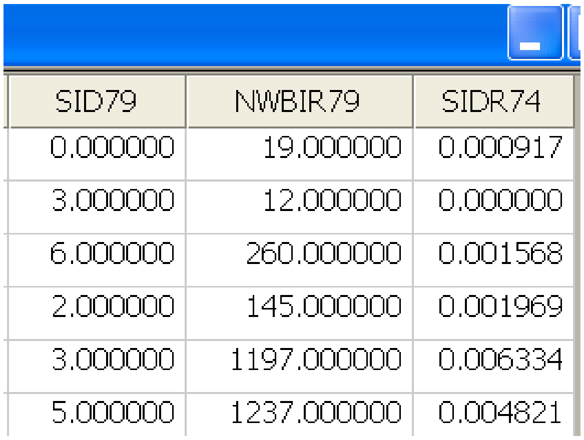
To calculate the rate, choose `Field Calculation` in the drop down menu (right click on the table) and click on the right hand tab (`Rate Operations`) in the `Field Calculation` dialog, as shown in Figure 3.7. This invokes a dialog specific to the computation of rates (including rate smoothing). For now, select the `Raw Rate` method and make sure to have `SIDR74` as the result, `SID74` as the `Event` and `BIR74` as the `base`, as illustrated in Figure 3.7. Click `OK` to have the new value added to the table, as shown in Figure 3.10 on p. 19.

As expressed in Figure 3.10, the rate may not be the most intuitive to interpret. For example, you may want to rescale it to show it in a more familiar form used by demographers and epidemiologists, with the rate expressed per 100,000 births. Invoke `Field Calculation` again, and, this time, select the second tab for `Binary Operations`. Rescale the variable `SIDR74` as `SIDR74 MULTIPLY 100,000` (simply type the 100,000 over the variable name `AREA`), as in Figure 3.11 on p. 20. To complete the operation, click on `OK` to replace



SID79	NWBIR79	SIDR74
0.000000	19.000000	
3.000000	12.000000	
6.000000	260.000000	
2.000000	145.000000	

Figure 3.9: Table with new empty column.



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

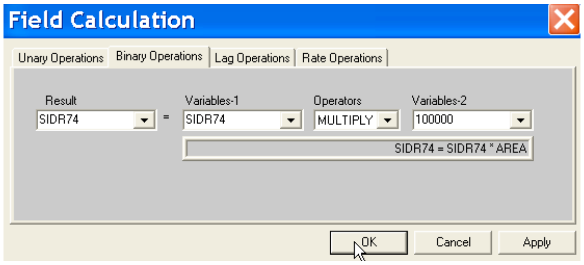
Figure 3.10: Computed SIDS death rate added to table.

the SIDS death rate by its rescaled value, as in Figure 3.12 on p. 20.

The newly computed values can immediately be used in all the maps and statistical procedures. However, it is important to remember that they are "temporary" and can still be removed (in case you made a mistake). This is accomplished by selecting `Refresh Data` from the `Table` menu or from the drop down menu in the table.

The new variables become permanent only after you save them to a shape file with a *different name*. This is carried out by means of the `Save to Shape File As` option.[3] The saved shape file will use the same map as

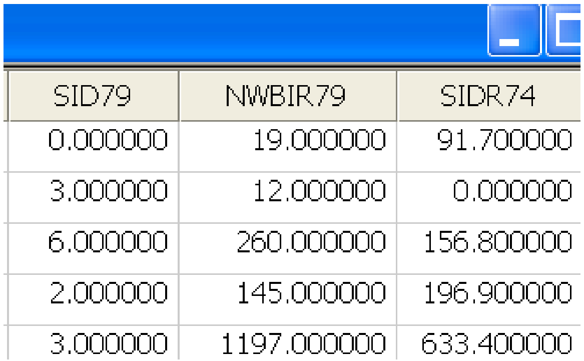
[3] This option only becomes active *after* some calculation or other change to the table has been carried out.



Result: SIDR74 = Variables-1: SIDR74 Operators: MULTIPLY Variables-2: 100000

SIDR74 = SIDR74 * AREA

Figure 3.11: 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 3.12: Rescaled SIDS death rate added to table.

the currently active shape file, but with the newly constructed table as its *dbf* file. If you dont care about the shape files, you can remove the new .shp and .shx files later and use the dbf file by itself (e.g., in a spreadsheet or statistics program).

Experiment with this procedure by creating a rate variable for `SIDR74` and `SIDR79` and saving the resulting table to a new file. Clear all windows and open the new shape file to check its contents.

3.5 Practice

Clear all windows and load the St. Louis sample data set with homicides for 78 counties (`stl_hom.shp` with `FIPSNO` as the `Key`). Create a choropleth map (e.g., quintile map or standard deviational map) to activate the table. Use the selection tools in the table to find out where particular counties are located (e.g., click on St. Louis county in the table and check where it is in the map). Sort the table to find out which counties has no homicides in the 84–88 period (`HC8488` $= 0$). Also use the range selection feature to find the counties with fewer than 5 homicides in this period (`HC8488` < 5).

Create a dummy variable for each selection (use a different name instead of the default `REGIME`). Using these new variables and the `Field Calculation` functions (*not* the `Range Selection`), create an additional selection for those counties with a nonzero homicide count less than 5. Experiment with different homicide count (or rate) variables (for different periods) and/or different selection ranges.

Finally, construct a homicide rate variable for a time period of your choice for the St. Louis data (`HCxxxx` and `POxxxx` are respectively the `Event` and `Base`). Compare your computed rates to the ones already in the table (`HRxxxx`). Rescale the rates to a different base and save the new table as a shape file under a different name. Clear all windows and load the new shape file. Check in the table to make sure that all the new variables are there. Experiment with some of the other calculation options as well.

Exercise 4

Creating a Point Shape File

4.1 Objectives

This exercise illustrates how you can create a point shape file from a text or dbf input file in situations where you do not have a proper ESRI formatted shape file to start out with. Since *GeoDa* requires a shape file as an input, there may be situations where this extra step is required. For example, many sample data sets from recent texts in spatial statistics are also available on the web, but few are in a shape file format. This functionality can be accessed without opening a project (which would be a logical contradiction since you don't have a shape file to load). It is available from the `Tools` menu.

At the end of the exercise, you should know how to:

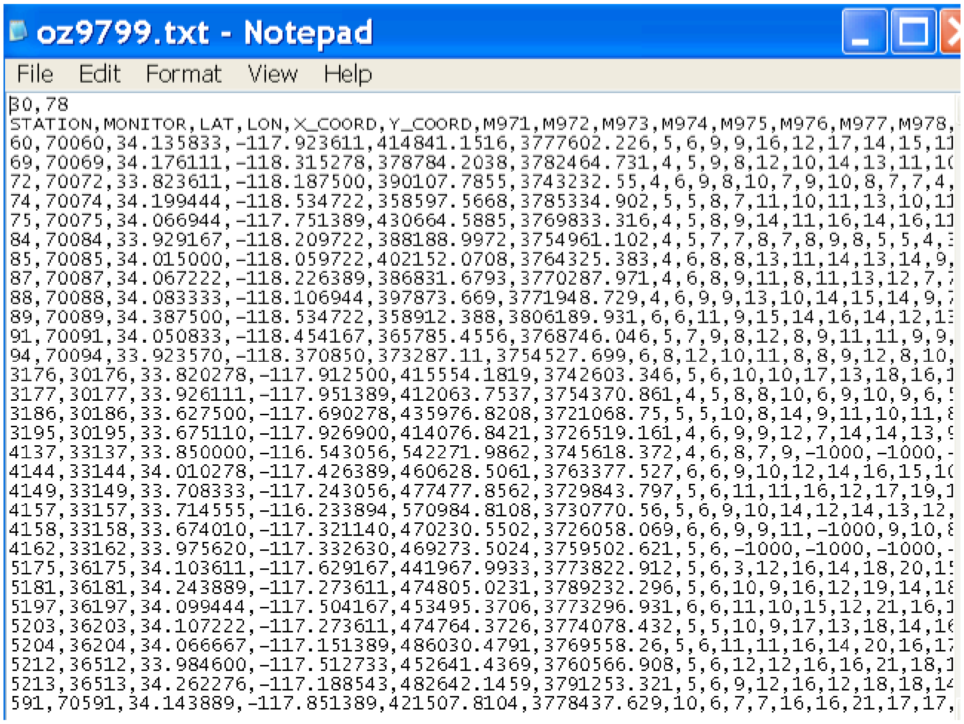
- format a text file for input into *GeoDa*
- create a point shape file from a text input file or dbf data file

More detailed information on these operations can be found in *Users's Guide* pp. 28–31.

4.2 Point Input File Format

The format for the input file to create a point shape file is very straightforward. The minimum contents of the input file are three variables: a unique identifier (integer value), the x-coordinate and the y-coordinate.[1] In a dbf

[1] Note that when latitude and longitude are included, the x-coordinate is the *longitude* and the y-coordinate the *latitude*.



```
30,78
STATION,MONITOR,LAT,LON,X_COORD,Y_COORD,M971,M972,M973,M974,M975,M976,M977,M978,
60,70060,34.135833,-117.923611,414841.1516,3777602.226,5,6,9,9,16,12,17,14,15,11
69,70069,34.176111,-118.315278,378784.2038,3782464.731,4,5,9,8,12,10,14,13,11,1
72,70072,33.823611,-118.187500,390107.7855,3743232.55,4,6,9,8,10,7,9,10,8,7,7,4,
74,70074,34.199444,-118.534722,358597.5668,3785334.902,5,5,8,7,11,10,11,13,10,11
75,70075,34.066944,-117.751389,430664.5885,3769833.316,4,5,8,9,14,11,16,14,16,11
84,70084,33.929167,-118.209722,388188.9972,3754961.102,4,5,7,7,8,7,8,9,8,5,5,4,
85,70085,34.015000,-118.059722,402152.0708,3764325.383,4,6,8,8,13,11,14,13,14,9,
87,70087,34.067222,-118.226389,386831.6793,3770287.971,4,6,8,9,11,8,11,13,12,7,
88,70088,34.083333,-118.106944,397873.669,3771948.729,4,6,9,9,13,10,14,15,14,9,
89,70089,34.387500,-118.534722,358912.388,3806189.931,6,6,11,9,15,14,16,14,12,1
91,70091,34.050833,-118.454167,365785.4556,3768746.046,5,7,9,8,12,8,9,11,11,9,9,
94,70094,33.923570,-118.370850,373287.11,3754527.699,6,8,12,10,11,8,8,9,12,8,10,
3176,30176,33.820278,-117.912500,415554.1819,3742603.346,5,6,10,10,17,13,18,16,1
3177,30177,33.926111,-117.951389,412063.7537,3754370.861,4,5,8,8,10,6,9,10,9,6,
3186,30186,33.627500,-117.690278,435976.8208,3721068.75,5,5,10,8,14,9,11,10,11,
3195,30195,33.675110,-117.926900,414076.8421,3726519.161,4,6,9,9,12,7,14,14,13,
4137,33137,33.850000,-116.543056,542271.9862,3745618.372,4,6,8,7,9,-1000,-1000,-
4144,33144,34.010278,-117.426389,460628.5061,3763377.527,6,6,9,10,12,14,16,15,1
4149,33149,33.708333,-117.243056,477477.8562,3729843.797,5,6,11,11,16,12,17,19,1
4157,33157,33.714555,-116.233894,570984.8108,3730770.56,5,6,9,10,14,12,14,13,12,
4158,33158,33.674010,-117.321140,470230.5502,3726058.069,6,6,9,9,11,-1000,9,10,
4162,33162,33.975620,-117.332630,469273.5024,3759502.621,5,6,-1000,-1000,-1000,-
5175,36175,34.103611,-117.629167,441967.9933,3773822.912,5,6,3,12,16,14,18,20,1
5181,36181,34.243889,-117.273611,474805.0231,3789232.296,5,6,10,9,16,12,19,14,1
5197,36197,34.099444,-117.504167,453495.3706,3773296.931,6,6,11,10,15,12,21,16,1
5203,36203,34.107222,-117.273611,474764.3726,3774078.432,5,5,10,9,17,13,18,14,1
5204,36204,34.066667,-117.151389,486030.4791,3769558.26,5,6,11,11,16,14,20,16,1
5212,36512,33.984600,-117.512733,452641.4369,3760566.908,5,6,12,12,16,16,21,18,1
5213,36513,34.262276,-117.188543,482642.1459,3791253.321,5,6,9,12,16,12,18,18,1
591,70591,34.143889,-117.851389,421507.8104,3778437.629,10,6,7,7,16,16,21,17,17,
```

Figure 4.1: Los Angeles ozone data set text input file with location coordinates.

format file, there are no further requirements.

When the input is a text file, the three required variables must be entered in a separate row for each observation, and separated by a comma. The input file must also contain two header lines. The first includes the number of observations and the number of variables, the second a list of the variable names. Again, all items are separated by a comma.

In addition to the identifier and coordinates, the input file can also contain other variables.[2] The text input file format is illustrated in Figure 4.1, which shows the partial contents of the OZ9799 sample data set in the text file `oz9799.txt`. This file includes monthly measures on ozone pollution taken at 30 monitoring stations in the Los Angeles basin. The first line gives the number of observations (30) and the number of variables (2 identifiers, 4 coordinates and 72 monthly measures over a three year period). The

[2] This is in contrast to the input files used to create polygon shape files in Exercise 5, where a two-step procedure is needed.

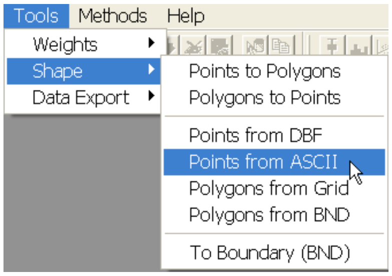
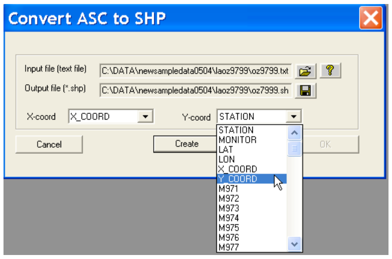


Figure 4.2: Creating a point shape file from ascii text input.



Input file (text file) C:\DATA\newsampledata0504\laoz9799\oz9799.txt

Output file (*.shp) C:\DATA\newsampledata0504\laoz9799\oz7999.sh

X-coord X_COORD

Y-coord STATION

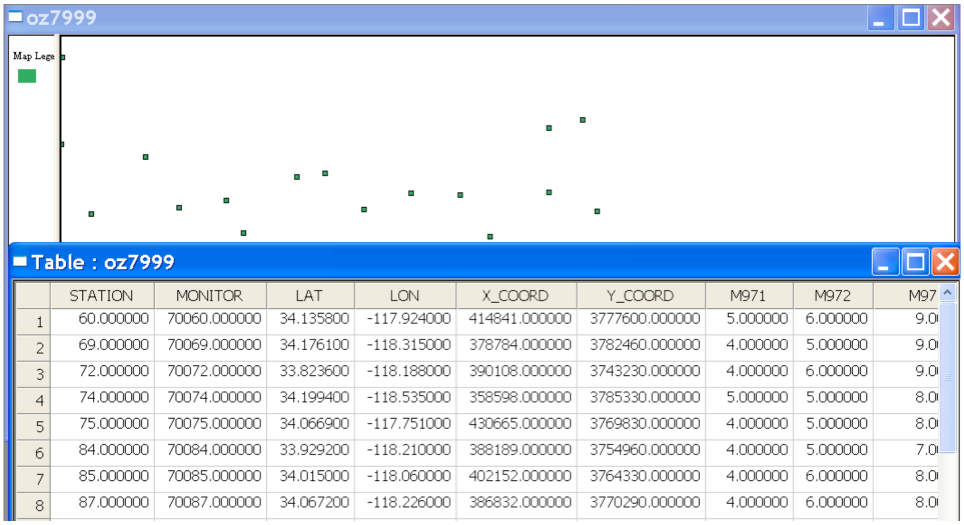
Figure 4.3: Selecting the x and y coordinates for a point shape file.

second line includes all the variable names, separated by a comma. Note that both the unprojected latitude and longitude are included as well as the projected `x, y` coordinates (UTM zone 11).

4.3 Converting Text Input to a Point Shape File

The creation of point shape files from text input is invoked from the `Tools` menu, by selecting `Shape > Points from ASCII`, as in Figure 4.2. When the input is in the form of a dbf file, the matching command is `Shape > Points from DBF`. This generates a dialog in which the path for the input text file must be specified as well as a file name for the new shape file. Enter `oz9799.txt` for the former and `oz9799` for the latter (the `shp` file extension will be added by the program). Next, the `X-coord` and `Y-coord` must be set, as illustrated in Figure 4.3 for the UTM projected coordinates in the `oz9799.txt` text file. Use either these same values, or, alternatively, select `LON` and `LAT`. Clicking on the `Create` button will generate the shape file. Finally, pressing `OK` will return to the main interface.

Check the contents of the newly created shape file by opening a new project (`File > Open Project`) and selecting the `oz7999.shp` file. The point map and associated data table will be as shown in Figure 4.4. Note that, in contrast to the ESRI point shape file standard, the coordinates for the points are included explicitly in the data table.



	STATION	MONITOR	LAT	LON	X_COORD	Y_COORD	M971	M972	M97
1	60.000000	70060.000000	34.135800	-117.924000	414841.000000	3777600.000000	5.000000	6.000000	9.0
2	69.000000	70069.000000	34.176100	-118.315000	378784.000000	3782460.000000	4.000000	5.000000	9.0
3	72.000000	70072.000000	33.823600	-118.188000	390108.000000	3743230.000000	4.000000	6.000000	9.0
4	74.000000	70074.000000	34.199400	-118.535000	358598.000000	3785330.000000	5.000000	5.000000	8.0
5	75.000000	70075.000000	34.066900	-117.751000	430665.000000	3769830.000000	4.000000	5.000000	8.0
6	84.000000	70084.000000	33.929200	-118.210000	388189.000000	3754960.000000	4.000000	5.000000	7.0
7	85.000000	70085.000000	34.015000	-118.060000	402152.000000	3764330.000000	4.000000	6.000000	8.0
8	87.000000	70087.000000	34.067200	-118.226000	386832.000000	3770290.000000	4.000000	6.000000	8.0

Figure 4.4: OZ9799 point shape file base map and data table.

4.4 Practice

The sample file BOSTON contains the classic Harrison and Rubinfeld (1978) housing data set with observations on 23 variables for 506 census tracts. The original data have been augmented with location coordinates for the tract centroids, both in unprojected latitude and longitude as well as in projected x, y (UTM zone 19). Use the `boston.txt` file to create a point shape file for the housing data. You can also experiment with the dbf files for some other point shape files in the sample data sets, such as BALTIMORE, JUVENILE and PITTSBURGH.

Exercise 5

Creating a Polygon Shape File

5.1 Objectives

This exercise illustrates how you can create a polygon shape file from text input for irregular lattices, or directly for regular grid shapes in situations where you do not have a proper ESRI formatted shape file. As in Exercise 4, this functionality can be accessed without opening a project. It is available from the `Tools` menu.

At the end of the exercise, you should know how to:

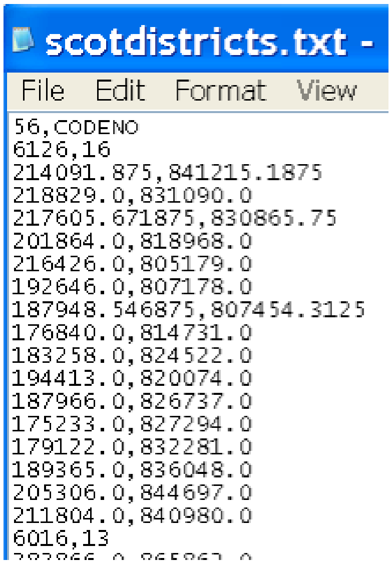
- create a polygon shape file from a text input file with the boundary coordinates
- create a polygon shape file for a regular grid layout
- join a data table to a shape file base map

More detailed information on these operations can be found in the *Release Notes*, pp. 13–17, and the *User's Guide*, pp. 63–64.

5.2 Boundary File Input Format

GeoDa currently supports one input file format for polygon boundary coordinates. While this is a limitation, in practice it is typically fairly straightforward to convert one format to another. The supported format, illustrated in Figure 5.1 on p. 27, consists of a header line containing the number of polygons and a unique polygon identifier, separated by a comma. For each polygon, its identifier and the number of points is listed, followed by the `x` and `y` coordinate pairs for each point (comma separated). This format is referred to as `1a` in the *User's Guide*. Note that it currently does *not* support multiple polygons associated with the same observation. Also, the first coordinate pair is *not* repeated as the last. The count of point coordinates for each polygon reflects this (there are 16 x, y pairs for the first polygon in Figure 5.1).

The boundary file in Figure 5.1 pertains to the classic Scottish lip cancer data used as an example in many texts (see, e.g., Cressie 1993, p. 537). The coordinates for the 56 districts were taken from the `scotland.map` boundaries included with the *WinBugs* software package, and exported to the S-Plus map format. The resulting file was then edited to conform to the *GeoDa* input format. In addition, duplicate coordinates were eliminated and sliver polygons taken out. The result is contained in the `scotdistricts.txt` file. Note that to avoid problems with multiple polygons, the island districts were simplified to a single polygon.



```
56,CODENO
6126,16
214091.875,841215.1875
218829.0,831090.0
217605.671875,830865.75
201864.0,818968.0
216426.0,805179.0
192646.0,807178.0
187948.546875,807454.3125
176840.0,814731.0
183258.0,824522.0
194413.0,820074.0
187966.0,826737.0
175233.0,827294.0
179122.0,832281.0
189365.0,836048.0
205306.0,844697.0
211804.0,840980.0
6016,13
```

Figure 5.1: Input file with Scottish districts boundary coordinates.

In contrast to the procedure followed for point shape files in Exercise 4, a two-step approach is taken here. First, a base map shape file is created (see Section 5.3). This file does not contain any data other than polygon identifiers, area and perimeter. In the second step, a data table must be joined to this shape file to add the variables of interest (see Section 5.4).

5.3 Creating a Polygon Shape File for the Base Map

The creation of the base map is invoked from the `Tools` menu, by selecting `Shape > Polygons from BND`, as illustrated in Figure 5.2. This generates the dialog shown in Figure 5.3, where the path of the input file and the name for the new shape file must be specified. Select `scotdistricts.txt` for the former and enter `scotdistricts` as the name for the base map shape file. Next, click `Create` to start the procedure. When the blue progress bar (see Figure 5.3) shows completion of the conversion, click on `OK` to return to the main menu.

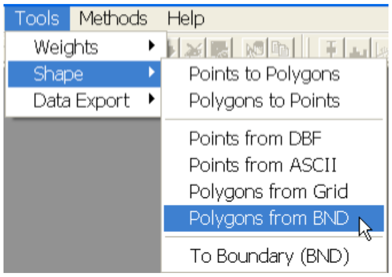
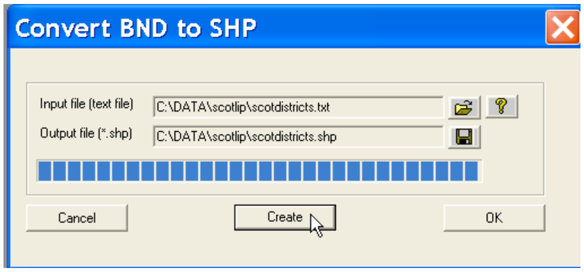


Figure 5.2: Creating a polygon shape file from ascii text input.


Convert BND to SHP

Input file (text file) C:\DATA\scotlip\scotdistricts.txt

Output file (*.shp) C:\DATA\scotlip\scotdistricts.shp

Figure 5.3: Specifying the Scottish districts input and output files.

The resulting base map is as in Figure 5.4 on p.29, which is created by means of the usual `Open project` toolbar button, followed by entering the file name and `CODENO` as the `Key` variable. Next, click on the `Table` toolbar button to open the corresponding data table. As shown in Figure 5.5 on p.29, this only contains identifiers and some geometric information, but no other useful data.

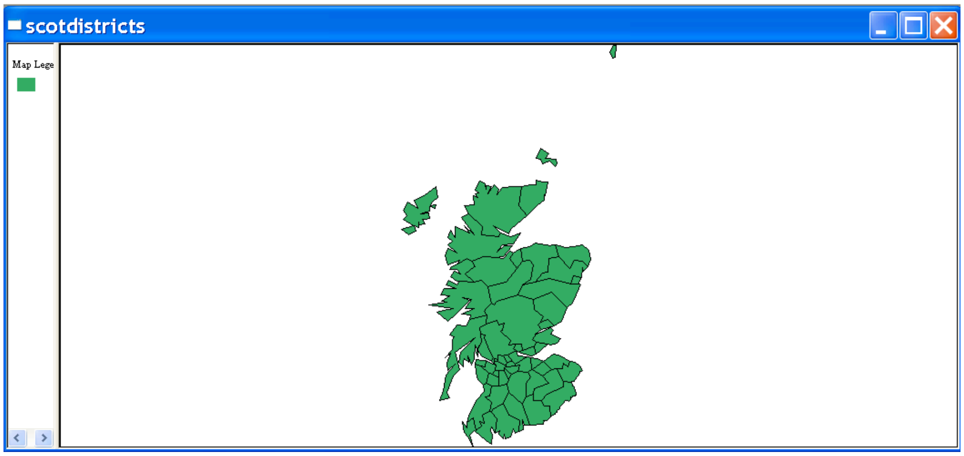
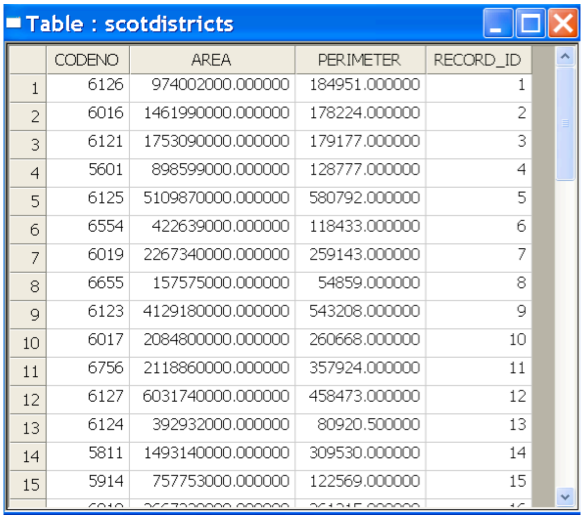


Figure 5.4: Scottish districts base map.



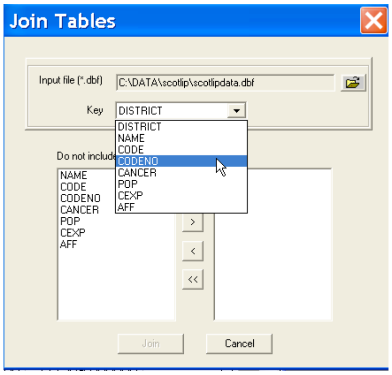
	CODENO	AREA	PERIMETER	RECORD_ID
1	6126	974002000.000000	184951.000000	1
2	6016	1461990000.000000	178224.000000	2
3	6121	1753090000.000000	179177.000000	3
4	5601	898599000.000000	128777.000000	4
5	6125	5109870000.000000	580792.000000	5
6	6554	422639000.000000	118433.000000	6
7	6019	2267340000.000000	259143.000000	7
8	6655	157575000.000000	54859.000000	8
9	6123	4129180000.000000	543208.000000	9
10	6017	2084800000.000000	260668.000000	10
11	6756	2118860000.000000	357924.000000	11
12	6127	6031740000.000000	458473.000000	12
13	6124	392932000.000000	80920.500000	13
14	5811	1493140000.000000	309530.000000	14
15	5914	757753000.000000	122569.000000	15

Figure 5.5: Scottish districts base map data table.

5.4 Joining a Data Table to the Base Map

In order to create a shape file for the Scottish districts that also contains the lip cancer data, a data table (dbf format) must be joined to the table for the base map. This is invoked using the `Table` menu with the `Join Tables` command (or by right clicking in the table and selecting `Join Tables` from the drop down menu, as in Figure 3.2 on p. 14).

This brings up a `Join Tables` dialog, as in Figure 5.6. Enter the file name for the input file as `scotlipdata.dbf`, and select `CODENO` for the `Key` variable, as shown in the Figure. Next, move all variables from the left hand side column over to the right hand side, by clicking on the `>>` button, as shown in Figure 5.7. Finally, click on the `Join` button to finish the operation. The resulting data table is as shown in Figure 5.8.

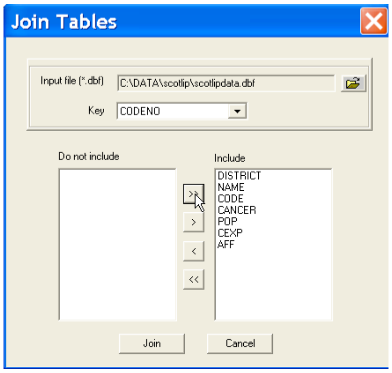


Input file (*.dbf): C:\DATA\scotlip\scotlipdata.dbf

Key: DISTRICT

Do not include: NAME, CODE, CODENO, CANCER, POP, CEXP, AFF

Figure 5.6: Specify join data table and key variable.



Input file (*.dbf): C:\DATA\scotlip\scotlipdata.dbf

Key: CODENO

Include: DISTRICT, NAME, CODE, CANCER, POP, CEXP, AFF

Figure 5.7: Join table variable selection.

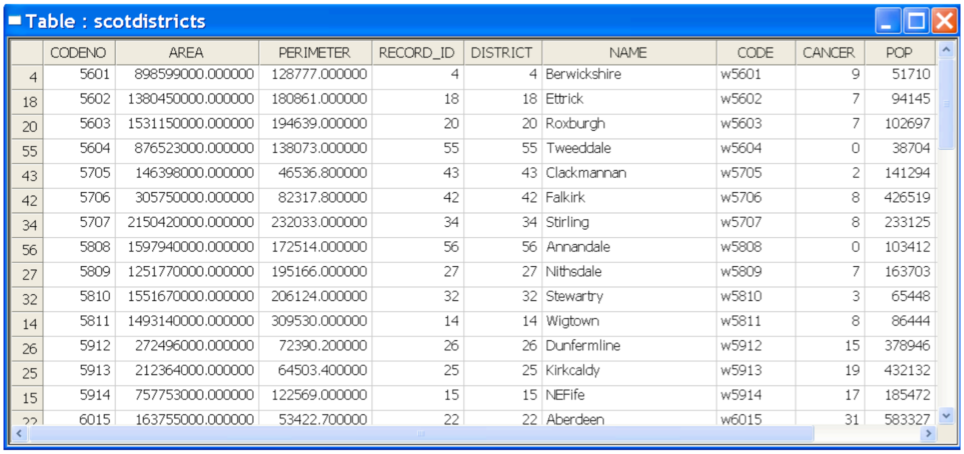
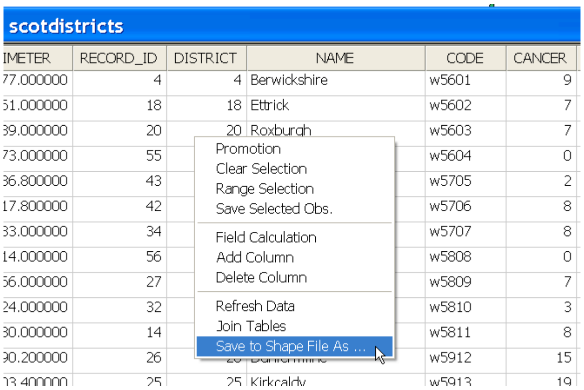


Table : scotdistricts

	CODENO	AREA	PERIMETER	RECORD_ID	DISTRICT	NAME	CODE	CANCER	POP
4	5601	898599000.000000	128777.000000	4	4	Berwickshire	w5601	9	51710
18	5602	1380450000.000000	180861.000000	18	18	Ettrick	w5602	7	94145
20	5603	1531150000.000000	194639.000000	20	20	Roxburgh	w5603	7	102697
55	5604	876523000.000000	138073.000000	55	55	Tweeddale	w5604	0	38704
43	5705	146398000.000000	46536.800000	43	43	Clackmannan	w5705	2	141294
42	5706	305750000.000000	82317.800000	42	42	Falkirk	w5706	8	426519
34	5707	2150420000.000000	232033.000000	34	34	Stirling	w5707	8	233125
56	5808	1597940000.000000	172514.000000	56	56	Annandale	w5808	0	103412
27	5809	1251770000.000000	195166.000000	27	27	Nithsdale	w5809	7	163703
32	5810	1551670000.000000	206124.000000	32	32	Stewartry	w5810	3	65448
14	5811	1493140000.000000	309530.000000	14	14	Wigtown	w5811	8	86444
26	5912	272496000.000000	72390.200000	26	26	Dunfermline	w5912	15	378946
25	5913	212364000.000000	64503.400000	25	25	Kirkcaldy	w5913	19	432132
15	5914	757753000.000000	122569.000000	15	15	NEFife	w5914	17	185472
22	6015	163755000.000000	53422.700000	22	22	Aberdeen	w6015	31	583327

Figure 5.8: Scottish lip cancer data base joined to base map.

At this point, all the variables contained in the table shown in Figure 5.8 are available for mapping and analysis. In order to make them permanent, however, the table (and shape file) must be saved to a file with a new name, as outlined in Section 3.4 on p. 19. This is carried out by using the `Save to Shape File As ...` function from the `Table` menu, or by right clicking in the table, as in Figure 5.9. Select this command and enter a new file name (e.g., scotdistricts) for the output shape file, followed by `OK`. Clear the project and load the new shape file to check that its contents are as expected.



scotdistricts

IMETER	RECORD_ID	DISTRICT	NAME	CODE	CANCER
77.000000	4	4	Berwickshire	w5601	9
51.000000	18	18	Ettrick	w5602	7
39.000000	20	20	Roxburgh	w5603	7
73.000000	55			w5604	0
36.800000	43			w5705	2
17.800000	42			w5706	8
33.000000	34			w5707	8
14.000000	56			w5808	0
56.000000	27			w5809	7
24.000000	32			w5810	3
30.000000	14			w5811	8
90.200000	26			w5912	15
03.400000	25	25	Kirkcaldy	w5913	19

Figure 5.9: Saving the joined Scottish lip cancer data to a new shape file.

5.5 Creating a Regular Grid Polygon Shape File

GeoDa contains functionality to create a polygon shape file for a regular grid (or lattice) layout without having to specify the actual coordinates of the boundaries. This is invoked from the `Tools` menu, using the `Shape > Polygons from Grid` function, as shown in Figure 5.10 on p. 32.

This starts up a dialog that offers many different options to specify the layout for the grid, illustrated in Figure 5.11 on p. 32. We will only focus on the simplest here (see the *Release Notes* for more details).

As shown in Figure 5.11, click on the radio button next to `Specify manually`, leave the `Lower-left corner` coordinates to the default setting of `0.0, 0.0`, and set the `Upper-right corner` coordinates to `49, 49`. In the text boxes for `Grid Size`, enter `7` for both the number of rows and the number of columns. Finally, make sure to specify a file name for the shape file, such as `grid77` (see Figure 5.11). Click on the `Create` button to proceed and `OK` to return to the main menu.

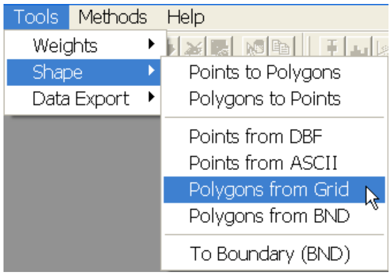
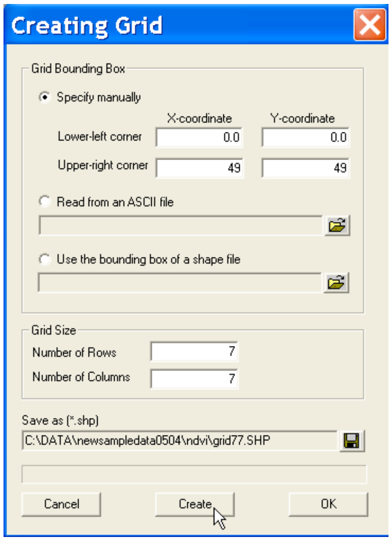


Figure 5.10: Creating a polygon shape file for a regular grid.


Creating Grid

Grid Bounding Box

● Specify manually

	X-coordinate	Y-coordinate
Lower-left corner	0.0	0.0
Upper-right corner	49	49

○ Read from an ASCII file

○ Use the bounding box of a shape file

Grid Size

Number of Rows	7
Number of Columns	7

Save as (*.shp)

C:\DATA\newsampledata0504\ndvi\grid77.SHP

Figure 5.11: Specifying the dimensions for a regular grid.

Check the resulting grid file with the usual `Open project` toolbar button and use `PolyID` as the `Key`. The shape will appear as in Figure 5.12 on p. 33. Use the `Table` toolbar button to open the associated data table. Note how it only contains the `POLYID` identifier and two geometric characteristics, as shown in Figure 5.13 on p. 33.

As in Section 5.4, you will need to join this table with an actual data table to get a meaningful project. Select the `Join Tables` function and specify the `ndvi.dbf` file as the `Input File`. This file contains four variables measured for a 7 by 7 square raster grids with 10 arcminute spacing from

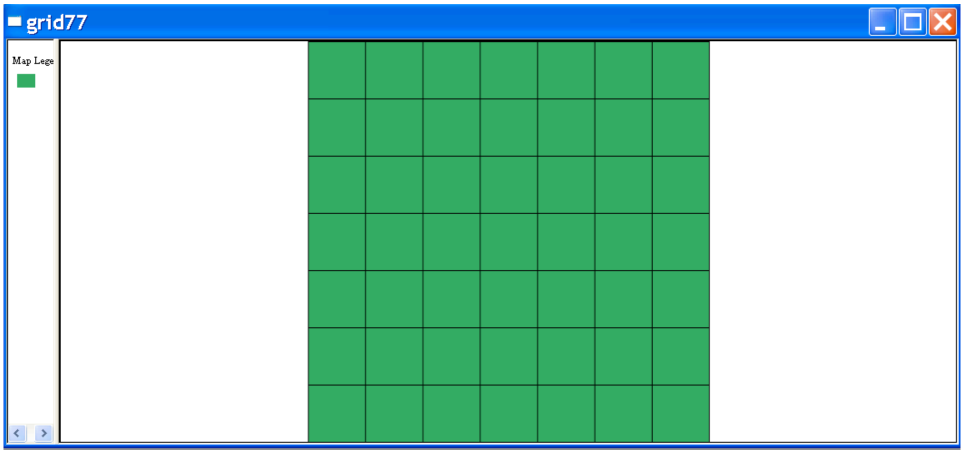
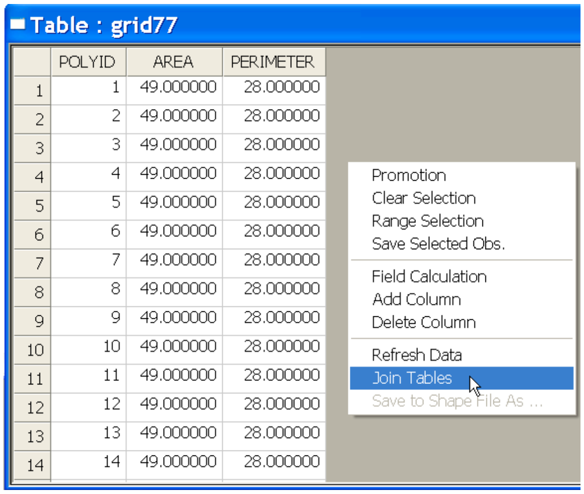


Figure 5.12: Regular square 7 by 7 grid base map.



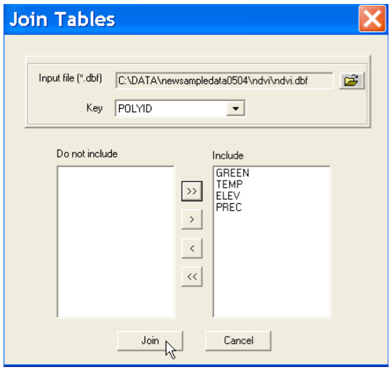
	POLYID	AREA	PERIMETER
1	1	49.000000	28.000000
2	2	49.000000	28.000000
3	3	49.000000	28.000000
4	4	49.000000	28.000000
5	5	49.000000	28.000000
6	6	49.000000	28.000000
7	7	49.000000	28.000000
8	8	49.000000	28.000000
9	9	49.000000	28.000000
10	10	49.000000	28.000000
11	11	49.000000	28.000000
12	12	49.000000	28.000000
13	13	49.000000	28.000000
14	14	49.000000	28.000000

Figure 5.13: Joining the NDVI data table to the grid base map.

a global change database. It was used as an illustration in Anselin (1993). The 49 observations match the layout for the regular grid just created.

In addition to the file name, select `POLYID` as the `Key` and move all four variables over to the right-hand side column, as in Figure 5.14 on p. 34. Finally, click on the `Join` button to execute the join. The new data table includes the four new variables, as in Figure 5.15 on p. 34. Complete the procedure by saving the shape file under a new file name, e.g., ndvigrid.

After clearing the screen, bring up the new shape file and check its contents.



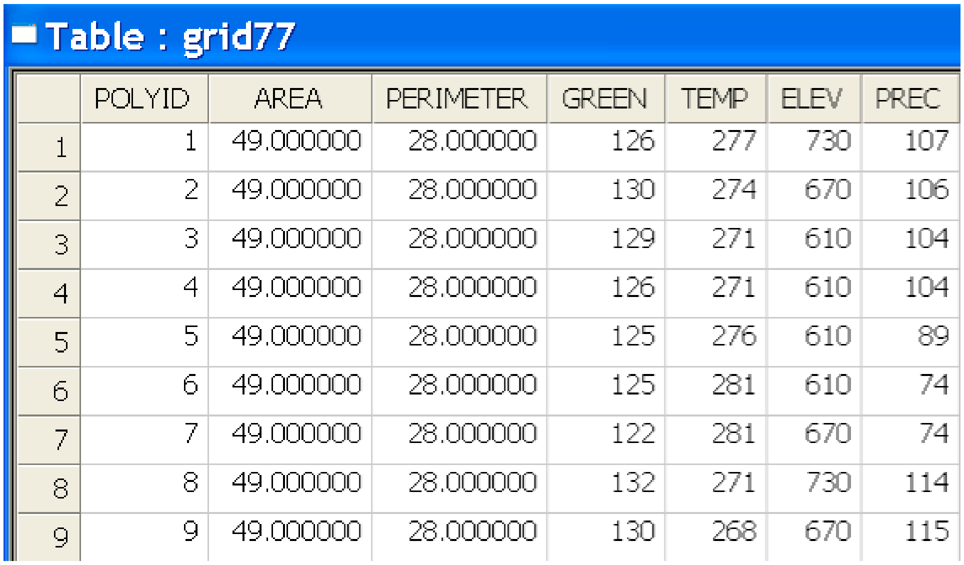
Input file (*.dbf) C:\DATA\newsampledata0504\ndvi\ndvi.dbf

Key POLYID

Do not include

Include
GREEN
TEMP
ELEV
PREC

Figure 5.14: Specifying the NDVI variables to be joined.



	POLYID	AREA	PERIMETER	GREEN	TEMP	ELEV	PREC
1	1	49.000000	28.000000	126	277	730	107
2	2	49.000000	28.000000	130	274	670	106
3	3	49.000000	28.000000	129	271	610	104
4	4	49.000000	28.000000	126	271	610	104
5	5	49.000000	28.000000	125	276	610	89
6	6	49.000000	28.000000	125	281	610	74
7	7	49.000000	28.000000	122	281	670	74
8	8	49.000000	28.000000	132	271	730	114
9	9	49.000000	28.000000	130	268	670	115

Figure 5.15: NDVI data base joined to regular grid base map.

5.6 Practice

The sample data sets include several files that can be used to practice the operations covered in this chapter. The OHIOLUNG data set includes the text file ohioutmbnd.txt with the boundary point coordinates for the 88 Ohio counties projected using UTM zone 17. Use this file to create a polygon shape file. Next, join this file with the "classic" Ohio lung cancer mortality data (Xia and Carlin 1998), contained in the ohdat.dbf file. Use FIPSNO as the Key, and create a shape file that includes all the variables.[1]

Alternatively, you can apply the Tools > Shape > To Boundary (BND) function to any polygon shape file to create a text version of the boundary coordinates in 1a format. This can then be used to recreate the original polygon shape file in conjunction with the dbf file for that file.

The GRID100 sample data set includes the file grid10x10.dbf which contains simulated spatially correlated random variables on a regular square 10 by 10 lattice. Create such a lattice and join it to the data file (the Key is POLYID). Save the result as a new shape file that you can use to map different patterns of variables that follow a spatial autoregressive or spatial moving average process.[2]

Alternatively, experiment by creating grid data sets that match the bounding box for one of the sample data sets. For example, use the COLUMBUS map to create a 7 by 7 grid with the Columbus data, or use the SIDS map to create a 5 by 20 grid with the North Carolina Sids data. Try out the different options offered in the dialog shown in Figure 5.11 on p. 32.

[1] The ohlung.shp/shx/dbf files contain the result.

[2] The grid100s.shp/shx/dbf files contain the result.

Exercise 6

Spatial Data Manipulation

6.1 Objectives

This exercise illustrates how you can change the representation of spatial observations between points and polygons by computing polygon centroids, and by applying a Thiessen polygon tessellation to points.[1] As in Exercises 4 and 5, this functionality can be accessed without opening a project. It is available from the `Tools` menu. Note that the computations behind these operations are only valid for properly *projected* coordinates, since they operate in a Euclidean plane. While they will work on lat-lon coordinates (*GeoDa* has no way of telling whether or not the coordinates are projected), the results will only be approximate and should not be relied upon for precise analysis.

At the end of the exercise, you should know how to:

- create a point shape file containing the polygon centroids
- add the polygon centroids to the current data table
- create a polygon shape file containing Thiessen polygons

More detailed information on these operations can be found in the *User's Guide*, pp. 19–28, and the *Release Notes*, pp. 20–21.

[1] More precisely, what is referred to in *GeoDa* as centroids are *central points*, or the average of the x and y coordinates in the polygon boundary.

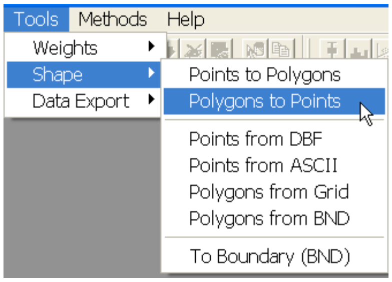
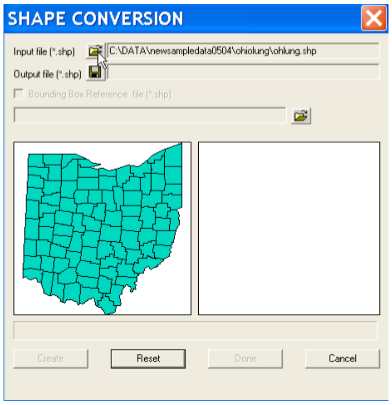


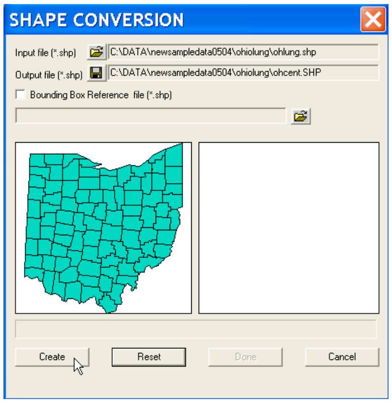
Figure 6.1: Creating a point shape file containing polygon centroids.



Input file (*.shp) C:\DATA\newsampledata0504\ohiolung\ohlung.shp

Output file (*.shp)

Figure 6.2: Specify the polygon input file.



Input file (*.shp) C:\DATA\newsampledata0504\ohiolung\ohlung.shp

Output file (*.shp) C:\DATA\newsampledata0504\ohiolung\ohcent.SHP

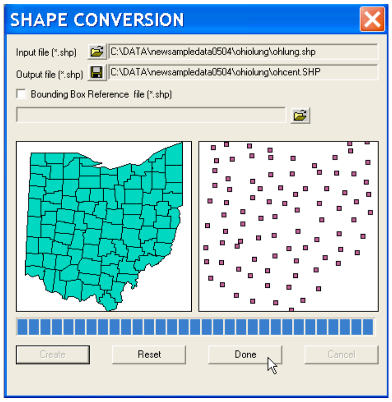
☐ Bounding Box Reference file (*.shp)

Figure 6.3: Specify the point output file.

6.2 Creating a Point Shape File Containing Centroid Coordinates

Centroid coordinates can be converted to a point shape file without having a *GeoDa* project open. From the `Tools` menu, select `Shape > Polygons to Points` (Figure 6.1) to open the `Shape Conversion` dialog. First, specify the filename for the polygon input file, e.g., `ohlung.shp` in Figure 6.2 (open the familiar file dialog by clicking on the file open icon). Once the file name is entered, a thumbnail outline of the 88 Ohio counties appears in the left hand pane of the dialog (Figure 6.2). Next, enter the name for the new shape file, e.g., `ohcent` in Figure 6.3 and click on the `Create` button.

After the new file is created, its outline will appear in the right hand pane of the dialog, as in Figure 6.4 on p. 38. Click on the `Done` button to



Input file (*.shp) C:\DATA\newsampledata0504\ohiolung\ohlung.shp

Output file (*.shp) C:\DATA\newsampledata0504\ohiolung\ohcent.SHP

□ Bounding Box Reference file (*.shp)

Figure 6.4: Centroid shape file created.

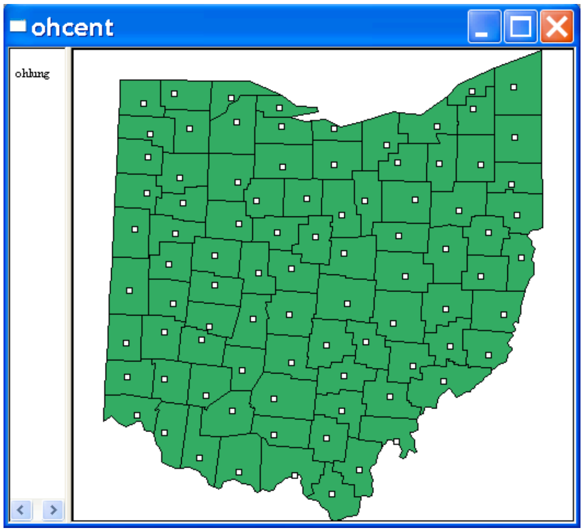


Figure 6.5: Centroid point shape file overlaid on original Ohio counties.

return to the main interface.

To check the new shape file, first open a project with the original Ohio counties (`ohlung.shp` using `FIPSNO` as the `Key`). Change the `Map` color to white (see the dialog in Figure 1.4 on p. 4). Next add a new layer (click on the `Add a layer` toolbar button, or use `Edit > Add Layer` from the menu) with the centroid shape file (`ohcent`, using `FIPSNO` as the `Key`). The original polygons with the centroids superimposed will appear as in Figure 6.5 on p. 38. The white map background of the polygons has been transferred to the "points." As the *top* layer, it receives all the properties specified for the map.

Check the contents of the data table. It is identical to the original shape file, except that the centroid coordinates have been added as variables.

6.2.1 Adding Centroid Coordinates to the Data Table

The coordinates of polygon centroids can be added to the data table of a polygon shape file without explicitly creating a new file. This is useful when you want to use these coordinates in a statistical analysis (e.g., in a trend surface regression, see Section 23.3 on p. 183).

This feature is implemented as one of the map options, invoked either from the `Options` menu (with a map as the active window), or by right clicking on the map and selecting `Add Centroids to Table`, as illustrated in Figure 6.6. Alternatively, there is a toolbar button that accomplishes the same function.

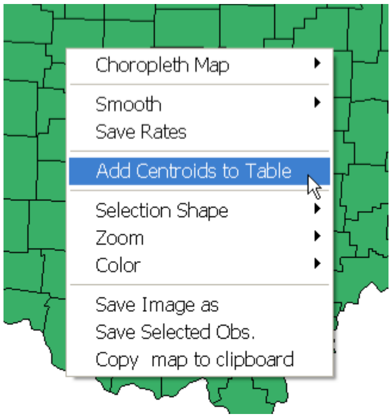
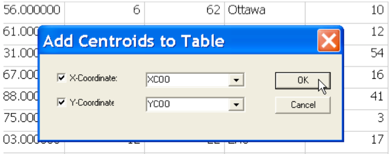


Figure 6.6: Add centroids from current polygon shape to data table.

Load (or reload) the Ohio Lung cancer data set (`ohlung.shp` with `FIPSNO` as the `Key`) and select the `Add Centroids to Table` option. This opens a dialog to specify the variable names for the x and y coordinates, as in Figure 6.7 on p. 40. Note that you don't need to specify both coordinates, one coordinate may be selected as well. Keep the defaults of `XCOO` and `YCOO` and click on `OK` to add the new variables. The new data table will appear as in

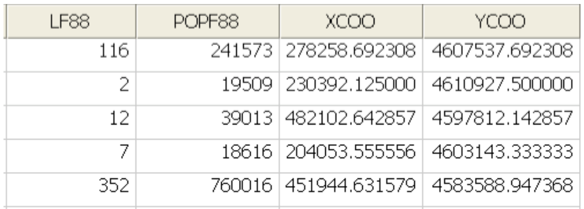


Add Centroids to Table

☑ X-Coordinate: XCOO

☑ Y-Coordinate: YCOO

Figure 6.7: Specify variable names for centroid coordinates.



LF88	POPF88	XCOO	YCOO
116	241573	278258.692308	4607537.692308
2	19509	230392.125000	4610927.500000
12	39013	482102.642857	4597812.142857
7	18616	204053.555556	4603143.333333
352	760016	451944.631579	4583588.947368

Figure 6.8: Ohio centroid coordinates added to data table.

Figure 6.8. As before, make sure to save this to a new shape file in order to make the variables a permanent addition.

6.3 Creating a Thiessen Polygon Shape File

Point shape files can be converted to polygons by means of a Thiessen polygon tessellation. The polygon representation is often useful for visualization of the spatial distribution of a variable, and allows the construction of spatial weights based on contiguity. This process is invoked from the `Tools` menu by selecting `Shape > Points to Polygons`, as in Figure 6.9.

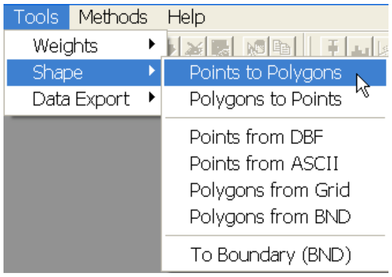
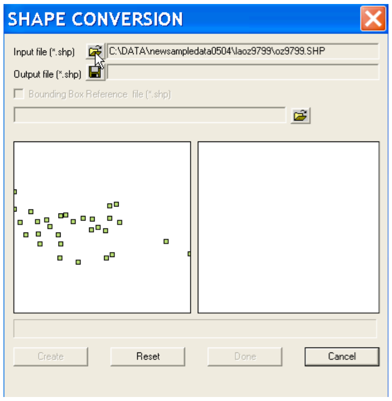


Figure 6.9: Creating a Thiessen polygon shape file from points.

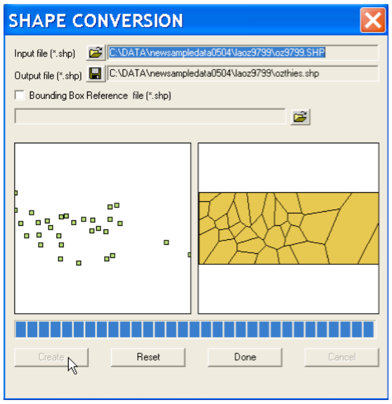


Input file (*.shp) C:\DATA\newsampledata0504\laoz9799\oz9799.SHP

Output file (*.shp)

□ Bounding Box Reference file (*.shp)

Figure 6.10: Specify the point input file.



Input file (*.shp) C:\DATA\newsampledata0504\laoz9799\oz9799.SHP

Output file (*.shp) C:\DATA\newsampledata0504\laoz9799\ozthies.shp

□ Bounding Box Reference file (*.shp)

Figure 6.11: Specify the Thiessen polygon output file.

This opens up a dialog, as shown in Figure 6.10. Specify the name of the input (point) shape file as `oz9799.shp`, the sample data set with the locations of 30 Los Angeles basin air quality monitors. As for the polygon to point conversion, specifying the input file name yields a thumbnail outline of the point map in the left hand panel of the dialog. Next, enter the name for the new (polygon) shape file, say `ozthies.shp`.

Click on `Create` and see an outline of the Thiessen polygons appear in the right hand panel (Figure 6.11). Finally, select `Done` to return to the standard interface.

Compare the layout of the Thiessen polygons to the original point pattern in the same way as for the centroids in Section 6.2. First, open the Thiessen polygon file (`ozthies` with `Station` as the `Key`). Change its map color to white. Next, add a layer with the original points (`oz9799` with `Station` as the `Key`). The result should be as in Figure 6.12 on p. 42. Check the contents of the data table. It is identical to that of the point coverage, with the addition of `Area` and `Perimeter`.

Note that the default for the Thiessen polygons is to use the bounding box of the original points as the bounding box for the polygons. If you take a close look at Figure 6.12, you will notice the white points on the edge of the rectangle. Other bounding boxes may be selected as well. For example, one can use the bounding box of an existing shape file. See the *Release Notes*, pp. 20–21.

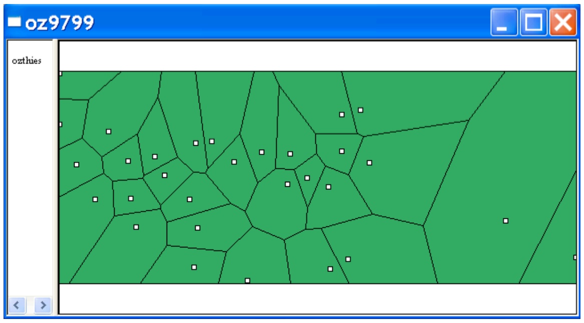


Figure 6.12: Thiessen polygons for Los Angeles basin monitors.

6.4 Practice

Use the SCOTLIP data set to create a point shape file with the centroids of the 56 Scottish districts. Use the points to generate a Thiessen polygon shape file and compare to the original layout. You can experiment with other sample data sets as well (but remember, the results for the centroids and Thiessen polygons are unreliable for unprojected lat-lon coordinates).

Alternatively, start with a point shape file, such as the 506 census tract centroids in the BOSTON data set (`Key` is `ID`) or the 211 house locations in the BALTIMORE sample data set (`Key` is `STATION`). These are both in projected coordinates. Turn them into a polygon coverage. Use the polygons to create a simple choropleth map for respectively the median house value (`MEDV`) or the house price (`PRICE`). Compare this to a choropleth map using the original points.

Exercise 7

EDA Basics, Linking

7.1 Objectives

This exercise illustrates some basic techniques for exploratory data analysis, or EDA. It covers the visualization of the non-spatial distribution of data by means of a histogram and box plot, and highlights the notion of *linking*, which is fundamental in *GeoDa*.

At the end of the exercise, you should know how to:

- create a histogram for a variable
- change the number of categories depicted in the histogram
- create a regional histogram
- create a box plot for a variable
- change the criterion to determine outliers in a box plot
- link observations in a histogram, box plot and map

More detailed information on these operations can be found in the *User's Guide*, pp. 65–67, and the *Release Notes*, pp. 43–44.

7.2 Linking Histograms

We start the illustration of traditional EDA with the visualization of the non-spatial distribution of a variable as summarized in a *histogram*. The histogram is a discrete approximation to the density function of a random

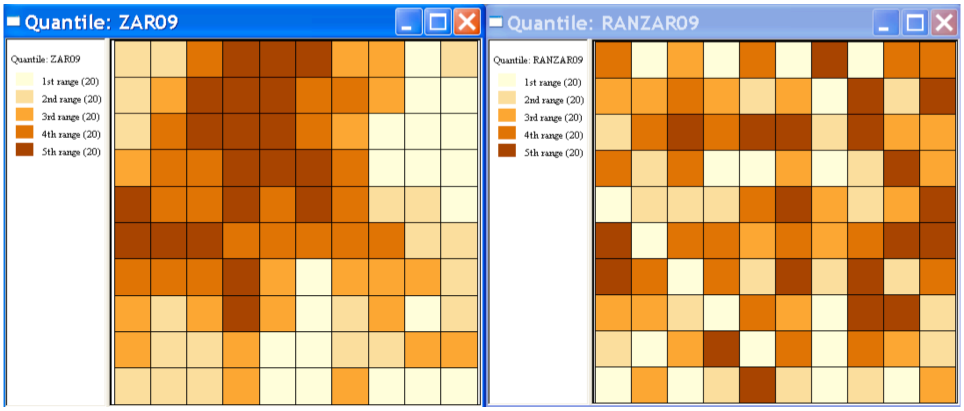


Figure 7.1: Quintile maps for spatial AR variables on 10 by 10 grid.

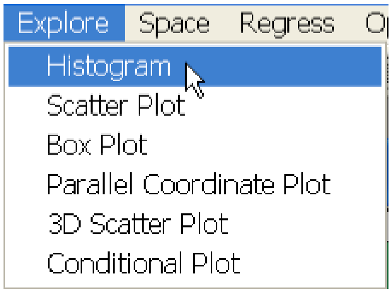


Figure 7.2: Histogram function.

variable and is useful to detect asymmetry, multiple modes and other peculiarities in the distribution.

Clear all windows and start a new project using the GRID100S sample data set (enter `grid100s` for the data set and `PolyID` as the `Key`). Start by constructing two quintile maps (`Map > Quantile` with 5 as the number of categories; for details, see Exercise 2), one for `zar09` and one for `ranzar09`.[1] The result should be as in Figure 7.1. Note the characteristic clustering associated with high positive spatial autocorrelation in the left-hand side panel, contrasted with the seeming random pattern on the right.

Invoke the histogram as `Explore > Histogram` from the menu (as in Figure 7.2) or by clicking the `Histogram` toolbar icon. In the `Variable`

[1] The first variable, `zar09`, depicts a spatial autoregressive process on a 10 by 10 square lattice with parameter 0.9. `ranzar09` is a randomly permuted set of the same values.

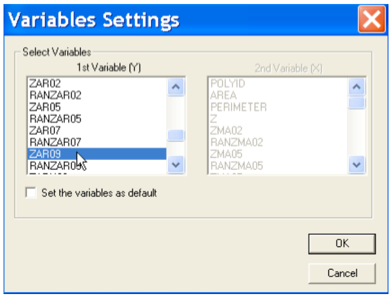


Figure 7.3: Variable selection for histogram.

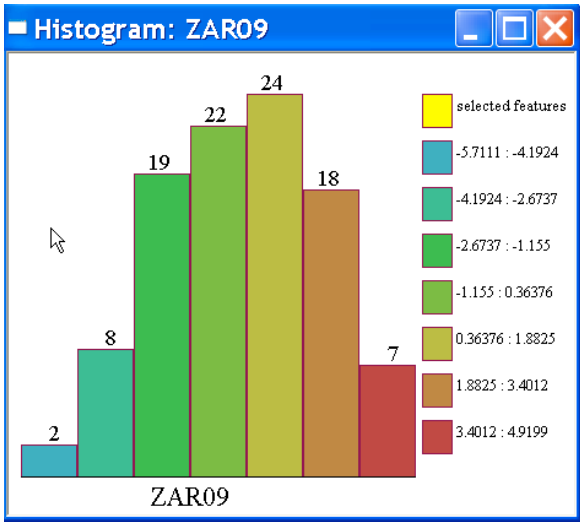


Figure 7.4: Histogram for spatial autoregressive random variate.

`Settings` dialog, select `zar09` as in Figure 7.3. The result is a histogram with the variables classified into 7 categories, as in Figure 7.4. This shows the familiar bell-like shape characteristic of a normally distributed random variable, with the values following a continuous color ramp. The figures on top of the histogram bars indicate the number of observations falling in each interval. The intervals themselves are shown on the right hand side.

Now, repeat the procedure for the variable `ranzar09`. Compare the result between the two histograms in Figure 7.5 on p. 46. Even though the maps in Figure 7.1 on p. 44 show strikingly different *spatial* patterns,

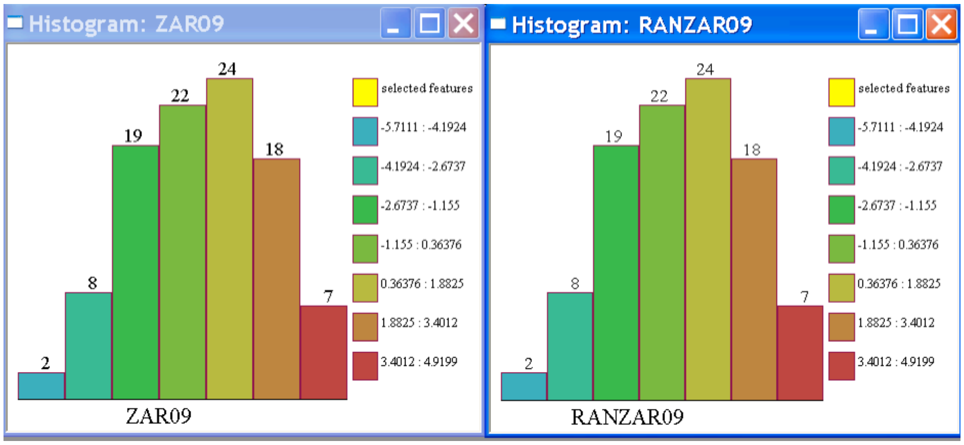


Figure 7.5: Histogram for SAR variate and its permuted version.

the histograms for the two variables are identical. You can verify this by comparing the number of observations in each category and the value ranges for the categories. In other words, the only aspect differentiating the two variables is *where* the values are located, not the non-spatial characteristics of the distribution.

This is further illustrated by *linking* the histograms and maps. Proceed by selecting (clicking on) the highest bar in the histogram of `zar09` and note how the distribution differs in the other histogram, as shown in Figure 7.6 on p. 47. The corrresponding observations are highlighted in the maps as well. This illustrates how the locations with the highest values for `zar09` are *not* the locations with the highest values for `ranzar09`.

Linking can be initiated in any window. For example, select a 5 by 5 square grid in the upper left map, as in Figure 7.7 on p. 48. The matching distribution in the two histograms is highlighted in yellow, showing a *regional* histogram. This depicts the distribution of a variable for a selected subset of locations on the map. Interest centers on the extent to which the regional distribution differs from the overall pattern, possibly suggesting the existence of *spatial heterogeneity*. One particular form is referred to as *spatial regimes*, which is the situation where subregions (regimes) show distinct distributions for a given variable (e.g., different means). For example, in the left-hand panel of Figure 7.7, the region selected yields values in the histogram (highlighted as yellow) concentrated in the upper half of the distribution. In contrast, in the panel on the right, the same selected locations yields values (the yellow subhistogram) that roughly follows the

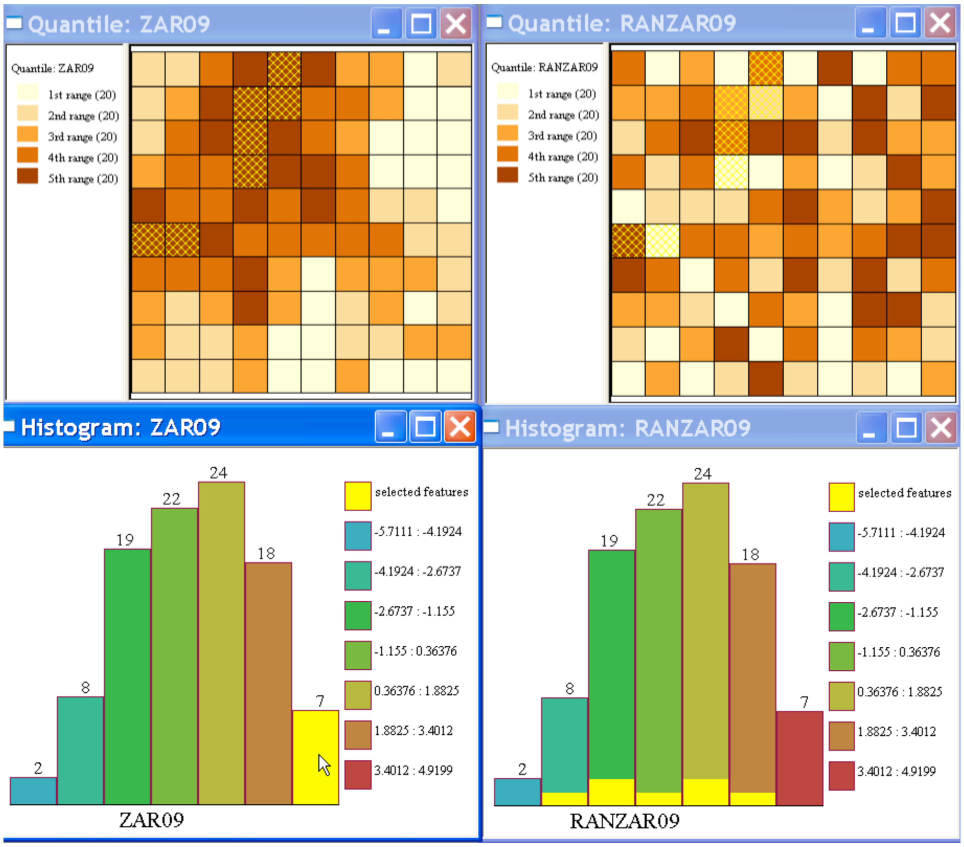


Figure 7.6: Linked histograms and maps (from histogram to map).

overall pattern. This would possibly suggest the presence of a spatial regime for `zar09`, but not for `ranzar09`.

The default number of categories of 7 can be changed by using `Option > Intervals` from the menu, or by right clicking on the histogram, as in Figure 7.8 on p. 48. Select this option and change the number of intervals to 12, as in Figure 7.9 on p. 48. Click on `OK` to obtain the histogram shown in Figure 7.10 on p. 49. The yellow part of the distribution still matches the subset selected on the map, and while it is now spread over more categories, it is still concentrated in the upper half of the distribution.

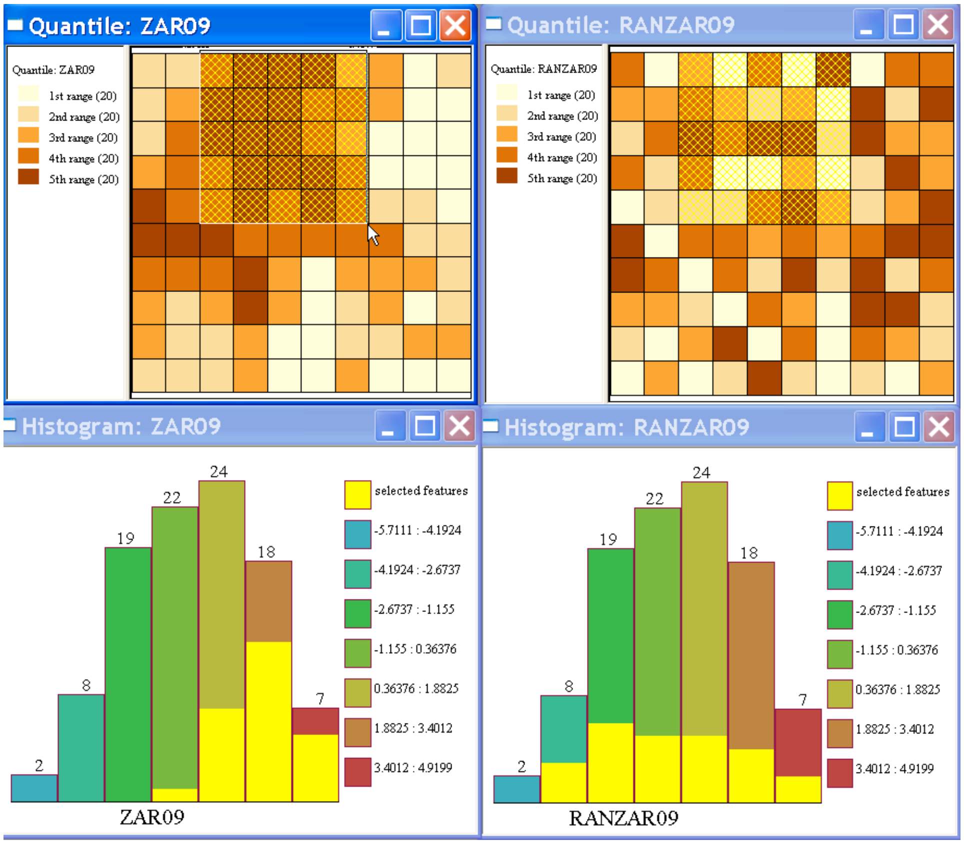


Figure 7.7: Linked histograms and maps (from map to histogram).

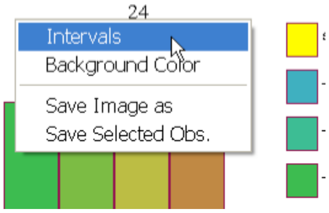
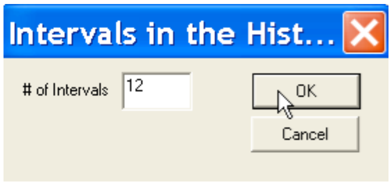


Figure 7.8: Changing the number of histogram categories.



# of Intervals 12

Figure 7.9: Setting the intervals to 12.

7.3 Linking Box Plots

The second basic EDA technique to depict the non-spatial distribution of a variable is the *box plot* (sometimes referred to as box and whisker plot). It

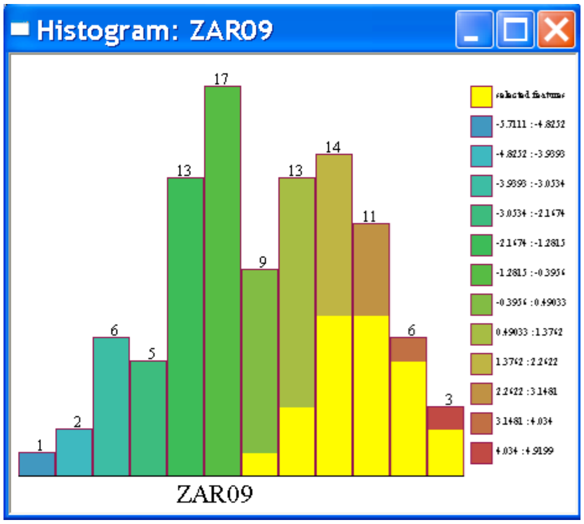


Figure 7.10: Histogram with 12 intervals.

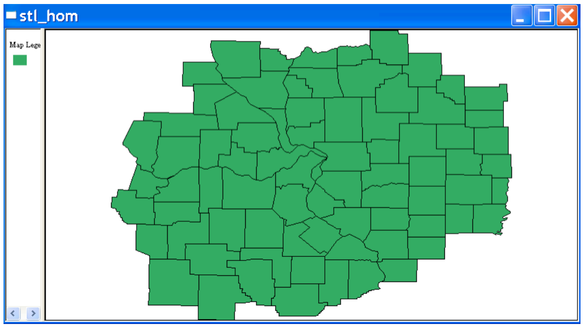


Figure 7.11: Base map for St. Louis homicide data set.

shows the median, first and third quartile of a distribution (the 50%, 25% and 75% points in the cumulative distribution) as well as a notion of *outlier*. An observation is classified as an outlier when it lies more than a given multiple of the interquartile range (the difference in value between the 75% and 25% observation) above or below respectively the value for the 75th percentile and 25th percentile. The standard multiples used are 1.5 and 3 times the interquartile range. *GeoDa* supports both values.

Clear all windows and start a new project using the `stl_hom.shp` homicide sample data set (use `FIPSNO` as the `Key`). The opening screen should

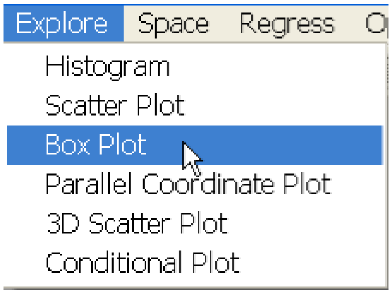


Figure 7.12: Box plot function.

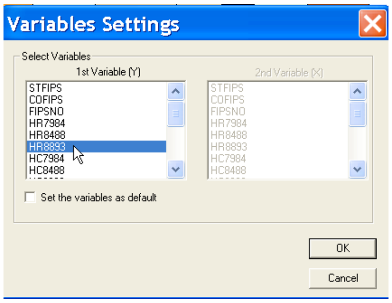


Figure 7.13: Variable selection in box plot.

show the base map with 78 counties as in Figure 7.11 on p. 49.

Invoke the box plot by selecting `Explore > Box Plot` from the menu (Figure 7.12), or by clicking on the `Box Plot` toolbar icon. Next, choose the variable `HR8893` (homicide rate over the period 1988–93) in the dialog, as in Figure 7.13. Click on `OK` to create the box plot, shown in the left hand panel of Figure 7.14 on p. 51. The rectangle represents the cumulative distribution of the variable, sorted by value. The value in parentheses on the upper right corner is the number of observations.

The red bar in the middle corresponds to the median, the dark part shows the interquartile range (going from the 25th percentile to the 75th percentile). The individual observations in the first and fourth quartile are shown as blue dots. The thin line is the *hinge*, here corresponding to the default criterion of 1.5. This shows six counties classified as outliers for this variable.

The hinge criterion determines how *extreme* observations need to be before they are classified as outliers. It can be changed by selecting `Option`

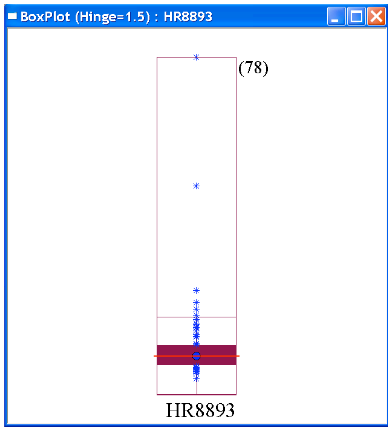


Figure 7.14: Box plot using 1.5 as hinge.

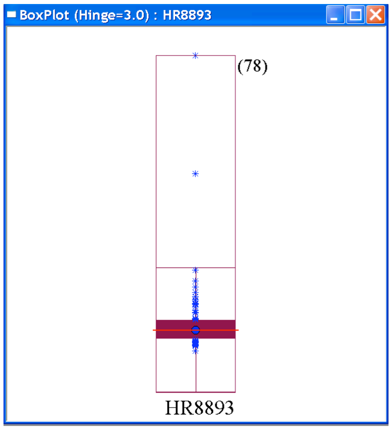


Figure 7.15: Box plot using 3.0 as hinge.

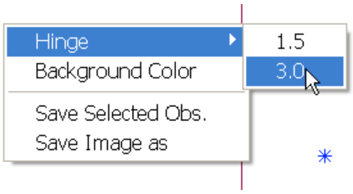
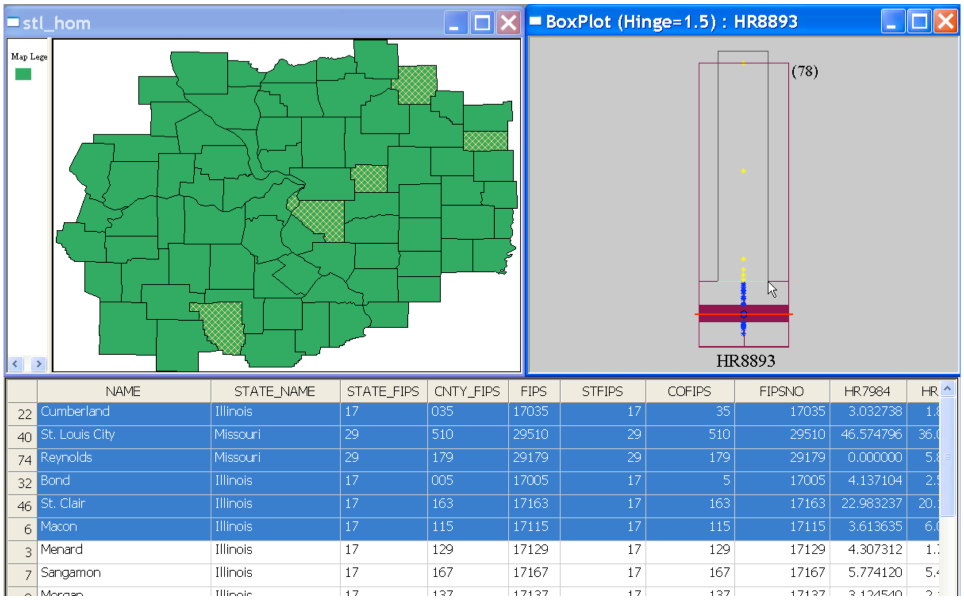


Figure 7.16: Changing the hinge criterion for a box plot.

`>` `Hinge` from the menu, or by right clicking in the box plot itself, as shown in Figure 7.16. Select 3.0 as the new criterion and observe how the number of outliers gets reduced to 2, as in the right hand panel of Figure 7.15.

Specific observations in the box plot can be selected in the usual fashion, by clicking on them, or by click-dragging a selection rectangle. The selection is immediately reflected in all other open windows through the linking mechanism. For example, make sure you have the table and base map open for the St. Louis data. Select the outlier observations in the box plot by dragging a selection rectangle around them, as illustrated in the upper left panel of Figure 7.17 on p. 52. Note how the selected counties are highlighted in the map and in the table (you may need to use the `Promotion` feature to get the selected counties to show up at the top of the table). Similarly, you can select rows in the table and see where they stack up in the box plot (or any other graph, for that matter).

Make a small selection rectangle in the box plot, hold down the `Control`



	NAME	STATE_NAME	STATE_FIPS	CNTY_FIPS	FIPS	STFIPS	COFIPS	FIPSNO	HR7984	HR
22	Cumberland	Illinois	17	035	17035	17	35	17035	3.032738	1.8
40	St. Louis City	Missouri	29	510	29510	29	510	29510	46.574796	36.0
74	Reynolds	Missouri	29	179	29179	29	179	29179	0.000000	5.8
32	Bond	Illinois	17	005	17005	17	5	17005	4.137104	2.5
46	St. Clair	Illinois	17	163	17163	17	163	17163	22.983237	20.
6	Macon	Illinois	17	115	17115	17	115	17115	3.613635	6.0
3	Menard	Illinois	17	129	17129	17	129	17129	4.307312	1.
7	Sangamon	Illinois	17	167	17167	17	167	17167	5.774120	5.4
	Morgan	Illinois	17	137	17137	17	137	17137	3.124540	2.

Figure 7.17: Linked box plot, table and map.

key and let go. The selection rectangle will blink. This indicates that you have started *brushing*, which is a way to change the selection dynamically. Move the brush slowly up or down over the box plot and note how the selected observations change in all the linked maps and graphs. We return to brushing in more detail in Exercise 8.

7.4 Practice

Use the St. Louis data set and histograms linked to the map to investigate the regional distribution (e.g., East vs West, core vs periphery) of the homicide rates (`HR****`) in the three periods. Use the box plot to assess the extent to which outlier counties are consistent over time. Use the table to identify the names of the counties as well as their actual number of homicides (`HC****`).

Alternatively, experiment with any of the other polygon sample data sets to carry out a similar analysis, e.g., investigating outliers in SIDS rates in North Carolina counties, or in the greenness index for the NDVI regular grid data.

Exercise 8

Brushing Scatter Plots and Maps

8.1 Objectives

This exercise deals with the visualization of the bivariate association between variables by means of a scatter plot. A main feature of *GeoDa* is the *brushing* of these scatter plots as well as maps.

At the end of the exercise, you should know how to:

- create a scatter plot for two variables
- turn the scatter plot into a correlation plot
- recalculate the scatter plot slope with selected observations excluded
- brush a scatter plot
- brush a map

More detailed information on these operations can be found in the *User's Guide*, pp. 68–76.

8.2 Scatter Plot

We continue using the same St. Louis homicide sample data set as in Exercise 7. Clear all windows if you have been working with a different data set and load `stl_hom` with `Fipsno` as the `Key`. Invoke the scatter plot functionality from the menu, as `Explore > Scatter Plot` (Figure 8.1 on p. 54), or

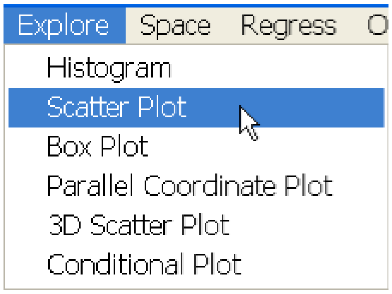


Figure 8.1: Scatter plot function.

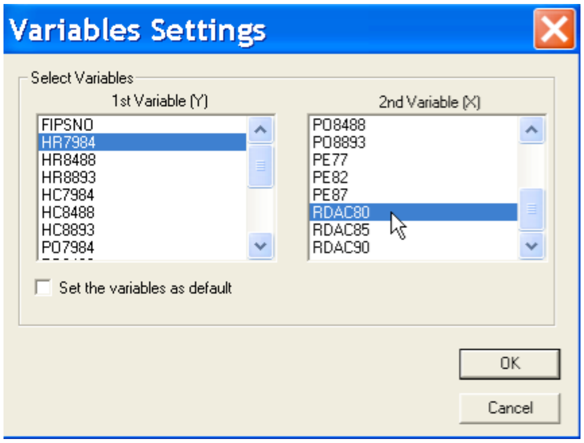


Figure 8.2: Variable selection for scatter plot.

by clicking on the `Scatter Plot` toolbar icon. This brings up the variables selection dialog, shown in Figure 8.2.

Select `HR7984` (the county homicide rate in the period 1979–84) in the left column as the `y` variable and `RDAC80` (a resource deprivation index constructed from census variables) in the right column as the `x` variable. Click on `OK` to bring up the basic scatter plot shown in Figure 8.3 on p. 55. The blue line through the scatter is the least squares regression fit, with the estimated slope shown at the top of the graph (`4.7957`). As expected, the relation is positive, with a higher degree of resource deprivation associated with a higher homicide rate. Since `RDAC80` has both positive and negative values, a vertical line is drawn at zero (when all variables take on only positive values, no such line appears).

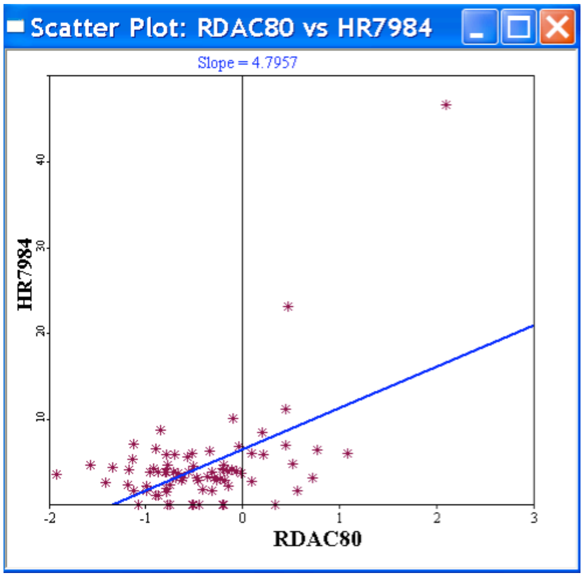


Figure 8.3: Scatter plot of homicide rates against resource deprivation.

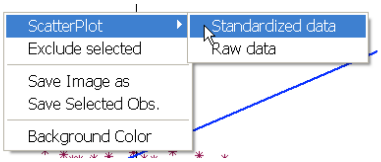


Figure 8.4: Option to use standardized values.

The scatter plot in *GeoDa* has two useful options. They are invoked by selection from the `Options` menu or by right clicking in the graph. Bring up the menu as shown in Figure 8.4 and choose `Scatter plot > Standardized data`. This converts the scatter plot to a correlation plot, in which the regression slope corresponds to the correlation between the two variables (as opposed to a bivariate regression slope in the default case).

The variables on both axes are rescaled to standard deviational units, so any observations beyond the value of 2 can be informally designated as outliers. Moreover, as shown in Figure 8.5 on p. 56, the plot is divided into four quadrants to allow a qualitative assessment of the association by type: high-high and low-low (relative to the mean) as positive correlation, and high-low and low-high as negative correlation. Select any of the points by

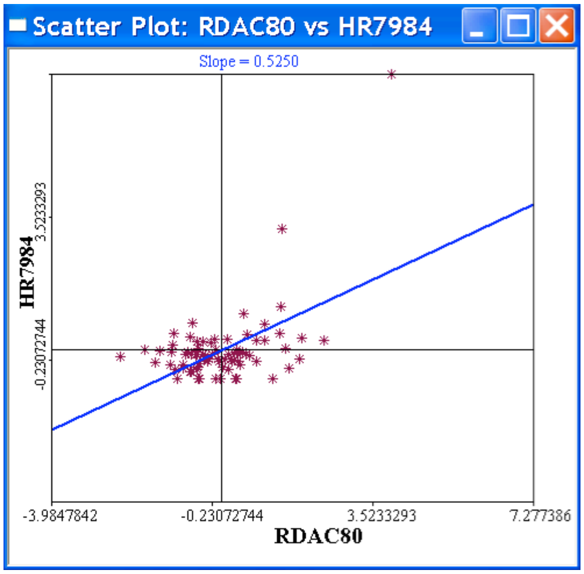


Figure 8.5: Correlation plot of homicide rates against resource deprivation.

clicking on them or by click-dragging a selection rectangle and note where the matching observations are on the St. Louis base map (we return to brushing and linking in more detail in section 8.2.2).

The correlation between the two variables is shown at the top of the graph (`0.5250`). Before proceeding with the next section, turn back to the default scatter plot view by right clicking in the graph and choosing `Scatter plot > Raw data`.

8.2.1 Exclude Selected

A second important option available in the scatter plot is the dynamic recalculation of the regression slope after *excluding* selected observations. This is particularly useful when brushing the scatter plot. Note that this option does *not* work correctly in the correlation plot and may yield seemingly incorrect results, such as correlations larger than 1. With selected observations excluded, the slope of the correlation scatter plot is no longer a correlation.[1]

In the usual fashion, you invoke this option from the `Options` menu or by right clicking in the graph. This brings up the options dialog shown in

[1] The reason for this problem is that *GeoDa* does not recenter the standardized variables, so that the data set without the selected observations is no longer standardized.

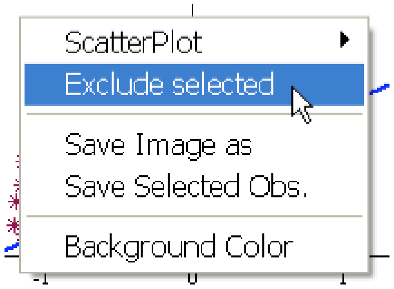


Figure 8.6: Option to use exclude selected observations.

Figure 8.6. Click on `Exclude selected` to activate the option.

In the scatter plot, select the two points in the upper right corner (the two observations with the highest homicide rate), as shown in Figure 8.7 on p. 58. Note how a new regression line appears (in brown), reflecting the association between the two variables *without* taking the selected observations into account. The new slope is also listed on top of the graph, to the right of the original slope. In Figure 8.7, the result is quite dramatic, with the slope dropping from `4.7957` to `0.9568`. This illustrates the strong *leverage* these two observations (St. Louis county, MO, and St. Clair county, IL) effect on the slope.[2]

8.2.2 Brushing Scatter Plots

The `Exclude selected` option becomes a really powerful exploratory tool when combined with a dynamic change in the selected observations. This is referred to as *brushing*, and is implemented in all windows in *GeoDa*.

To get started with brushing in the scatter plot, first make sure the `Exclude selected` option is on. Next, create a small selection rectangle and hold down the `Control` key. Once the rectangle starts to blink, the *brush* is activated, as shown in Figure 8.8 on p. 58. From now on, as you move the brush (rectangle), the selection changes: some points return to their original color and some new ones turn yellow. As this process continues, the regression line is recalculated on the fly, reflecting the slope for the data set *without* the current selection.

[2] Note that this analysis of influential observations does *not* include any reference to *significance*, but is purely exploratory at this point. For further elaboration and substantive interpretation, see Messner and Anselin (2004).

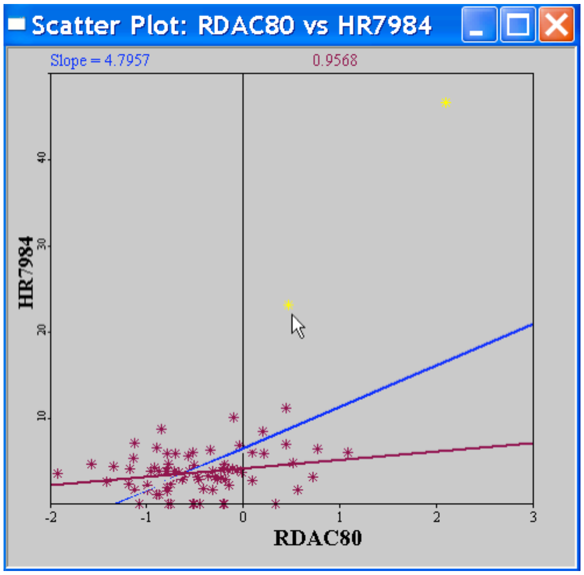


Figure 8.7: Scatter plot with two observations excluded.

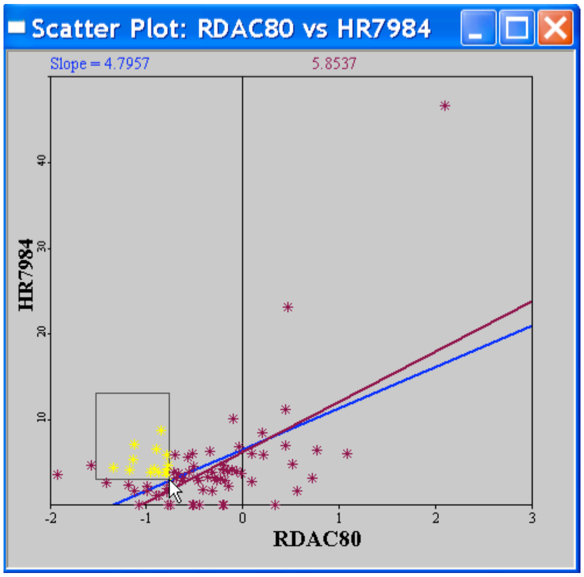


Figure 8.8: Brushing the scatter plot.

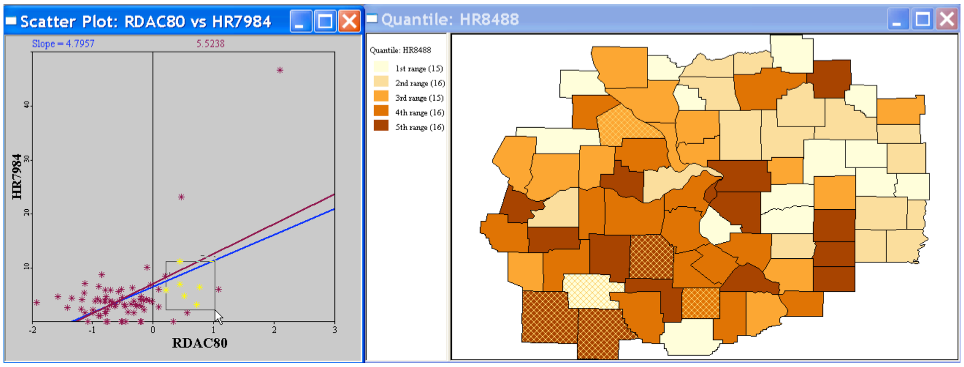


Figure 8.9: Brushing and linking a scatter plot and map.

The full power of brushing becomes apparent when combined with the *linking* functionality. For example, in Figure 8.9 on p. 59, the scatter plot is considered together with a quintile map of the homicide rate in the next period (`HR8488`). Create this choropleth map by clicking on the "green" base map and using the `Map > Quantile` function (or right clicking on the map with the same effect). As soon as the map appears, the selection from the scatter plot will be reflected in the form of cross-hatched counties.

The *brushing and linking* means that as you move the brush in the scatter plot, the selected counties change, as well as the slope in the scatter plot. This can easily be extended to other graphs, for example, by adding a histogram or box plot. In this manner, it becomes possible to start exploring the associations among variables in a multivariate sense. For example, histograms could be added with homicide rates and resource deprivation measures for the different time periods, to investigate the extent to which high values and their location in one year remain consistent over time.

8.3 Brushing Maps

In Figure 8.9, the brushing is initiated in the scatter plot, and the selection in the map changes as a result of its link to the scatter plot. Instead, the logic can be reversed.

Click anywhere in the scatter plot to close the brush. Now, make the map active and construct a brush in the same fashion as above (create a selection rectangle and hold down the `Control` key). The result is illustrated in Figure 8.10.

After a short time, the rectangle will start to blink, signaling that the

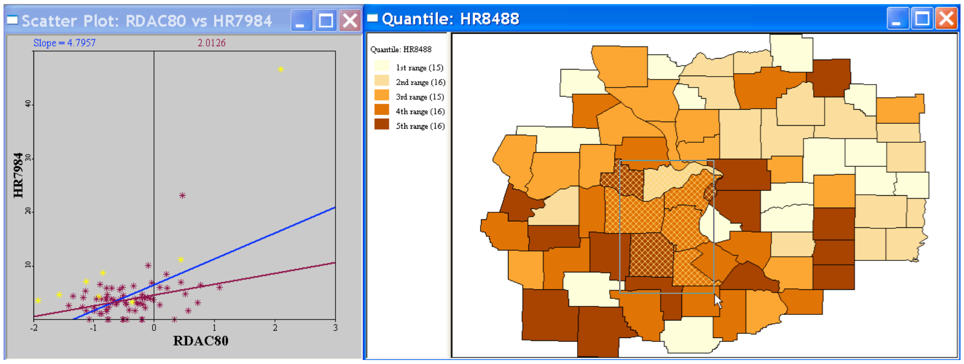


Figure 8.10: Brushing a map.

brush is ready. As you now move the brush over the map, the selection changes. This happens not only on the map, but in the scatter plot as well. In addition, the scatter plot slope is recalculated on the fly, as the brush moves across the map.

Similarly, the brushing can be initiated in any statistical graph to propagate its selection throughout all the graphs, maps and table in the project.

8.4 Practice

Continue exploring the associations between the homicide and resource deprivation variables using brushing and linking between maps, scatter plots, histograms and box plots. Compare this association to that between homicide and police expenditures (`PE**`). Alternatively, consider the Atlanta (`atl_hom.shp` with `FIPSNO` as the `Key`) and Houston (`hou_hom.shp` with `FIPSNO` as the `Key`) sample data sets, which contain the same variables.

Exercise 9

Multivariate EDA basics

9.1 Objectives

This exercise deals with the visualization of the association between multiple variables by means of a scatter plot matrix and a parallel coordinate plot.

At the end of the exercise, you should know how to:

- arrange scatter plots and other graphs into a scatter plot matrix
- brush the scatter plot matrix
- create a parallel coordinate plot
- rearrange the axes in a parallel coordinate plot
- brush a parallel coordinate plot

More detailed information on these operations can be found in the *Release Notes*, pp. 29–32.

9.2 Scatter Plot Matrix

We will explore the association between police expenditures, crime and unemployment for the 82 Mississippi counties in the POLICE sample data set (enter `police.shp` for the file name and `FIPSNO` for the `Key`). The beginning base map should be as in Figure 9.1 on p. 62.

The first technique considered is a scatter plot matrix. Even though *GeoDa* does *not* include this functionality per se, it is possible (though a little tedious) to construct this device for brushing and linking. The matrix

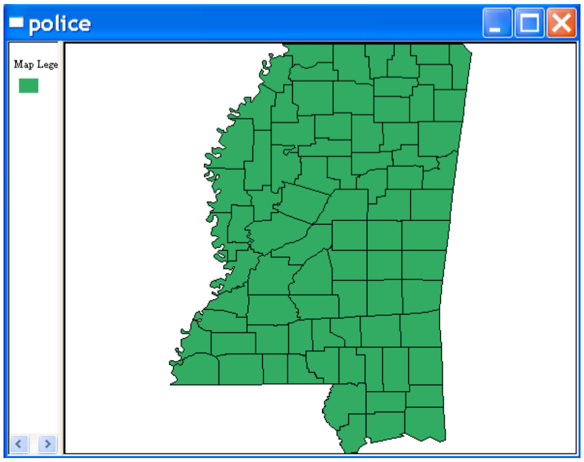


Figure 9.1: Base map for the Mississippi county police expenditure data.

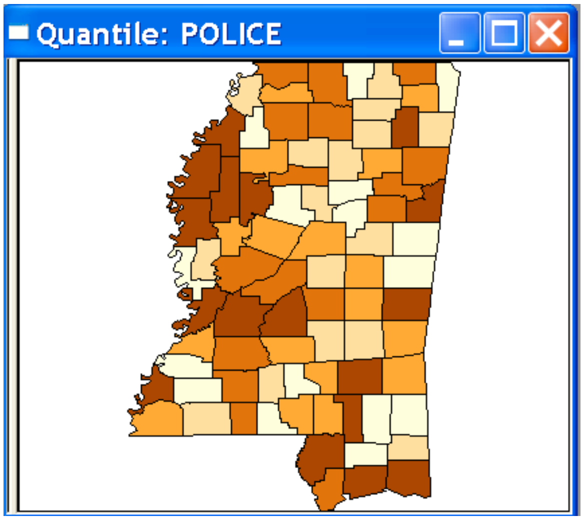


Figure 9.2: Quintile map for police expenditures (no legend).

consists of a set of pairwise scatter plots, arranged such that all the plots in a row have the same variable on the y-axis. The diagonal blocks can be left empty, or used for any univariate statistical graph.

For example, start by creating a quintile map for the `police` variable (`Map > Quantile` with 5 categories, see Exercise 2). Move the vertical sep-

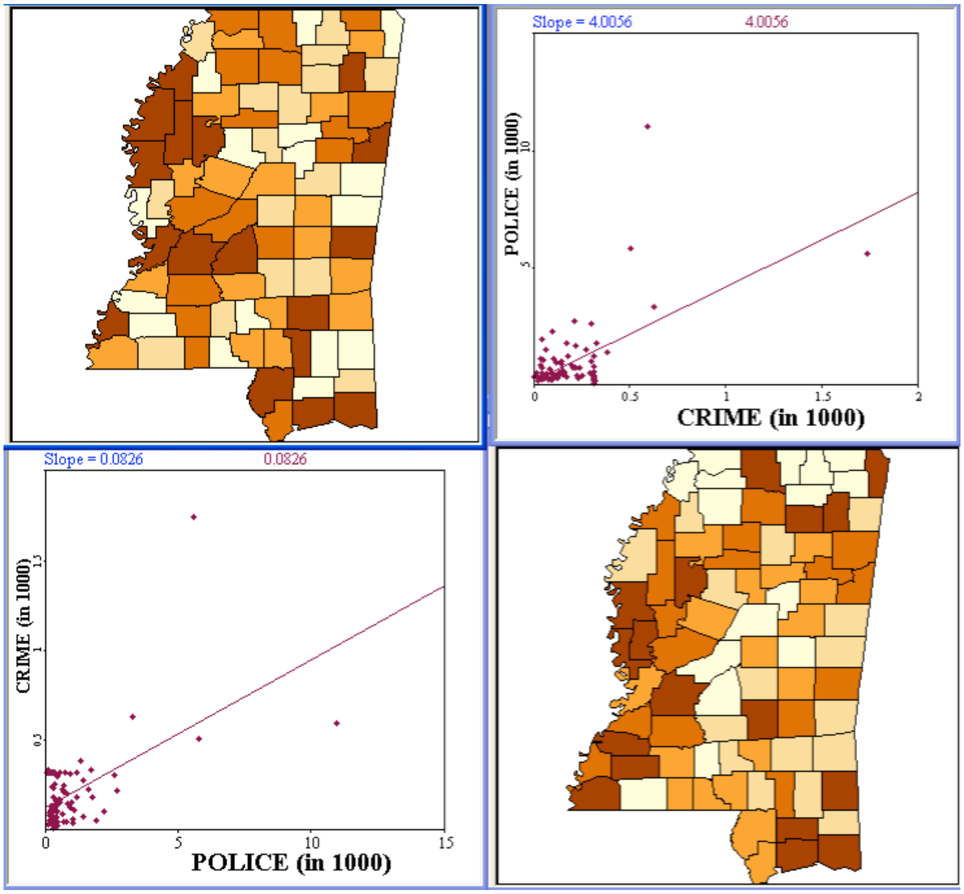


Figure 9.3: Two by two scatter plot matrix (police, crime).

arator between the legend and the map to the left, such that the legend disappears, as in Figure 9.2 on p. 62. Repeat this process for the variables `crime` and `unemp`. Next, create two bivariate scatter plots for the variables `police` and `crime` (see Section 8.2 on p. 53 for specific instructions), with each variable in turn on the y-axis. Arrange the scatter plots and the two matching quintile maps in a two by two matrix, as shown in Figure 9.3.

Continue this operation (this is the tedious part) for the remaining scatter plots between `police` and `unemp`, and `unemp` and `crime`. Make sure to turn the `Exclude selected` option on in each scatter plot (since this option is specific to a graph, it must be set for each scatter plot individually). To facilitate viewing the selected points, you may also want to turn the background color to grey, or, alternatively, change the color of the selection.

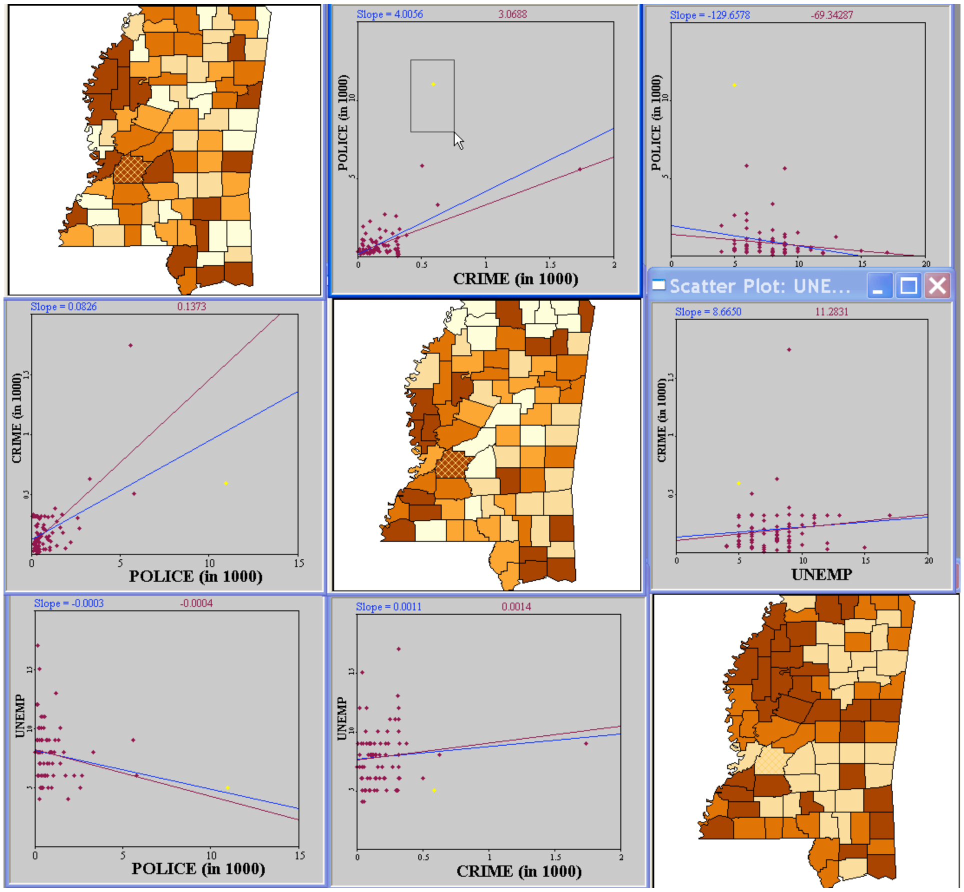


Figure 9.4: Brushing the scatter plot matrix.

After rearranging the various graphs, the scatter plot matrix should look as in Figure 9.4.

You can now explore the multivariate association between the variables by carefully examining the pairwise slopes, selecting interesting observations, and brushing the scatter plots. For example, in Figure 9.4, police expenditures are positively related to crime (as is to be expected), but negatively related to unemployment. On the other hand, crime is positively related to unemployment, suggesting a more complex interaction between these three variables than a simple pairwise association would reveal. Assess the sensitivity of the slopes to specific observations by brushing (only one brush can be active at a time). For example, in Figure 9.4, the brush includes the observation with the highest police expenditures in the `police-crime` scatter

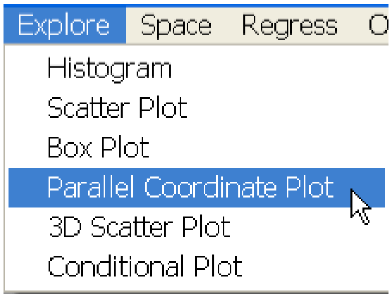


Figure 9.5: Parallel coordinate plot (PCP) function.

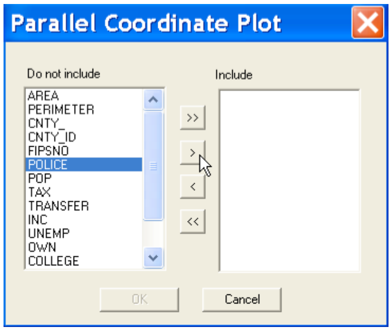


Figure 9.6: PCP variable selection.

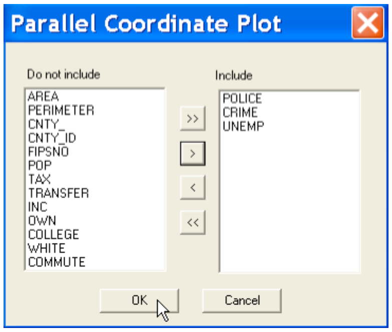


Figure 9.7: Variables selected in PCP.

plot, resulting in a much lower slope when excluded. Since all the graphs are linked, the location of the selected observations is also highlighted in the three maps on the diagonal of the matrix.

9.3 Parallel Coordinate Plot (PCP)

An alternative to the scatter plot matrix is the parallel coordinate plot (PCP). Each variable considered in a multivariate comparison becomes a *parallel* axis in the graph (this contrasts with a scatter plot, where the axes are orthogonal). On each axis, the values observed for that variable are shown from the lowest (left) to the highest (right). Consequently, an observation with multiple variables is represented by a series of line segments, connecting its position on each axis. This collection of line segments is the counterpart of a point in a multivariate (multidimensional) scatter plot.

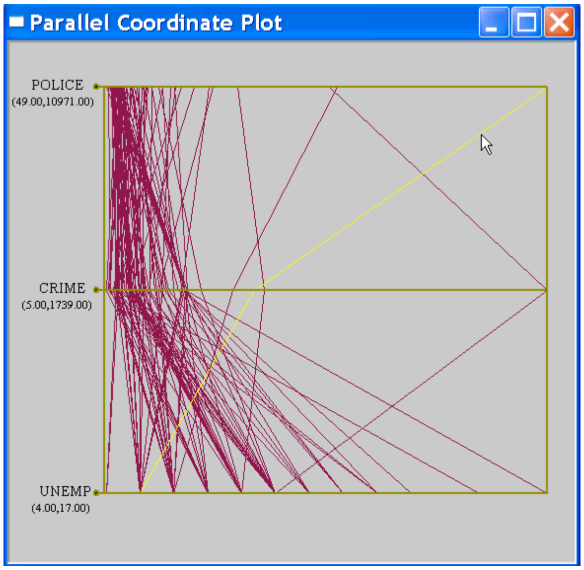


Figure 9.8: Parallel coordinate plot (police, crime, unemp).

The PCP is started by selecting `Explore > Parallel Coordinate Plot` from the menu (see Figure 9.5 on p. 65) or by clicking its toolbar icon. This brings up a variable selection dialog, as in Figure 9.6 on p. 65.

For example, move the three same variables as in the scatter plot matrix (`police`, `crime` and `unemp`) from the left hand column to the right, by selecting them and clicking on the `>` button (as in Figure 9.6). When the relevant variables are all in the right hand column, click `OK` (Figure 9.7 on p. 65) to create the graph. The result should be as in Figure 9.8.

You can select individual observations by clicking on the corresponding line, or by using a selection rectangle (any intersecting line will be selected). For example, click on the line with the highest value for `crime`. It will turn yellow (as in Figure 9.8) indicating its selection. To make it easier to distinguish the selected lines, the background color of the graph has been changed from the default white to a grey.

The selected observation corresponds to Hinds county, which includes Jackson, MS, the capital of the state. Note how this location is both high (to the right of the axis) in police expenditures and crime (though in terms of value much more to the center of the distribution than for police expenditures, hence the negative slope), but on the low end in terms of unemployment (to the left of the axis).

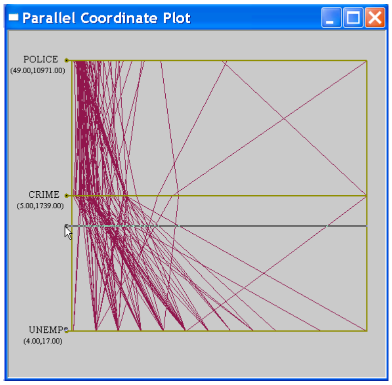


Figure 9.9: Move axes in PCP.

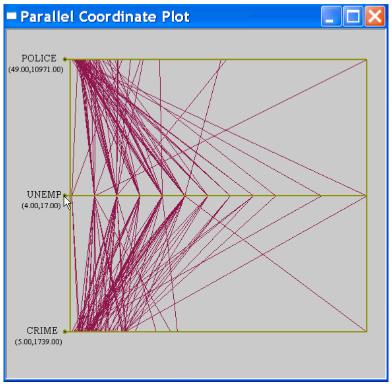


Figure 9.10: PCP with axes moved.

It is important to keep in mind that each variable has been rescaled, such that the mininum value is on the left end point and the maximum on the right hand side. The observations are sorted by increasing magnitude from left to right and positioned relative to the range of values observed (the difference between maximum and minimum).

You can change the order of the axes to focus more specifically on the association between two variables. Click on the small dot next to `unemp` in the graph and move it upward, as shown in Figure 9.9. When you reach the position of the middle axis, let go: the variable axes for `crime` and `unemp` will have switched places, as shown in Figure 9.10.

A major application of the PCP is to identify observations that *cluster* in multivariate space. This is reflected in a similar signature of the lines on the graph. Brush the graph to find line segments that show a similar pattern and check their location on the map. You can also obtain the county name in the data table (you may have to use the `Promotion` option in the `Table` to find the selected observations more easily). For example, in Figure 9.11 on p. 68, the selection brush is moved along the `police` axis. While some of the lines follow similar patterns, many do not. Explore potential clusters by brushing along the other axes as well (and switch the axis position, if desired).

Since the PCP is linked to all the other graphs in the project, it is possible to assess the extent to which multivariate clusters correspond to *spatial* clusters. In Figure 9.11 (on p. 68), the PCP is shown together with the quintile map for police expenditures, with the selected (brushed) observations in the PCP also highlighted on the map. Conversely, try to assess the

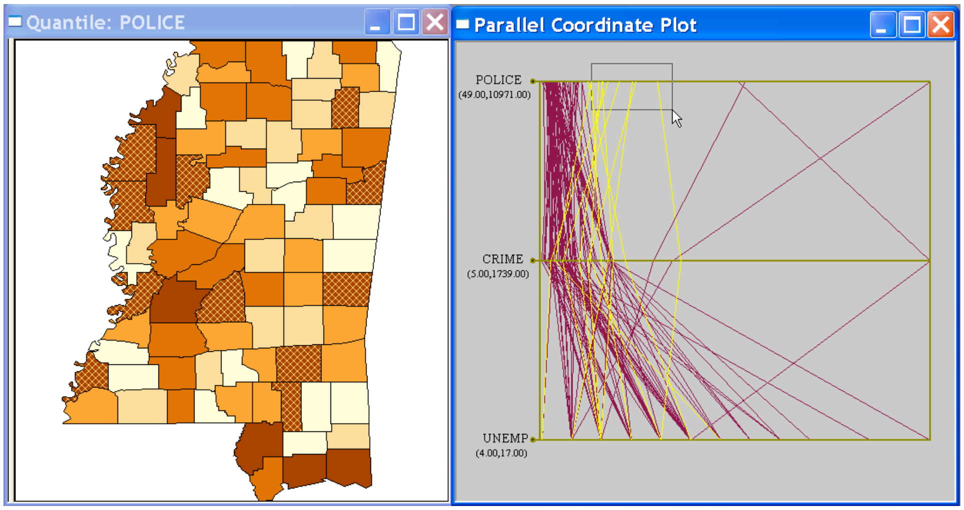


Figure 9.11: Brushing the parallel coordinate plot.

extent to which subregions in the map correspond to multivariate clusters by originating the brush in the map.

9.4 Practice

You can experiment further with these techniques by considering the association between police expenditures, crime and the other variables in the POLICE data set. Compare your impressions to the results in the regression analysis of Kelejian and Robinson (1992).

Alternatively, consider revisiting the St. Louis homicide data set, or the homicide data sets for Atlanta and Houston, and explore the multivariate association between the homicide rate, resource deprivation (RDAC**) and police expenditures (PE**) that was addressed more informally in Exercise 8.

Exercise 10

Advanced Multivariate EDA

10.1 Objectives

This exercise deals with more advanced methods to explore the multivariate relationships among variables, by means of conditional plots and a three-dimensional scatter plot.

At the end of the exercise, you should know how to:

- create conditional histograms, box plots and scatter plots
- change the conditioning intervals in the conditional plots
- create a three-dimensional scatter plot
- zoom and rotate the 3D scatter plot
- select observations in the 3D scatter plot
- brush the 3D scatter plot

More detailed information on these operations can be found in the *Release Notes*, pp. 33–43.

10.2 Conditional Plots

We continue the exploration of multivariate patterns with the POLICE data set from Exercise 9. Before proceeding, make sure that the county centroids are added to the data table (follow the instructions in Section 6.2.1 on p. 39). In the example, we will refer to these centroids as `XCOO` and `YCOO`.

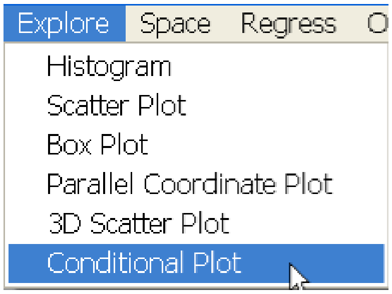
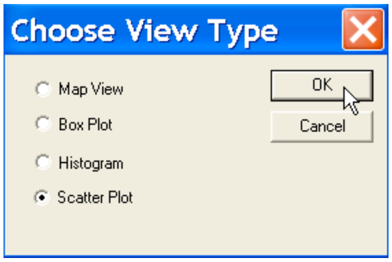


Figure 10.1: Conditional plot function.



○ Map View
○ Box Plot
○ Histogram
● Scatter Plot

Figure 10.2: Conditional scatter plot option.

A conditional plot consists of 9 micro plots, each computed for a subset of the observations. The subsets are obtained by conditioning on two other variables. Three intervals for each variable (for a total of 9 pairs of intervals) define the subsets.

Start the conditional plot by clicking the associated toolbar icon, or by selecting `Explore > Conditional Plot` from the menu, as in Figure 10.1. This brings up a dialog with a choice of four different types of plots, shown in Figure 10.2. Two of the plots are univariate (histogram and box plot, covered in Exercise 7), one bivariate (scatter plot, covered in Exercise 8), and one is a map.[1] Select the radio button next to `Scatter Plot` and click `OK`.

Next, you need to specify the conditioning variables as well as the variables of interest. In the variable selection dialog, shown in Figure 10.3 on p. 71, select a variable from the drop down list and move it to the text boxes on the right hand side by clicking on the `>` button. Specifically, choose `XCOO`

[1] Conditional maps are covered in Excercise 12.

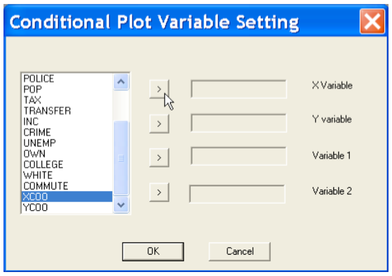
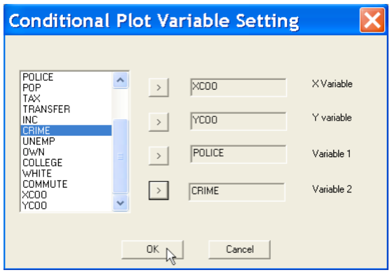


Figure 10.3: Conditional scatter plot variable selection.



XCOO X Variable
YCOO Y variable
POLICE Variable 1
CRIME Variable 2

Figure 10.4: Variables selected in conditional scatter plot.

for the first conditioning variable (the `X variable` in the dialog), `YCOO` for the second conditioning variable (`Y variable`), `POLICE` as the variable for the y-axis in the scatter plot (`Variable 1`), and `CRIME` as the variable for the x-axis (`Variable 2`).[2] The complete setup should be as in Figure 10.4. Click on `OK` to create the conditional scatter plot, shown in Figure 10.5 on p. 72.

Consider the figure more closely. On the horizontal axis is the first conditioning variable, which we have taken to be the location of the county centroid in the West-East dimension. The counties are classified in three "bins," depending on whether their centroid X coordinate (`XCOO`) falls in the range -91.44603 to -90.37621, -90.37621 to -89.30639, or -89.30639 to -88.23657. The second conditioning variable is on the vertical axis and corresponds to the South-North dimension, again resulting in three intervals. Consequently, the nine micro plots correspond to subsets of the counties arranged by their geographic location from the southwestern corner to the northeastern corner.[3]

The scatter plots suggest some strong regional differences in the slope of the regression of police expenditures on crime. Note that this is still *exploratory* and should be interpreted with caution. Since the number of observations in each of the plots differs, the precision of the estimated slope coefficient will differ as well. This would be taken into account in a more rigorous comparison, such as in an analysis of variance. Nevertheless, the plots confirm the strong effect of the state capital on this bivariate relationship (the middle plot in the left hand column), which yields by far the steepest

[2] The dialog is similar for the other conditional plot, except that only one variable can be specified in addition to the conditioning variables.

[3] The conditioning variables do not have to be geographic, but can be any dimension of interest.

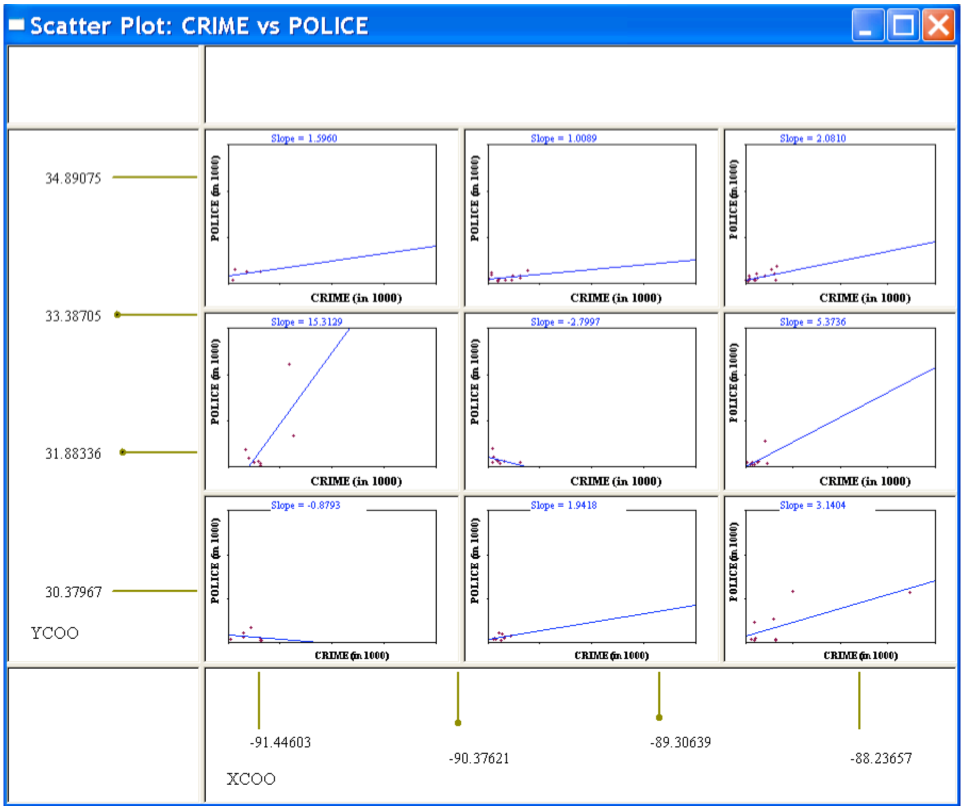


Figure 10.5: Conditional scatter plot.

slope. In two of the plots, the slope is even slightly negative.

The categories can be adjusted by moving the "handles" sideways or up and down. The handles are the small circles on the two interval bars. To convert the matrix of plots to a two by two format by collapsing the east-most and northern-most categories together, pull the right-most handle to the right, as in Figure 10.6 on p. 73. Similarly, pull the top-most handle to the top. The two by two classification still suggests a difference for the western-most counties. Experiment with moving the classification handles to get a better sense for how the plots change as the definition of the subsets is altered. Also, try using different variables (such as `tax` and `white`) as the conditioning variables.

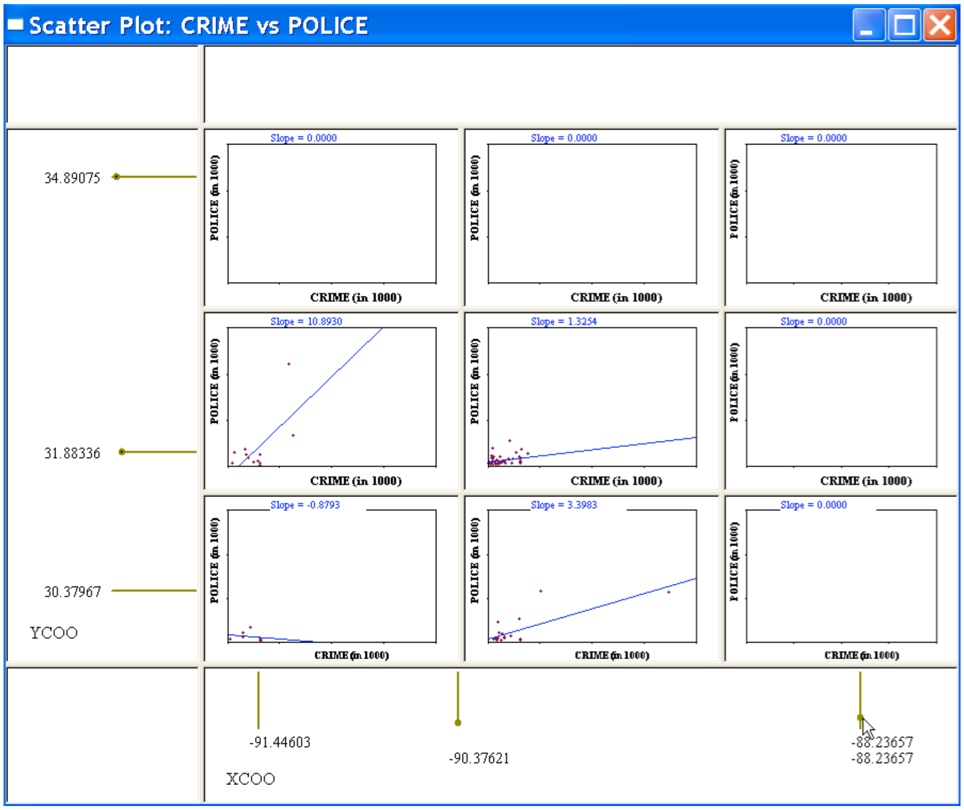


Figure 10.6: Moving the category breaks in a conditional scatter plot.

10.3 3-D Scatter Plot

The final technique we consider to explore multivariate associations consists of a three dimensional scatter plot (or cube). Clear the conditional plot and start up the 3D scatter plot interface by clicking on the toolbar icon or selecting `Explore > 3D Scatter Plot` from the menu (Figure 10.7 on p. 74).

This starts the variable selection dialog, shown in Figure 10.8 on p. 74. Select the variable `CRIME` in the drop down list for the `X Variable`, `UNEMP` for the `Y Variable` and `POLICE` for the `Z Variable`, as in Figure 10.9 on p. 74. Note that the order of the variables does not really matter, since the data cube can easily be rotated (for example, switching the x-axis from horizontal to vertical). Finally, click the `OK` button to generate the initial 3D view, shown in Figure 10.10 on p. 74.

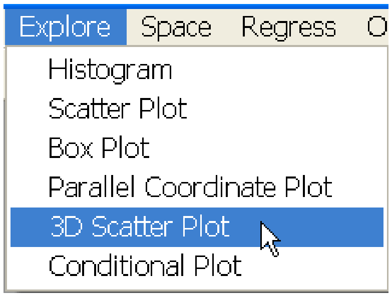


Figure 10.7: Three dimensional scatter plot function.

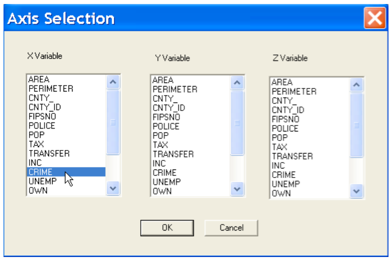


Figure 10.8: 3D scatter plot variable selection.

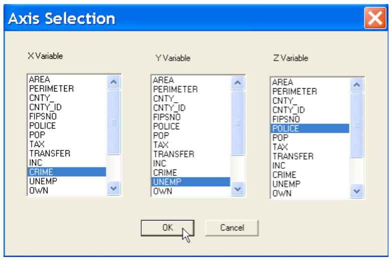


Figure 10.9: Variables selected in 3D scatter plot.

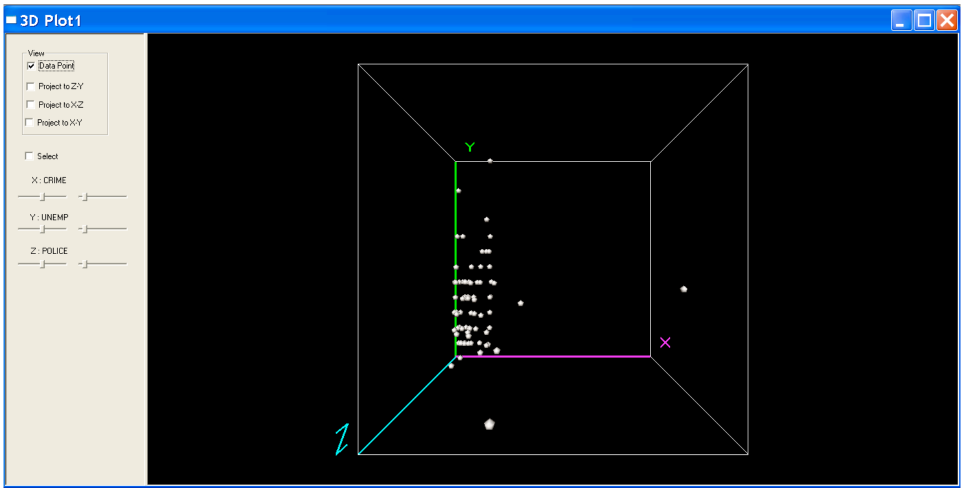


Figure 10.10: Three dimensional scatter plot (police, crime, unemp).

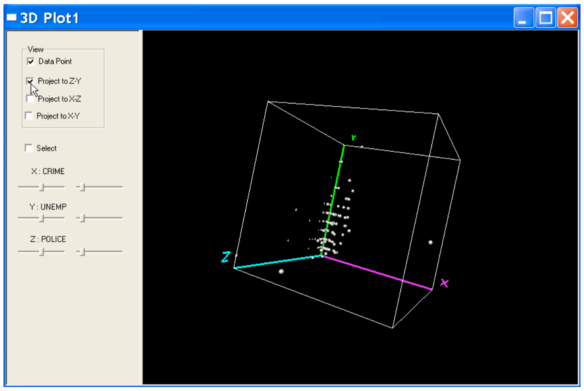


Figure 10.11: 3D scatter plot rotated with 2D projection on the zy panel.

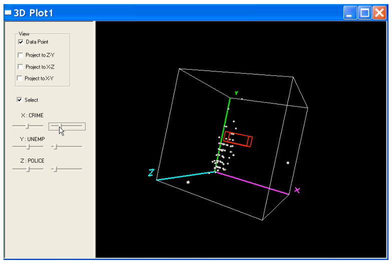


Figure 10.12: Setting the selection shape in 3D plot.

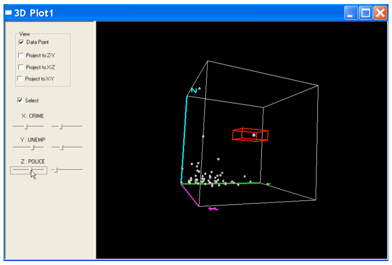


Figure 10.13: Moving the selection shape in 3D plot.

The 3D plot can be further manipulated in a number of ways. For example, click anywhere in the plot and move the mouse to rotate the plot. Right click to zoom in and out. More importantly, a number of options are available in the left hand panel of the window. At the top of the panel are three check boxes to toggle projection of the 3D point cloud on the side panels. For example, in Figure 10.11 the cube has been rotated and the projection on the `z-y` panel turned on.

The options in the bottom half of the left hand panel define the selection shape and control brushing. Check the `Select` box and a red outline of a cube will appear in the graph. Now move the slider to the right and below each variable name in the panel to change the size of the selection box along that dimension. For example, in Figure 10.12, the side of the box along the X dimension (`CRIME`) is increased as the slider is moved to the right. Manipulate the sliders for each of the three variables until the box has a

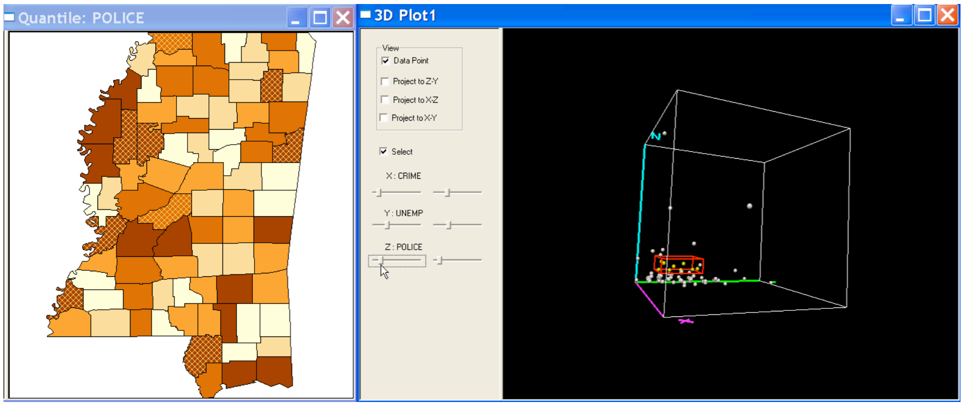


Figure 10.14: Brushing the 3D scatter plot.

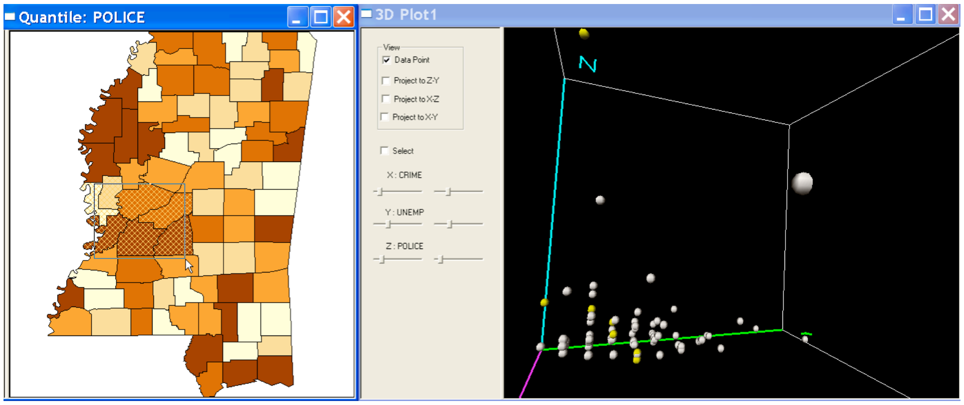
sizeable dimension. You can rotate the cube to get a better feel of where in the 3D space your selection box is situated.

The slider to the left and below each variable in the left panel of the graph is used to move the selection box along the corresponding dimension. As shown in Figure 10.13 on p. 75, dragging the bottom left slider will move the selection box along the Z axis, which corresponds to the POLICE variable. Selected observations are shown as yellow.

Experiment with changing the selection shape and moving the box around. You may find it helpful to rotate the cube often, so that you can see where the box is in relation to the point cloud. Also, you can move the selection box directly by holding down the Control key while clicking with the left mouse button.

The 3D scatter plot is linked to all the other maps and graphs. However, the update of the selection is implemented slightly differently from the two dimensional case. In contrast to the standard situation, where the updating is continous, the selection in the 3D plot is updated each time the mouse stops moving. The yellow points in the cloud plot will be matched to the corrresponding observations in all the other graphs. For example, in Figure 10.14, the selected points in the cloud plot are highlighted on the Mississippi county map. In all other respect, brushing is similar to the two dimensional case, although it takes some practice to realize *where* the selection box is in the three dimensional space.

Brushing also works in the other direction, but only with the Select check box turned off. To see this, create a brush in the county map and start moving it around. Each time the brush stops, the matching selection



View
☑ Data Point
☐ Project to Z-Y
☐ Project to X-Z
☐ Project to X-Y

☐ Select

X : CRIME
Y : UNEMP
Z : POLICE

Figure 10.15: Brushing a map linked to the 3D scatter plot.

in the 3D plot will be shown as yellow points, as shown in Figure 10.15. In practice, you may find that it often helps to zoom in and out and rotate the cube frequently.

10.4 Practice

Apply the conditional plots and the 3D scatter plot in an exploration of the relation between median house value (`CMEDV`) and other variables in the BOSTON sample data set (`boston.shp` with `ID` as the `Key`). For example, consider the relation between house value and air quality (`NOX`) conditioned by geographical location (`X` and `Y`) in a conditional scatter plot. Also, explore the associations between house value, air quality and crime (`CRIM`) in a 3D scatter plot. Experiment with brushing on a Thiessen polygon map created from the tract centroids.

You should now be at a point where you can pull together the various traditional EDA tools in conjunction with a map to explore both non-spatial and spatial patterns in most of the sample data sets.

Exercise 11

ESDA Basics and Geovisualization

11.1 Objectives

This exercise begins to deal with the exploration of data where the *spatial* aspects are explicitly taken into account. We focus primarily on map making and geovisualization at this point. More advanced techniques are covered in Exercise 12.

At the end of the exercise, you should know how to:

- create a percentile map
- create a box map
- change the hinge option in a box map
- create a cartogram
- change the hinge option in a cartogram

More detailed information on these operations can be found in the *User's Guide*, pp. 39–40, and *Release Notes*, pp. 23–26.

11.2 Percentile Map

We will illustrate the basic mapping functions with the sample data set BUENOSAIRES, containing results for 209 precincts in the city of Buenos Aires (Argentine) covering the 1999 national elections for the Argentine

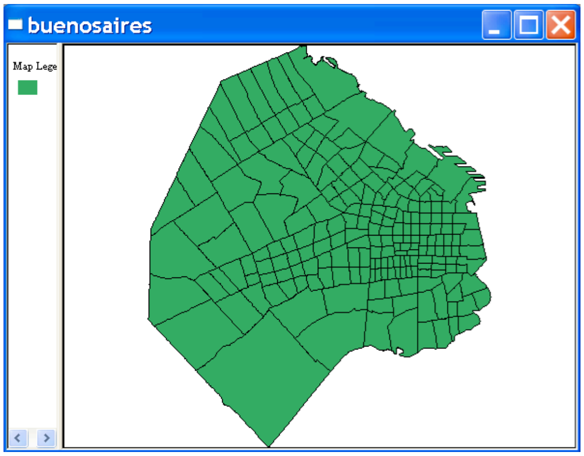


Figure 11.1: Base map for the Buenos Aires election data.

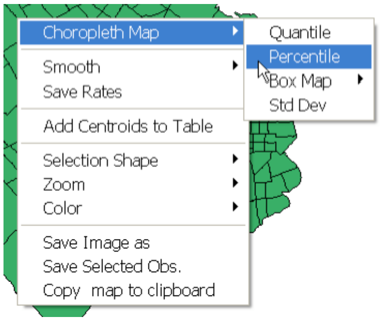


Figure 11.2: Percentile map function.

Congress.[1] The shape file is `buenosaires.shp` with `INDRANO` as the `Key` variable. Open up a new project with this shape file. The base map should be as in Figure 11.1.

Invoke the percentile map function from the menu, by selecting `Map > Percentile` or by right clicking in the base map. The latter will bring up the menu shown in Figure 11.2. Select `Choropleth Map > Percentile` to bring up the variable settings dialog. Alternatively, you can click on the toolbar icon.

[1] A more elaborate discussion of the data and substantive context can be found in Calvo and Escobar (2003).

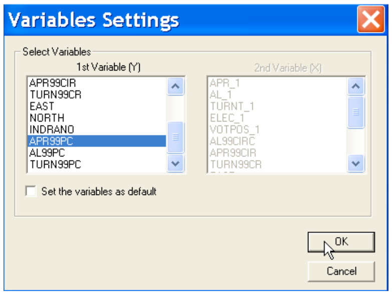


Figure 11.3: Variable selection in mapping functions.

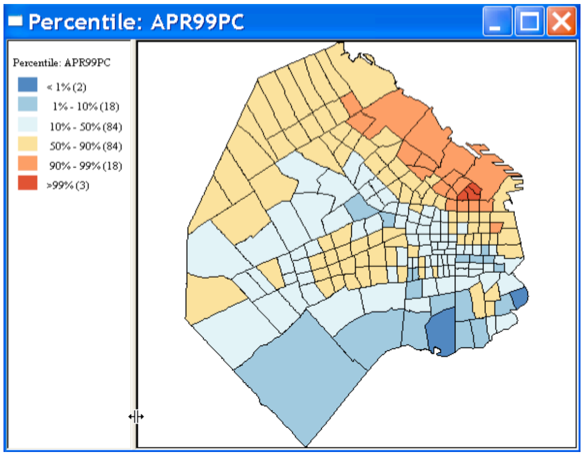


Figure 11.4: Percentile map for APR party election results, 1999.

Choose the variable APR99PC (the electoral results for the center right party APR, "Action por la Republica"), as in Figure 11.3, and click OK to bring up the map. The percentile map shown in Figure 11.4 emphasizes the importance of the very small (lowest percentile) and very high (highest percentile) values. Note how the three highest returns for this party are concentrated in three small (in area) precincts (colored in red). Also note how the broad classification greatly *simplifies* the spatial pattern in the map. Experiment with a percentile map for the other party (AL99PC, for the centrist "Alianza") and for the vote turnout (TURN99PC). Make a mental note of the general patterns depicted in these maps.

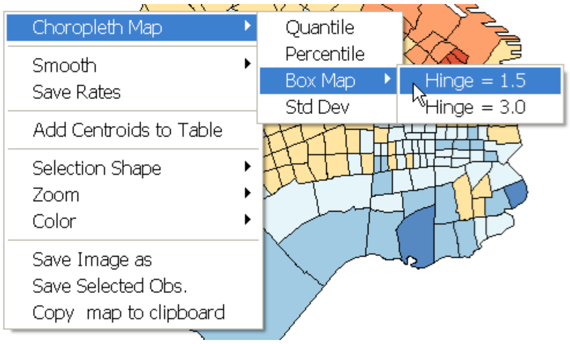


Figure 11.5: Box map function.

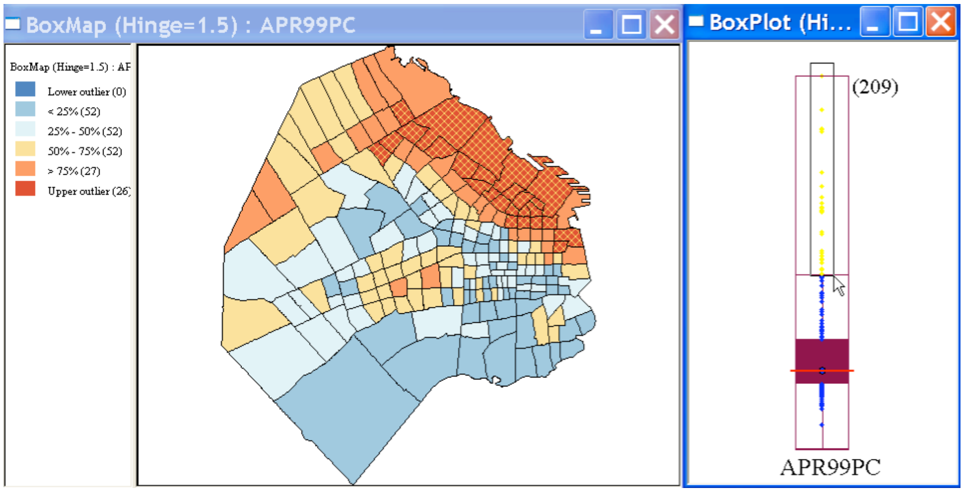


Figure 11.6: Box map for APR with 1.5 hinge.

11.3 Box Map

A box map is an enhanced version of a quartile map, in which the outliers in the first and fourth quartile are highlighted separately. The classification in the box map is thus identical to that used in a box plot. The box map is invoked from the main `Map` menu, by right clicking in an existing map (in Figure 11.5), or by clicking on the toolbar icon. Right click on the current map (or, first create a duplicate of the base map) and select the function `Choropleth Map > Box Map > Hinge = 1.5` to bring up the variable selection dialog (see Figure 11.3 on p. 80). Again, select `APR99PC` and click `OK`

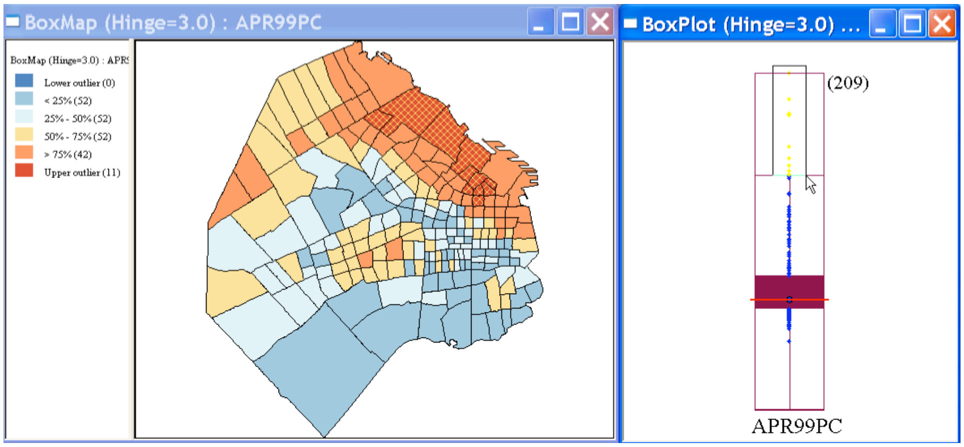


Figure 11.7: Box map for APR with 3.0 hinge.

to bring up the box map shown in Figure 11.6 on p. 81.

To confirm the classification of the outliers, bring up a regular box plot for the same variable (see Section 7.3 on p. 48), using the default hinge of 1.5, and select the upper outliers, as in Figure 11.6. Note how the selected points in the box plot correspond exactly to the dark red locations in the box map. In addition to showing the high values, the map also suggests that these may be clustered in space, something which the standard box plot is unable to do.

As with the standard box plot, the criterion to define outliers in the box map can be set to either 1.5 or 3.0. Create a new box map for APR99PC using the hinge of 3.0 and change the option in the box plot to that value as well. Again, check where the outliers are in the map by selecting them in the box plot, as illustrated in Figure 11.7. Note how the spatial clustering of outliers is even more pronounced in this map. As in the previous section, experiment with a box map for AL99PC and TURN99PC. Compare the locations of the outliers and the extent to which they suggest spatial clustering.

11.4 Cartogram

A cartogram is yet a third way to highlight extreme values on a map. *GeoDa* implements a version of a circular cartogram, in which the original spatial units are replaced by circles. The area of the circle is proportional to the value of a selected variable. The circles themselves are aligned as closely as possible to the original location of the matching spatial units by means of a

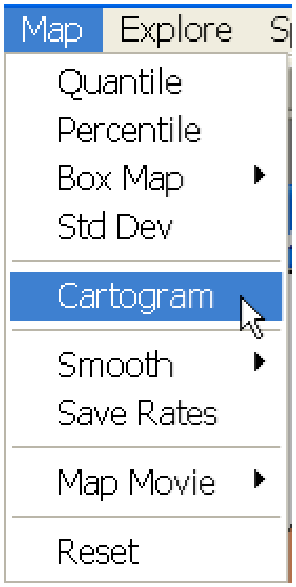


Figure 11.8: Cartogram map function.

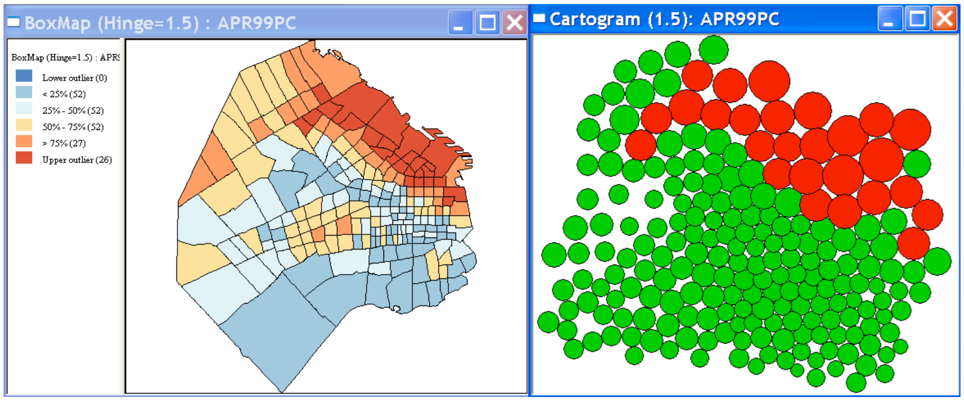


Figure 11.9: Cartogram and box map for APR with 1.5 hinge.

nonlinear optimization routine.

As with the other maps, the cartogram can be invoked from the `Map` menu (as `Cartogram`) (as in Figure 11.8), or from the context menu that appears when right clicking in any open map, or by clicking on the toolbar icon.

This opens up the usual variable selection dialog (see Figure 11.3 on

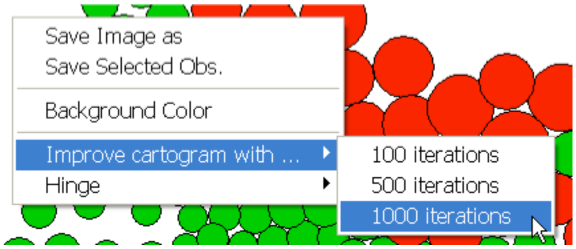


Figure 11.10: Improve the cartogram.

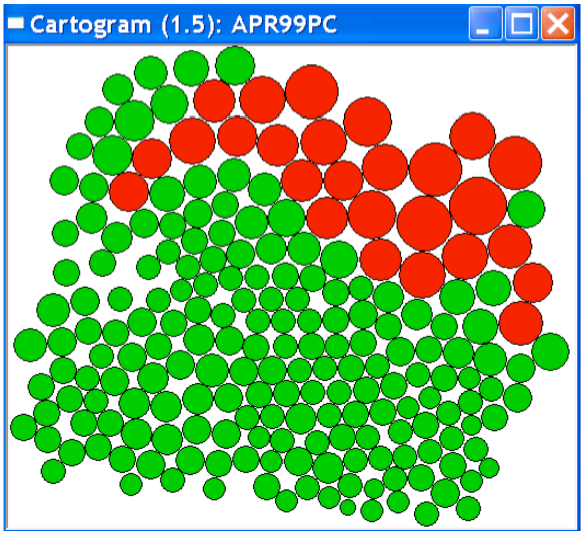


Figure 11.11: Improved cartogram.

p. 80). Select `APR99PC` to create the cartogram shown in Figure 11.9 on p. 83. The cartogram also highlights outliers in a different color from the rest of the circles (upper outliers are in red, lower outliers in blue). Note the general similarity between the cartogram and the box map in Figure 11.9.

Since the location of the circles is the result of an iterative nonlinear procedure, it can be refined if deemed necessary. Right click in the cartogram and select the option `Improve cartogram with ... > 1000 iterations`, as in Figure 11.10. After a brief delay, the circles will appear to jump, resulting in a slightly different alignment, as illustrated in Figure 11.11.

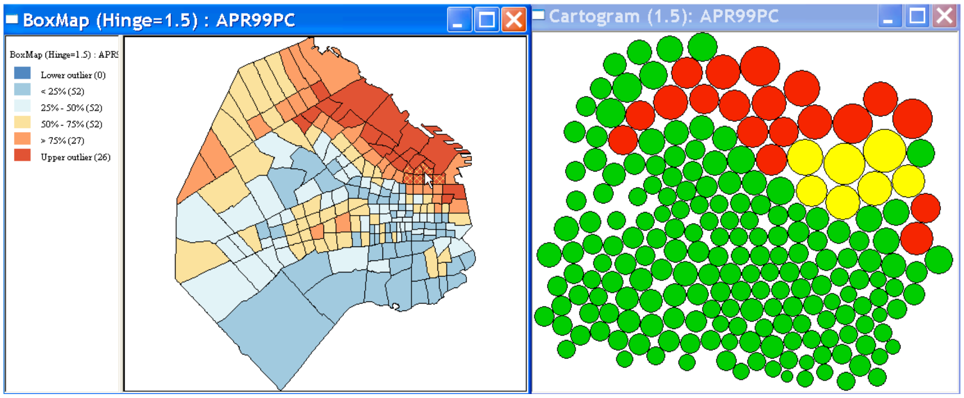


Figure 11.12: Linked cartogram and box map for APR.

Note how one of the other options for the cartogram pertains to the `Hinge`. Since the cartogram highlights outliers similar to a box map (and box plot), you can change the hinge criterion with this option. For example, change the `Hinge` to 3 and compare the result to the box map shown in Figure 11.7 (p. 82).

The cartogram is linked to all the other maps and graphs, in the usual fashion. This is useful to make the connection between the actual spatial layout of the observations and the idealized one presented in the cartogram. Select the small precincts on the southern and eastern edge of the outlier "cluster" in the box map and note where they are in the cartogram. As shown in Figure 11.12, they figure prominently in the cartogram, whereas they are barely noticeable in the standard choropleth map. Experiment further with a cartogram for `AL99PC` and `TURN99PC`.

11.5 Practice

For a change in topic, use the `rosas2001.shp` sample data set (with `ID` as `Key`) for 1705 measures on corn yield and relevant input variables in a precision farming experiment in Cordoba, Argentina.[2] Create percentile maps, box maps and cartograms for the `yield` variable as well as `BV`, an indicator for low organic matter. Compare the suggested patterns between the two. You may also want to map `N`, but what do you observe?[3]

[2] The data are part of the LasRosas file on the SAL data samples site.

[3] Hint: this is an agricultural experiment. For more on this data set, see Anselin et al. (2004a).

Exercise 12

Advanced ESDA

12.1 Objectives

This exercise illustrates some more advanced visualization techniques in ESDA in the form of map animation and conditional maps.

At the end of the exercise, you should know how to:

- create and control a map movie
- create a conditional map
- change conditioning categories in a conditional map

More detailed information on these operations can be found in the *User's Guide*, pp. 40–41, and *Release Notes*, pp. 26–28, and 38–40.

12.2 Map Animation

We continue to use the BUENOSAIRES sample data set. If this is not in your current project, clear all windows and load the file `buenosaires.shp` with `Key` variable `INDRANO`. The simple form of map animation implemented in *GeoDa* consists of automatically moving through all observations for a given variable, from the lowest value to the highest value. The matching observations are shown on a base map, either one at a time (`Single`), or cumulatively (`Cumulative`).

Invoke this function by choosing `Map > Map Movie > Cumulative` from the menu, as in Figure 12.1 (p. 87), or by clicking the toolbar icon. This brings up the familiar variables setting dialog (see Figure 11.3 on p. 80).

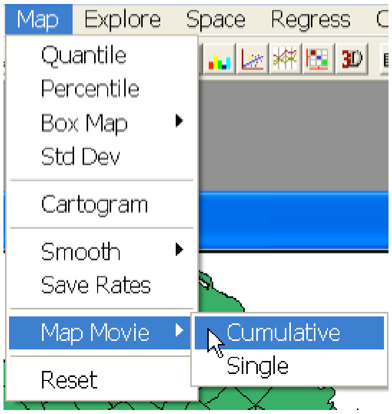


Figure 12.1: Map movie function.

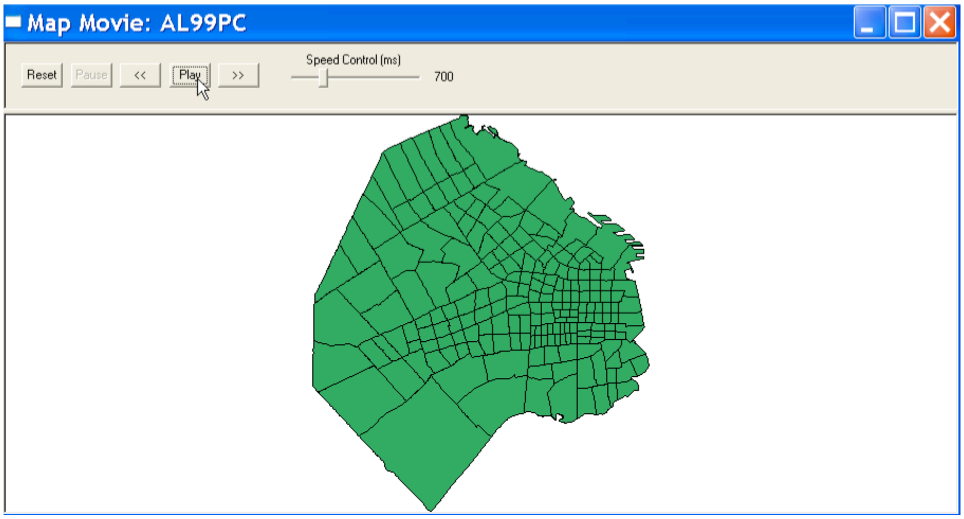


Figure 12.2: Map movie initial layout.

Select `AL99PC` (for the "Alianza" party election results) and click `OK` to bring up the initial map movie interface, shown in Figure 12.2.

Click on the `Play` button to start the movie. The polygons on the map will gradually be filled out in a pink shade, going from the lowest value to the highest. Note how the map movie is linked to all other graphs and maps in the project, such that the selection in the movie becomes the selection in all other windows. You can stop the movie at any time, by pressing the

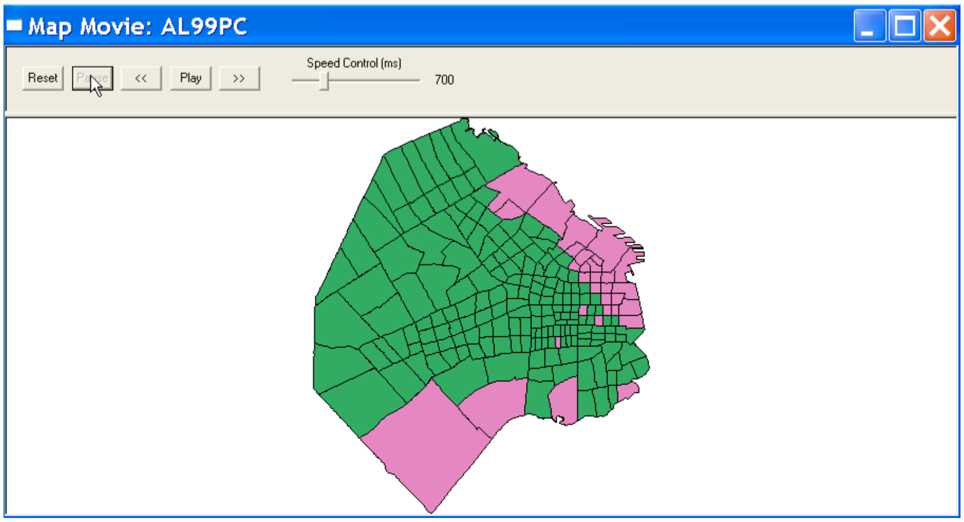


Figure 12.3: Map movie for AL vote results – pause.

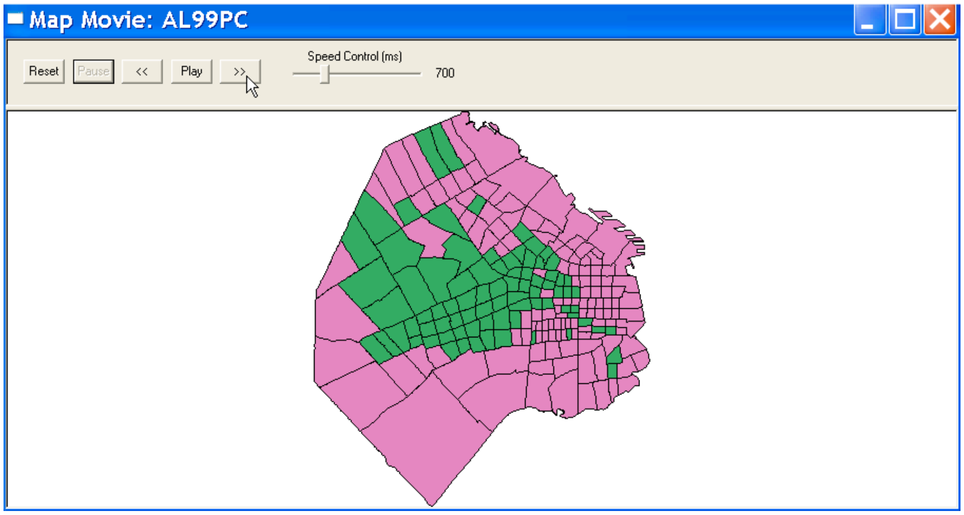
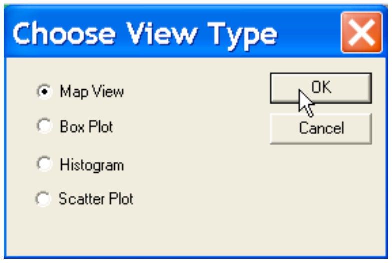


Figure 12.4: Map movie for AL vote results – stepwise.

`Pause` button, as in Figure 12.3. Clicking on `Reset` will wipe out all selected polygons and start over with a blank base map. You can affect the speed of the movie with the slider bar (`Speed Control`): positioning the slider bar more to the *left* will increase the speed.

Once the movie is paused, it can be advanced (or moved back) one



Choose View Type

● Map View

○ Box Plot

○ Histogram

○ Scatter Plot

Figure 12.5: Conditional plot map option.

observation at a time, by using the >> (or, <<) key. This is illustrated in Figure 12.4 (p. 88) for the map at a later stage in the animation process.

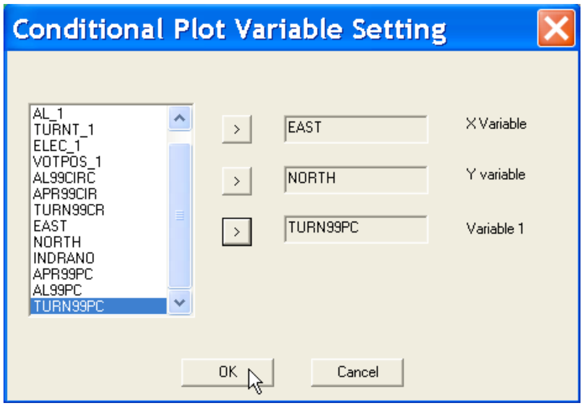
The purpose of the map animation is to assess the extent to which similar values occur in similar locations. For example, in Figures 12.3 and 12.4, the low values for AL99PC systematically start in the north eastern precincts and move *around* along the periphery, leaving the higher values in the city core. This is very different from a *random* pattern, where the values would jump all over the map. An example of such a random pattern can be seen in the grid100s.shp sample data set. Check this out for any of the randomly permuted variables (the variables starting with ranz).

12.3 Conditional Maps

The conditional map is a special case of the conditional plots considered in Section 10.2 on p. 69. Start the conditional plot function as before (see Figure 10.1 on p. 70) and select the radio button next to Map View in the view type dialog, as in Figure 12.5. Click on OK to bring up the variable selection dialog, Figure 12.6 on p. 90.

Select EAST for the X Variable, and NORTH for the Y Variable. Take the variable of interest (Variable 1) as TURN99PC. As in the previous examples, this is an illustration of *geographic* conditioning according to the location of the precincts, grouped into 9 subregions. Any other two conditioning variables could be selected, as in the generic conditional plot example. The interval ranges for the conditioning variables are changed by moving the associated handle to the left or right.

Finally, click on OK to bring up the conditional map, shown in Figure 12.7 on p. 90. The map employs a continuous color ramp, going from blue-green at the low end to brown-red at the high end. The color ramp is shown at


Conditional Plot Variable Setting

AL_1
TURNT_1
ELEC_1
VOTPOS_1
AL99CIRC
APR99CIR
TURN99CR
EAST
NORTH
INDRANO
APR99PC
AL99PC
TURN99PC

X Variable: EAST

Y variable: NORTH

Variable 1: TURN99PC

Figure 12.6: Conditional map variable selection.

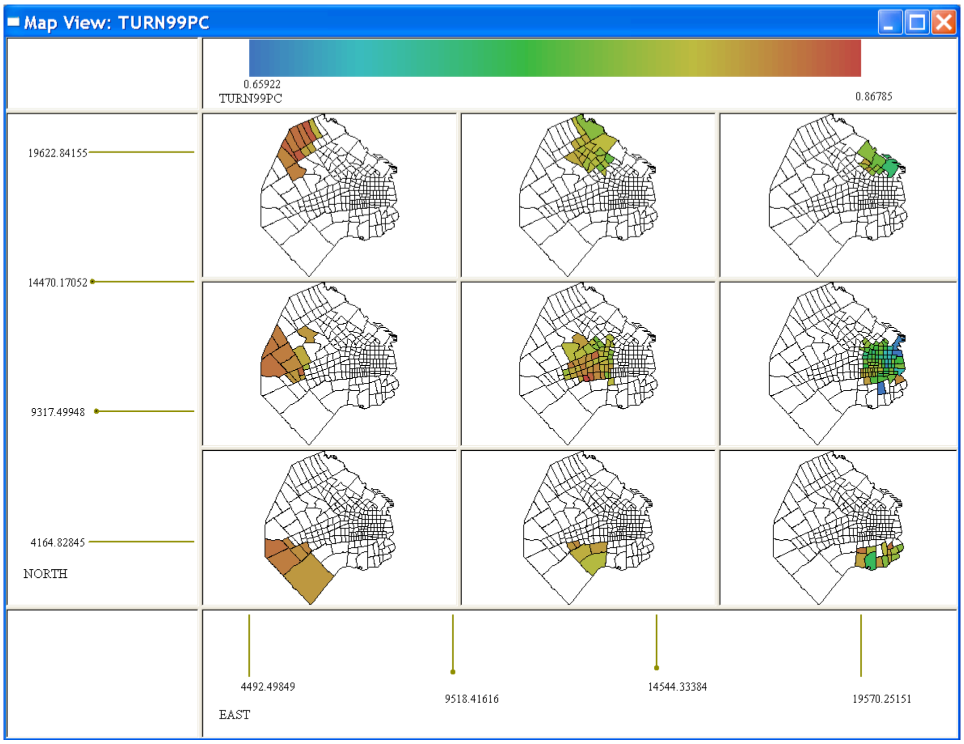


Figure 12.7: Conditional map for AL vote results.

the top of the graph, with the range for the variable TURN99PC indicated.

The purpose of conditioning is to assess the extent to which there is a suggestion of systematic differences in the variable distribution among the subregions. The maps in Figure 12.7 seem to indicate that the higher turnout precincts are on the west side and the lower turnout precincts on the east side. In a more elaborate exploration, other variables would be investigated whose spatial distribution may show similar patterns, in order to begin to construct a set of hypotheses that eventually lead to regression specifications.

12.4 Practice

Consider the three election variables for Buenos Aires more closely (`APR99PC`, `AL99PC`, and `TURN99PC`) to assess the extent to which they show similar or contrasting geographic patterns. Alternatively, revisit the `rosas2001.shp` corn yield example, or any of the other sample data sets you may have considered in Exercise 11.

Exercise 13

Basic Rate Mapping

13.1 Objectives

This exercise illustrates some basic concepts that arise when mapping rates or proportions.

At the end of the exercise, you should know how to:

- create a map for rates constructed from events and population at risk
- save the computed rates to the data table
- create an excess risk map

More detailed information on these operations can be found in the *User's Guide*, pp. 47–49, 51–53.

13.2 Raw Rate Maps

We consider rate maps for the lung cancer data in the 88 counties of the state Ohio that are a commonly used example in recent texts covering disease mapping and spatial statistics.[1] Clear the current project window and load the `ohlung.shp` sample data set, with `FIPSNO` as the `Key`. This brings up the Ohio county base map, shown in Figure 13.1 on p. 93.

The rate maps are special cases of choropleth maps, with a distinct interface. Instead of selecting a rate variable from the data set in the usual

[1] For more extensive discussion and illustration of advanced spatial statistical analyses of this data set, see Waller et al. (1997), Xia and Carlin (1998) and Lawson et al. (2003).

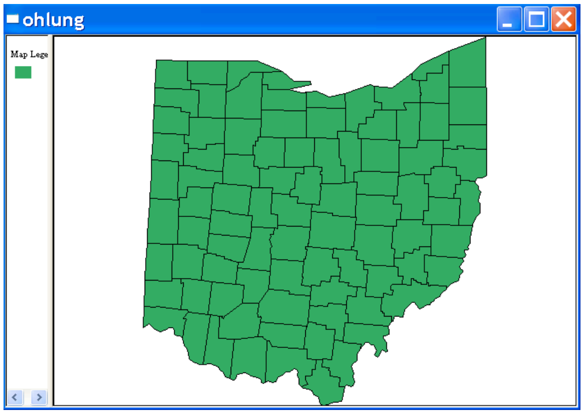


Figure 13.1: Base map for Ohio counties lung cancer data.

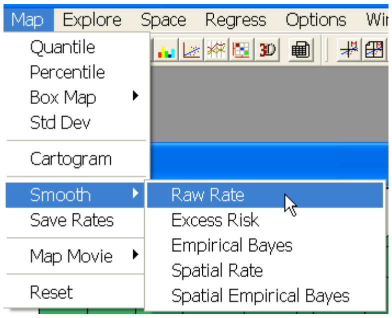
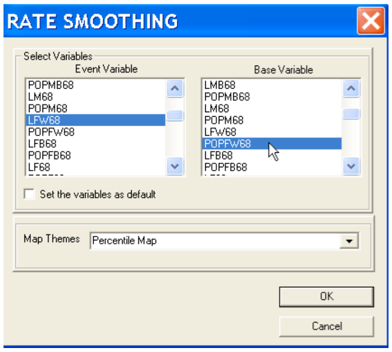


Figure 13.2: Raw rate mapping function.

variable settings dialog, both the `Event` and population at risk (`Base`) are specified and the rate is calculated on the fly.

Select this function from the menu as `Map > Smooth > Raw Rate`, shown in Figure 13.2. Alternatively, right click in any window with the base map and select `Smooth > Raw Rate`. There currently is no matching toolbar icon for rate maps.

The `Rate Smoothing` dialog appears, with a column of candidate `Event` variables and a column of candidate `Base` variables, as in Figure 13.3 on p. 94. Choose `LFW68` as the event (total lung cancer deaths for white females in 1968) and `POPFW68` as the population at risk (total white female population in 1968). Next, make sure to select the proper type of map from the drop down list shown in Figure 13.4 on p. 94. The default is `Percentile`



RATE SMOOTHING

Select Variables

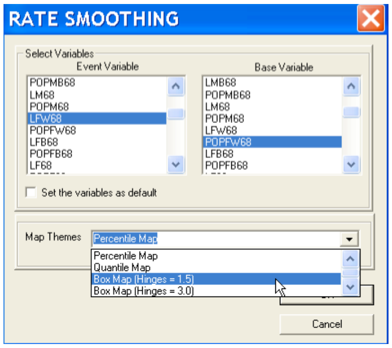
Event Variable: POPMB68, LM68, POPM68, LFW68 (selected), POPFW68, LFB68, POPFB68, LF68

Base Variable: LMB68, POPMB68, LM68, POPM68, LFW68, POPFW68 (selected), LFB68, POPFB68

☐ Set the variables as default

Map Themes: Percentile Map

Figure 13.3: Selecting variables for event and base.



RATE SMOOTHING

Select Variables

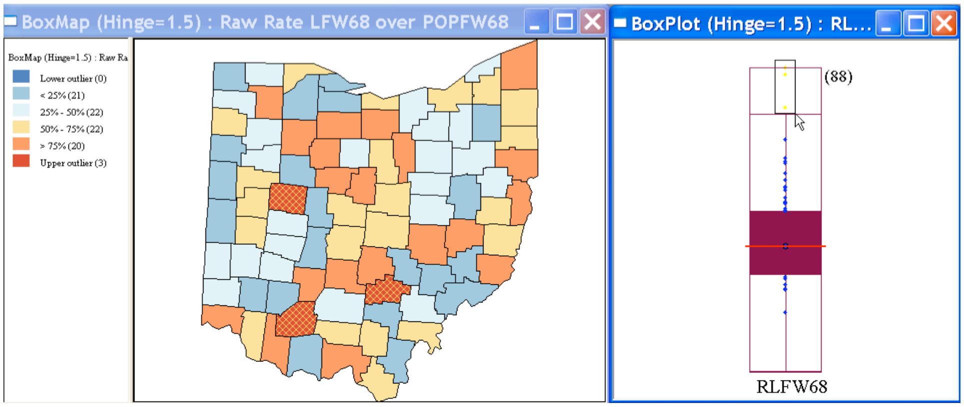
Event Variable: POPMB68, LM68, POPM68, LFW68 (selected), POPFW68, LFB68, POPFB68, LF68

Base Variable: LMB68, POPMB68, LM68, POPM68, LFW68, POPFW68 (selected), LFB68, POPFB68

☐ Set the variables as default

Map Themes: Percentile Map

Figure 13.4: Selecting the type of rate map.



BoxMap (Hinge=1.5) : Raw Rate LFW68 over POPFW68
BoxMap (Hinge=1.5) : Raw Ra
Lower outlier (0)
< 25% (21)
25% - 50% (22)
50% - 75% (22)
> 75% (20)
Upper outlier (3)
BoxPlot (Hinge=1.5) : RL...
(88)
RLFW68

Figure 13.5: Box map for Ohio white female lung cancer mortality in 1968.

Map, but that would not be appropriate in this example (Ohio has 88 counties, which is less than the 100 required for a meaningful percentile map). Instead, select Box Map with a hinge of 1.5, as in Figure 13.4. Finally, click on OK to bring up the box map shown in the left panel of Figure 13.5. Three counties appear as upper outliers with elevated mortality rates. However, due to the inherent variance instability of the rates, these may be spurious. We return to this in Exercise 14.

Even though we have a map for the lung cancer mortality rates, the rates themselves are not available for any other analyses. However, this can be easily accomplished. Right click in the map to bring up the options menu, shown in Figure 13.6 on p. 95. Select Save Rates to create the rate variable. The dialog in Figure 13.7 (p. 95) lets you specify a name for the

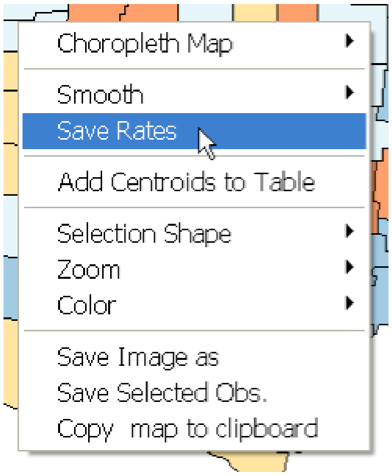
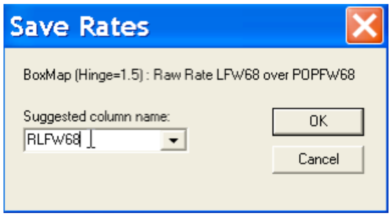


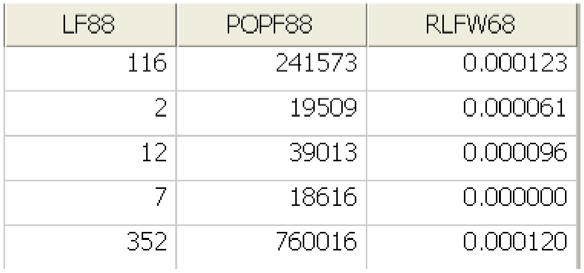
Figure 13.6: Save rates to data table.



BoxMap (Hinge=1.5) : Raw Rate LFW68 over POPFW68

Suggested column name: RLFW68

Figure 13.7: Variable name for saved rates.



LF88	POPF88	RLFW68
116	241573	0.000123
2	19509	0.000061
12	39013	0.000096
7	18616	0.000000
352	760016	0.000120

Figure 13.8: Raw rates added to data table.

variable other than the default. Overwrite the default with `RLFW68` (or any other variable name you can easily recognize) and click on `OK` to add the new variable to the table. Bring the data table to the foreground and verify that a new column has been added, as in Figure 13.8.

Bring up the box plot function (`Explore > Box plot`, or click on the toolbar icon) and select `RLFW68` as the variable name. The result will be the graph shown in the right panel of Figure 13.5 on p. 94. Select the three outliers to check their location in the box map. You can also bring up the data table to find the names of the three counties (Logan, Highland and Hocking) and check whether or not they have unusually small base populations (`POPFW68`).

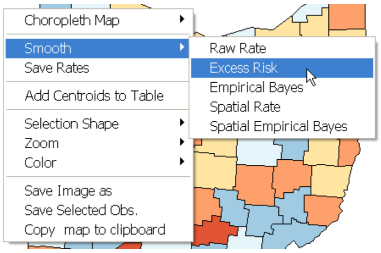


Figure 13.9: Excess risk map function.

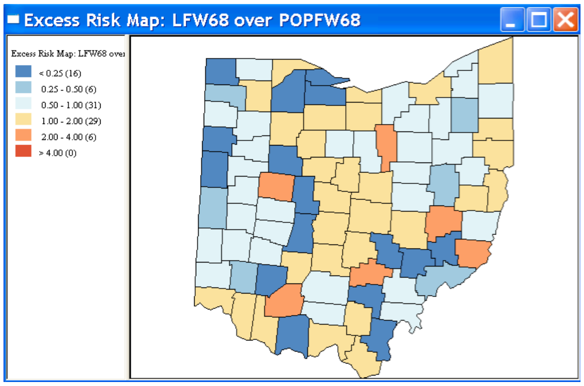


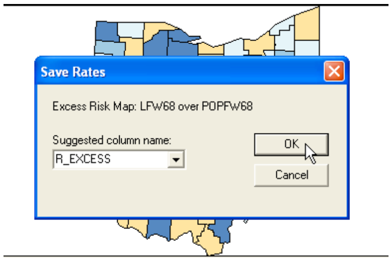
Figure 13.10: Excess risk map for Ohio white female lung cancer mortality in 1968.

13.3 Excess Risk Maps

A commonly used notion in public health analysis is the concept of a standardized mortality rate (SMR), or, the ratio of the observed mortality rate to a national (or regional) standard. *GeoDa* implements this in the form of an `Excess Risk` map as part of the `Map > Smooth` functionality (see Figure 13.9).

The excess risk is the ratio of the observed rate to the average rate computed for all the data. Note that this average is *not* the average of the county rates. Instead, it is calculated as the ratio of the total sum of all events over the total sum of all populations at risk (e.g., in our example, all the white female deaths in the state over the state white female population).

Start this function by selecting it from the `Map` menu, or right clicking on any base map and choosing `Smooth > Excess Risk`. Again, use `LFW68`



Save Rates

Excess Risk Map: LFW68 over POPFW68

Suggested column name: R_EXCESS

Figure 13.11: Save standardized mortality rate.

as the `Event` and `POPFW68` as the `Base` in the variable selection dialog (Figure 13.3 on p. 94). Click on `OK` to bring up the map shown in Figure 13.10 on p. 96. The legend categories in the map are hard coded, with the blue tones representing counties where the risk is less than the state average (excess risk ratio < 1) and the red tones those counties where the risk is higher than the state average (excess risk ratio > 1). The three outlier counties from Figure 13.5 on p. 94 have an excess risk rate between 2 and 4.

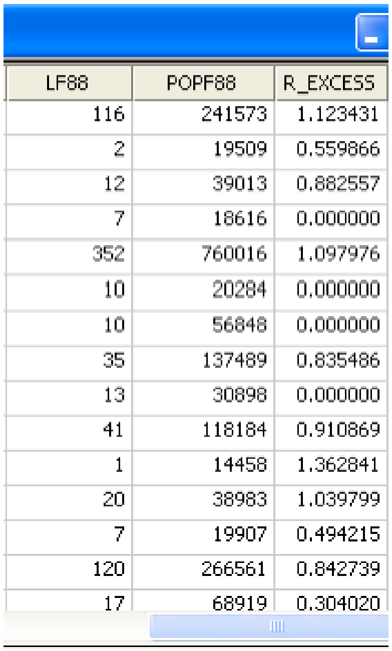
One feature that may throw you off the first time is that this type of map is hard coded and the usual map options (box map, percentile map, as in Figure 13.4 on p. 94) are *ignored*. To construct one of the familiar map types for the excess risk rates (or, standardized mortality ratios), you must first add the computed rates to the data table. Right click on the map and select `Save Rates` as the option, as was illustrated in Figure 13.6 on p. 95 for the raw rates. However, in this instance, the suggested name for the new variable is `R_Excess`, as shown in Figure 13.11. Click `OK` to add the excess risk rate to the data table, shown in Figure 13.12 on p. 98.

The standardized rates are now available for any type of analysis, graph or map. For example, Figure 13.13 (p. 98) illustrates a box map of the excess risk rates (`R_Excess`) just computed. Compare this to the box map in Figure 13.5 on p. 94. What do you think is going on?[2]

13.4 Practice

Experiment with the rate computation and standardized mortality rates for other population categories and/or years in the Ohio lung cancer data set. Alternatively, consider lip cancer death rates contained in the famous data set for 56 Scottish districts (load `scotlip.shp` with `CODENO` as the `Key`),

[2] Note that the excess rate is nothing but a rescaled raw rate.



LF88	POPF88	R_EXCESS
116	241573	1.123431
2	19509	0.559866
12	39013	0.882557
7	18616	0.000000
352	760016	1.097976
10	20284	0.000000
10	56848	0.000000
35	137489	0.835486
13	30898	0.000000
41	118184	0.910869
1	14458	1.362841
20	38983	1.039799
7	19907	0.494215
120	266561	0.842739
17	68919	0.304020

Figure 13.12: SMR added to data table.

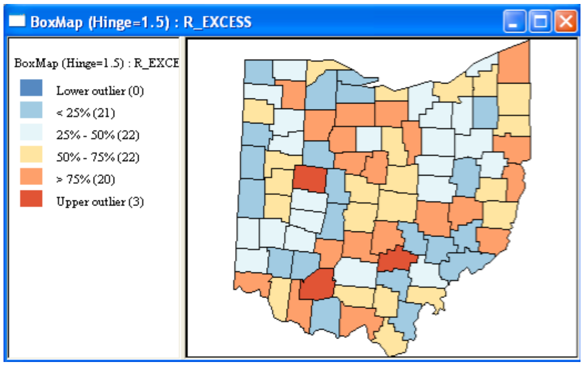


Figure 13.13: Box map for excess risk rates.

or the equally famous SIDS data for 100 North Carolina counties (load `sids.shp` with `FIPSNO` as the `Key`). The homicide data sets and Buenos Aires election results also lend themselves well to this type of analysis.

Exercise 14

Rate Smoothing

14.1 Objectives

This exercise illustrates some techniques to smooth rate maps to correct for the inherent variance instability of rates.

At the end of the exercise, you should know how to:

- create a map with rates smoothed by the Empirical Bayes method
- create a k-nearest neighbors spatial weights file
- create a map with spatially smoothed rates
- save the computed rates to the data table

More detailed information on these operations can be found in the *User's Guide*, pp. 49–50.

14.2 Empirical Bayes Smoothing

We continue to use the Ohio lung cancer example. If you tried a different data set to practice a previous exercise, first clear all windows and load the shape file with `ohlung.shp` (use `FIPSNO` as the `Key`). The first smoothing technique uses an Empirical Bayes (EB) approach, whereby the raw rates are "shrunk" towards the overall statewide average. In essense, the EB technique consists of computing a weighted average between the raw rate for each county and the state average, with weights proportional to the underlying

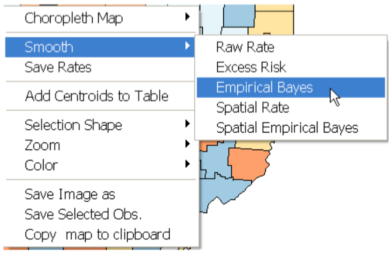
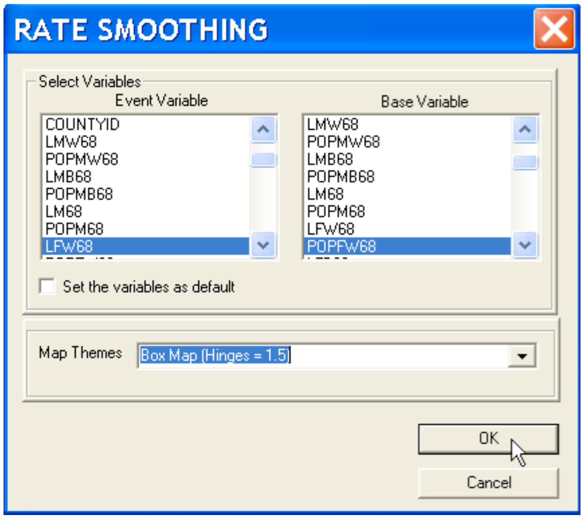


Figure 14.1: Empirical Bayes rate smoothing function.


RATE SMOOTHING

Select Variables

Event Variable

COUNTYID
LMW68
POPMW68
LMB68
POPMB68
LM68
POPM68
LFW68

Base Variable

LMW68
POPMW68
LMB68
POPMB68
LM68
POPM68
LFW68
POPFW68

□ Set the variables as default

Map Themes: Box Map (Hinges = 1.5)

OK

Cancel

Figure 14.2: Empirical Bayes event and base variable selection.

population at risk. Simply put, small counties (i.e., with a small population at risk) will tend to have their rates adjusted considerably, whereas for larger counties the rates will barely change.[1]

Invoke this function from the `Map` menu, or by right clicking in a current map and selecting `Smooth > Empirical Bayes` (Figure 14.1). This brings up the same variable selection dialog as in the previous rate mapping exercises, shown in Figure 14.2.

Select `LFW68` as the `Event` and `POPFW68` as the `Base`, and choose a `Box`

[1] For methodological details and further illustrations, see Bailey and Gatrell (1995), pp. 303-308, and Anselin et al. (2004b).

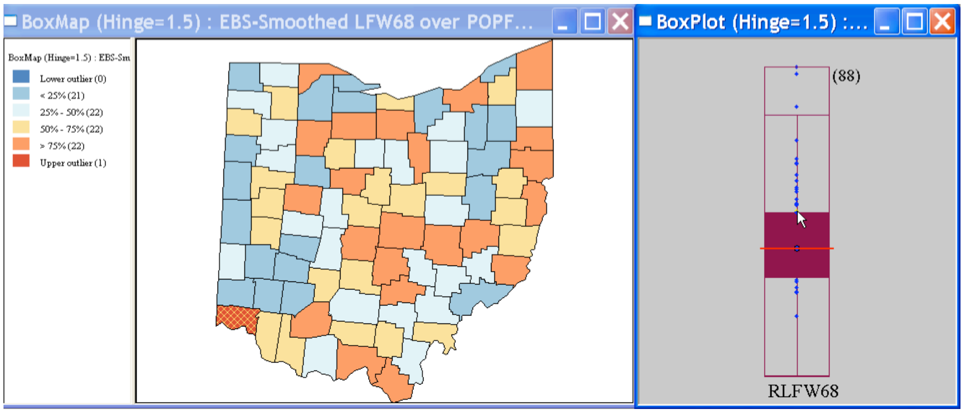


Figure 14.3: EB smoothed box map for Ohio county lung cancer rates.

Map with hinge 1.5, as illustrated in Figure 14.2. Click OK to bring up the smoothed box map shown in the left panel of Figure 14.3. Compare this to the original box map in Figure 13.5 on p. 94. Note how none of the original outliers survive the smoothing, whereas a new outlier appears in Hamilton county (in the southwestern corner).

Create a box plot for the raw rate RLFW68, computed in Exercise 13 (if you did not save the rate variable at the time, create a box map for the raw rate and subsequently save the rate). Click on the outlier in the box map and locate its position in the box plot. As shown by the arrow in the right panel of Figure 14.3, this observations is around the 75 percentile in the raw rate map. Since many of the original outlier counties have small populations at risk (check in the data table), their EB smoothed rates are quite different (lower) from the original. In contrast, Hamilton county is one of the most populous counties (it contains the city of Cincinnati), so that its raw rate is barely adjusted. Because of that, it percolates to the top of the distribution and becomes an outlier. You can systematically select observations in the box plot for the raw rates and compare their position in the cumulative distribution to the one for the smoothed rates to see which observations are affected most. Use the table to verify that they have small populations.

14.3 Spatial Rate Smoothing

Spatial rate smoothing consists of computing the rate in a moving window centered on each county in turn. The moving window includes the county

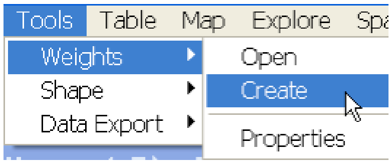


Figure 14.4: Spatial weights creation function.

as well as its *neighbors*. In *GeoDa* the neighbors are defined by means of a spatial weights file. This is discussed in more detail in Exercises 15 and 16. However, to be able to illustrate the spatial smoothing, a quickstart on creating spatial weights is provided next.

14.3.1 Spatial Weights Quickstart

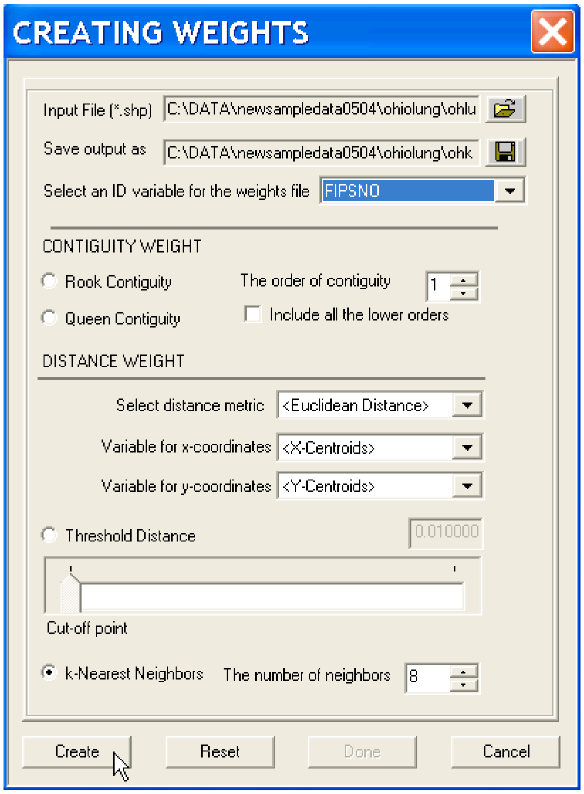
We will construct a simple spatial weights file consisting of the 8 nearest neighbors for each county. Click on the `Create weights` icon in the toolbar, or invoke the function from the menu as `Tools > Weights > Create` (see Figure 14.4).

This brings up the weights creation dialog, shown in Figure 14.5 on p. 103. Enter the path to the `ohlung.shp` file as the input file, `ohk8` as the output file (a file extension of `GWT` will be added by the program), and select `FIPSNO` in the drop down list for the `ID` variable. Leave all the options under `Distance Weight` to their default values (the program will compute the centroids for the Ohio counties to calculate the needed distances).[2] Finally, check the radio button next to `k-Nearest Neighbors` and change the number of neighbors to `8`.

Click `Create` to start the process and `Done` when the progress bar (the blue bar in the `shp -> gwt` window) is complete. You now have a weights file ready to use.

Before starting the spatial smoothing, you must *load* the spatial weights to make them available to the program. Click on the `Load weights` toolbar button, or, from the menu, select `Tools > Weights > Open`, as in Figure 14.6 (p. 103). Next, click the radio button next to `Select from file` in the select weight dialog (Figure 14.7, p. 103), and enter the file name for the weights file (`ohk8.GWT`). Click `OK` to load the weights file. You are now ready to go.

[2] Since the `ohlung.shp` file is in projected (UTM) coordinates, the centroids will be calculated properly and the distance metric can be kept as `Euclidean distance`.



CREATING WEIGHTS

Input File (*.shp) C:\DATA\newsampledata0504\ohiolung\ohlu

Save output as C:\DATA\newsampledata0504\ohiolung\ohk

Select an ID variable for the weights file FIPSNO

CONTIGUITY WEIGHT

○ Rook Contiguity The order of contiguity 1

○ Queen Contiguity □ Include all the lower orders

DISTANCE WEIGHT

Select distance metric <Euclidean Distance>

Variable for x-coordinates <X-Centroids>

Variable for y-coordinates <Y-Centroids>

○ Threshold Distance 0.010000

Cut-off point

● k-Nearest Neighbors The number of neighbors 8

Figure 14.5: Spatial weights creation dialog.

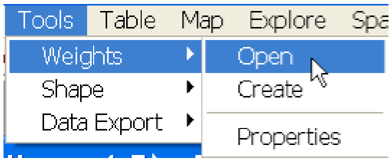
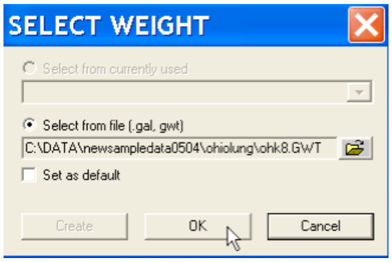


Figure 14.6: Open spatial weights function.



SELECT WEIGHT

○ Select from currently used

● Select from file (.gal, gwt)

C:\DATA\newsampledata0504\ohiolung\ohk8.GWT

□ Set as default

Figure 14.7: Select spatial weight dialog.

14.3.2 Spatially Smoothed Maps

With a spatial weights file loaded, invoke the spatial smoothing from the `Map` menu, or by right clicking in any map and choosing `Smooth > Spatial`

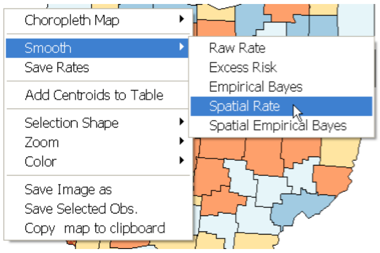


Figure 14.8: Spatial rate smoothing function.

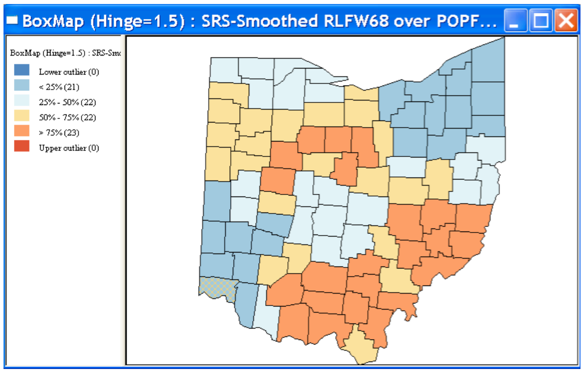


Figure 14.9: Spatially smoothed box map for Ohio county lung cancer rates.

`Rate`, as in Figure 14.8. If there is no currently loaded spatial weights file, an error message will appear at this point. If all is well, select `LFW68` as the event and `POPFW68` as the base in the dialog (see Figure 14.2 on p. 100). As before, select the `Box Map` with a hinge of 1.5 from the drop down list and click `OK` to create the map. The smoothed map appears as in Figure 14.9.

A spatially smoothed maps emphasizes broad regional patterns in the map. Note how there are no more outliers. Moreover, due to the averaging with the 8 neighbors, Hamilton county (the outlier in the EB smoothed map) is part of a region of second quartile counties.

14.4 Practice

As in Excercise 13, further explore the differences and similarities in spatial patterns between raw rates and smoothed rates for various population categories and/or years in the Ohio lung cancer data set. Alternatively, use any of the other sample data sets employed previously. Focus on the difference in outliers between the raw rate map and EB smoothed map and on the broad regional patterns that emerge from the spatially smoothed map.

Exercise 15

Contiguity-Based Spatial Weights

15.1 Objectives

This exercise begins the illustration of spatial weights manipulation with the construction of contiguity-based spatial weights, where the definition of neighbor is based on sharing a common boundary.

At the end of the exercise, you should know how to:

- create a first order contiguity spatial weights file from a polygon shape file, using both rook and queen criteria
- analyze the connectivity structure of the weights in a histogram
- turn a first order contiguity weights file into higher order contiguity

More detailed information on these operations can be found in the *User's Guide*, pp. 78–83, 86–87, and *Release Notes*, pp. 18–20.

15.2 Rook-Based Contiguity

Start this exercise by opening the sample data shape file containing census variables for 403 census tracts in Sacramento, CA (use `sacramentot2.shp` with `POLYID` as the `Key`). The initial base map should look like Figure 15.1 on p. 107.

Invoke the weights construction functionality from the menu, by selecting `Tools > Weights > Create`, as in Figure 15.2 on p. 107. This can also be

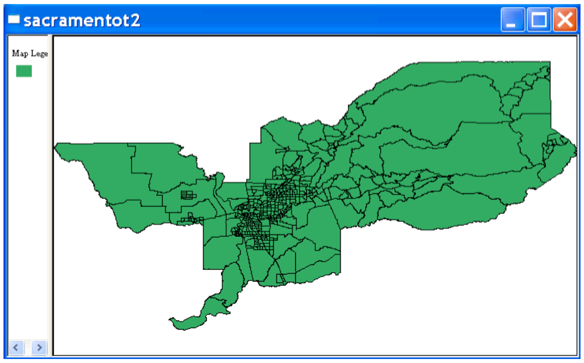


Figure 15.1: Base map for Sacramento census tract data.

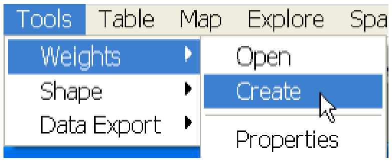
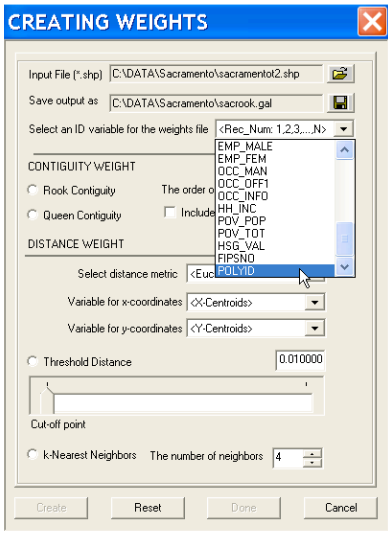


Figure 15.2: Create weights function.

executed without having a current project. In other words, it is not necessary to load the shape file in order to create the weights. Alternatively, from within a project, this function can be started by clicking on the matching toolbar button.

The weights creation function generates a dialog, in which the relevant options are specified. First are the names for the input shape file (for contiguity weights this must be a *polygon* shape file) and the name for the weights file. Enter `sacramentot2.shp` for the former, and `sacrook` for the latter, as shown in Figure 15.3 on p. 108. A file extension of `GAL` will be added to the weights file by the program. It is also very important to specify the `Key` variable: enter `POLYID`, as in Figure 15.3 on p. 108. While this is not *absolutely* necessary, it ensures a complete match between the data in the table and their corresponding contiguity entries in the weights file.

The only other action needed for a rook contiguity weights file is to check the radio button next to `Rook Contiguity`, as in Figure 15.4 on p. 108. Next, click on `Create` to start the construction of the weights. A progress bar will appear, as in Figure 15.5 on p. 109, indicating when the process is completed (this is typically done in a very short time). Finish the procedure



Input File (*.shp) C:\DATA\Sacramento\sacramentot2.shp

Save output as C:\DATA\Sacramento\sacrook.gal

Select an ID variable for the weights file <Rec_Num: 1,2,3,...,N>

Select distance metric <Euc

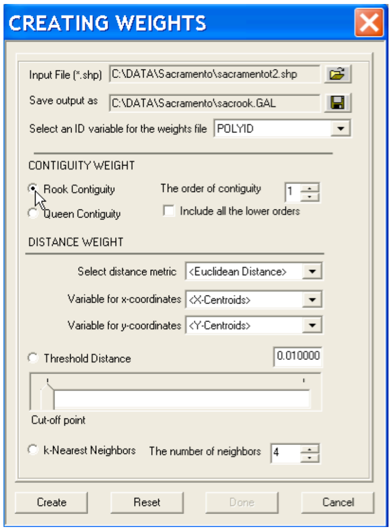
Variable for x-coordinates <X-Centroids>

Variable for y-coordinates <Y-Centroids>

○ Threshold Distance 0.010000

○ k-Nearest Neighbors The number of neighbors 4

Figure 15.3: Weights creation dialog.



Input File (*.shp) C:\DATA\Sacramento\sacramentot2.shp

Save output as C:\DATA\Sacramento\sacrook.GAL

Select an ID variable for the weights file POLYID

● Rook Contiguity

○ Queen Contiguity

The order of contiguity 1

□ Include all the lower orders

Select distance metric <Euclidean Distance>

Variable for x-coordinates <X-Centroids>

Variable for y-coordinates <Y-Centroids>

○ Threshold Distance 0.010000

○ k-Nearest Neighbors The number of neighbors 4

Figure 15.4: Rook contiguity.

by clicking on Done (see Figure 15.5 on p. 109) to return to the standard interface.

The resulting GAL format spatial weights file is a simple text file that

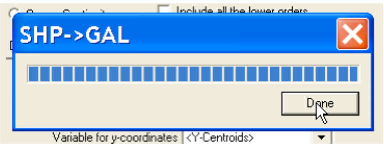
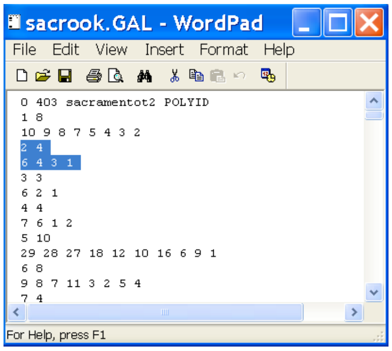


Figure 15.5: GAL shape file created.



```
0 403 sacramentot2 POLYID
1 8
10 9 8 7 5 4 3 2
2 4
6 4 3 1
3 3
6 2 1
4 4
7 6 1 2
5 10
29 28 27 18 12 10 16 6 9 1
6 8
9 8 7 11 3 2 5 4
7 4
```

Figure 15.6: Contents of GAL shape file.

can be edited with any text editor or word processor (make sure to save it as a *text* file). For example, for the Sacramento census tracts, using `POLYID` as the `Key`, the partial contents of the file `sacrook.gal` are shown in Figure 15.6. The first line of this file is a header line, that contains `0` (a flag reserved for future use), the number of observations (`403`), the name of the polygon shape file from which the contiguity structure was derived (`sacramentot2`), and the name of the `Key` variable (`POLYID`).

Note that *GeoDa* will generate an error message when the shape file and the weights file are not in the same directory. It is always possible to edit this header line if either the shape file name or `Key` variable would have changed in the course of an analysis.[1]

Open the file `sacrook.gal` in any text editor and focus on the neighbors for the observation with `POLYID 2` (`FIPS` code `6061020106`). Open the table (click on the `Table` toolbar icon) and select the tract. The selected location should be the one identified by the pointer in the base map in Figure 15.7 on p. 110. It has `4` neighbors, indicated by the second entry in the first

[1] This is quite common and often a source of confusion, since the error message refers to insufficent memory, but not to the mismatch in the header line.

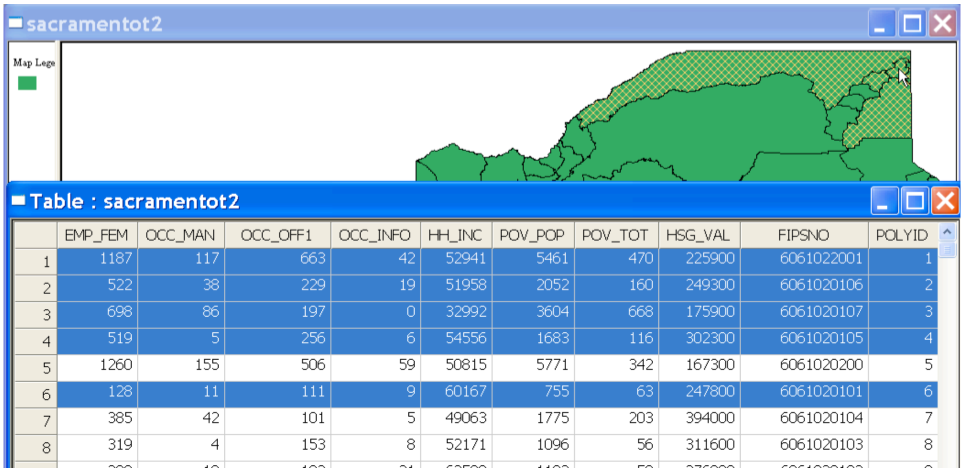


Table : sacramentot2

	EMP_FEM	OCC_MAN	OCC_OFF1	OCC_INFO	HH_INC	POV_POP	POV_TOT	HSG_VAL	FIPSNO	POLYID
1	1187	117	663	42	52941	5461	470	225900	6061022001	1
2	522	38	229	19	51958	2052	160	249300	6061020106	2
3	698	86	197	0	32992	3604	668	175900	6061020107	3
4	519	5	256	6	54556	1683	116	302300	6061020105	4
5	1260	155	506	59	50815	5771	342	167300	6061020200	5
6	128	11	111	9	60167	755	63	247800	6061020101	6
7	385	42	101	5	49063	1775	203	394000	6061020104	7
8	319	4	153	8	52171	1096	56	311600	6061020103	8

Figure 15.7: Rook contiguity structure for Sacramento census tracts.

highlighted line in Figure 15.6. The values for the `Key` variable (`POLYID`) of these neigbhors are listed in the second line. Select these tracts in the table (use `shift-click`), as in Figure 15.7, and the corresponding locations will be highlighted in the map. Try this for some other locations as well. Note how in some cases the *rook* criterion eliminates *corner* neighbors, i.e., tracts that do not have a full boundary segment in common. Such locations will be included by the *queen* criterion, covered in Section 15.4 (p. 112).

15.3 Connectivity Histogram

Select `Tools > Weights > Properties` from the menu (Figure 15.8 on p. 111) to create a histogram that reflects the connectivity distribution for the census tracts in the data set. Alternatively, click on the matching toolbar button.

The histogram is very important to detect *strange* features of this distribution, which may affect spatial autocorrelation statistics and spatial regression specifications. Two features in particular warrant some attention. One is the occurrence of *islands*, or unconnected observations, the other a bimodal distribution, with some locations having very few (such as one) and others very many neighbors.

Selecting the weights properties function brings up a dialog to specify the weights file, as in Figure 15.9 on p. 111. Enter `sacrook.gal` as the file and click `OK`.

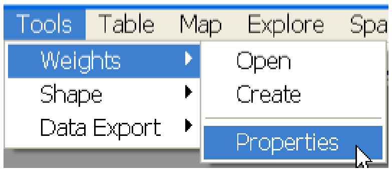
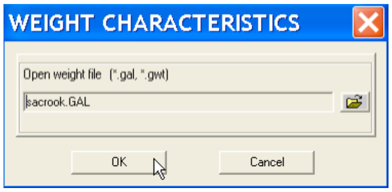


Figure 15.8: Weights properties function.



Open weight file (*.gal, *.gwt)

sacrook.GAL

Figure 15.9: Weights properties dialog.

The resulting histogram is as in Figure 15.10 on p. 112, shown next to the Sacramento tract base map. It describes the distribution of locations (the number of observations in each category is shown at the top of the corresponding bar) by number of neighbors (shown in the legend). For example, the right most bar corresponds to a tract with 14 neighbors. Click on the histogram bar to find the location of the tract in the map, as illustrated in Figure 15.10. Alternatively, select a location in the map and find out from the histogram how many neighbors it has. Use the map zoom feature (right click on the map, choose `zoom` and create a rectangular selection shape around a tract) to see the neighbor structure more clearly. Experiment by selecting several tracts and comparing their connectivity histogram to the overall distribution.

To illustrate a connectivity structure with islands, load the data set for the 56 Scottish districts (load `scotlip.shp` with `CODENO` as the `Key`) and create a rook contiguity file (e.g., `scotrook.gal`). Construct the connectivity histogram and select the left most bar, corresponding to zero neighbors (or islands). As shown in Figure 15.11 on p. 112, these are indeed the three island districts in the data set.

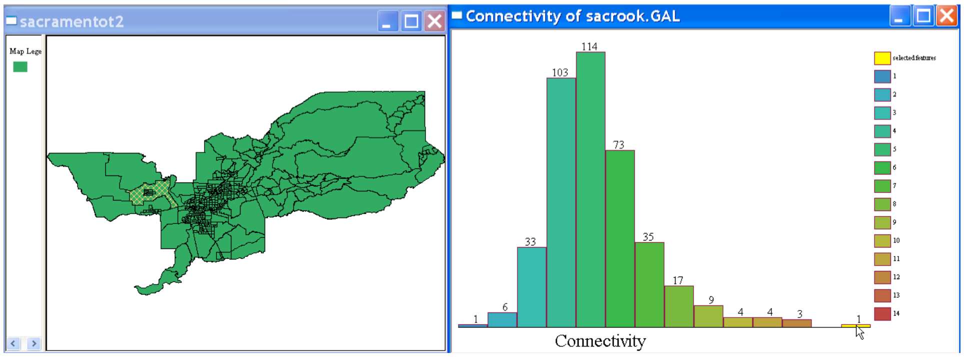


Figure 15.10: Rook contiguity histogram for Sacramento census tracts.

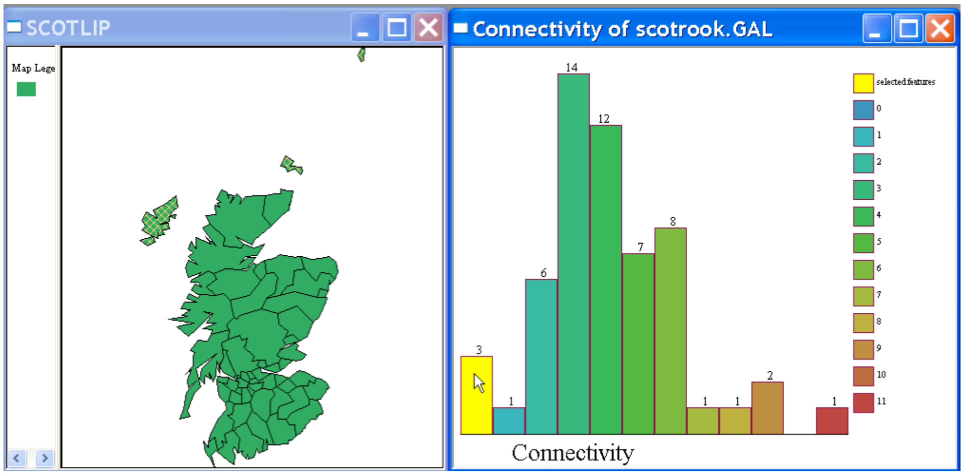
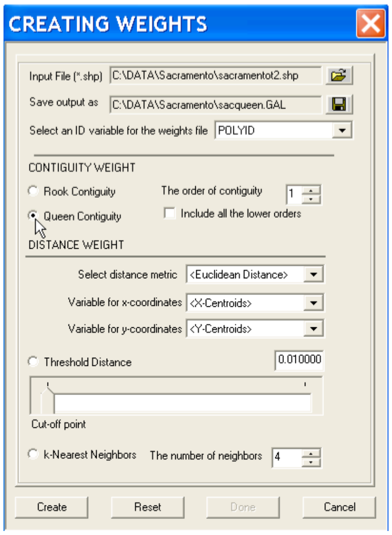


Figure 15.11: Islands in a connectivity histogram.

15.4 Queen-Based Contiguity

The creation of a contiguity weights file that uses the *queen* criterion to define neigbhors proceeds in the same fashion as for the rook criterion.[2] As before, select `Tools > Weights > Create` (see Figure 15.2 on p. 107) to bring up the weights creation dialog shown in Figure 15.12 on p. 113. Select `sacramentot2.shp` as the input shape file and specify `sacqueen` as the name

[2] The *queen* criterion determines neighboring units as those that have *any* point in common, including both common boundaries and common corners. Therefore, the number of neighbors for any given unit according to the queen criterion will be equal to or greater than that using the rook criterion.



Input File (*.shp) C:\DATA\Sacramento\sacramentot2.shp

Save output as C:\DATA\Sacramento\sacqueen.GAL

Select an ID variable for the weights file POLYID

○ Rook Contiguity

● Queen Contiguity

The order of contiguity 1

□ Include all the lower orders

Select distance metric <Euclidean Distance>

Variable for x-coordinates <X-Centroids>

Variable for y-coordinates <Y-Centroids>

○ Threshold Distance 0.010000

○ k-Nearest Neighbors The number of neighbors 4

Figure 15.12: Queen contiguity.

for the output file. Make sure to set the ID variable to POLYID. Select the radio button next to Queen Contiguity, as shown in Figure 15.12, and click on Create. The same progress bar as before will appear (Figure 15.5 on p. 109). Click on Done to complete the process.

Compare the connectivity structure between the rook and queen criterion for the Sacramento data (see Section 15.3). The two histograms are shown in Figure 15.13 on p. 114. Click on the bar corresponding to five neighbors for the rook criterion and note how the distribution for queen has five or *more* neighbors. Check the selection in the map to find the tracts where the difference occurs.

15.5 Higher Order Contiguity

Proceed in the same fashion to construct spatial weights files for higher order contiguity. The weights creation dialog is identical. Select the radio button for either rook or queen contiguity and the order of contiguity. For example, in Figure 15.14 on p. 114, second order contiguity is chosen for a rook criterion.

Note the check box under the order of contiguity. *GeoDa* allows two defnitions for higher order contiguity. One is *pure* and does not include

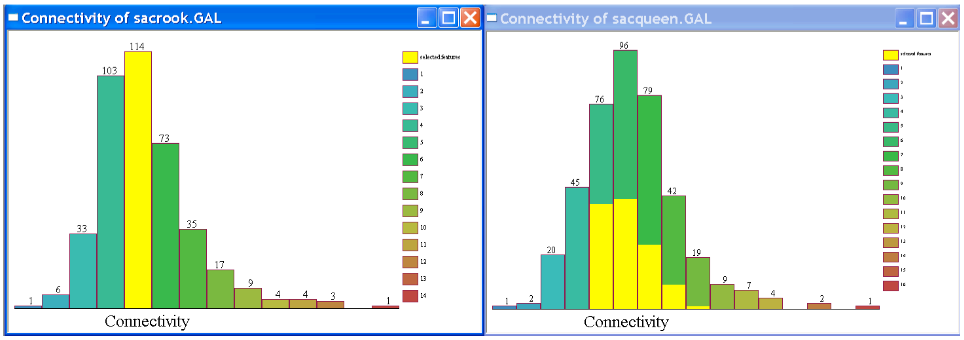
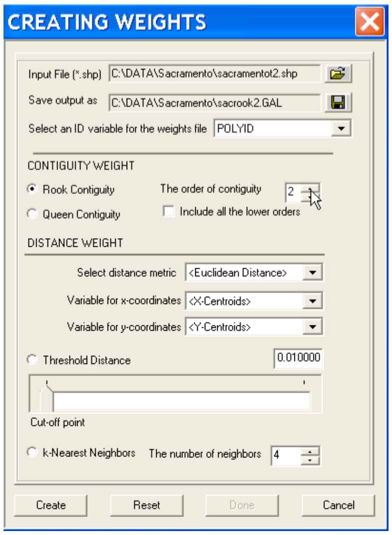


Figure 15.13: Comparison of connectedness structure for rook and queen contiguity.



Input File (*.shp) C:\DATA\Sacramento\sacramentot2.shp

Save output as C:\DATA\Sacramento\sacrook2.GAL

Select an ID variable for the weights file POLYID

CONTIGUITY WEIGHT

● Rook Contiguity The order of contiguity 2

○ Queen Contiguity □ Include all the lower orders

DISTANCE WEIGHT

Select distance metric <Euclidean Distance>

Variable for x-coordinates <X-Centroids>

Variable for y-coordinates <Y-Centroids>

○ Threshold Distance 0.010000

Cut-off point

○ k-Nearest Neighbors The number of neighbors 4

Figure 15.14: Second order rook contiguity.

locations that were also contiguous of a lower order (this is the textbook definition of higher order contiguity). The other is *cumulative*, and includes all lower order neighbors as well. Experiment with this function by creating a pure and a cumulative second order weights file for Sacramento. Compare the connectedness structure between the two. Figures 15.15 and 15.16 on p. 115 show the connectivity histograms for second order rook contiguity in the two cases, with the same observations highlighted as before (locations with 5 first order rook neighbors).

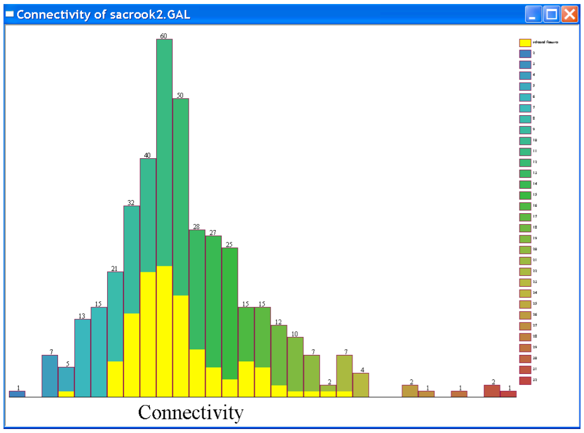


Figure 15.15: Pure second order rook connectivity histogram.

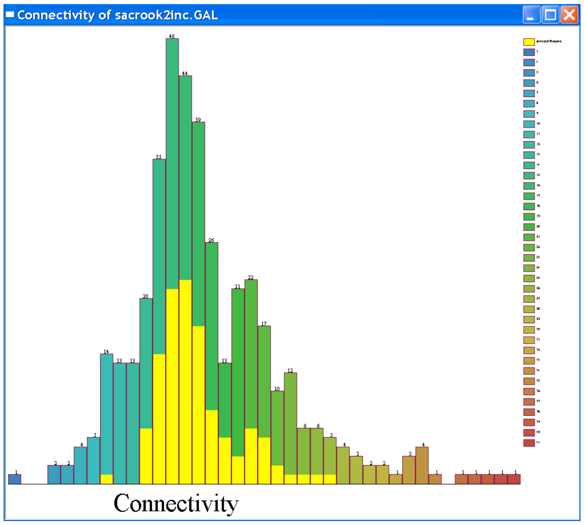


Figure 15.16: Cumulative second order rook connectivity histogram.

15.6 Practice

Practice creating rook and queen contiguity weights as well as higher order weights for any of the *polygon* shape files contained in the sample data set collection. These operations will be needed over and over again in the analysis of spatial autocorrelation and the estimation of spatial regression models. Create some higher order contiguity weights as well. Check the connectedness structure and use the linking functionality to find the number of neighbors for selected locations in the map.

Exercise 16

Distance-Based Spatial Weights

16.1 Objectives

This exercise illustrates the construction of distance-based spatial weights, where the definition of neighbor is based on the distance between points, or between polygon centroids.

At the end of the exercise, you should know how to:

- create a distance-based spatial weights file from a point shape file, by specifying a distance band
- adjust the critical distance
- create a spatial weights file based on a k-nearest neighbor criterion

More detailed information on these operations can be found in the *User's Guide*, pp. 83–85, and *Release Notes*, pp. 18–19.

16.2 Distance-Band Weights

Begin this exercise by loading the point shape file with the centroids for 506 census tracts of the Boston housing sample data set (enter `boston.shp` for the shape file and `ID` as the `Key` variable). The resulting base map should be as in Figure 16.1. As for contiguity weights, the process of constructing distance-based weights can be started without having a project loaded, directly from the `Tools` menu. Within a project, this same approach can

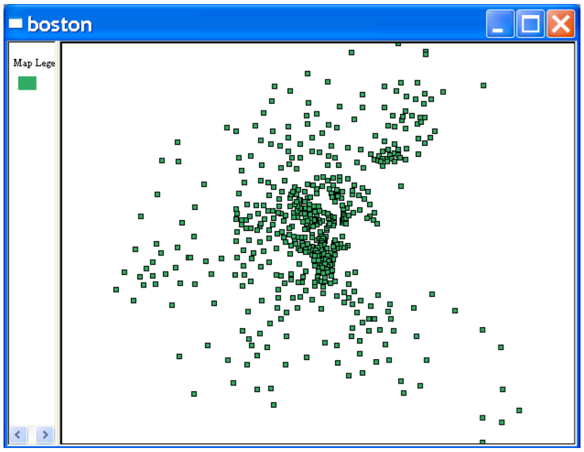


Figure 16.1: Base map for Boston census tract centroid data.

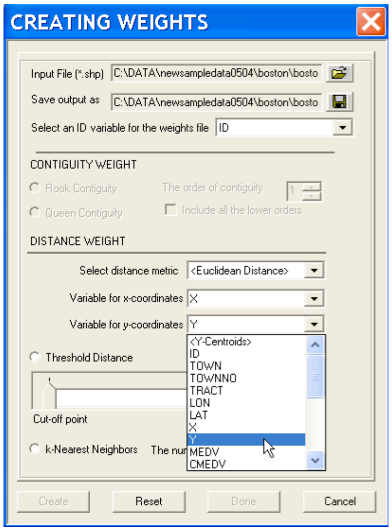
be used (see Figure 15.2 on p. 107), or, alternatively, the matching toolbar button can be clicked.

Select `Tools > Weights > Create` to open the weights creation dialog shown in Figure 16.2 on p. 119. Enter `boston.shp` for the input file, `bostondist` for the name of the spatial weights file (a file extension of `GWT` will be added by the program), and specify `ID` as the `ID` variable. Next, move to the part of the dialog that pertains to `Distance Weight`. Leave the default to `<Euclidean Distance>`, since the Boston data set contains the coordinates in UTM projection. If the points were in latitude and longitude, you would need to select the `<Arc Distance>` option.

Next, specify the variable names for the x and y coordinates as `X` and `Y`, as illustrated in Figure 16.2. Note that, in contrast to pure contiguity weights, distance-based spatial weights can be calculated for both point shape files as well as polygon shape files. For the latter, if no coordinate variables are specified, the polygon centroids will be calculated and used as the basis for the distance calculation.[1]

Proceed by checking the radio button next to `Threshold distance`, as in Figure 16.3 on p. 119. Note how the value in the text box changes to `3.972568`. This is the minimum distance required to ensure that each location has *at least* one neighbor. If the threshold distance is set to a smaller value, *islands* will result. Typically, some experimentation (and checking of

[1] However, for unprojected maps, the resulting centroids will not be correct, only approximate. For proper computation of centroids in *GeoDa*, the map must be projected.



Input File (*.shp) C:\DATA\newsampledata0504\boston\bosto

Save output as C:\DATA\newsampledata0504\boston\bosto

Select an ID variable for the weights file ID

CONTIGUITY WEIGHT

○ Rook Contiguity The order of contiguity 1

○ Queen Contiguity □ Include all the lower orders

DISTANCE WEIGHT

Select distance metric <Euclidean Distance>

Variable for x-coordinates X

Variable for y-coordinates Y

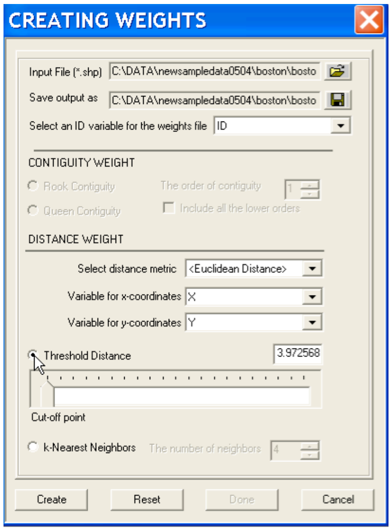
<Y-Centroids>
ID
TOWN
TOWNNO
TRACT
LON
LAT
X
Y
MEDV
CMEDV

○ Threshold Distance

Cut-off point

○ k-Nearest Neighbors The nur

Figure 16.2: Distance weights dialog.



Input File (*.shp) C:\DATA\newsampledata0504\boston\bosto

Save output as C:\DATA\newsampledata0504\boston\bosto

Select an ID variable for the weights file ID

CONTIGUITY WEIGHT

○ Rook Contiguity The order of contiguity 1

○ Queen Contiguity □ Include all the lower orders

DISTANCE WEIGHT

Select distance metric <Euclidean Distance>

Variable for x-coordinates X

Variable for y-coordinates Y

● Threshold Distance 3.972568

Cut-off point

○ k-Nearest Neighbors The number of neighbors 4

Figure 16.3: Threshold distance specification.

the connecitivity structure) is needed to specify useful values larger than the minimum threshold.

Click on `Create` to start the process. A progress bar will appear, as in

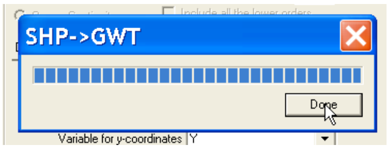
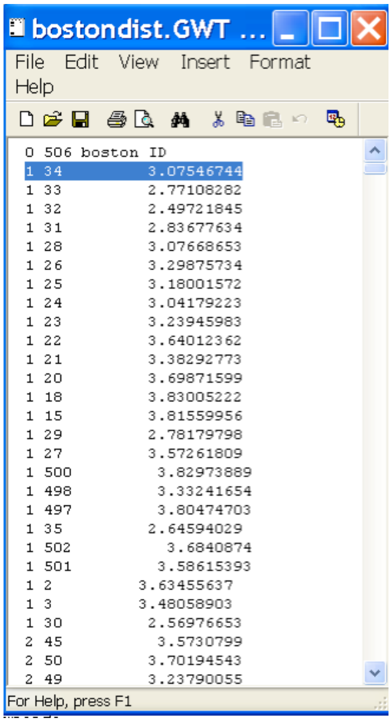


Figure 16.4: GWT shape file created.



```
0 506 boston ID
1 34          3.07546744
1 33          2.77108282
1 32          2.49721845
1 31          2.83677634
1 28          3.07668653
1 26          3.29875734
1 25          3.18001572
1 24          3.04179223
1 23          3.23945983
1 22          3.64012362
1 21          3.38292773
1 20          3.69871599
1 18          3.83005222
1 15          3.81559956
1 29          2.78179798
1 27          3.57261809
1 500          3.82973889
1 498          3.33241654
1 497          3.80474703
1 35          2.64594029
1 502           3.6840874
1 501          3.58615393
1 2          3.63455637
1 3          3.48058903
1 30          2.56976653
2 45           3.5730799
2 50          3.70194543
2 49          3.23790055
```

Figure 16.5: Contents of GWT shape file

Figure 16.4. Click on `Done` to return to the standard interface.

The `GWT` format file that contains the spatial weights information is a standard text file, just as the previously considered `GAL` file. Its format is slightly different, as illustrated in Figure 16.5. Open the file `bostondist.gwt` you just created with any text editor or word processor and check its contents. As in Figure 15.6 (p. 109), the first line in the file is a header line, that contains the same information as for the `GAL` file (a place holder, the number of observations, file name for the shape file and `ID` variable). The

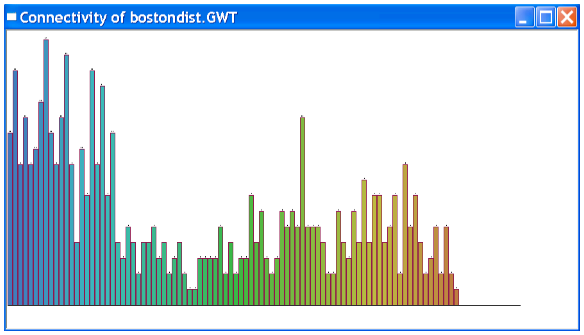


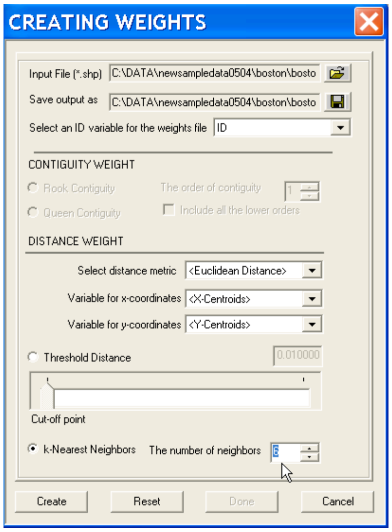
Figure 16.6: Connectivity for distance-based weights.

remainder of the file contains, for each defined neighbor pair, the “origin” `ID`, the “destination” `ID` and the *distance* between the two points. Note that this third entry is *not* the value of the spatial weights. In the current version, this is ignored by *GeoDa* and only the existence of a neighbor relation is taken into account.

Check the connectivity structure for the `bostondist.gwt` weights using the techniques outlined in Section 15.3 (p. 110). The result should be as in Figure 16.6. Note how the distribution has a much wider range, compared to the contiguity-based weights. In practice, this is typical for distance-based weights when the points have an irregular distribution (i.e., some points are clustered, whereas others are far apart). In such a situation, the minimum threshold distance needed to avoid islands may be too large for many (most) of the locations in the data set. This minimum distance is driven by the pair of points that are the furthest apart, which may not be representative for the rest of the distribution. In such cases, care is needed in the specification of the distance threshold, and the use of k-nearest neighbor weights may be more appropriate.

16.3 k-Nearest Neighbor Weights

Start the process of constructing a k-nearest neighbor weights file in the same fashion as for the other weights (`Tools > Weights > Create`). This brings up a weights dialog, shown in Figure 16.7 on p. 122. In the weights create dialog, enter `boston.shp` for the input file, `bostonk6` for the output file, and `ID` as the `ID` variable. Move down in the dialog and check the radio button next to `k-nearest neighbors`, as in Figure 16.7. Move the value for



CREATING WEIGHTS

Input File (*.shp) C:\DATA\newsampledata0504\boston\bosto

Save output as C:\DATA\newsampledata0504\boston\bosto

Select an ID variable for the weights file ID

CONTIGUITY WEIGHT

○ Rook Contiguity The order of contiguity 1

○ Queen Contiguity □ Include all the lower orders

DISTANCE WEIGHT

Select distance metric <Euclidean Distance>

Variable for x-coordinates <X-Centroids>

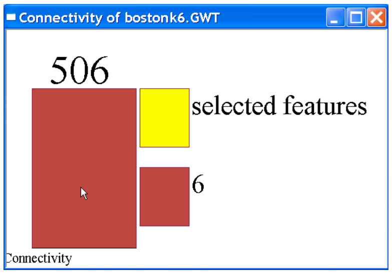
Variable for y-coordinates <Y-Centroids>

○ Threshold Distance 0.010000

Cut-off point

● k-Nearest Neighbors The number of neighbors 6

Figure 16.7: Nearest neighbor weights dialog.



Connectivity of bostonk6.GWT

Figure 16.8: Nearest neighbor connectivity property.

the number of neighbors to 6, and click Create to create the weights. Click on the Done button to return to the standard interface when the progress bar indicates completion.

The k-nearest neighbor criterion ensures that each observation has exactly the same number (k) of neighbors. Inspect the GWT file you just created to check that this is the case. Alternatively, check the weights properties. The connectivity histogram is not very meaningful, as in Figure 16.8, but it confirms that each location has exactly 6 neighbors. While not useful as such, this may be handy to make sure the number of neighbors is correct, since there are currently no metadata for spatial weights files in *GeoDa*.

16.4 Practice

Practice the construction of distance-based spatial weights using one of the other point shape files from the sample data sets, such as the 30 Los Angeles air quality monitoring stations in `oz9799.shp` (with `STATION` as the `Key`), or the 211 locations of house sales transactions in the `baltimore.shp` point file (also with `STATION` as the `Key`). Alternatively, check out the default feature for polygon shape files. For example, use the `ohlung.shp` polygon shape file for the 88 counties in Ohio (with `FIPSNO` as the `Key`) to create distance-band and k-nearest neighbor weights.

In the later exercises, you will need spatial weights as an essential input to spatial autocorrelation analysis, so any files you create now will not need to be created at that point, but can then simply be "opened."

Exercise 17

Spatially Lagged Variables

17.1 Objectives

This exercise illustrates the construction of a spatially lagged variable and its use in a Moran scatter plot.

At the end of the exercise, you should know how to:

- create a spatially lagged variable for a specified weights file
- use the spatial lag to construct a Moran scatter plot "by hand"

More detailed information on these operations can be found in the *User's Guide*, pp. 61–62.

17.2 Spatial Lag Construction

Spatially lagged variables are an essential part of the computation of spatial autocorrelation tests and the specification of spatial regression models. *GeoDa* typically computes these variables on the fly, but in some instances it may be useful to calculate the spatially lagged variables explicitly. For example, this is handy if one wants to use these variables in other statistical packages.

The spatial lag computation is part of the `Table` functionality in *GeoDa* (see Exercise 3, and especially Section 3.4 on p. 17). Begin by loading the sample data shape file for the 403 census tracts in Sacramento, CA (use `sacramentot2.shp` with `POLYID` as the `Key`). The base map should be as in Figure 15.1 on p. 107.

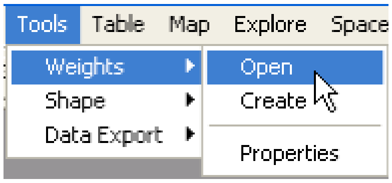
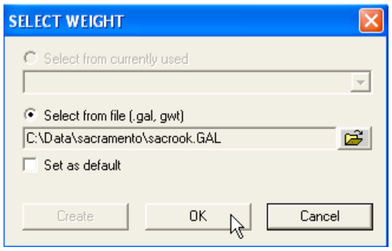


Figure 17.1: Open spatial weights file.



C:\Data\sacramento\sacrook.GAL

Figure 17.2: Select spatial weights file.

Before starting the `Table` operations, make sure that a spatial weights file has been *opened*. If no such file is present, the spatial lag computation will generate an error message. Open the weights from the menu, using `Tools > Weights > Open`, as in Figure 17.1, or by clicking on the matching toolbar button. Specify `sacrook.GAL` as the file name in the weights dialog, as shown in Figure 17.2.

Everything should now be in place to start the lag computation. Open the data table (click on the `Table` toolbar button) and right click to select `Field Calculation` from the menu (Figure 17.3 on p. 126). Next, select the `Lag Operations` tab in the `Field Calculation` dialog, as in Figure 17.4 (p. 126). Overwrite the entry in the left most text box with `W_INC`, as in Figure 17.5 (p. 126), leave the weights file to the default, and select `HH_INC` (census tract median household income) as the variable to be lagged. Click on `OK` to compute the new variable. It will be added to the data table in a new column, as illustrated in Figure 17.6 on p. 127.

Recall from Figure 15.7 (p. 110) that the tract with `POLYID 2` had four neighbors. The relevant tracts are highlighted in the map and table of Figure 17.6.

For a contiguity weights file, such as `sacrook.GAL`, the spatially lagged variable amounts to a simple average of the values for the neighboring units.

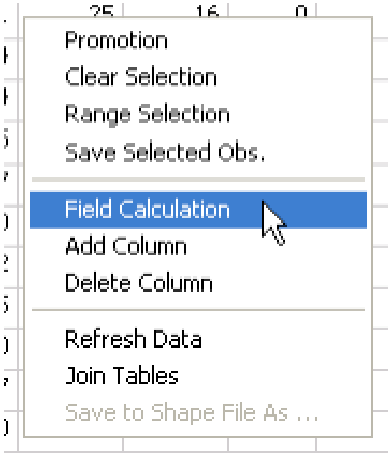
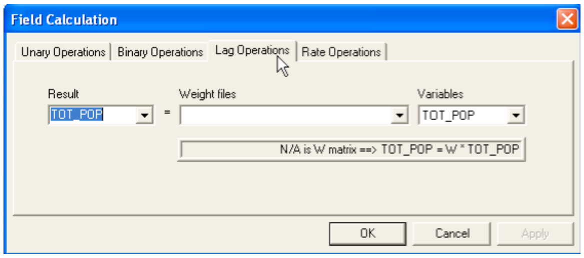


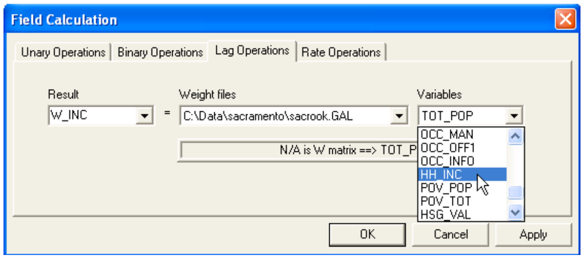
Figure 17.3: Table field calculation option.



Result: TOT_POP = Weight files: Variables: TOT_POP

N/A is W matrix ==> TOT_POP = W * TOT_POP

Figure 17.4: Spatial lag calculation option tab in table.



Result: W_INC = Weight files: C:\Data\sacramento\sacrook.GAL Variables: TOT_POP

N/A is W matrix ==> TOT_P

OCC_MAN
OCC_OFF1
OCC_INFO
HH_INC
POV_POP
POV_TOT
HSG_VAL

Figure 17.5: Spatial lag dialog for Sacramento tract household income.

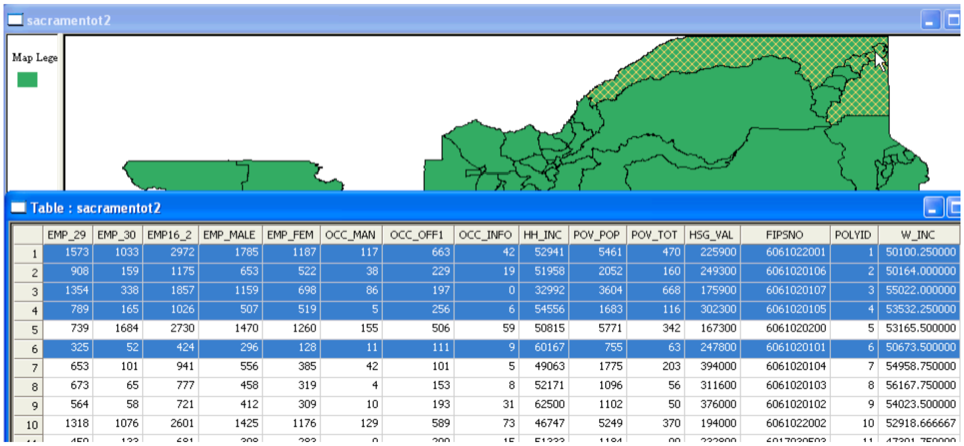


Table : sacramentot2

	EMP_29	EMP_30	EMP16_2	EMP_MALE	EMP_FEM	OCC_MAN	OCC_OFF1	OCC_INFO	HH_INC	POV_POP	POV_TOT	HSG_VAL	FIPSNO	POLYID	W_INC
1	1573	1033	2972	1785	1187	117	663	42	52941	5461	470	225900	6061022001	1	50100.250000
2	908	159	1175	653	522	38	229	19	51958	2052	160	249300	6061020106	2	50164.000000
3	1354	338	1857	1159	698	86	197	0	32992	3604	668	175900	6061020107	3	55022.000000
4	789	165	1026	507	519	5	256	6	54556	1683	116	302300	6061020105	4	53532.250000
5	739	1684	2730	1470	1260	155	506	59	50815	5771	342	167300	6061020200	5	53165.500000
6	325	52	424	296	128	11	111	9	60167	755	63	247800	6061020101	6	50673.500000
7	653	101	941	556	385	42	101	5	49063	1775	203	394000	6061020104	7	54958.750000
8	673	65	777	458	319	4	153	8	52171	1096	56	311600	6061020103	8	56167.750000
9	564	58	721	412	309	10	193	31	62500	1102	50	376000	6061020102	9	54023.500000
10	1318	1076	2601	1425	1176	129	589	73	46747	5249	370	194000	6061022002	10	52918.666667

Figure 17.6: Spatial lag variable added to data table.

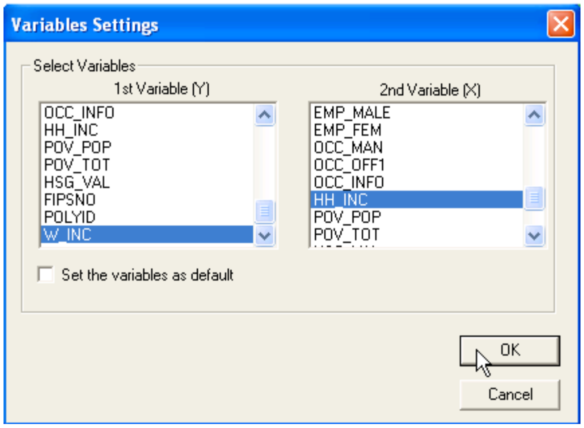
Check in Figure 17.6 how the value for `W_INC` in row `2` (`50164`) is the average of the values of `HH_INC` in rows `1, 3, 4` and `6`.

17.3 Spatial Autocorrelation

A Moran scatter plot is a plot with the variable of interest on the x-axis and the spatial lag on the y-axis (for details, see Section 18.2.2 on p. 131). Since the just computed spatial lag is immediately available for any analysis, you can now "manually" construct a Moran scatter plot using `W_INC` for the spatial lag and `HH_INC` for the variable on the x-axis in a regular scatter plot (see Section 8.2 on p. 53).

Start the scatter plot function as `Explore > Scatter plot` from the menu or by clicking the matching toolbar button. In the variable selection dialog, specify `W_INC` in the left side column and `HH_INC` in the right side column, as in Figure 17.7 on p. 128. Next, click `OK` to generate the plot shown in Figure 17.8 on p. 128.

The slope of the regression line (`0.5632`) is the Moran's I statistic for `HH_INC`, using a rook contiguity weights definition. Feel free to run ahead and follow the instructions in Section 18.2.2 (but with the relevant Sacramento data and weights substituted) to check that this is indeed the case. Of course, since Figure 17.8 is a non-spatial scatter plot, it does not contain any means to assess the significance of the Moran's I.



Variables Settings

Select Variables

1st Variable (Y)

OCC_INFO
HH_INC
POV_POP
POV_TOT
HSG_VAL
FIPSNO
POLYID
W_INC

2nd Variable (X)

EMP_MALE
EMP_FEM
OCC_MAN
OCC_OFF1
OCC_INFO
HH_INC
POV_POP
POV_TOT

□ Set the variables as default

OK

Cancel

Figure 17.7: Variable selection of spatial lag of income and income.

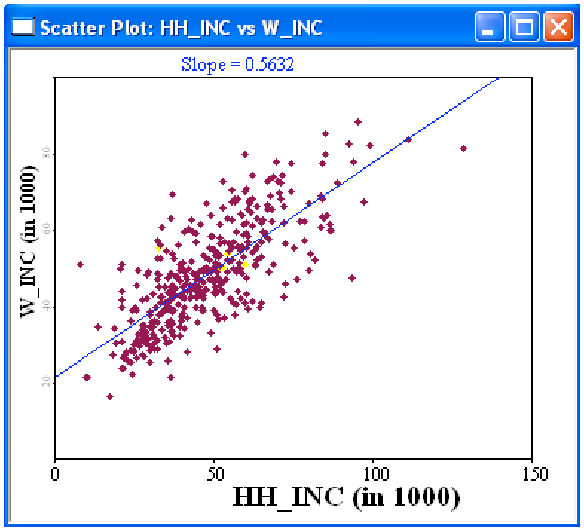


Figure 17.8: Moran scatter plot constructed as a regular scatter plot.

17.4 Practice

Use any of the spatial weights constructed as part of Exercises 15 or 16 to create spatially lagged variables and construct a Moran scatter plot by hand.

Exercise 18

Global Spatial Autocorrelation

18.1 Objectives

This exercise begins the illustration of the analysis of spatial autocorrelation with the univariate case and the Moran scatter plot. For methodological background on these methods, see Anselin (1995, 1996).

At the end of the exercise, you should know how to:

- create a Moran scatter plot for the description of univariate spatial autocorrelation
- perform a significance assessment by means of a permutation test
- construct significance envelopes
- brush the Moran scatter plot
- save the spatial lag and standardized variable

More detailed information on these operations can be found in the *User's Guide*, pp. 88–94.

18.2 Moran Scatter Plot

18.2.1 Preliminaries

We will work with the lip cancer data for 56 Scottish districts. Load this shape file as `scotlip.shp` with `CODENO` as the `Key`. The resulting base map should be as in Figure 18.1 on p. 130.

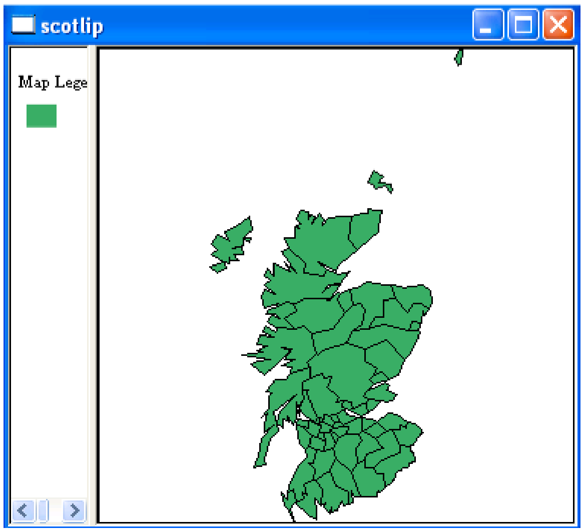


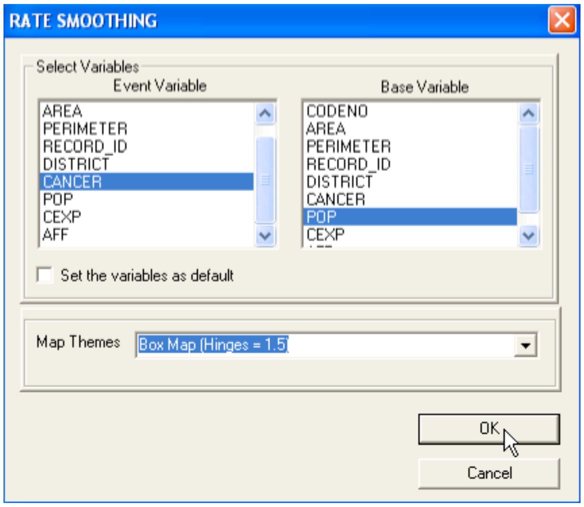
Figure 18.1: Base map for Scottish lip cancer data.

In order to be able to compare the Moran scatter plot for “raw” rates to that for EB standardized rates in Exercise 20, make sure to have a variable with the raw rates in the data set. If you did not compute this previously, an easy way to proceed is to create a box map using `Map > Smooth > Raw Rate` from the menu. Select `Cancer` as the `Event` and `Pop` as the `Base` variable, as illustrated in Figure 18.2 on p. 131. Make sure to set the map type to `Box Map`. Click on `OK` to yield the map shown in Figure 18.3 on p. 131.

Select the option to add the raw rate to the data table by right clicking in the box map and choosing `Save Rates` (see Figure 13.6 on p. 95). For simplicity, leave the variable name to the default specification of `R_RAWRATE`.

Finally, make sure there is a spatial weights file for the Scottish districts shape file. In the example, we will use 5 nearest neighbors. Create such a weights file as `scot5k.GWT` if you haven't done so in a previous exercise (see Section 16.3 on p. 121 for details).[1]

[1] Do *not* use a weights file based on simple contiguity in this example, since there are three *islands*, see Figure 15.11 on p. 112.



RATE SMOOTHING

Select Variables

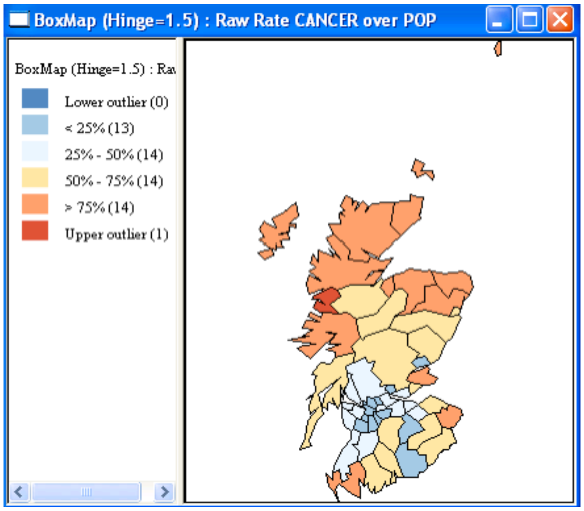
Event Variable: AREA, PERIMETER, RECORD_ID, DISTRICT, CANCER (selected), POP, CEXP, AFF

Base Variable: CODENO, AREA, PERIMETER, RECORD_ID, DISTRICT, CANCER, POP (selected), CEXP

□ Set the variables as default

Map Themes: Box Map (Hinges = 1.5)

Figure 18.2: Raw rate calculation for Scottish lip cancer by district.



BoxMap (Hinge=1.5) : Raw Rate CANCER over POP

BoxMap (Hinge=1.5) : Ra

- Lower outlier (0)
- < 25% (13)
- 25% - 50% (14)
- 50% - 75% (14)
- > 75% (14)
- Upper outlier (1)

Figure 18.3: Box map with raw rates for Scottish lip cancer by district.

18.2.2 Moran scatter plot function

With all the preliminaries completed, start the Moran scatter plot function by clicking its toolbar button, or proceed from the menu by selecting `Space`

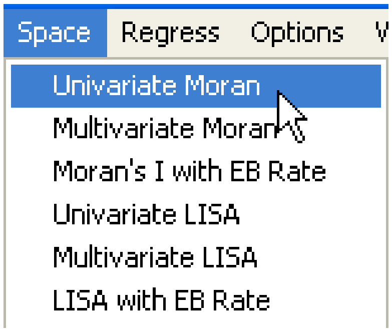


Figure 18.4: Univariate Moran scatter plot function.

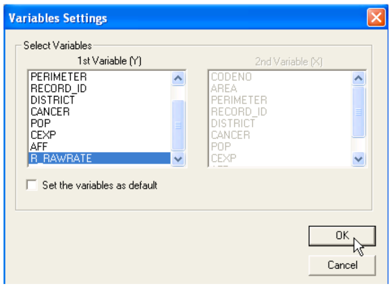
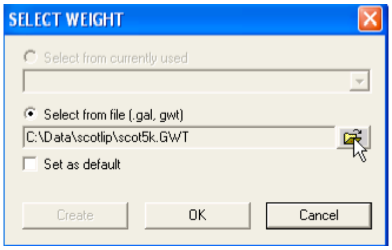


Figure 18.5: Variable selection dialog for univariate Moran.

`> Univariate Moran` (Figure 18.4). This brings up the variable selection dialog shown in Figure 18.5.

Select `R_RAWRATE` as the variable and click `OK`. Next, select `scot5k.GWT` as the weights file. In Figure 18.6 (p. 133), this is illustrated for the case where the spatial weights file name needs to be specified (e.g., `Select from file`). If the file had already been *opened*, the dialog would be slightly different (for example, see Figure 20.3 on p. 150). Click `OK` to create the Moran scatter plot shown in Figure 18.7 on p. 133.

Note how the y-axis has been specified as `W_R_RAWRATE` without the need for an explicit calculation of a spatial lag. The `R_RAWRATE` is on the x-axis and has been standardized such that the units correspond to standard deviations (any observations beyond 2 standard deviations are typically categorized as *outliers*). The scatter plot figure has also been centered on the mean with the axes drawn such that the four quadrants are clearly shown. Each quadrant corresponds to a different type of spatial autocorrelation: high-high and low-


SELECT WEIGHT

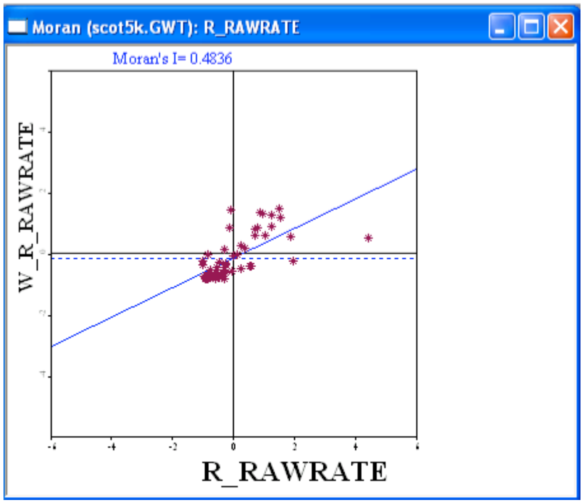
○ Select from currently used

● Select from file (.gal, gwt)

C:\Data\scotlip\scot5k.GWT

□ Set as default

Figure 18.6: Spatial weight selection dialog for univariate Moran.


Moran (scot5k.GWT): R_RAWRATE

Moran's I= 0.4836

Figure 18.7: Moran scatter plot for Scottish lip cancer rates.

low for positive spatial autocorrelation; low-high and high-low for negative spatial autocorrelation. Use the selection tool to investigate which locations correspond to each of the types of spatial autocorrelation (through linking with the map).

The value listed at the top of the graph (`0.4836`) is the Moran's I statistic. Since the graph is a special case of a scatter plot, the `Exclude Selected` option may be applied. Try this out (invoke the option in the usual way by right clicking in the graph, or from the `Options` menu) and assess how the spatial autocorrelation coefficient (the slope of the regression line) changes as specific locations are *excluded* from the calculation. Similarly, you can brush the Moran scatter plot in the same way as any other scatter plot.

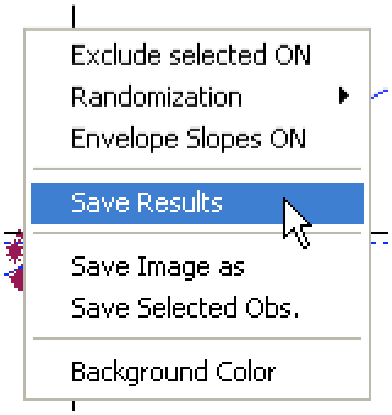
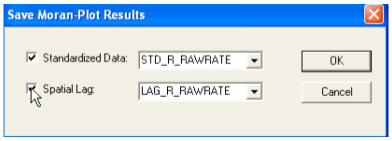


Figure 18.8: Save results option for Moran scatter plot.


Save Moran-Plot Results

☑ Standardized Data: STD_R_RAWRATE

☑ Spatial Lag: LAG_R_RAWRATE

Figure 18.9: Variable dialog to save results in Moran scatter plot.

The intermediate calculations used to create the plot may be saved to the current data table. Right click on the graph to bring up the `Options` menu, as in Figure 18.8, and select `Save Results` to generate the variable selection dialog shown in Figure 18.9. This time, you cannot keep the default variable names, since *GeoDa* currently only supports variable names less than 12 characters in length.[2] Edit the variable names accordingly (e.g., use `W_RAWRATE`) to add the standardized values for `R_RAWRATE` and the corresponding spatial lag to the table.

18.3 Inference

Inference for Moran's I is based on a random permutation procedure, which recalculates the statistic many times to generate a reference distribution. The obtained statistic is then compared to this reference distribution and a *pseudo significance* level is computed.

[2] For example, `LAG_R_RAWRATE` is one character too long.

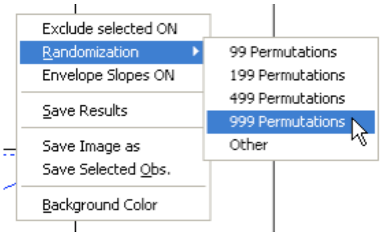
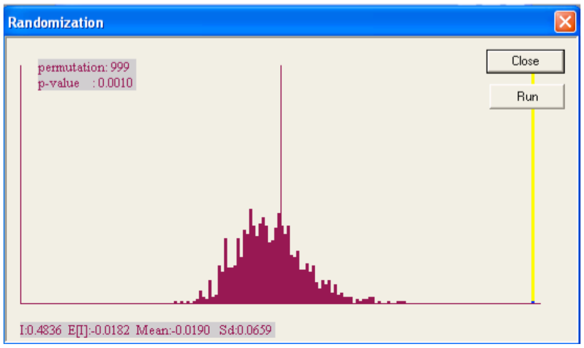


Figure 18.10: Randomization option dialog in Moran scatter plot.



```
permutation: 999
p-value    : 0.0010
```

```
I:0.4836  E[I]:-0.0182  Mean:-0.0190  Sd:0.0659
```

Figure 18.11: Permutation empirical distribution for Moran's I.

The inference computation is started by right clicking on the scatter plot to invoke the options menu, as in Figure 18.10. Select `Randomization > 999 permutations` to bring up the histogram shown in Figure 18.11.

In addition to the reference distribution (in brown) and the statistic (as a yellow bar), this graph lists the number of permutations and the pseudo significance level in the upper left corner, as well as the value of the statistic (`0.4836`), its theoretical mean (`E[I] = -0.0182`), and the mean and standard deviation of the empirical distribution. In Figure 18.11, these values are `-0.0191` and `0.0659` respectively (these values depend on the particular random permutation and will typically differ slightly between permutations). Click on the `Run` button to assess the sensitivity of the results to the particular random permutation. Typically, with 999 permutations, these results will not vary much, but for a smaller number of permutations, such as 99, there may be quite substantial differences. Also note that the *most significant* p-level depends directly on the number of permutations. For example, for 99 permutations, this will be $p = 0.01$, and for 999, $p = 0.001$.

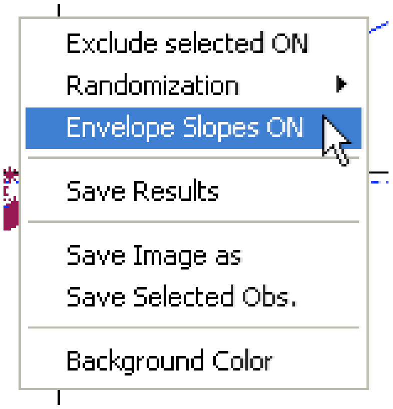


Figure 18.12: Envelope slopes option for Moran scatter plot.

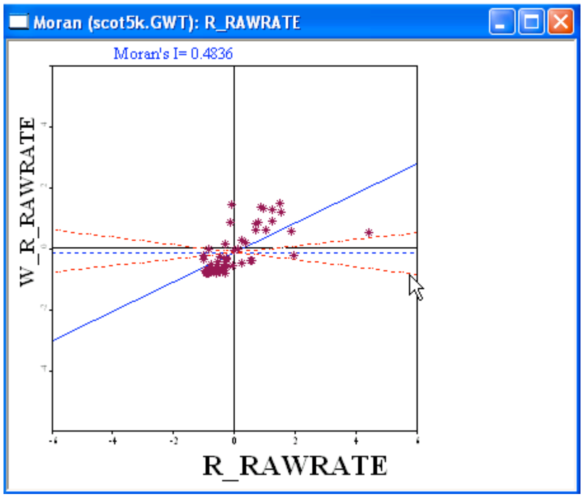


Figure 18.13: Envelope slopes added to Moran scatter plot.

A slightly different way to visualize the significance of the Moran's I statistic is to draw randomization envelopes on the graph. These slopes correspond to the 2.5 and 97.5 percentiles of the reference distribution, and thus contain 95% of the distribution of the Morans' I statistics computed in spatially random data sets.

Right click on the graph to select turn this option on (as `Exclude Selected`, this is a toggle option), as illustrated in Figure 18.12. Subsequently, two dashed lines will appear in the plot, as shown in Figure 18.13 on p. 136. Note how the actual Moran scatter plot slope is well outside the range corresponding to the randomly permuted data.

18.4 Practice

Several of the sample data sets are appropriate to analyze spatial autocorrelation in rates, allowing the comparison between the statistic for raw rates and for EB standardized rates in Exercise 20. This includes the lung cancer data for 88 Ohio counties in `ohlung.shp` (with `FIPSNO` as the `Key`) and the SIDS death rates for 100 North Carolina counties in `sids.shp` (with `FIPSNO` as the `Key`) as classic public health applications. The 209 electoral districts for Buenos Aires in `buenosaires.shp` (with `INDRANO` as the `Key`) allow for a political science example, while the 78 county homicide data in `stl_hom` (with `FIPSNO` as the `Key`) illustrate a criminological application. In each of these sample files, different variables can be analyzed and the sensitivity of spatial autocorrelation assessed to the choice of the spatial weights.

Exercise 19

Local Spatial Autocorrelation

19.1 Objectives

This exercise focuses on the notion of local spatial autocorrelation, and the local Moran statistic in particular. Methodological background can be found in Anselin (1995).

At the end of the exercise, you should know how to:

- compute the local Moran statistic and associated significance map and cluster map
- assess the sensitivity of the cluster map to the number of permutations and significance level
- interpret the notion of spatial cluster and spatial outlier

More detailed information on these operations can be found in the *User's Guide*, pp. 99–105.

19.2 LISA Maps

19.2.1 Basics

To illustrate the local spatial autocorrelation functionality, load the file for the homicide data in 78 counties surrounding St Louis, MO (`stl_hom.shp` with `FIPSNO` as the `Key`). The base map should be as in Figure 19.1 on p. 139. You will also need a spatial weights file for this data set. If you haven't already done so, create a rook weights file (call the file `stlrook.GAL`) before embarking on the analysis. Start the Local Moran function from the

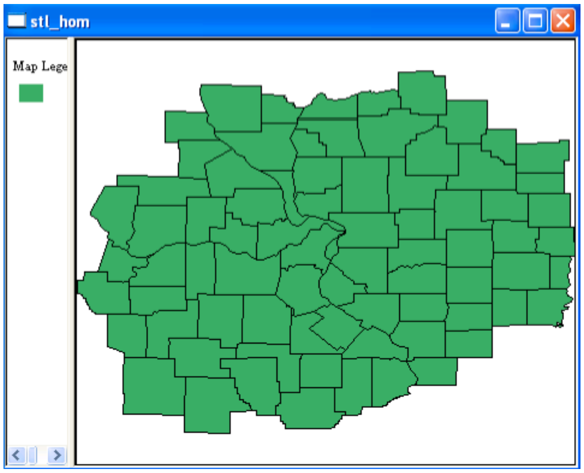


Figure 19.1: St Louis region county homicide base map.

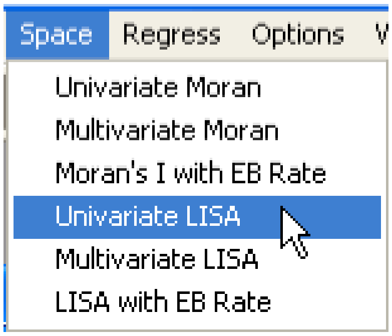
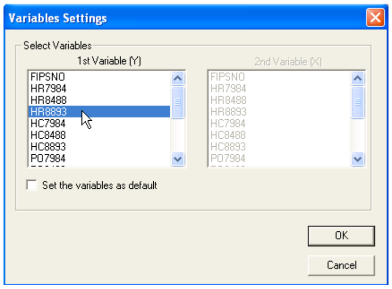


Figure 19.2: Local spatial autocorrelation function.

menu, by invoking `Space > Univariate LISA` (as in Figure 19.2), or click the matching toolbar button. This brings up the by now familiar variable selection dialog. Select `HR8893` for the variable, as in Figure 19.3 (p. 140). Next, click `OK` to bring up the weight selection dialog. Specify `stlrook.GAL` as the weight file, as in Figure 19.4 (p. 140), and proceed with `OK`. As a final step follows the results options dialog, shown in Figure 19.5 on p. 141.

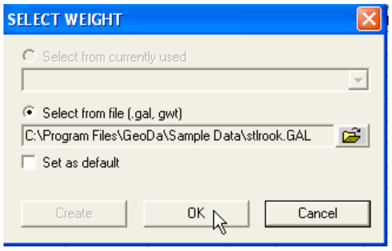
Four different types of result graphs and maps are available: a significance map, a cluster map, a box plot and a Moran scatter plot. For now, check all four boxes, as in Figure 19.5, but this is by no means necessary. Click `OK` to bring up all four graphs.



FIPSNO
HR7984
HR8488
HR8893
HC7984
HC8488
HC8893
PO7984

□ Set the variables as default

Figure 19.3: Variable selection dialog for local spatial autocorrelation.



○ Select from currently used

● Select from file (.gal, gwt)

C:\Program Files\GeoDa\Sample Data\stlrook.GAL

□ Set as default

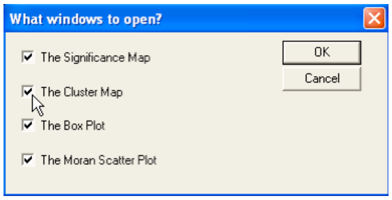
Figure 19.4: Spatial weights selection for local spatial autocorrelation.

19.2.2 LISA Significance Map

The significance map, illustrated in Figure 19.6 on p. 141, shows the locations with significant local Moran statistics in different shades of green (the corresponding p values are given in the legend). Your results may be slightly different, since the first maps are based on the default of 99 permutations (the map shown here results after a number of runs for 9999 permutations to avoid too great a sensitivity on the particular randomization). We return to inference in Section 19.3. It should be noted that the results for $p = 0.05$ are somewhat unreliable, since they likely ignore problems associated with multiple comparisons (as a consequence, the *true* p-value is likely well above 0.05).

19.2.3 LISA Cluster Map

Arguably the most useful graph is the so-called *LISA cluster map*, shown in Figure 19.7 on p. 142. This provides essentially the same information as the significance map (in fact, the two maps are synchronized), but with



What windows to open?

☑ The Significance Map

☑ The Cluster Map

☑ The Box Plot

☑ The Moran Scatter Plot

Figure 19.5: LISA results option window.

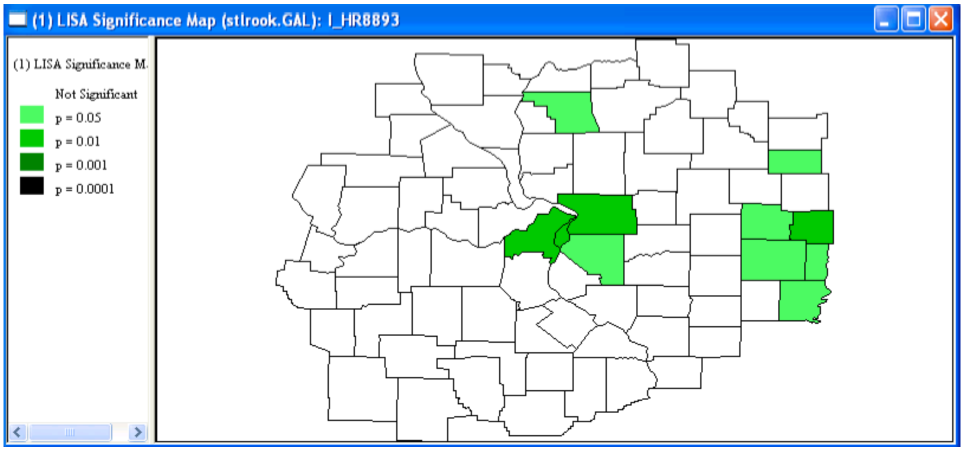


Figure 19.6: LISA significance map for St Louis region homicide rates.

the significant locations color coded by type of spatial autocorrelation. The four codes are shown in the legend: dark red for high-high, dark blue for low-low, pink for high-low, and light blue for low-high (there is no low-high location in the sample map). These four categories corrrespond to the four quadrants in the Moran scatter plot. Check this by selecting the counties with the same color and noting their position in the Moran scatter plot.

19.2.4 Other LISA Result Graphs

The next result is a box plot for the distribution of the local Moran statistics across observations. This is primarily of technical interest, suggesting potential locations that show very different local autocorrelation patterns.

The global Moran's I statistic is the mean of the local Moran statistics. Hence, if the distribution of these local statistics is highly asymmetric, or dominated by a few large values (as in Figure 19.8 on p. 143), the overall

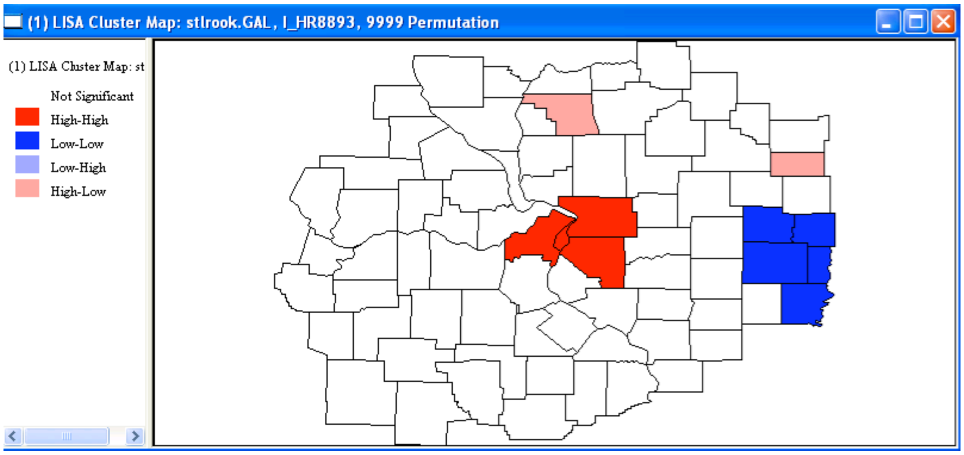


Figure 19.7: LISA cluster map for St Louis region homicide rates.

indication may be spurious or overly sensitive to a few observations. Brush the box plot from high to low and locate the corresponding counties on the map. Note that positive values for the local Moran may be associated with *either* high-high (the four highest) or low-low patterns (the next few).

The final result is the familiar Moran scatter plot, shown in Figure 19.9 (see Exercise 18 on p. 129 for details).

19.2.5 Saving LISA Statistics

Several intermediate results can be added to the data table. Right click on any of the maps or select `Options > Save Results` from the `Options` menu, in the usual fashion. A dialog appears that suggests variable names for the local Moran statistics or `Lisa Indices` (`I_HR8893`), an indicator for the type of cluster for significant locations only (`CL_HR8893`), and the p-values from the permutation routine (`P_HR8893`). Check the check boxes next to the default, as in Figure 19.10, and click on `OK` to add these new variables to the data table. Verify that they have been added, as illustrated in Figure 19.11 on p. 145.

19.3 Inference

The pair of significance map and cluster map first generated is based on very quick calculations, using only 99 permutations and a default significance level of $p = 0.05$. In most applications, this is a good first approximation,

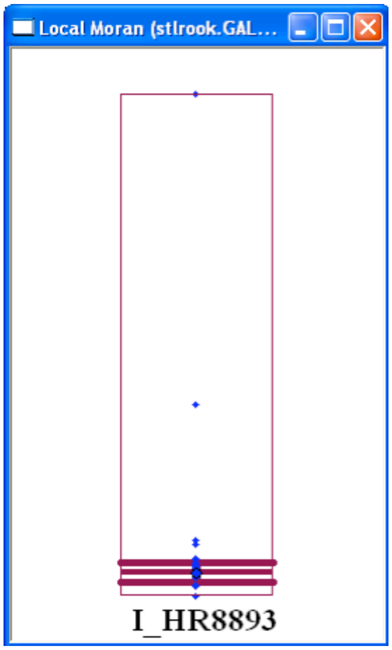
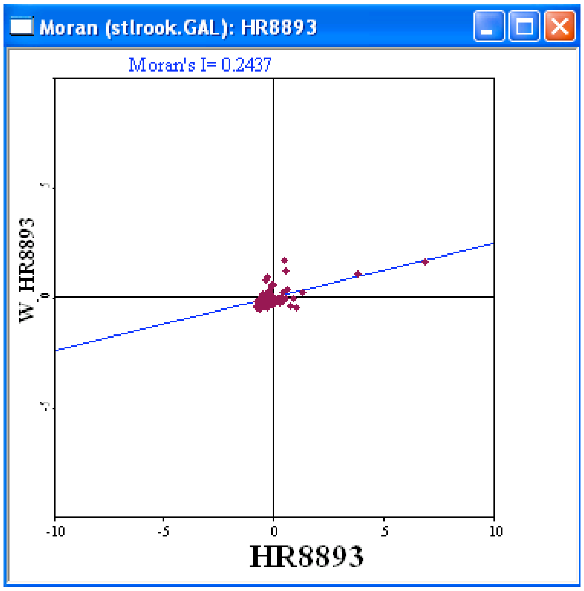


Figure 19.8: LISA box plot.

but it also tends to be somewhat sensitive to the particular randomization. In order to obtain more robust results, it is good practice (for reasonably sized data sets) to increase the number of permutations to 999 or even 9999, and to carry out several runs until the results stabilize.

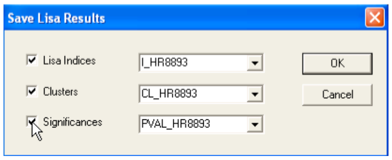
Right click on either significance map or cluster map to bring up the `Options` menu, shown in Figure 19.12 on p. 145. Select `Randomization > Other` to enter a custom number of permutations as 9999, illustrated in Figure 19.13 (p. 145). Click `OK` and note some (slight) changes in the counties that are significant. Typically, this will affect only marginally significant (at $p = 0.05$) counties, but it may lead to quite drastic changes in the overall layout of the results. In the current example, it seems to primarily affect the presence of the spatial outlier in Morgan county, IL (click on the outlier in the north central part of the map and locate its name in the table), and the spatial extent of the low-low cluster.

The sensitivity of the results can be further assessed by changing the significance cut-off value for the maps. Right click on either one of the maps to bring up the `Options` menu, but now select `Significance Filter > 0.01`, as in Figure 19.14. Note how the locations that were significant at



Moran's I= 0.2437

Figure 19.9: LISA Moran scatter plot.

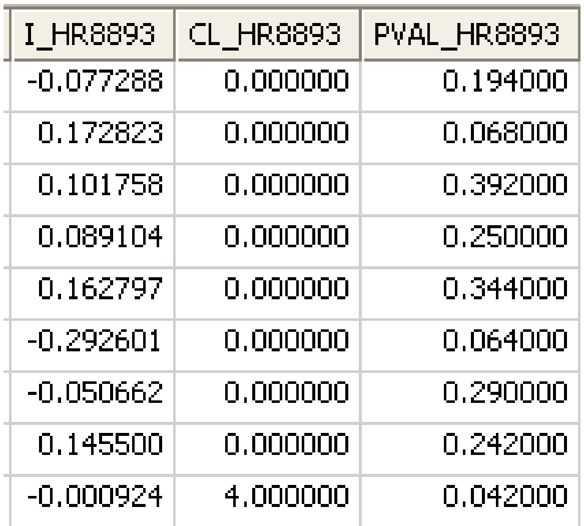


☑ Lisa Indices I_HR8893
☑ Clusters CL_HR8893
☑ Significances PVAL_HR8893

Figure 19.10: Save results option for LISA.

$p < 0.05$ but not $p < 0.01$ disappear from both maps.

For example, in Figure 19.15 on p. 146, the resulting LISA cluster map lacks significant high-low locations and has the low-low cluster reduced to one county. Assess the effect of tightening the significance level even more, to $p = 0.001$. Locations that are consistently significant, even at such demanding levels, are fairly robust indictions of spatial clusters or outliers.



I_HR8893	CL_HR8893	PVAL_HR8893
-0.077288	0.000000	0.194000
0.172823	0.000000	0.068000
0.101758	0.000000	0.392000
0.089104	0.000000	0.250000
0.162797	0.000000	0.344000
-0.292601	0.000000	0.064000
-0.050662	0.000000	0.290000
0.145500	0.000000	0.242000
-0.000924	4.000000	0.042000

Figure 19.11: LISA statistics added to data table.

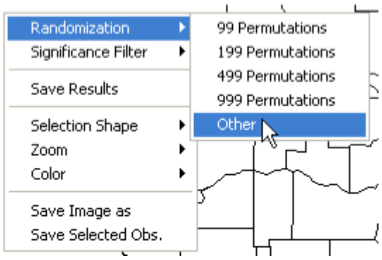
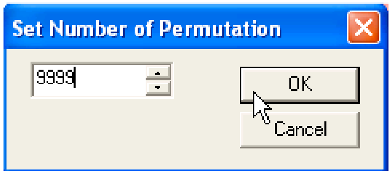


Figure 19.12: LISA randomization option.



9999

Figure 19.13: Set number of permutations.

19.4 Spatial Clusters and Spatial Outliers

The high-high and low-low locations (positive local spatial autocorrelation) are typically referred to as *spatial clusters*, while the high-low and low-high locations (negative local spatial autocorrelation) are termed *spatial outliers*. While outliers are single locations by definition, this is not the case for clusters.

It should be kept in mind that the so-called spatial clusters shown on the LISA cluster map only refer to the *core* of the cluster. The cluster is classified as such when the value at a location (either high or low) is

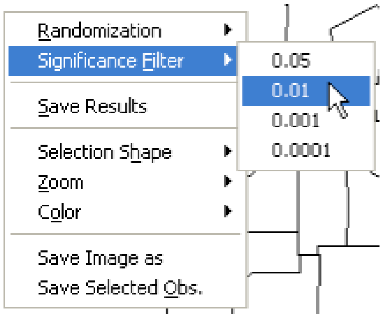


Figure 19.14: LISA significance filter option.

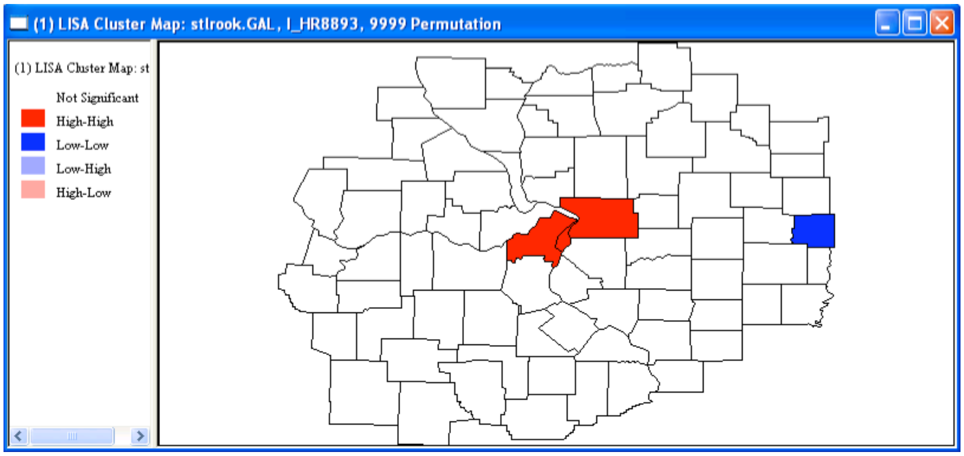


Figure 19.15: LISA cluster map with p < 0.01.

more similar to its neighbors (as summarized by the weighted average of the neighboring values, the *spatial lag*) than would be the case under spatial randomness. Any location for which this is the case is labeled on the cluster map. However, the *cluster* itself likely extends to the neighbors of this location as well.

For example, in Figure 19.16 on p. 147, the neighbors of the cluster counties (for $p = 0.01$) have been cross-hatched, to better illustrate the spatial extent of the clusters. Several of the neighbors are overlapping, which is not suprising. Overall, the impression of the spatial extent of the clusters is quite larger than suggested by their cores alone. However, it is not that different from the cores suggested by the more liberal significance of $p = 0.05$.

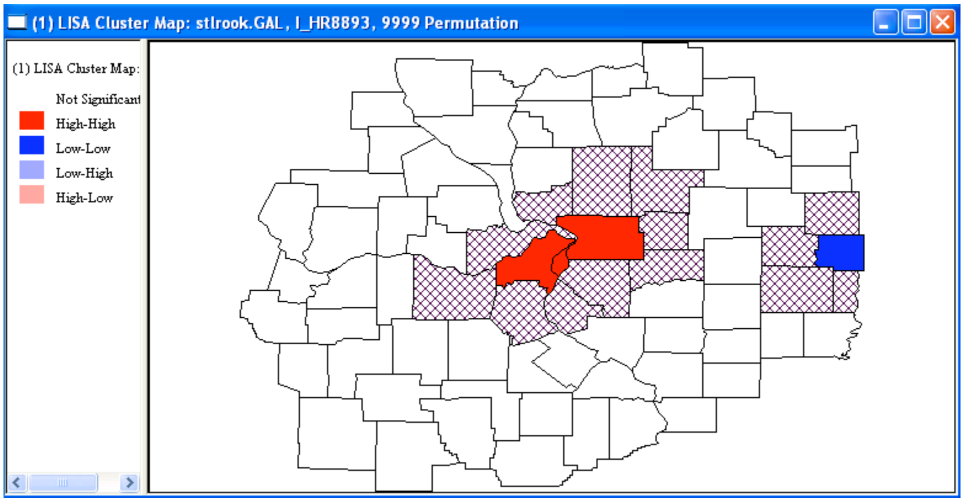


Figure 19.16: Spatial clusters.

19.5 Practice

Revisit the global spatial autocorrelation analysis you carried out in Practice Session 18.4 from a local perspective. Try to assess which results are particularly sensitive to the choice of significance level, the number of permutations or the spatial weights chosen. Based on this careful sensitivity analysis, identify spatial clusters and outliers that are fairly robust to these factors.

Exercise 20

Spatial Autocorrelation Analysis for Rates

20.1 Objectives

This exercise illustrates the analysis of spatial autocorrelation with an adjustment for the variance instability of rates. For methodological background on these methods, see Assunção and Reis (1999).

At the end of the exercise, you should know how to:

- create a Moran scatter plot for the rates
- use the empirical Bayes (EB) adjustment to take into account variance instability of rates in the Moran scatter plot
- use the empirical Bayes (EB) adjustment to take into account variance instability of rates in local spatial autocorrelation analysis

More detailed information on these operations can be found in the *User's Guide*, pp. 97–98, 105.

20.2 Preliminaries

When the Moran's I statistic is computed for rates or proportions, the underlying assumption of stationarity may be violated by the intrinsic variance instability of rates. The latter follows when the population at risk (the `Base`) varies considerably across observations. The variance instability may lead to spurious inference for Moran's I. To correct for this, *GeoDa* implements

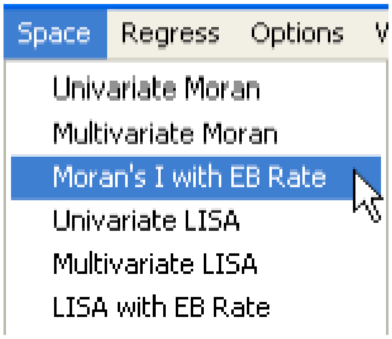


Figure 20.1: Empirical Bayes adjusted Moran scatter plot function.

the Empirical Bayes (EB) standardization suggested by Assunção and Reis (1999). This is not the same as computing Moran's I for EB-smoothed rates, but is a direct standardization of the variable, using a similar (but not identical) rationale. In *GeoDa*, this is implemented for both global (Moran scatter plot) and local spatial autocorrelation statistics.

In order to be able to compare the results to the non-standardized rates used in Exercises 18 and 19, we will use the same data sets and weights files. For the global measure, we will use the Scottish lip cancer data set and associated spatial weights file.[1]

For the local analysis, we will use the St Louis homicide data and first order rook weights file.[2]

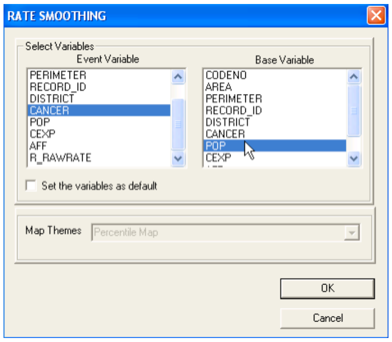
20.3 EB Adjusted Moran Scatter Plot

With the Scottish lip data loaded, invoke the EB-adjusted Morans' I from the menu as `Space > Morans' I with EB Rate`, as shown in Figure 20.1, or click the matching toolbar button. This brings up the variable selection dialog, which follows the same format as that used in the smoothing operations. Select `Cancer` as the `Event` and `Pop` as the `Base` variable, as in Figure 20.2 on p. 150.

Click `OK` to bring up the weights selection dialog. If you have already loaded the file `scot5k.GWT`, the dialog will be as in Figure 20.3 on p. 150

[1] The shape file for the 56 districts is `scotlip.shp` with `CODENO` as the `Key`. The weights file is `scot5k.GWT`. See Section 18.2.1 for details.

[2] The shape file for the 78 counties is `stl_hom.shp` with `FIPSNO` as the `Key`. The weights file is `stlrook.GAL`. Details are given in Section 19.2.1.



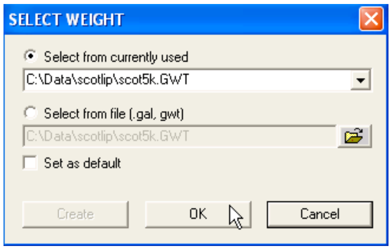
Event Variable: CANCER

Base Variable: POP

□ Set the variables as default

Map Themes: Percentile Map

Figure 20.2: Variable selection dialog for EB Moran scatter plot.



● Select from currently used

C:\Data\scotlip\scot5k.GWT

○ Select from file (.gal, gwt)

C:\Data\scotlip\scot5k.GWT

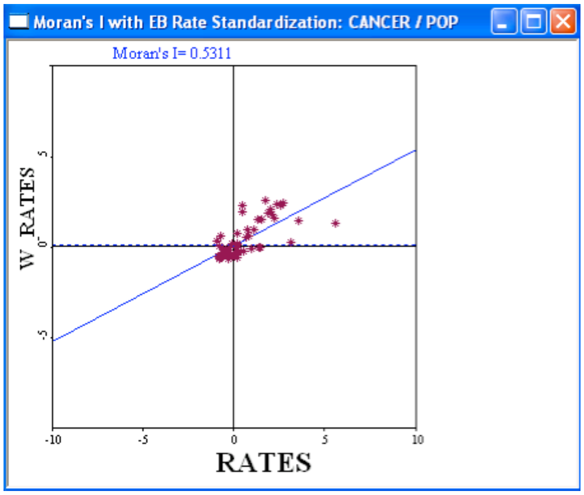
□ Set as default

Figure 20.3: Select current spatial weights.

(`Select from currently used`).

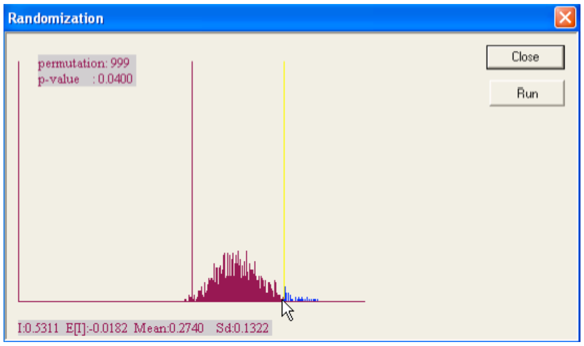
Click `OK` again to generate the Moran scatter plot shown in Figure 20.4 on p. 151. Note how the value for Moran's I of `0.5311` differs somewhat from the statistic for the unstandardized rates (`0.4836` in Figure 18.7 on p. 133).

More important is to assess whether or not inference is affected. As before, right click in the graph to bring up the `Options` menu, and select `Randomization > 999 Permutations`. The resulting permutation empirical distribution in Figure 20.5 on p. 151 still suggests a highly significant statistic, although the pseudo-significance level is lower, at $p = 0.04$ (your results may vary slightly due to the random permutation). Click on `Run` a few times to assess the robustness of this result.



Moran's I= 0.5311

Figure 20.4: Empirical Bayes adjusted Moran scatter plot for Scottish lip cancer rates.



```
permutation: 999
p-value    : 0.0400
```

```
I:0.5311  E[I]:-0.0182  Mean:0.2740   Sd:0.1322
```

Figure 20.5: EB adjusted permutation empirical distribution.

20.4 EB Adjusted LISA Maps

In *GeoDa*, the EB standardization has been implemented for the Local Moran statistics as well. We will repeat the analysis using the homicide rate computed from the homicide count and county population as in Exercise 19. With the St Louis homicide data loaded, invoke the EB adjusted local Moran by clicking the toolbar button, or as `Space > LISA with EB Rate` from the menu (Figure 20.6 on p. 152).

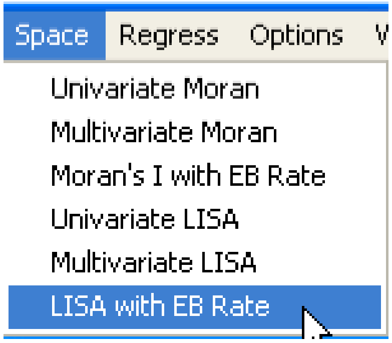


Figure 20.6: EB adjusted LISA function.

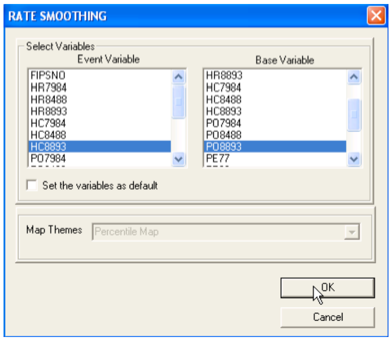


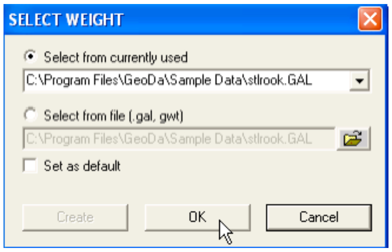
Figure 20.7: Variable selection dialog for EB LISA.

As before, specify the `Event` variable as `HC8893` and the `Base` variable as `PO8893` in the variable selection dialog, as shown in Figure 20.7. Next, select `stlrook.GAL` as the weights file (Figure 20.8 on p. 153).

Finally, check the `Cluster Map` option in the results window dialog, shown in Figure 20.9 on p. 153. Click `OK` to generate the map.

Right click to bring up the `Options` dialog and set the `Randomization` to 9999. Run several permutations until the pattern shown stabilizes to that on the right hand panel of Figure 20.10 on p. 154. Note the slight differences with the cluster map for the raw rates, reproduced in the left hand panel of Figure 20.10.

Focus on the spatial outlier in Morgan county (click on the county in the left hand map and identify it in the table; it is in row 9). Select its neighbors in the map, as in Figure 20.11 on p. 154 (use `shift-click`



SELECT WEIGHT

● Select from currently used

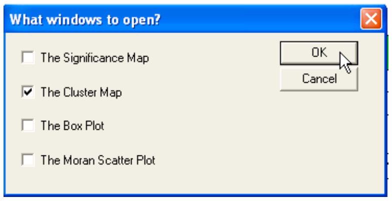
C:\Program Files\GeoDa\Sample Data\stlrook.GAL

○ Select from file (.gal, gwt)

C:\Program Files\GeoDa\Sample Data\stlrook.GAL

□ Set as default

Figure 20.8: Spatial weights selection for EB LISA.



What windows to open?

□ The Significance Map

☑ The Cluster Map

□ The Box Plot

□ The Moran Scatter Plot

Figure 20.9: LISA results window cluster map option.

to add to the selection) and promote the selected counties in the table. Consider the values for `HR8893`, `HC8893` and `PO8893` in the table shown in Figure 20.12 on p. 154. Note how Morgan county has a fairly high homicide rate (`4.581251`), but not the highest in the group (that is `6.029489` for Sangamon county, IL – `FIPS 17167`). More importantly, however, the two lowest values in the group, counties with zero homicides over the period, also have the smallest population sizes (`33911` and `35051` respectively). The EB standardization will tend to pull their homicide rate estimates up, hence lessening the difference with Morgan county, and removing the suggestion of a spatial outlier.

Go through a similar exercise to assess what happened for the county in the low-low cluster that was eliminated after EB standardization.

20.5 Practice

Use the same data sets and rate variables as in Excercises 18 and 19 to assess the sensitivity of your inference to the variance instability of rates.

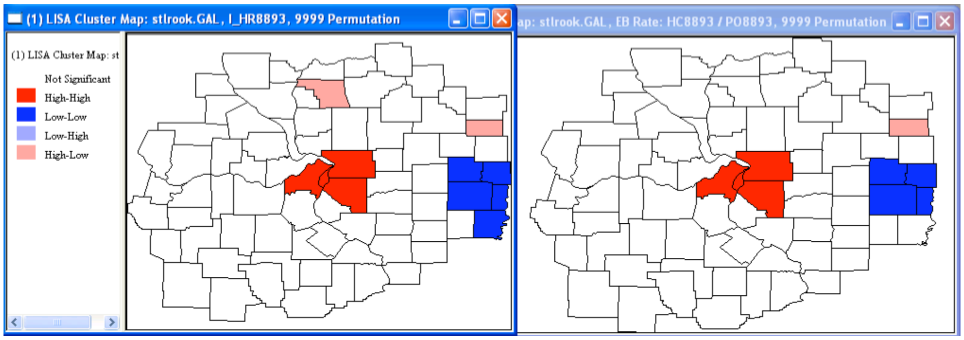

(1) LISA Cluster Map: stlrook.GAL , I_HR8893, 9999 Permutation
(1) LISA Cluster Map: st
Not Significant
🟥 High-High
🟦 Low-Low
Low-High
High-Low
ap: stlrook.GAL , EB Rate: HC8893 / PO8893, 9999 Permutation

Figure 20.10: LISA cluster map for raw and EB adjusted rates.

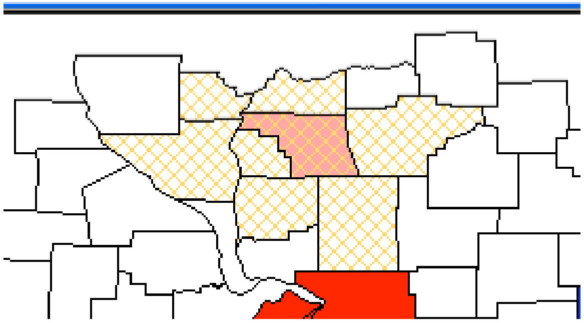
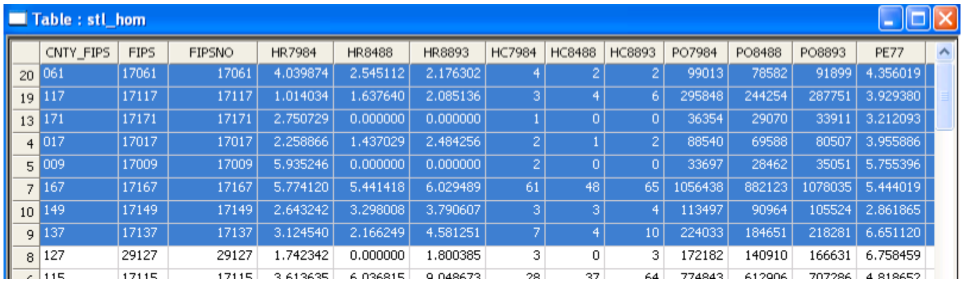


Figure 20.11: Sensitivity analysis of LISA rate map: neighbors.



	CNTY_FIPS	FIPS	FIPSNO	HR7984	HR8488	HR8893	HC7984	HC8488	HC8893	PO7984	PO8488	PO8893	PE77
20	061	17061	17061	4.039874	2.545112	2.176302	4	2	2	99013	78582	91899	4.356019
19	117	17117	17117	1.014034	1.637640	2.085136	3	4	6	295848	244254	287751	3.929380
13	171	17171	17171	2.750729	0.000000	0.000000	1	0	0	36354	29070	33911	3.212093
4	017	17017	17017	2.258866	1.437029	2.484256	2	1	2	88540	69588	80507	3.955886
5	009	17009	17009	5.935246	0.000000	0.000000	2	0	0	33697	28462	35051	5.755396
7	167	17167	17167	5.774120	5.441418	6.029489	61	48	65	1056438	882123	1078035	5.444019
10	149	17149	17149	2.643242	3.298008	3.790607	3	3	4	113497	90964	105524	2.861865
9	137	17137	17137	3.124540	2.166249	4.581251	7	4	10	224033	184651	218281	6.651120
8	127	29127	29127	1.742342	0.000000	1.800385	3	0	3	172182	140910	166631	6.758459
[illegible]	115	17115	17115	3.613635	6.036815	9.048673	28	37	64	774843	612906	707286	4.818652

Figure 20.12: Sensitivity analysis of LISA rate map: rates.

Exercise 21

Bivariate Spatial Autocorrelation

21.1 Objectives

This exercise focuses on the extension of the Moran scatter plot and LISA maps to bivariate spatial association, of which space-time association is a special case. Methodological details can be found in Anselin et al. (2002).

At the end of the exercise, you should know how to:

- create and interpret a bivariate Moran scatter plot
- construct a Moran scatter plot matrix
- interpret the various forms of space-time association
- create and interpret a bivariate LISA map

More detailed information on these operations can be found in the *User's Guide*, pp. 94–96, 105.

21.2 Bivariate Moran Scatter Plot

Start the exercise by loading the polygon shape file with the Thiessen polygons for the 30 Los Angeles air quality monitors from the OZ9799 example data set (use `ozthies.shp` with `STATION` as the `Key`). The base map should be as in Figure 21.1 on p. 156. If you don't have the Thiessen polygons yet, follow the instructions in Section 6.3 on p. 40 to create the necessary shape file.

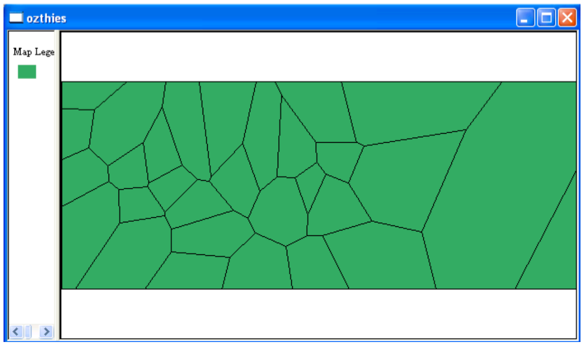


Figure 21.1: Base map with Thiessen polygons for Los Angeles monitoring stations.

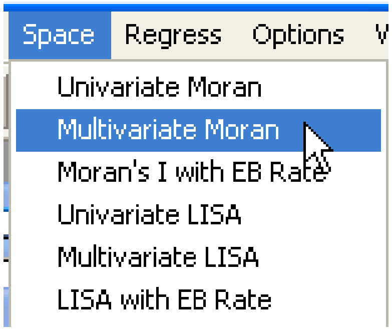
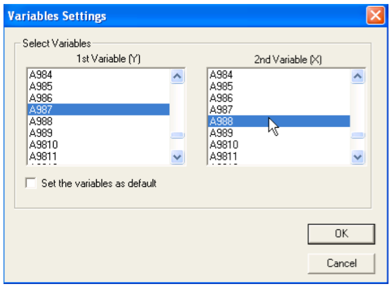


Figure 21.2: Bivariate Moran scatter plot function.

You will also need a rook contiguity weights file for the Thiessen polygons (`ozrook.GAL`). Again, create this weights file if needed (see Section 15.2 on p. 106 for detailed instructions).

Invoke the bivariate Moran scatter plot function from the menu, as `Space > Multivariate Moran` (Figure 21.2), or by clicking the matching toolbar button. This brings up the variable settings dialog, shown in Figure 21.3 on p. 157. Note that there are two columns for variables. The one on the left (`Y`) is for the *spatially lagged variable*, the one on the right (`X`) for the non-lagged variable. Specify `A987` (average 8 hour ozone measure for July 1998) for the lag, and `A988` (average 8 hour ozone measure for August 1998) for the x-variable, as in Figure 21.3, and click `OK` to bring up the weights selection dialog, shown in Figure 21.4 on p. 157.



Variables Settings

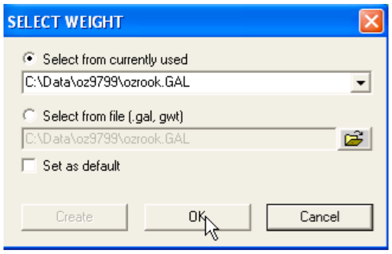
Select Variables

1st Variable (Y)
A984
A985
A986
A987
A988
A989
A9810
A9811

2nd Variable (X)
A984
A985
A986
A987
A988
A989
A9810
A9811

□ Set the variables as default

Figure 21.3: Variable selection for bivariate Moran scatter plot.



SELECT WEIGHT

● Select from currently used
C:\Data\oz9799\ozrook.GAL

○ Select from file (.gal, gwt)
C:\Data\oz9799\ozrook.GAL

□ Set as default

Figure 21.4: Spatial weights selection for bivariate Moran scatter plot.

Select `ozrook.GAL` as the weights file and click on `OK` to generate the bivariate Moran scatter plot, shown in Figure 21.5 on p. 158. Note how the proper interpretation of this graph refers to a generalization of Moran's I to assess the extent to which the value at a location for the x-variable (`A988`) is correlated with the weighted average of another variable (`A987`), with the average computed over the neighboring locations. As a Moran scatter plot, all the standard options are implemented, such as randomization, randomization envelopes and saving the intermediate variables. Also, as a scatter plot, the `Exclude Selected` option can be set to facilitate linking and brushing (see Exercise 18, p. 129 for a more extensive discussion of these standard options).

21.2.1 Space-Time Correlation

The particular example used in Figure 21.5 pertains to the same variable (average 8 hour ozone measurement) observed at two different points in time. Space-time correlation is thus considered as a special case of general bivariate

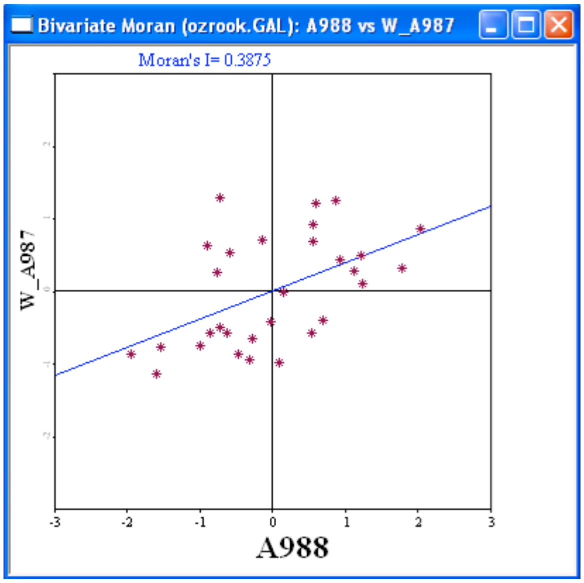


Figure 21.5: Bivariate Moran scatter plot: ozone in 988 on neighbors in 987.

spatial correlation. However, the association depicted in Figure 21.5 is not the only interesting one. It is perfectly reasonable, for example, to switch the roles of the x-variable and spatial lag.

Invoke the bivariate Moran scatter plot function with `A988` as the y-variable (spatial lag) and `A987` as the x-variable. The result should be as in Figure 21.6 on p. 159. This depicts the correlation of ozone in July 98 at a location with the average for its neighbors in August 1998. One can think of the bivariate correlation in Figure 21.5 as focusing on an inward *diffusion* (from the neighbors now to the core in the future), whereas the bivariate correlation in Figure 21.6 refers to an outward diffusion (from the core now to the neighbors in the future). Each is a slightly different view of space-time correlation.

The space-time correlation can be decomposed into a pure spatial autocorrelation, as in the Moran scatter plots shown in Figure 21.7 on p. 159, and a pure serial (time-wise) correlation, as in the correlation plot of Figure 21.8 on p. 160.

More interesting as an alternative to the Moran's I analogue is to use a space-time regression. Since the spatial lag and the original variable pertain to different time periods, it is perfectly legitimate to explain the latter by

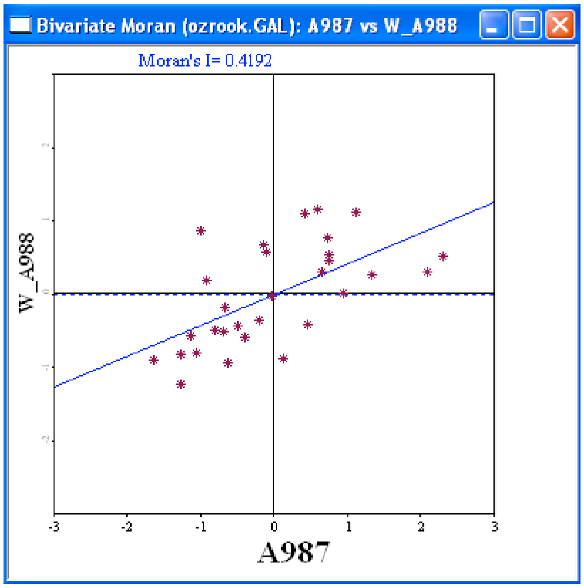


Figure 21.6: Bivariate Moran scatter plot: ozone in 987 on neighbors in 988.

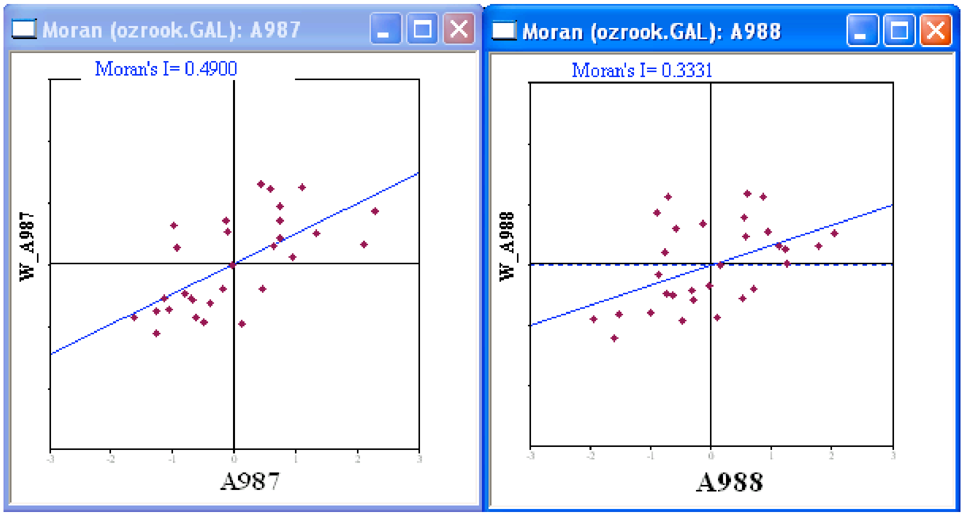


Figure 21.7: Spatial autocorrelation for ozone in 987 and 988.

the former. In a pure cross-sectional context, this is invalid, due to the endogeneity (simultaneous equation bias) of the spatial lag, but in a space-

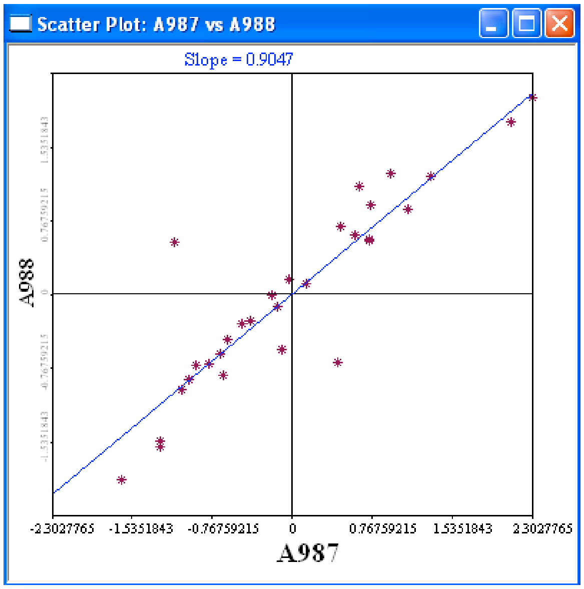


Figure 21.8: Correlation between ozone in 987 and 988.

time setting there is no such problem.

For example, construct the spatial lag for `A987`, using the techniques outlined in Exercise 17 (p. 124). Next, create a scatter plot with `A988` as y-variable and the spatial lag of `A987` as the x-variable. The result should be as in Figure 21.9 on p. 161. The slope in this scatter plot (`0.5749`) is the space-time regression coefficient for the ozone variable. It can be compared to the two space-time Moran's I coefficients. In addition, its sensitivity to particular high leverage observations can be assessed in the usual fashion, using brushing and linking.

21.3 Moran Scatter Plot Matrix

A combination of the different perspectives offered by the cross-sectional spatial autocorrelation and the space-time coefficient can be implemented in an extension of the idea of a scatter plot matrix (see Section 9.2 on p. 61). A so-called *Moran scatter plot matrix* consists of the cross-sectional Moran scatter plot on the main diagonal and the space-time plots on the off-diagonal positions.

Use the two space-time Moran scatter plots and the cross-sectional plots

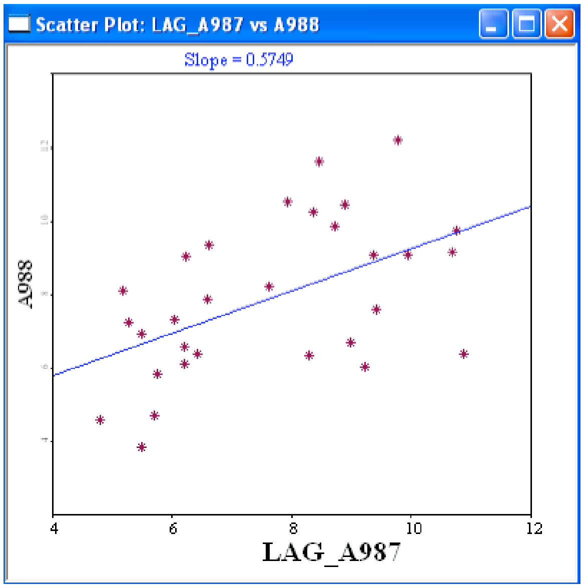


Figure 21.9: Space-time regression of ozone in 988 on neighbors in 987.

just created to arrange them in a matrix, as in Figure 21.10. Make sure to set the `Exclude Selected` option to `ON` in each of the graphs (this must be done one at a time). Then use brushing to explore the associations between the different measures of spatial and space-time correlation, and to identify influential observations (locations).

21.4 Bivariate LISA Maps

The bivariate LISA is a straightforward extension of the LISA functionality to two different variables, one for the location and another for the average of its neighbors. Invoke this function from the menu as `Space > Multivariate LISA`, as in Figure 21.11 on p. 163, or click the matching toolbar button. This brings up the same variable settings dialog as before. Select `A987` as the y-variable and `A988` as the x-variable, in the same way as depicted in Figure 21.3 (p. 157). Note that the y-variable is the one with the spatial lag (i.e., the average for the neighbors). As before, select `ozrook.GAL` as the spatial weights file (see Figure 21.4 on p. 157).

In the results window dialog, check `Cluster Map`, as in Figure 21.12 on p. 163. Click `OK` to generate the bivariate LISA cluster map, shown

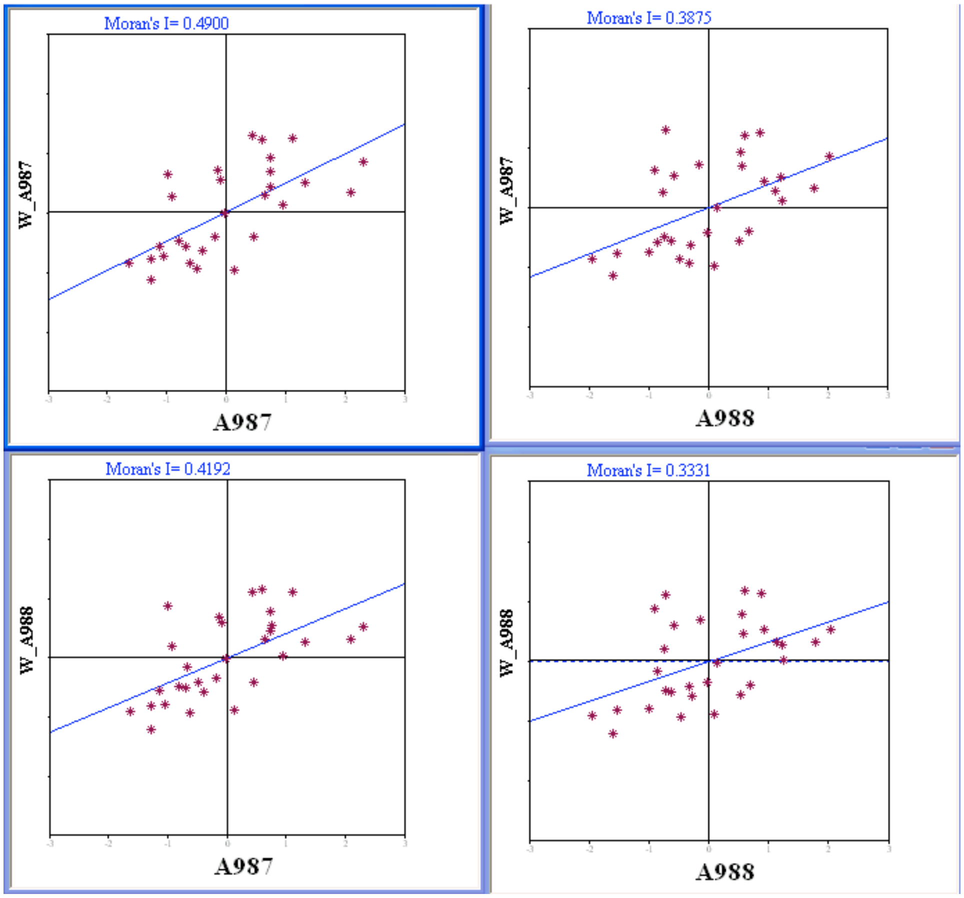


Figure 21.10: Moran scatter plot matrix for ozone in 987 and 988.

in Figure 21.13 on p. 164. Note that this shows local patterns of spatial correlation at a location between ozone in August 1998 and the average for its neighbors in July 1998. Switching the selection of y and x variables in the variables settings dialog would create a LISA map of ozone at a location in July 1998 and the average for its neighbors in August 1998. The interpretation is similar to that of a space-time scatter plot. Compare the two bivariate LISA maps to their cross-sectional counterparts.

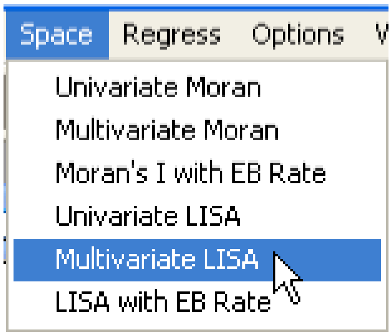
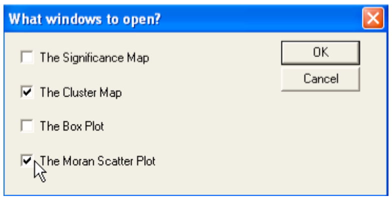


Figure 21.11: Bivariate LISA function.


What windows to open?

☐ The Significance Map

☑ The Cluster Map

☐ The Box Plot

☑ The Moran Scatter Plot

Figure 21.12: Bivariate LISA results window options.

21.5 Practice

Several of the sample data sets contain variables observed at multiple points in time. Apart from the many measures included in the Los Angeles ozone data set, this includes the St Louis homicide data (`stl_hom.shp` with `FIPSNO` as the `Key`), the SIDS data (`sids2.shp` with `FIPSNO` as the `Key`), and the Ohio lung cancer data (`ohlung.shp` with `FIPSNO` as the `Key`). Alternatively, consider a bivariate analysis between two variables that do not contain a time dimension.

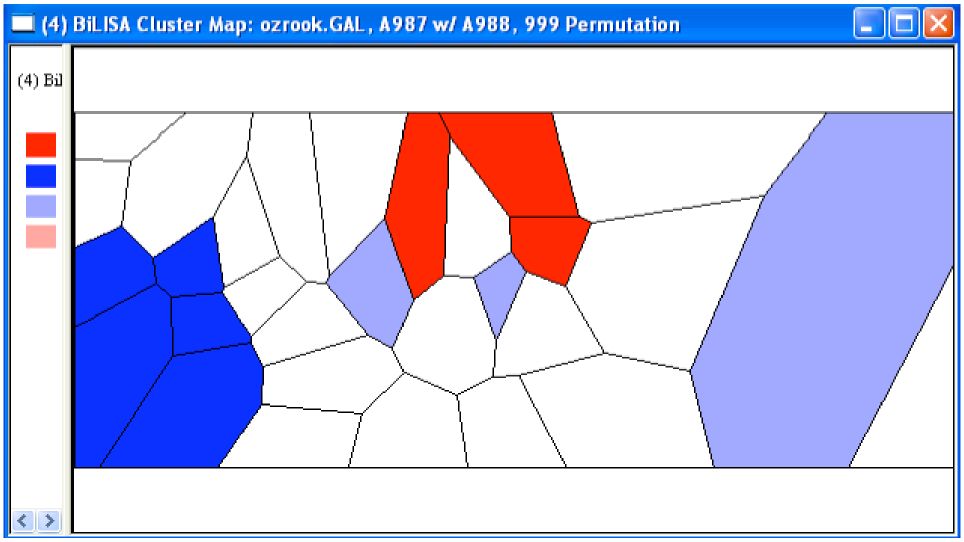

(4) BiLISA Cluster Map: ozrook.GAL, A987 w/ A988, 999 Permutation
(4) Bil

Figure 21.13: Bivariate LISA cluster map for ozone in 988 on neighbors in 987.

Exercise 22

Regression Basics

22.1 Objectives

This exercise begins the review of spatial regression functionality in *GeoDa*, starting with basic concepts. Methodological background for multivariate regression analysis can be found in many econometrics texts and will not be covered here. The discussion of regression diagnostics and specific spatial models is left for Exercises 23 to 25.

At the end of the exercise, you should know how to:

- set up the specification for a linear regression model
- run ordinary least squares estimation (OLS)
- save OLS output to a file
- add OLS predicted values and residuals to the data table
- create maps with predicted values and residuals

More detailed information on these operations can be found in the *Release Notes*, pp. 45–56.

22.2 Preliminaries

In this exercise, we will use the *classic* Columbus neighborhood crime data (Anselin 1988, pp. 188–190) contained in the `columbus.shp` sample data set (use `POLYID` as the `Key`). The base map should be as in Figure 22.1 on p. 166.

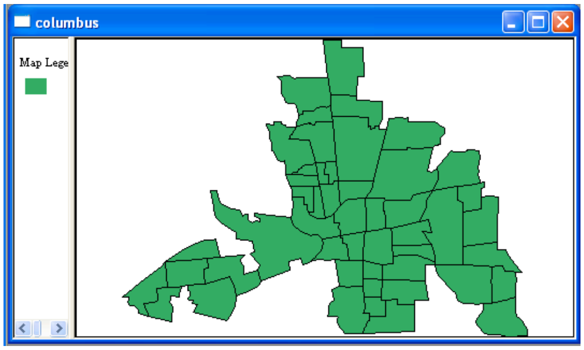


Figure 22.1: Columbus neighborhood crime base map.

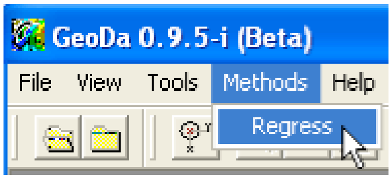


Figure 22.2: Regression without project.

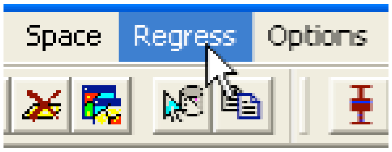
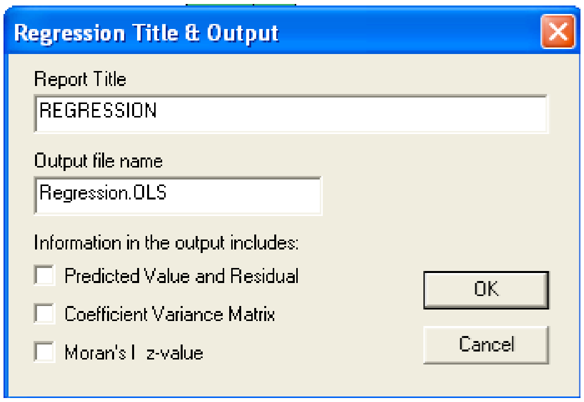


Figure 22.3: Regression inside a project.

In *GeoDa*, the regression functionality can be invoked without opening a project. This is particularly useful in the analysis of large data sets (10,000 and more) when it is better to avoid the overhead of linking and brushing the data table. To start a regression from the *GeoDa* opening screen (Figure 1.1 on p. 2), select `Methods > Regress`, as in Figure 22.2. Alternatively, when a project is open (i.e., after a shape file has been loaded), invoke the `Regress` command directly from the main menu bar, as in Figure 22.3. This brings up the default regression title and output dialog, shown in Figure 22.4 on p. 167.

There are two important aspects to this dialog, the output file name and the options for output. The `Report Title` can be safely ignored as it is not currently used. The file specified in `Output file name` will contain the regression results in a rich text format (RTF) file in the current working directory.[1] The default is `Regression.OLS`, which is usually not very mean-

[1] A rich text format file is a text file with additional formatting commands. It is often used as a file interchange format for Microsoft Word documents. It can be opened by many simple text editors as well, such as Wordpad, but *not* Notepad.



Report Title: REGRESSION

Output file name: Regression.OLS

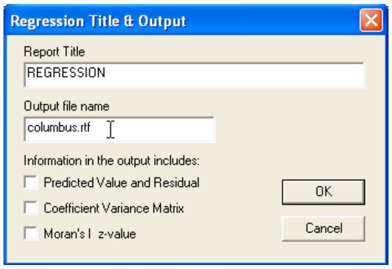
Information in the output includes:

□ Predicted Value and Residual

□ Coefficient Variance Matrix

□ Moran's I z-value

Figure 22.4: Default regression title and output dialog.



Report Title: REGRESSION

Output file name: columbus.rtf

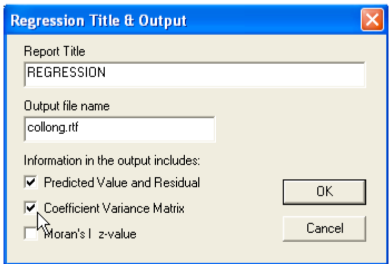
Information in the output includes:

□ Predicted Value and Residual

□ Coefficient Variance Matrix

□ Moran's I z-value

Figure 22.5: Standard (short) output option.



Report Title: REGRESSION

Output file name: collong.rtf

Information in the output includes:

☑ Predicted Value and Residual

☑ Coefficient Variance Matrix

□ Moran's I z-value

Figure 22.6: Long output options.

ingful as a file name. Instead, enter something that gives a hint about the type of analysis, such as `columbus.rtf`, shown in Figures 22.5 and 22.6.[2]

The dialog also contains a number of check boxes to specify long output options. The default is to leave them unchecked, as in Figure 22.5. Long output is created by checking the respective boxes, such as `Predicted Value and Residual` and `Coefficient Variance Matrix` in Figure 22.6.[3]

The option for predicted values and residuals should be used with caution, especially for large data sets. It adds two vectors to the regression output window and file whose length equals the number of observations. This can quickly get out of hand, even for medium sized data sets.

[2] When you run several regressions, make sure to specify a *different* output file for each analysis, otherwise all the results will be written to the same file (usually the default).

[3] The option for `Moran's I z-value` will be discussed in Exercise 23.

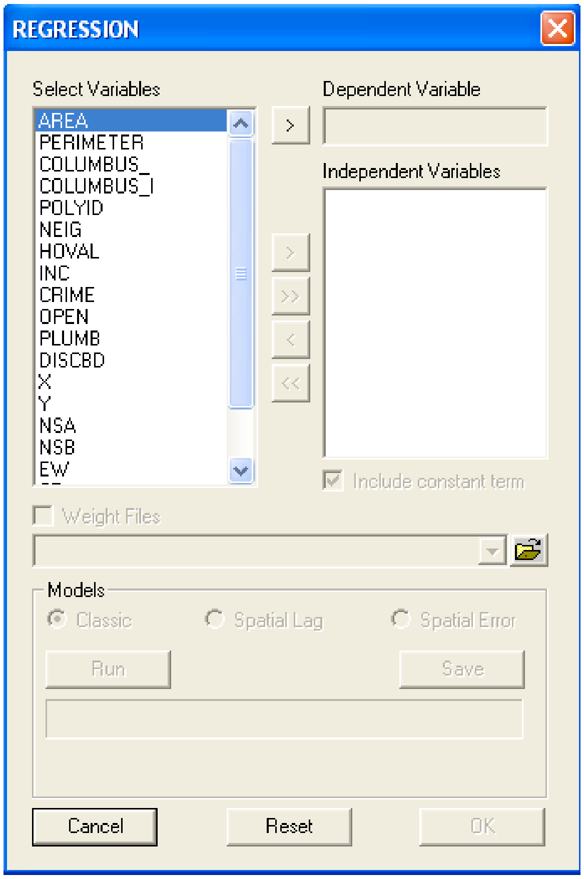
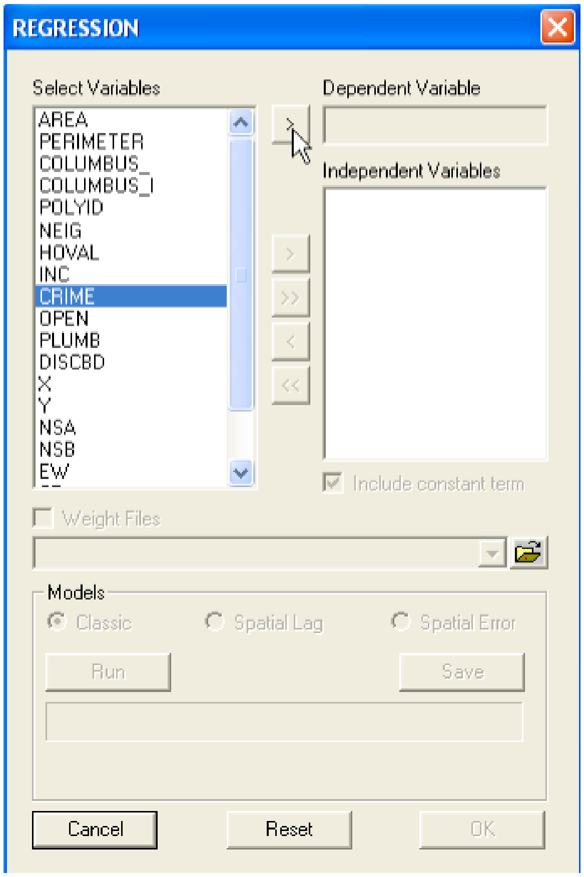


Figure 22.7: Regression model specification dialog.

The coefficient variance matrix provides not only the variance of the estimates (on the diagonal) but also all covariances. This matrix can be used to carry out customized tests of contraints on the model coefficients outside of *GeoDa*. If this is not of interest, this option can safely be left *off*. Also, it is important to note that the long output options do *not* need to be checked in order to add predicted values or residuals to the data table (see Section 22.4.1). They only affect what is listed in the output window and file.

Click on the `OK` button in the title and output dialog to bring up the regression model specification dialog, shown in Figure 22.7.



AREA
PERIMETER
COLUMBUS_
COLUMBUS_I
POLYID
NEIG
HOVAL
INC
CRIME
OPEN
PLUMB
DISCBD
X
Y
NSA
NSB
EW

☑ Include constant term

□ Weight Files

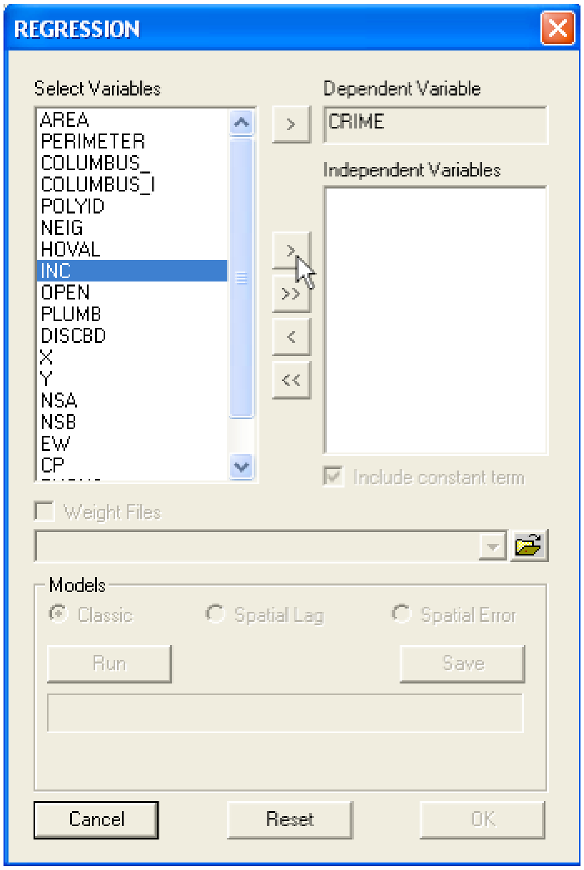
● Classic ○ Spatial Lag ○ Spatial Error

Figure 22.8: Selecting the dependent variable.

22.3 Specifying the Regression Model

The regression dialog shown in Figure 22.7 on p. 168 is the place where the dependent and explanatory variables are selected, as well as the spatial weights file and the type of regression to carry out. For now, we will limit our attention to estimation with the `Classic` option only (the default), and not specify a spatial weights file.

First, select `CRIME` as the dependent variable, by clicking on the variable name in the `Select Variables` column and on the `>` button next to `Dependent Variable`, as in Figure 22.8. Only the `Cancel` button is active before a dependent variable is selected. This remains the case until at least



Dependent Variable
CRIME

☑ Include constant term

□ Weight Files

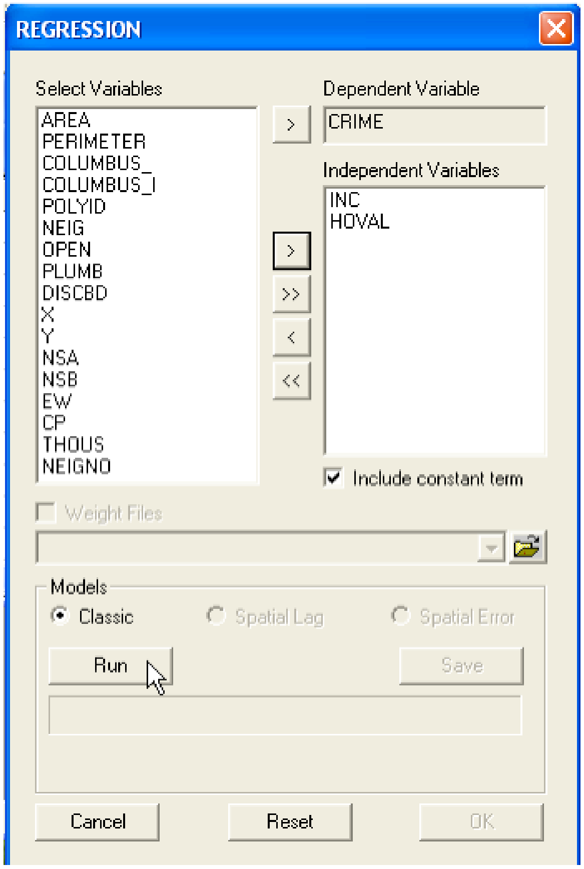
● Classic ○ Spatial Lag ○ Spatial Error

Figure 22.9: Selecting the explanatory variables.

one `Independent Variable` has been selected as well. The latter are specified in the same fashion. Select `INC` in the `Select Variables` column and move it to the `Independent Variables` list by clicking on the `>` button, as shown in Figure 22.9. Repeat this process for `HOVAL` and the basic regression model specification is complete, as shown in Figure 22.10 on p. 171. Note how the `Run` and `Reset` buttons are now active as well. Use the latter if you want to respecify the model before running it.[4]

Note how the `Include constant term` option is checked *by default*.

[4] In the current version of *GeoDa*, there is no way to save a model specification for a future run. The interactive process of entering dependent and explanatory variables must be repeated for each analysis.



Dependent Variable: CRIME

Independent Variables: INC, HOVAL

☑ Include constant term

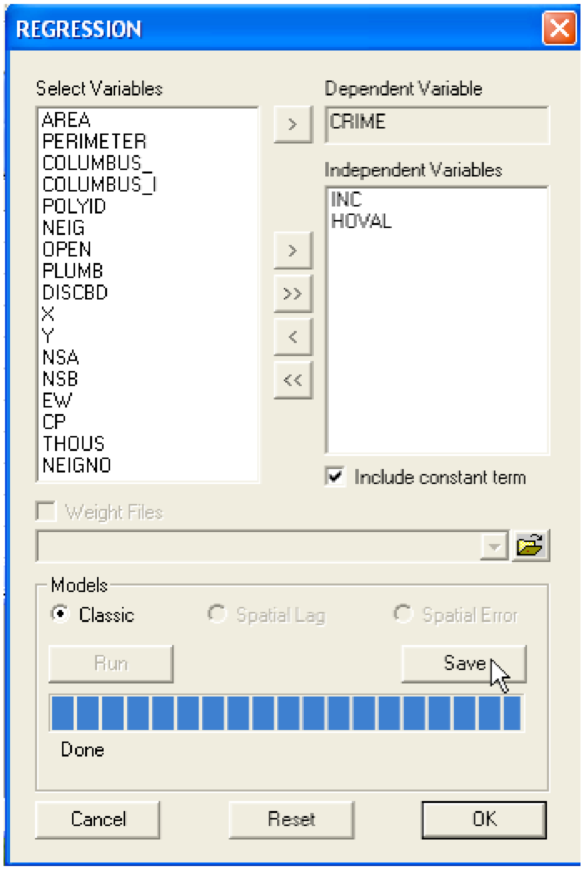
Models: ● Classic

Figure 22.10: Run classic (OLS) regression.

Uncheck this only if you have a very good reason to run a model without a constant term.

To run the regression, click the `Run` button. A progress bar will appear and show when the estimation process has finished (for OLS this is typically a very short time), as in Figure 22.11 on p. 172.

At this point, the regression results window can be brought up by clicking on `OK`. However, if you want to add predicted values and/or residuals to the data table, you must select `Save` *first*. In the current version of *GeoDa*, once the output window is open, the regression dialog closes, and it is no longer possible to go back and save these items.



Dependent Variable: CRIME

Independent Variables: INC, HOVAL

☑ Include constant term

□ Weight Files

Models: ● Classic ○ Spatial Lag ○ Spatial Error

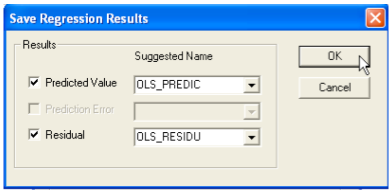
Figure 22.11: Save predicted values and residuals.

22.4 Ordinary Least Squares Regression

22.4.1 Saving Predicted Values and Residuals

If you want to add the predicted values and/or residuals to the current data table, do *not* select the `OK` button, but instead, click on `Save`, as in Figure 22.11. This brings up a dialog to specify the variable names for residuals and/or predicted values, as shown in Figure 22.12 on p. 173.

In this dialog, you can check the box next to `Predicted Value` and/or `Residual` and either keep the default variable names (`OLS_PREDIC` for the predicted value and `OLS_RESIDU` for the residuals), or replace them with more meaningful names (simply overwrite the defaults). Click `OK` to add the



Save Regression Results

Results

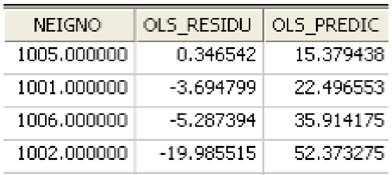
Suggested Name

☑ Predicted Value OLS_PREDIC

□ Prediction Error

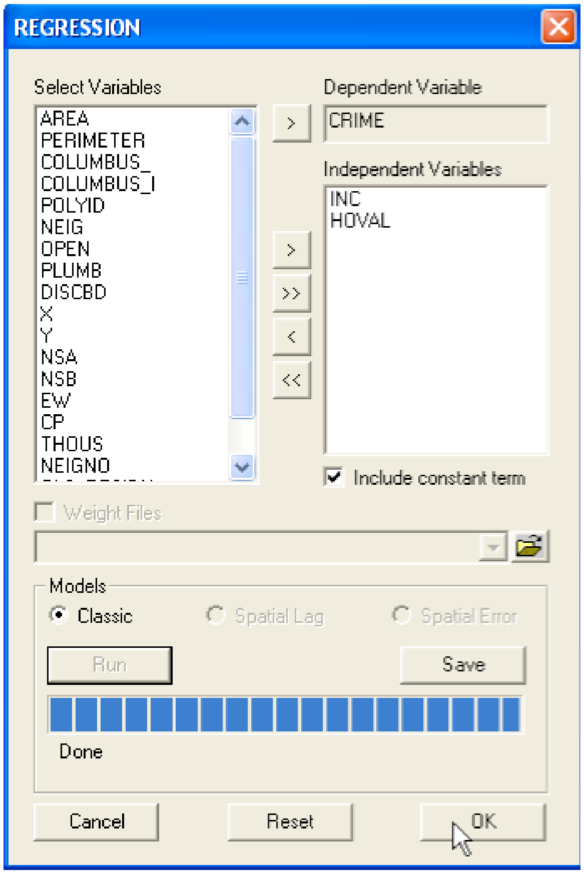
☑ Residual OLS_RESIDU

Figure 22.12: Predicted values and residuals variable name dialog.



NEIGNO	OLS_RESIDU	OLS_PREDIC
1005.000000	0.346542	15.379438
1001.000000	-3.694799	22.496553
1006.000000	-5.287394	35.914175
1002.000000	-19.985515	52.373275

Figure 22.13: Predicted values and residuals added to table.



REGRESSION

Select Variables

AREA
PERIMETER
COLUMBUS_
COLUMBUS_I
POLYID
NEIG
OPEN
PLUMB
DISCBD
X
Y
NSA
NSB
EW
CP
THOUS
NEIGNO

Dependent Variable: CRIME

Independent Variables:
INC
HOVAL

☑ Include constant term

□ Weight Files

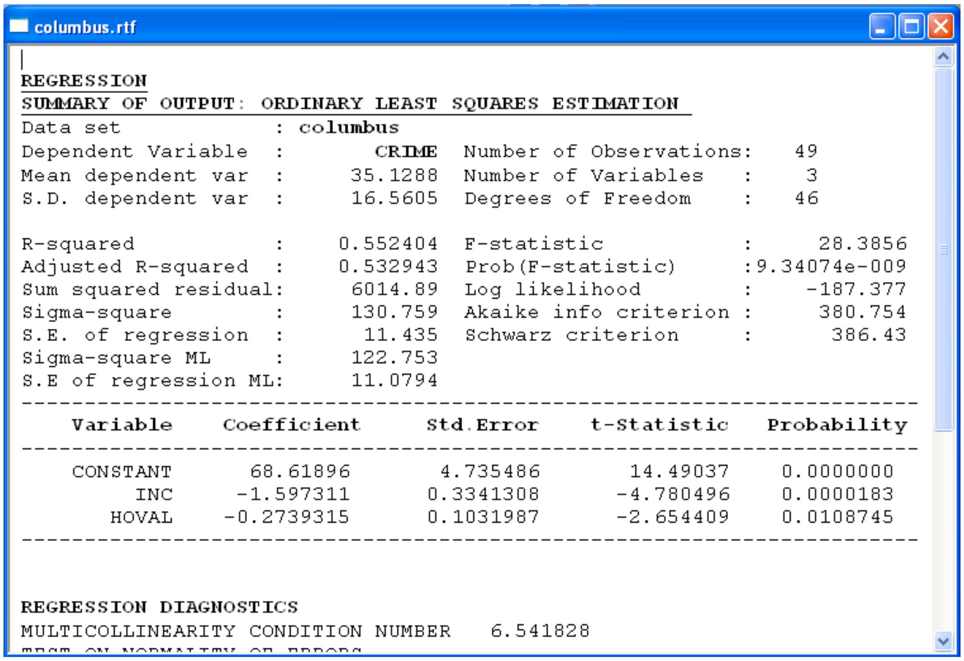
Models

● Classic ○ Spatial Lag ○ Spatial Error

Done

Figure 22.14: Showing regression output.

selected columns to the data table. The result, with both options checked,



```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : columbus
Dependent Variable  :        CRIME  Number of Observations:   49
Mean dependent var  :      35.1288  Number of Variables   :    3
S.D. dependent var  :      16.5605  Degrees of Freedom    :   46

R-squared           :     0.552404  F-statistic           :     28.3856
Adjusted R-squared  :     0.532943  Prob(F-statistic)     :9.34074e-009
Sum squared residual:      6014.89  Log likelihood        :    -187.377
Sigma-square        :      130.759  Akaike info criterion :     380.754
S.E. of regression  :       11.435  Schwarz criterion     :      386.43
Sigma-square ML     :      122.753
S.E of regression ML:      11.0794
-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT       68.61896       4.735486       14.49037     0.0000000
         INC      -1.597311      0.3341308      -4.780496     0.0000183
       HOVAL     -0.2739315      0.1031987      -2.654409     0.0108745
-----------------------------------------------------------------------


REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   6.541828
TEST ON NORMALITY OF ERRORS
```

Figure 22.15: Standard (short) OLS output window.

is as in Figure 22.13 on p. 173. Remember to save the shapefile under a different file name to make the new variables permanently part of the dbf file.

22.4.2 Regression Output

Click on `OK` in the regression variable dialog (Figure 22.14 on p. 173) to bring up the results window, shown in Figure 22.15. The top part of the window contains several summary characteristics of the model as well as measures of fit. This is followed by a list of variable names, with associated coefficient estimates, standard error, t-statistic and probability (of rejecting the null hypothesis that $\beta = 0$).[5] Next are given a list of model diagnostics, a discussion of which is left for Exercise 23.

The summary characteristics of the model listed at the top include the name of the data set (`columbus`), the dependent variable (`CRIME`), its mean (`35.1288`) and standard deviation (`16.5605`). In addition, the number of

[5] A small probability, such as $p < 0.05$, suggests rejection of the null.

observations are listed (49), the number of variables included in the model (inclusive of the constant term, so, the value is 3), and the degrees of freedom (46).

In the left hand column of the standard output are traditional measures of fit, including the R^2 (0.552404) and adjusted R^2 (0.532943), the sum of squared residuals (6014.89), and the residual variance and standard error estimate, both with adjustment for a loss in degrees of freedom (`Sigma-square` and `S.E. of regression`) as well as without (`Sigma-square ML` and `S.E. of regression ML`).[6]

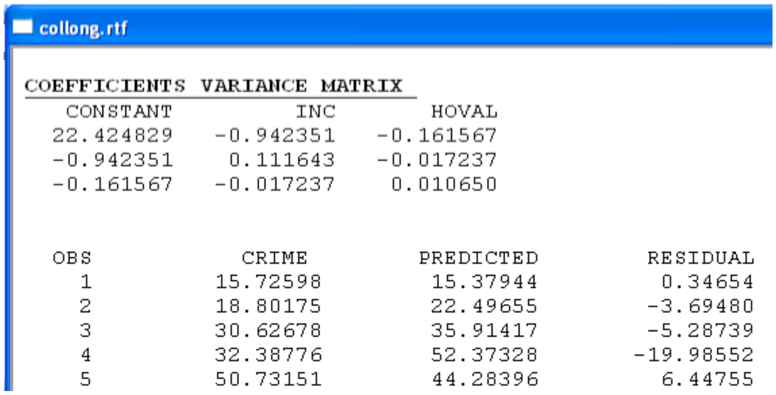
In the right hand column are listed the F-statistic on the null hypothesis that all regression coefficients are jointly zero (28.3856), and the associated probability (9.34074e-009). This test statistic is included for completeness sake, since it typically will reject the null hypothesis and is therefore not that useful.

Finally, this column contains three measures that are included to maintain comparability with the fit of the spatial regression models, treated in Exercises 24 and 25. They are the log likelihood (-187.377), the Akaike information criterion (380.754) and the Schwarz criterion (386.43). These three measures are based on an assumption of multivariate normality and the corresponding likelihood function for the standard regression model. The higher the log-likelihood, the better the fit (high on the real line, so less negative is better). For the information criteria, the direction is opposite, and the lower the measure, the better the fit.[7]

When the long output options are checked in the regression title dialog, as in Figure 22.6 on p. 167, an additional set of results is included in the output window. These are the full covariance matrix for the regression coefficient estimates, and/or the predicted values and residuals for each observation. These results are listed after the diagnostics and are illustrated in Figure 22.16 on p. 176. The variable names are given on top of the columns of the covariance matrix (this matrix is symmetric, so the rows match the columns). In addition, for each observation, the observed dependent variable is listed, as well as the predicted value and residual (observed less predicted).

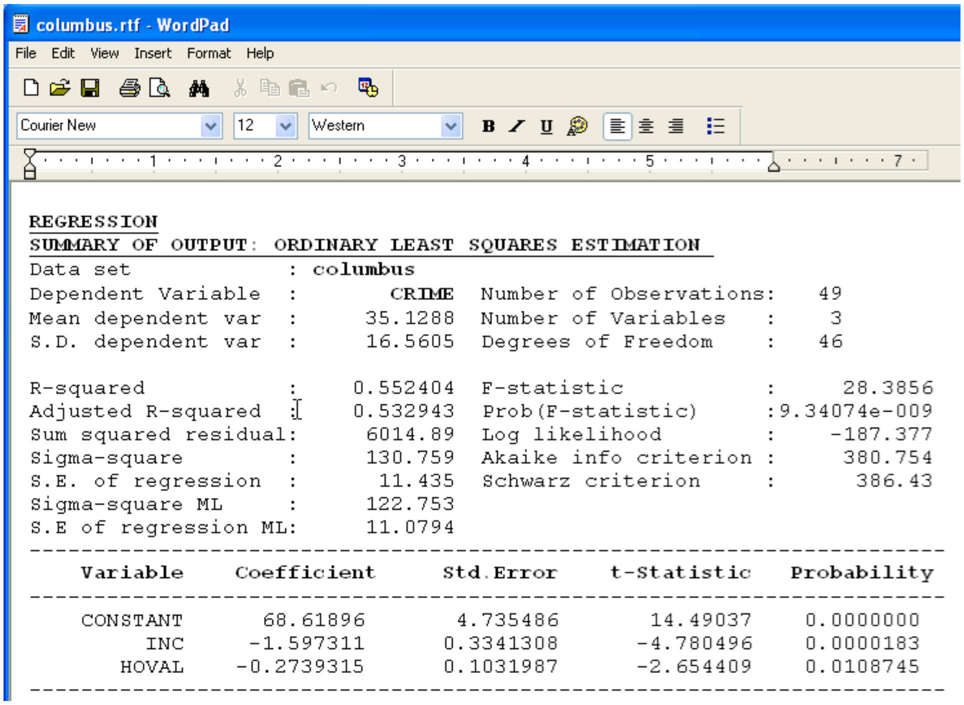
[6] The difference between the two is that the first divides the sum of squared residuals by the degrees of freedom (46), the second by the total number of observations (49). The second measure will therefore always be smaller than the first, but for large data sets, the difference will become negligible.

[7] The AIC $= -2L+2K$, where L is the log-likelihood and K is the number of parameters in the model, here 3. Hence, in the Columbus example: AIC $= -2 \times (-187.377) + 2 \times 3 = 380.754$. The SC $= -2L + K.\ln(N)$, where ln is the natural logarithm. As a result, in the Columbus example: SC $= -2 \times (-187.377) + 3 \times 3.892 = 386.43$.



```
COEFFICIENTS VARIANCE MATRIX
   CONSTANT          INC        HOVAL
  22.424829    -0.942351    -0.161567
  -0.942351     0.111643    -0.017237
  -0.161567    -0.017237     0.010650


  OBS             CRIME          PREDICTED          RESIDUAL
    1          15.72598           15.37944           0.34654
    2          18.80175           22.49655          -3.69480
    3          30.62678           35.91417          -5.28739
    4          32.38776           52.37328         -19.98552
    5          50.73151           44.28396           6.44755
```

Figure 22.16: OLS long output window.

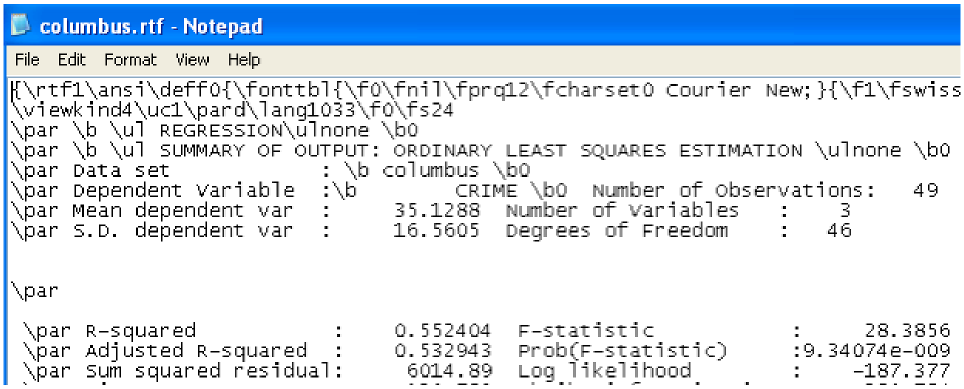


```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : columbus
Dependent Variable  :       CRIME  Number of Observations:   49
Mean dependent var  :     35.1288  Number of Variables   :    3
S.D. dependent var  :     16.5605  Degrees of Freedom    :   46

R-squared           :    0.552404  F-statistic           :     28.3856
Adjusted R-squared  :    0.532943  Prob(F-statistic)     :9.34074e-009
Sum squared residual:     6014.89  Log likelihood        :    -187.377
Sigma-square        :     130.759  Akaike info criterion :     380.754
S.E. of regression  :      11.435  Schwarz criterion     :      386.43
Sigma-square ML     :     122.753
S.E of regression ML:     11.0794
-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT        68.61896       4.735486        14.49037     0.0000000
         INC       -1.597311      0.3341308       -4.780496     0.0000183
       HOVAL      -0.2739315      0.1031987       -2.654409     0.0108745
-----------------------------------------------------------------------
```

Figure 22.17: OLS rich text format (rtf) output file in Wordpad.

22.4.3 Regression Output File

The results of the regression are also written to a file in the current working directory, with a file name specified in the title dialog (Figure 22.6 on



```
{\rtf1\ansi\deff0{\fonttbl{\f0\fnil\fprq12\fcharset0 Courier New;}{\f1\fswiss
\viewkind4\uc1\pard\lang1033\f0\fs24
\par \b \ul REGRESSION\ulnone \b0
\par \b \ul SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION \ulnone \b0
\par Data set             : \b columbus \b0
\par Dependent Variable   :\b         CRIME \b0  Number of Observations:   49
\par Mean dependent var   :     35.1288  Number of Variables    :    3
\par S.D. dependent var   :     16.5605  Degrees of Freedom     :   46


\par

 \par R-squared            :     0.552404  F-statistic            :     28.3856
 \par Adjusted R-squared   :     0.532943  Prob(F-statistic)      :9.34074e-009
 \par Sum squared residual:      6014.89  Log likelihood         :    -187.377
```

Figure 22.18: OLS rich text format (rtf) output file in Notepad.

p. 167). In the current example, this is `columbus.rtf`. The file is in rich text format and opens readily in Wordpad, Word and other word processors, allowing you to cut and past results to other documents. The file contents are illustrated in Figure 22.17 on p. 176.

Note that when you attempt to open this file in a simple text editor like Notepad, the result is as in Figure 22.18, revealing the formatting codes.

22.5 Predicted Value and Residual Maps

When the predicted value and regression residuals are saved to the data table, they become available as variables to any exploratory function in *GeoDa*, including mapping. Such maps, referred to as predicted value maps and residual maps, are useful for a *visual* inspection of patterns. The predicted value map can be thought of as a *smoothed* map, in the sense that random variability, due to factors other than those included in the model has been smoothed out. A residual map may give an indication of systematic over- or underprediction in particular regions, which could be evidence of spatial autocorrelation (to be assessed more rigorously by means of a hypothesis test).

For example, with the Columbus base map loaded (as in Figure 22.1 on p. 166), a quantile map of predicted values (using 6 quantiles) is readily obtained. Select `Map > Quantile` from the menu and take `OLS_PREDIC` as the variable (assuming you added it to the data table, as in Section 22.4.1). Next, change the number of classes from the default `4` to `6`. The resulting

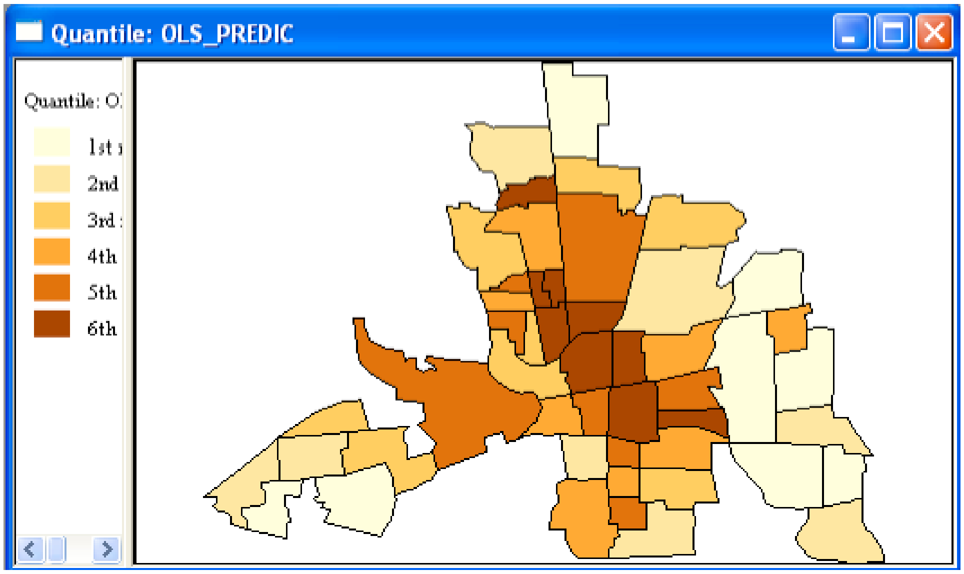


Figure 22.19: Quantile map (6 categories) with predicted values from CRIME regression.

map should be as in Figure 22.19.

Using `Map > St.Dev` and selecting `OLS_RESIDU` yields a standard deviational map for the residuals, as in Figure 22.20 on p. 179. This map does suggest that similarly colored areas tend to be in similar locations, which could indicate positive spatial autocorrelation (a Moran's I test for residual spatial autocorrelation is positive and highly significant). Also, it indicates a tendency to overpredict (negative residuals) in the outlying areas and a tendency to underpredict (positive residuals) in the core, suggesting the possible presence of spatial heterogeneity in the form of spatial regimes. Two large outliers (one positive and one negative) may also warrent further attention.

22.6 Practice

Several of the sample data sets included on the SAL site contain variables that allow the replication of published spatial regression studies. This includes the BOSTON data with the variables from Harrison and Rubinfeld (1978) (see also Gilley and Pace 1996, Pace and Gilley 1997), the POLICE

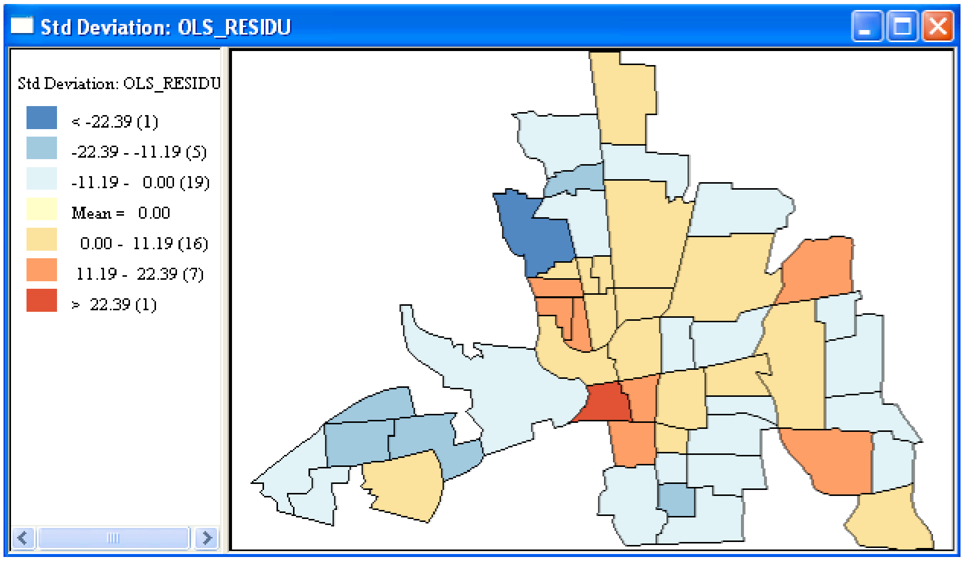


Figure 22.20: Standard deviational map with residuals from CRIME regression.

data set with the variables needed to replicate Kelejian and Robinson (1992), the BALTIMORE data with the variables to rerun the regression in Dubin (1992), the NDVI data with the variables for the example in Anselin (1993), the SOUTH data set with the variables from Baller et al. (2001), and the LASROSAS data set with the variables from Anselin et al. (2004a).

For example, load the POLICE data set and rerun the regression from Kelejian and Robinson (1992), using `POLICE` as the dependent variable, and `TAX`, `INC`, `CRIME`, `UNEMP`, `OWN`, `COLLEGE`, `WHITE` and `COMMUTE` as explanatory variables (see Kelejian and Robinson 1992, pp. 323–324). Save the predicted values and residuals and compare a quantile map of the observed police expenditures to that of the predicted ones. Create a standard deviational map of the residuals and “visually” assess any possible patterns. Alternatively, carry out a similar exercise for any of the other sample data sets listed above.

Exercise 23

Regression Diagnostics

23.1 Objectives

This exercise continues the discussion of spatial regression functionality in *GeoDa*, now focusing on regression diagnostics. Methodological background for standard regression diagnostics, such as the multicollineartity condition number and the test statistics for normality and heteroskedasticity are covered in most econometrics texts and will not be discussed in detail here. Technical aspects of spatial regression diagnostics are reviewed in Anselin (1988), Anselin and Bera (1998), and Anselin (2001), among others.

At the end of the exercise, you should know how to:

- set up the specification for a trend surface regression model
- construct and interpret regression diagnostic plots
- interpret standard regression diagnostics for multicollinearity, non-normality and heteroskedasticity
- interpret regression diagnostics for spatial autocorrelation
- choose an alternative spatial regression model based on the results for the spatial autocorrelation diagnostics

More detailed information on these operations can be found in the *Release Notes*, pp. 48–52.

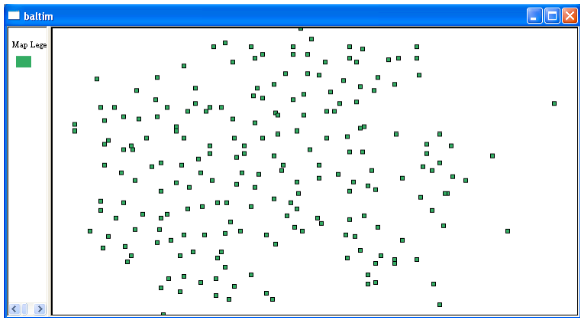


Figure 23.1: Baltimore house sales point base map.

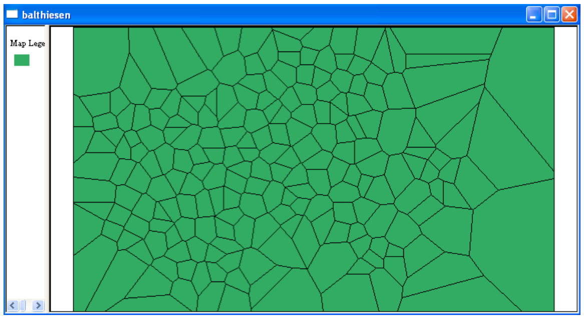
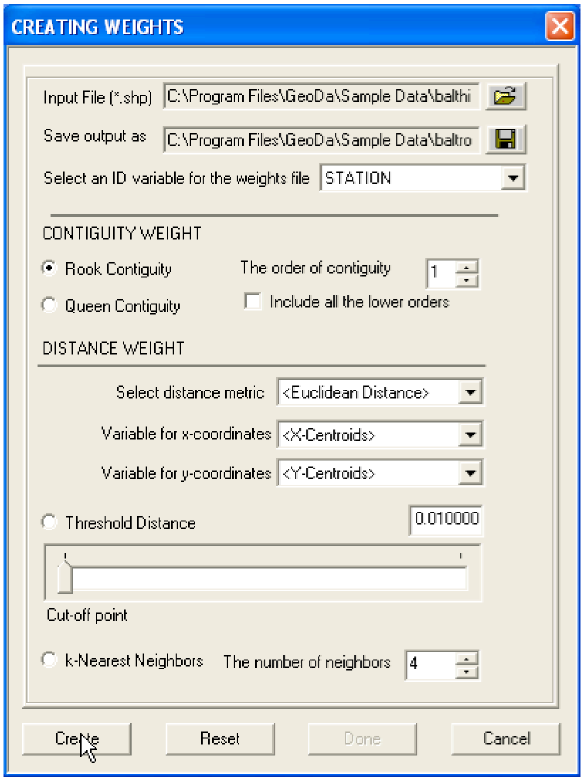


Figure 23.2: Baltimore house sales Thiessen polygon base map.

23.2 Preliminaries

Load the Baltimore sample data set with point data for 211 observations on house sales price and hedonic variables (`baltim.shp` with `STATION` as the `Key`). The base map should be as in Figure 23.1. Also create a Thiessen polygon shape file from these points, say `balthiesen.shp`, with the same `Key` (follow the instructions in Section 6.3). The result should be as in Figure 23.2. Finally, if you have not already done so, construct a rook contiguity spatial weights file (`baltrook.GAL`) from the Thiessen polygons, as illustrated in Figure 23.3 on p. 182 (see also the extensive instructions in Section 15.2).



CREATING WEIGHTS

Input File (*.shp) C:\Program Files\GeoDa\Sample Data\balthi

Save output as C:\Program Files\GeoDa\Sample Data\baltro

Select an ID variable for the weights file STATION

CONTIGUITY WEIGHT

● Rook Contiguity — The order of contiguity 1

○ Queen Contiguity — □ Include all the lower orders

DISTANCE WEIGHT

Select distance metric <Euclidean Distance>

Variable for x-coordinates <X-Centroids>

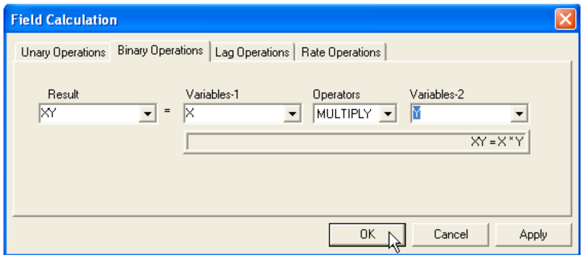
Variable for y-coordinates <Y-Centroids>

○ Threshold Distance 0.010000

Cut-off point

○ k-Nearest Neighbors — The number of neighbors 4

Figure 23.3: Rook contiguity weights for Baltimore Thiessen polygons.

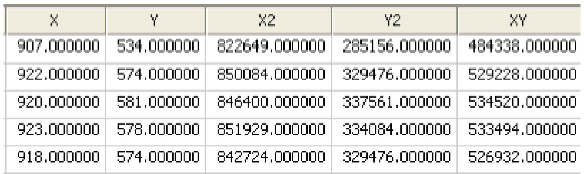


Field Calculation

Result XY = Variables-1 X Operators MULTIPLY Variables-2 Y

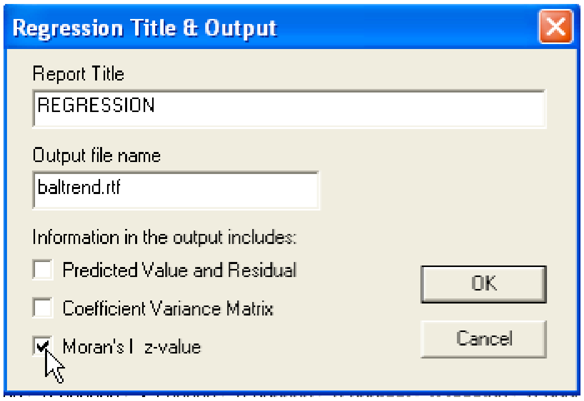
XY = X * Y

Figure 23.4: Calculation of trend surface variables.



X	Y	X2	Y2	XY
907.000000	534.000000	822649.000000	285156.000000	484338.000000
922.000000	574.000000	850084.000000	329476.000000	529228.000000
920.000000	581.000000	846400.000000	337561.000000	534520.000000
923.000000	578.000000	851929.000000	334084.000000	533494.000000
918.000000	574.000000	842724.000000	329476.000000	526932.000000

Figure 23.5: Trend surface variables added to data table.



Regression Title & Output

Report Title: REGRESSION

Output file name: baltrend.rtf

Information in the output includes:

□ Predicted Value and Residual

□ Coefficient Variance Matrix

☑ Moran's I z-value

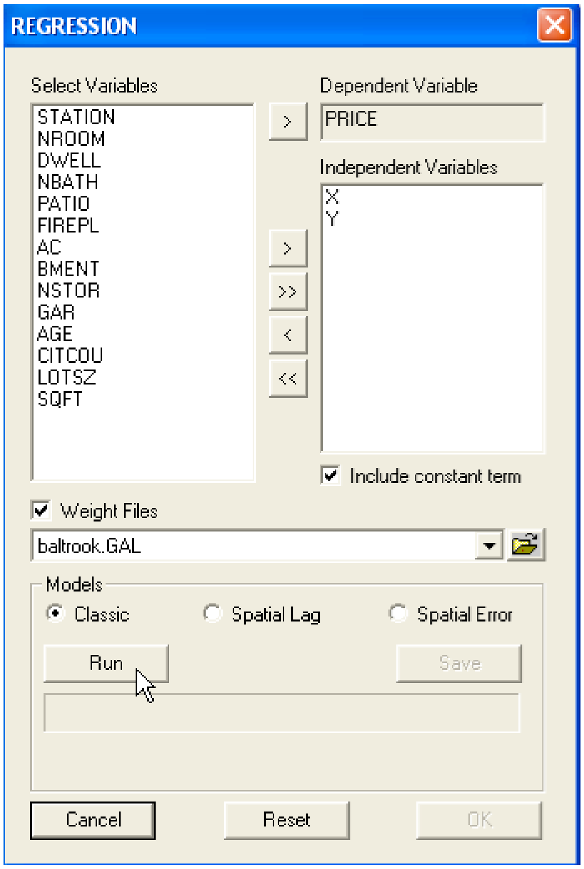
Figure 23.6: Linear trend surface title and output settings.

23.3 Trend Surface Regression

We will analyze regression diagnostics for trend surface regression models of the house price (the variable `PRICE`). Trend surface models are specifications in which the explanatory variables consist of polynomials in the x and y coordinates of the observations. We will consider both a linear trend surface (x and y as explanatory variables) and a quadratic trend surface (x, y, x^2, y^2, and xy as explanatory variables).

23.3.1 Trend Surface Variables

Of the explanatory variables needed in the trend surface regressions, only `X` and `Y` are present in the data set. The quadratic powers and the cross-product must be constructed using the `Table` functionality (see Section 3.4 for details). Select `Field Calculation` from the `Table` menu item and check the tab for `Binary Operations` to compute the squared coordinates and the cross-product. For example, Figure 23.4 on p. 182 shows the



Select Variables
STATION
NROOM
DWELL
NBATH
PATIO
FIREPL
AC
BMENT
NSTOR
GAR
AGE
CITCOU
LOTSZ
SQFT

Dependent Variable: PRICE

Independent Variables:
X
Y

☑ Include constant term

☑ Weight Files: baltrook.GAL

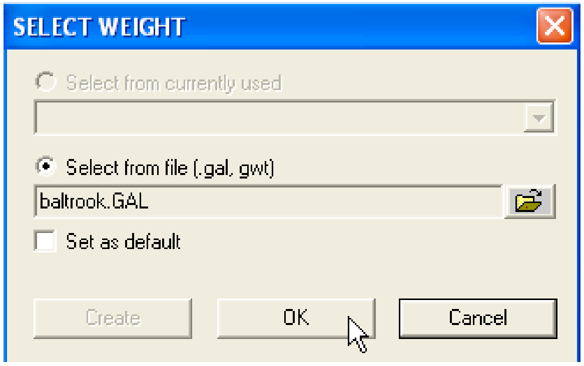
Models: ● Classic ○ Spatial Lag ○ Spatial Error

Figure 23.7: Linear trend surface model specification.

setup for the calculation of the cross-product, with `XY` as the `Result`, `X` as `Variables-1`, `MULTIPLY` as the `Operators` and `Y` as `Variables-2`. Click on `OK` to add the cross product to the data table. At the end of these operations, you should have three additional variables in the data table, as in Figure 23.5 on p. 183.

23.3.2 Linear Trend Surface

Invoke the regression title dialog (select `Regress` on the main menu) and make sure to check the box next to `Moran's I z-value`, as in Figure 23.6 on p. 183. Optionally, you can also specify an output file other than the default, for example, `baltrend.rtf` in Figure 23.6.



SELECT WEIGHT

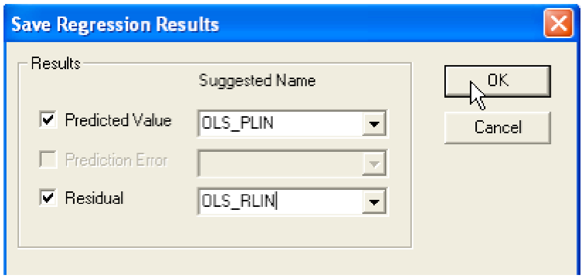
○ Select from currently used

● Select from file (.gal, gwt)

baltrook.GAL

□ Set as default

Figure 23.8: Spatial weights specification for regression diagnostics.



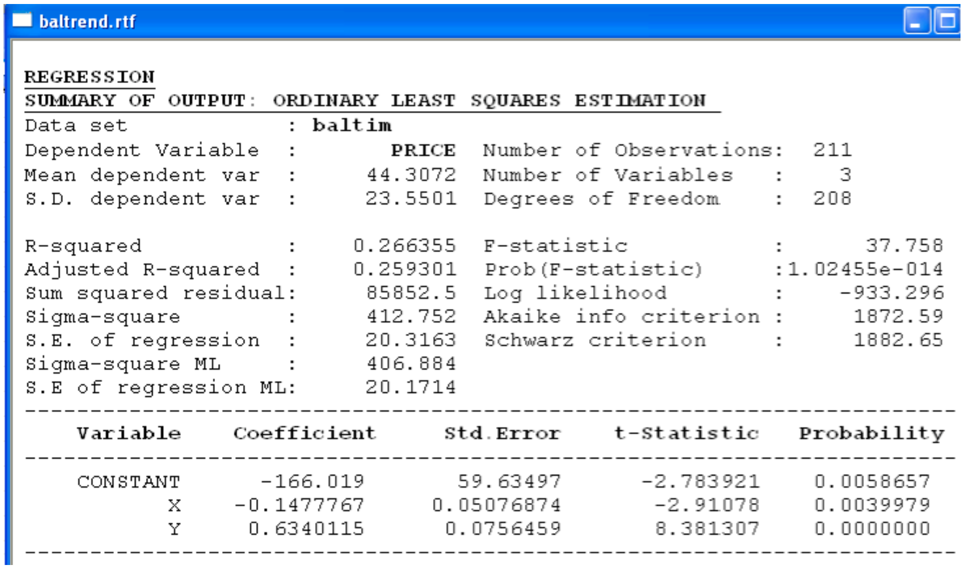
Save Regression Results

Results

	Suggested Name
☑ Predicted Value	OLS_PLIN
□ Prediction Error	
☑ Residual	OLS_RLIN

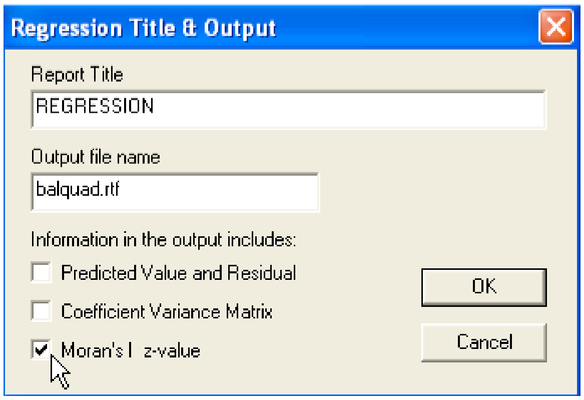
Figure 23.9: Linear trend surface residuals and predicted values.

Click OK to bring up the regression specification dialog. As the dependent variable, select PRICE, and as independent variables, choose X and Y, as in Figure 23.7 on p. 184. Since we will be focusing on the diagnostics for spatial autocorrelation, make sure a weights file is selected in the dialog before running the regression. For example, in Figure 23.7, this is the baltrook.GAL file you just created. Click on the file open icon in the dialog and choose Select from file in the select weights dialog, shown in Figure 23.8. Choose the baltrook.GAL file in the dialog. Next, click on Run to carry out the estimation. Before checking the actual regression results, make sure to select the Save button and to specify variable names for the predicted values and residuals. They will then be added to the data table and can be used in residual maps and other diagnostic plots. In Figure 23.9, the respective variables are OLS_PLIN and OLS_RLIN. Finally, click on OK to bring up the results window, shown in Figure 23.10 on p. 186.



```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : baltim
Dependent Variable  :        PRICE  Number of Observations:  211
Mean dependent var  :      44.3072  Number of Variables   :    3
S.D. dependent var  :      23.5501  Degrees of Freedom    :  208

R-squared           :     0.266355  F-statistic           :      37.758
Adjusted R-squared  :     0.259301  Prob(F-statistic)     :1.02455e-014
Sum squared residual:      85852.5  Log likelihood        :    -933.296
Sigma-square        :      412.752  Akaike info criterion :     1872.59
S.E. of regression  :      20.3163  Schwarz criterion     :     1882.65
Sigma-square ML     :      406.884
S.E of regression ML:      20.1714
-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT       -166.019        59.63497       -2.783921     0.0058657
           X     -0.1477767      0.05076874        -2.91078     0.0039979
           Y      0.6340115       0.0756459        8.381307     0.0000000
-----------------------------------------------------------------------
```

Figure 23.10: Linear trend surface model output.



Report Title: REGRESSION

Output file name: balquad.rtf

Information in the output includes:

☐ Predicted Value and Residual

☐ Coefficient Variance Matrix

☑ Moran's I z-value

Figure 23.11: Quadratic trend surface title and output settings.

The regression output shows a decent fit, with an adjusted R^2 of `0.26`. Both dimensions are strongly significant, but with different signs. The sign of the `X` variable is negative, suggesting a declining trend from West to East. In contrast, the sign of `Y` is positive, indicating an increase from South to North.

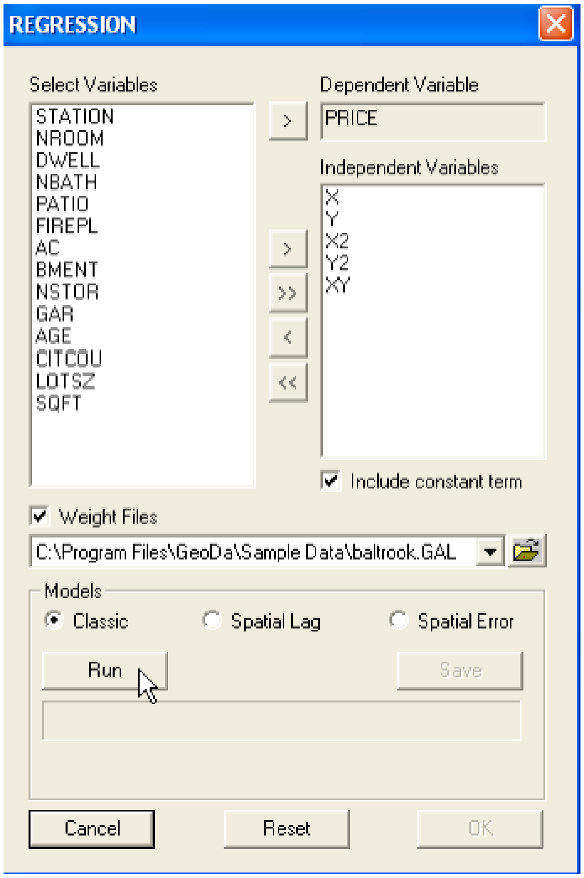


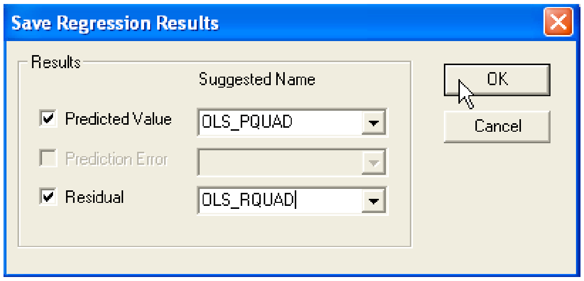
Figure 23.12: Quadratic trend surface model specification.

23.3.3 Quadratic Trend Surface

In addition to the linear trend, we will also consider the diagnostics for the quadratic trend surface model. Start the regression title dialog (`Regress`), enter a file name (such as `balquad.rtf`), and make sure to check the box next to `Moran's I z-value`, as in Figure 23.11 on p. 186. Click `OK` to bring up the regression specification dialog in Figure 23.12.

As for the linear trend, specify `PRICE` as the dependent variable. Select `X`, `Y`, `X2`, `Y2` and `XY` as the independent variables (see Figure 23.12). Make sure to select `baltrook.GAL` as the weights file and click `Run` to start the estimation.

Before inspecting the regression results, select `Save` to specify the vari-



Results

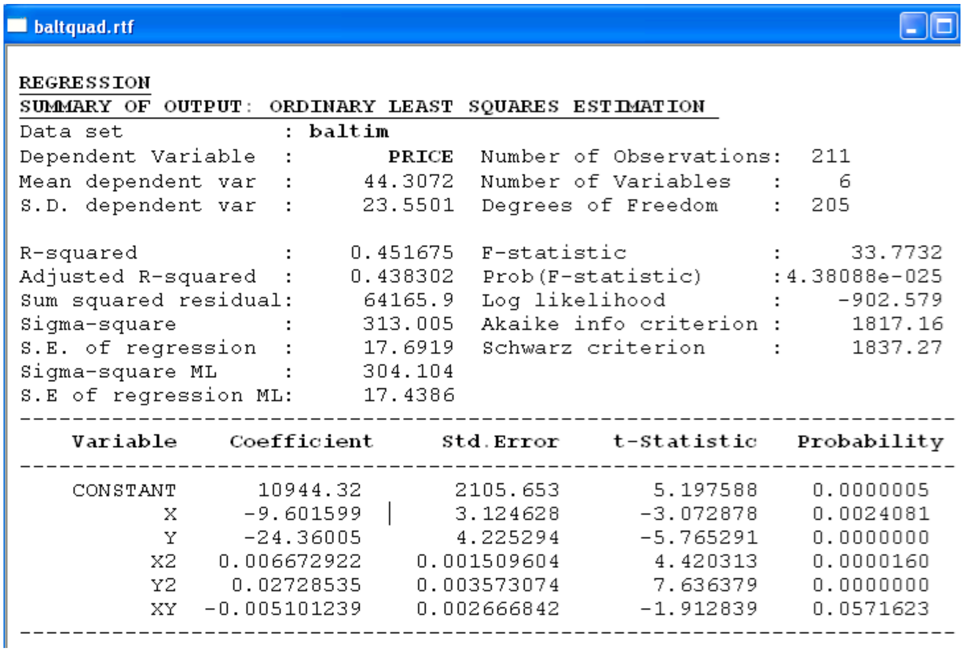
Suggested Name

☑ Predicted Value OLS_PQUAD

☐ Prediction Error

☑ Residual OLS_RQUAD

Figure 23.13: Quadratic trend surface residuals and predicted values.



```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : baltim
Dependent Variable  :         PRICE  Number of Observations:  211
Mean dependent var  :       44.3072  Number of Variables   :    6
S.D. dependent var  :       23.5501  Degrees of Freedom    :  205

R-squared           :      0.451675  F-statistic           :     33.7732
Adjusted R-squared  :      0.438302  Prob(F-statistic)     :4.38088e-025
Sum squared residual:       64165.9  Log likelihood        :    -902.579
Sigma-square        :       313.005  Akaike info criterion :     1817.16
S.E. of regression  :       17.6919  Schwarz criterion     :     1837.27
Sigma-square ML     :       304.104
S.E of regression ML:       17.4386
-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error     t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT        10944.32       2105.653        5.197588     0.0000005
           X       -9.601599       3.124628       -3.072878     0.0024081
           Y       -24.36005       4.225294       -5.765291     0.0000000
          X2     0.006672922    0.001509604        4.420313     0.0000160
          Y2      0.02728535    0.003573074        7.636379     0.0000000
          XY    -0.005101239    0.002666842       -1.912839     0.0571623
-----------------------------------------------------------------------
```

Figure 23.14: Quadratic trend surface model output.

able names for the predicted value (OLS_PQUAD) and residuals (OLS_RQUAD), as illustrated in Figure 23.13. Click OK on this dialog to add the variables to the data table, and OK again in the regression dialog to bring up the results window, shown in Figure 23.14.

Adding the squared and interaction terms to the regression improves the adjusted R^2 to 0.44. However, the interaction term XY turns out not to be

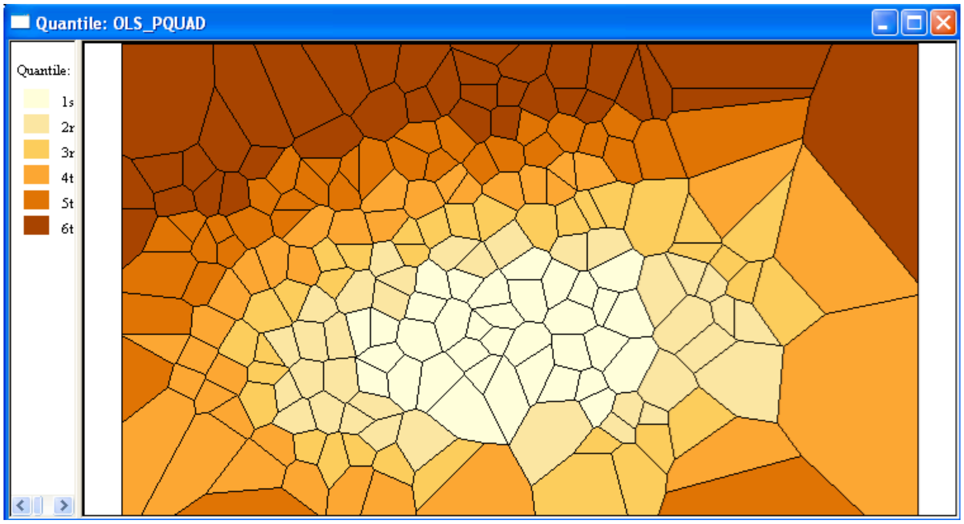


Figure 23.15: Quadratic trend surface predicted value map.

significant, but all the other coefficients are. The model suggests a bowl-like shape for the trend surface, with the lower house prices in the middle and increasing to the outside.

To illustrate this pattern, construct a quantile map with the predicted values. Make sure the active window is the base map with the Thiessen polygons (`balthiesen.shp`). Next, select `Map > Quantile` from the menu, choose `OLS_PQUAD` as the variable and change the number of quantiles to `6`. The result is as in Figure 23.15, with the darker shades corresponding to higher house prices.

23.4 Residual Maps and Plots

When the predicted values and residuals are saved to the data table as additional variables, they become available to all the exploratory functionality of *GeoDa*. This is particularly useful for the construction of diagnostic maps and plots. In the following examples, we will use the residual and predicted value of the quadratic trend surface regression, `OLS_PQUAD` and `OLS_RQUAD`. It is straightforward to replicate these examples for the residuals (`OLS_RLIN`) and predicted values (`OLS_PLIN`) of the linear trend model as well.

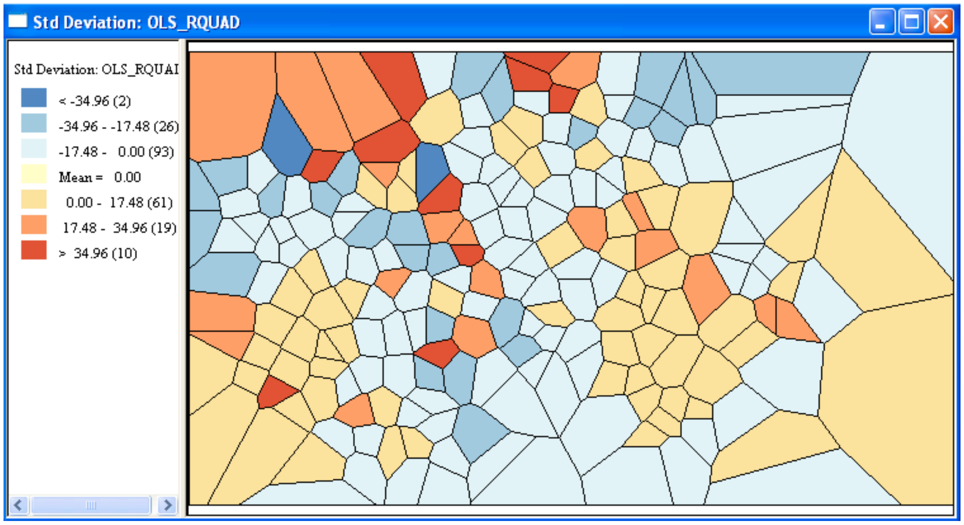


Figure 23.16: Residual map, quadratice trend surface.

23.4.1 Residual Maps

The most useful residual map is probably a standard deviational map, since it clearly illustrates patterns of over- or under-prediction, as well as the magnitude of the residuals, especially those greater than two standard deviational units.

Select `Map > St.Dev` and choose `OLS_RQUAD` as the variable. The resulting map should be as in Figure 23.16. Note the broad patterns in overprediction (negative residuals, or blue tones) and underprediction (positive residuals, or brown tones). This "visual inspection" would suggest the presence of spatial autocorrelation, but this requires a formal test before it can be stated more conclusively.

Also note several very large residuals (the very dark brown and blue). This is not surprising, since the "model" only contains location as a variable and no other distinguishing characteristics of the houses were considered. The outliers suggest the existence of transactions where location alone was not sufficient to explain the price. Selecting these locations and linking with other graphs or maps (e.g., some of the multivariate EDA tools) might shed light on which variables should be included in an improved regression specification.

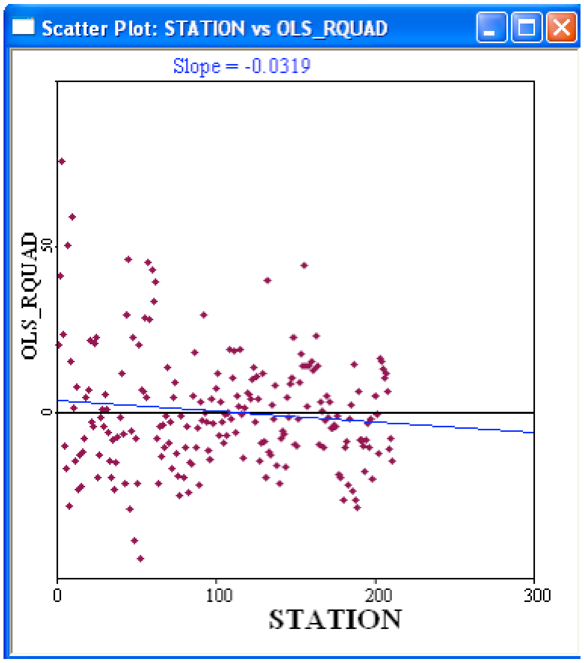


Figure 23.17: Quadratic trend surface residual plot.

23.4.2 Model Checking Plots

A simple plot of the model residuals is often revealing in that it *should not* suggest any type of patterning. While *GeoDa* currently does not have simple plot functions, it is possible to use a scatter plot to achieve the same goal. For example, plot the residuals of the quadratic trend surface model against a simple list of observation numbers, such as contained in the variable `STATION`.

Start with `Explore > Scatter Plot` and select `OLS_RQUAD` as the first variable (y-axis) and `STATION` as the second variable (x-axis). The resulting plot should be as in Figure 23.17.

The plot confirms the existence of several very large residuals. Selecting these on the graph and linking with a map or with other statistical graphs (describing other variables) may suggest systematic relationships with "ignored" variables and improve upon the model.

A different focus is taken with a plot of the residuals against the predicted values. Here, the interest lies in detecting patterns of *heteroskedasticity*, or a change in the variance of the residuals with another variable. As before,

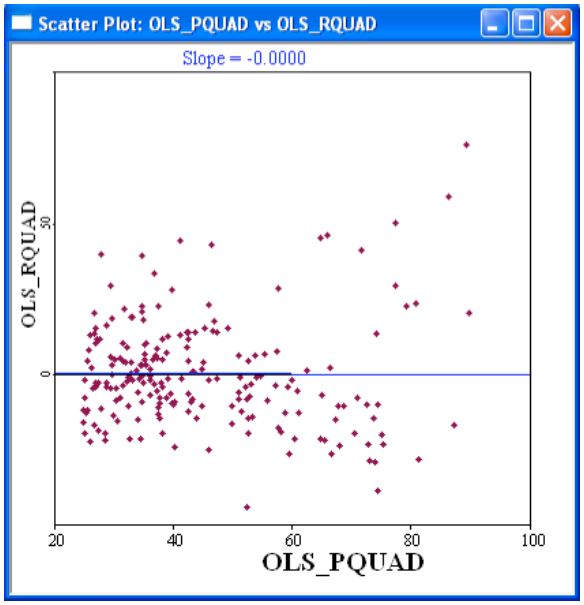


Figure 23.18: Quadratic trend surface residual/fitted value plot.

select `Explore > Scatter Plot` and choose `OLS_RQUAD` as the first variable (y-axis) and `OLS_PQUAD` as the second variable (x-axis). The resulting plot should be as in Figure 23.18.

In this graph, one tries to find evidence of funnel-like patterns, suggesting a relation between the spread of the residuals and the predicted value. There is some slight evidence of this in Figure 23.18, but insufficient to state a strong conclusion. Formal testing for heteroskedasticity will need to supplement the visual inspection.

Instead of the predicted values, other variables may be selected for the x-axis as well, especially when there is strong suspicion that they may "cause" heteroskedasticity. Often, such variables are related to *size*, such as area or total population.

23.4.3 Moran Scatter Plot for Residuals

Spatial patterns in the residuals can be analyzed more formally by means of a Moran scatter plot. In the usual fashion, select `Space > Univariate Moran` from the menu, choose `OLS_RQUAD` as the variable, and `baltrook.GAL` as the spatial weights file.

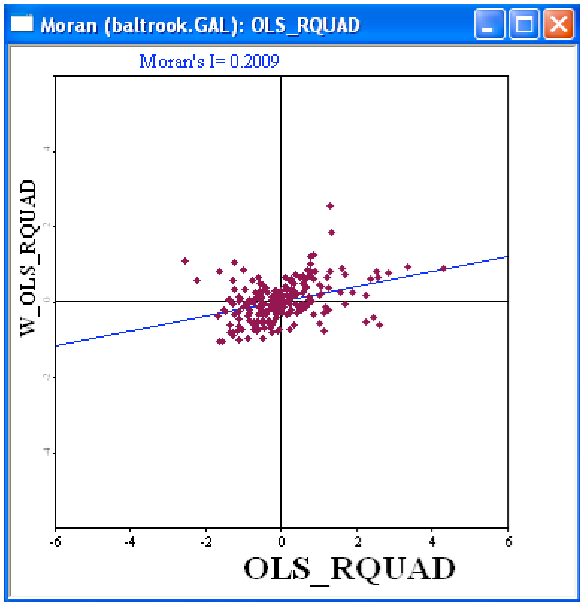


Figure 23.19: Moran scatter plot for quadratic trend surface residuals.

The resulting graph should be as in Figure 23.19, indicating a Moran's I for the residuals of `0.2009`. Note that this measure is purely descriptive, and while it allows for linking and brushing, it is *not appropriate* to use the permutation approach to assess significance.[1] For the same reason, it is also not appropriate to construct LISA maps for the residuals.

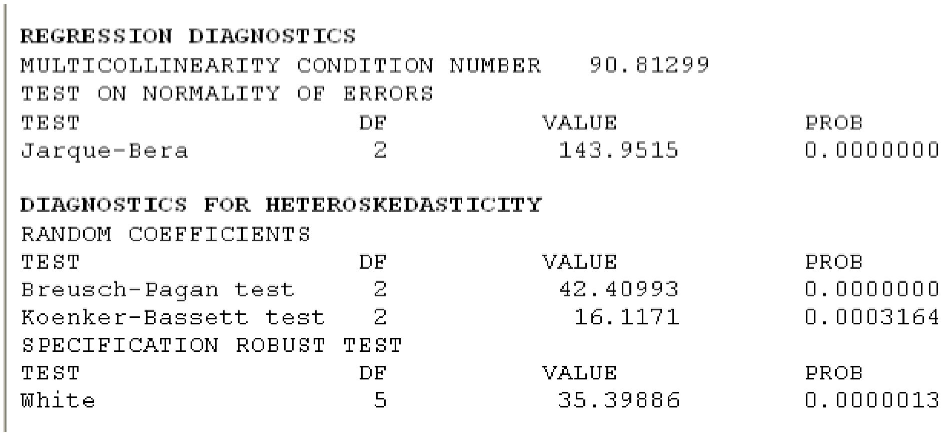
23.5 Multicollinearity, Normality and Heteroskedasticity

The first set of diagnostics provided in the regression output window consists of three traditional measures: the multicollinearity condition number, a test for non-normality (Jarque-Bera), and three diagnostics for heteroskedasticity (Breusch-Pagan, Koenker-Bassett, and White).[2]

The results for the linear and quadratic trend surface model are listed

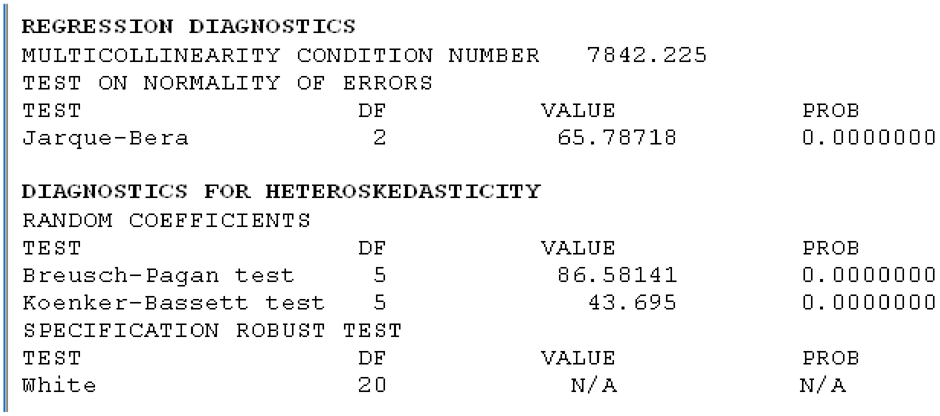
[1] This is because OLS residuals are already correlated by construction and the permutation approach ignores this fact.

[2] For methodological details on these diagnostics, see most intermediate Econometrics texts.



```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   90.81299
TEST ON NORMALITY OF ERRORS
TEST                    DF          VALUE            PROB
Jarque-Bera              2           143.9515        0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF          VALUE            PROB
Breusch-Pagan test       2           42.40993        0.0000000
Koenker-Bassett test     2            16.1171        0.0003164
SPECIFICATION ROBUST TEST
TEST                    DF          VALUE            PROB
White                    5           35.39886        0.0000013
```

Figure 23.20: Regression diagnostics – linear trend surface model.



```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   7842.225
TEST ON NORMALITY OF ERRORS
TEST                    DF          VALUE            PROB
Jarque-Bera              2           65.78718        0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                    DF          VALUE            PROB
Breusch-Pagan test       5           86.58141        0.0000000
Koenker-Bassett test     5             43.695        0.0000000
SPECIFICATION ROBUST TEST
TEST                    DF          VALUE            PROB
White                   20             N/A           N/A
```

Figure 23.21: Regression diagnostics – quadratic trend surface model.

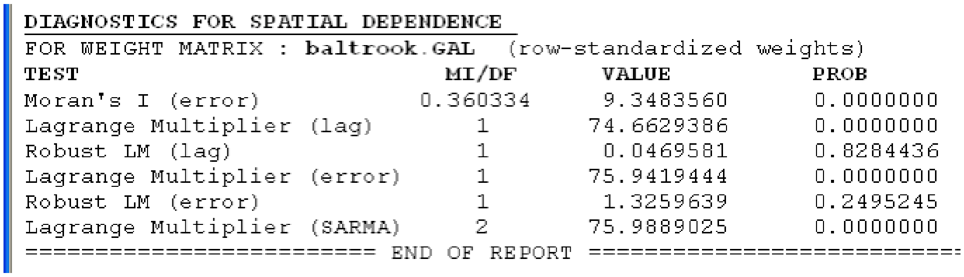
in, respectively, Figure 23.20 and Figure 23.21. First, consider the multicollinearity condition number. This is not a test statistic per se, but a diagnostic to suggest problems with the stability of the regression results due to multicollinearity (the explanatory variables are too correlated and provide insufficient separate information). Typically, an indicator over 30 is suggestive of problems. In trend surface regressions, this is very common, since the explanatory variables are simply powers and cross products of each other. In our example, the linear model has a value of 90.8, but the quadratic model reaches as high as 7842.2!

The Jarque-Bera test on normality of the errors is distributed as a χ^2 statistic with 2 degrees of freedom. In both cases, there is strong suggestion of non-normality of the errors. In and of itself, this may not be too serious a problem, since many properties in regression analysis hold *asymptotically* even without assuming normality. However, for *finite sample* (or exact) inference, normality is essential and the current models clearly violate that assumption.

The next three diagnostics are common statistics to detect heteroskedasticity, i.e., a non-constant error variance. Both the Breusch-Pagan and Koenker-Bassett tests are implemented as tests on *random coefficients*, which assumes a specific functional form for the heteroskedasticity.[3] The Koenker-Bassett test is essentially the same as the Breusch-Pagan test, except that the residuals are *studentized*, i.e., they are made robust to non-normality. Both test statistics indicate serious problems with heteroskedasticity in each of the trend surface specifications.

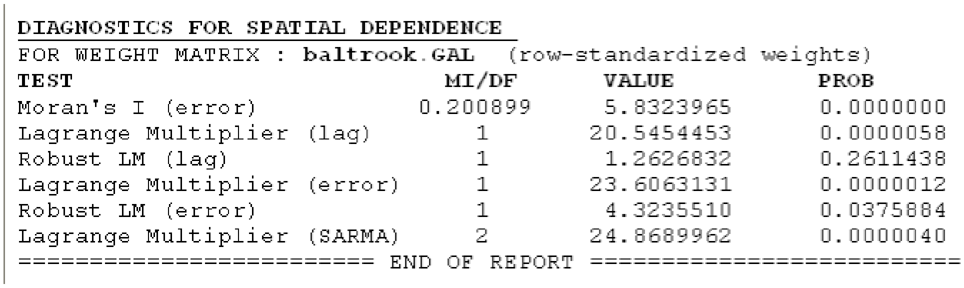
The White test is a so-called *specification-robust* test for heteroskedasticity, in that it does not assume a specific functional form for the heteroskedasticity. Instead, it approximates a large range of possibilities by all square powers and cross-products of the explanatory variables in the model. In some instances, this creates a problem when a cross-product is already included as an interaction term in the model. This is the case for the quadratic trend surface, which already included the squares and cross-product of the x and y variables. In such an instance, there is perfect multicollinearity. Currently, *GeoDa* is not able to correct for this and reports N/A instead, as in Figure 23.21. In the linear trend model, the White statistics is 35.4, which supports the evidence of heteroskedasticity provided by the other two tests. This result does not necessarily always hold, since it may be that the random coefficient assumption implemented in the Breusch-Pagan and Koenker-Bassett tests is not appropriate. In such an instance, the White test may be significant, but the other two may not be. It is important to keep in mind that the White test is against a more general form of heteroskedasticity.

[3] Specifically, the heteroskedasticity is a function of the *squares* of the explanatory variables. This is the form implemented in *GeoDa*. In some other econometric software, the x-values themselves are used, and not the squares, which may give slightly different results.



```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : baltrook.GAL  (row-standardized weights)
TEST                            MI/DF      VALUE          PROB
Moran's I (error)            0.360334     9.3483560      0.0000000
Lagrange Multiplier (lag)        1       74.6629386      0.0000000
Robust LM (lag)                  1        0.0469581      0.8284436
Lagrange Multiplier (error)      1       75.9419444      0.0000000
Robust LM (error)                1        1.3259639      0.2495245
Lagrange Multiplier (SARMA)      2       75.9889025      0.0000000
========================= END OF REPORT ==========================
```

Figure 23.22: Spatial autocorrelation diagnostics – linear trend surface model.



```
DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : baltrook.GAL  (row-standardized weights)
TEST                            MI/DF      VALUE          PROB
Moran's I (error)            0.200899     5.8323965      0.0000000
Lagrange Multiplier (lag)        1       20.5454453      0.0000058
Robust LM (lag)                  1        1.2626832      0.2611438
Lagrange Multiplier (error)      1       23.6063131      0.0000012
Robust LM (error)                1        4.3235510      0.0375884
Lagrange Multiplier (SARMA)      2       24.8689962      0.0000040
========================= END OF REPORT =========================
```

Figure 23.23: Spatial autocorrelation diagnostics – quadratic trend surface model.

23.6 Diagnostics for Spatial Autocorrelation

The final collection of model diagnostics consists of tests against spatial autocorrelation. A total of six test statistics are reported, as shown in Figure 23.22 for the linear trend model, and in Figure 23.23 for the quadratic trend. Currently, the tests are all computed for the same weights matrix, here `baltrook.GAL`, as listed in the diagnostics output. Consequently, if you want to test the residuals using several spatial weights, you need to re-run the regression analysis.[4]

23.6.1 Morans' I

The first statistic is Moran's I, which gives the same value as in the Moran scatter plot in Figure 23.19 on p. 193 (e.g., `0.200899` for the quadratic trend

[4] This *feature* will likely change in future versions of *GeoDa*.

model). When the `Moran's I z-value` box was checked in the regression title dialog (e.g., as in Figure 23.11 on p. 186), a z-value and associated p-value will be reported in the diagnostics output.[5] When this box is not checked, only the statistic is reported.[6]

In both the linear and quadratic trend surface models (with respective z-values of `9.35` and `5.83`), the Moran statistic is highly significant, suggesting a problem with spatial autocorrelation. While Moran's I statistic has great power in detecting misspecifications in the model (and not only spatial autocorrelation), it is less helpful in suggesting which alternative specification should be used. To this end, we use the Lagrange Multiplier test statistics.

23.6.2 Lagrange Multiplier Test Statistics

Five Lagrange Multiplier test statistics are reported in the diagnostic output. The first two (LM-Lag and Robust LM-Lag) pertain to the spatial lag model as the alternative. The next two (LM-Error and Robust LM-Error) refer to the spatial error model as the alternative. The last test, LM-SARMA, relates to the higher order alternative of a model with both spatial lag and spatial error terms. This test is only included for the sake of completeness, since it is not that useful in practice. More specifically, in addition to detecting the higher order alternative for which it is designed, the test also has high power against the one-directional alternatives. In other words, it will tend to be significant when *either* the error *or* the lag model are the proper alternatives, but not necessarily the higher order alternative.

All one-directional test statistics are distributed as χ^2 with one degree of freedom (the LM-SARMA test statistics has two degrees of freedom). To guide the specification search, the test statistics should be considered in a given sequence, which is elaborated upon in Section 23.6.3. The important issue to remember is to only consider the Robust versions of the statistics when the standard versions (LM-Lag or LM-Error) *are significant*. When they are not, the properties of the robust versions may no longer hold.[7]

For both trend surface models, the LM-Lag and LM-Error statistics are highly significant, with the latter slightly more so. The rejection of the null

[5] The z-value is based on a normal approximation and takes into account the fact that these are residuals. See Anselin and Bera (1998) for details.

[6] For large data sets, the matrix manipulations required to compute the z-value for Moran's I become quite time consuming. Consequently, inference for this statistic is an *option* in *GeoDa*. Moreover, the specification search outlined in Section 23.6.3 is based on the Lagrange Multiplier statistics and not on Moran's I.

[7] For technical details, see Anselin et al. (1996) and Anselin and Florax (1995).

hypothesis by both LM test statistics is a situation commonly encountered in practice, and it requires the consideration of the Robust forms of the statistics. Note also how the LM-SARMA statistic is significant. However, its value is only slightly higher than that of the one-directional statistics, suggesting it is probably picking up the single alternative rather than a true higher order model.

In the linear trend model (Figure 23.22), the results for the Robust tests are pathological. While the standard tests are significant, the Robust form is not, suggesting there is no spatial autocorrelation problem. This is clearly wrong (see the Moran's I). It indicates that other misspecification problems are present that invalidate the asymptotic results on which the Robust LM test statistics are based. This is not surprising, since the linear trend specification is extremely simplistic. Luckily, this type of result for the test statistics occurs only rarely in practice.

The more common result is the one for the quadratic trend surface, shown in Figure 23.23. Here, the Robust LM-Error statistic is significant (with $p < 0.04$), while the Robust LM-Lag statistic is not (with $p = 0.26$). This suggests that a spatial error specification should be estimated next.

23.6.3 Spatial Regression Model Selection Decision Rule

The array of test statistics for spatial autocorrelation may seem bewildering at first, but there is a fairly intuitive way to proceed through the results towards a spatial regression specification. This process is summarized in Figure 23.24 on p. 199.

Begin the process at the top of the graph and consider the standard (i.e., *not* the robust forms) LM-Error and LM-Lag test statistics. If neither rejects the null hypothesis, stick with the OLS results. It is likely that in this case, the Moran's I test statistic will not reject the null hypothesis either.[8] If one of the LM test statistics rejects the null hypothesis, and the other does not, then the decision is straightforward as well: estimate the alternative spatial regression model that matches the test statistic that rejects the null. So, if LM-Error rejects the null, but LM-lag does not, estimate a spatial error model, and vice versa.

When both LM test statistics reject the null hypothesis, proceed to the bottom part of the graph and consider the Robust forms of the test statistics. Typically, only one of them will be significant (as in Figure 23.23), or one

[8] If it does, i.e., if there is a conflict between the indication given by Moran's I and that given by the LM test statistics, it is likely due to the Moran's I power against other alternatives than spatial autocorrelation, such as heteroskedasticity or non-normality.

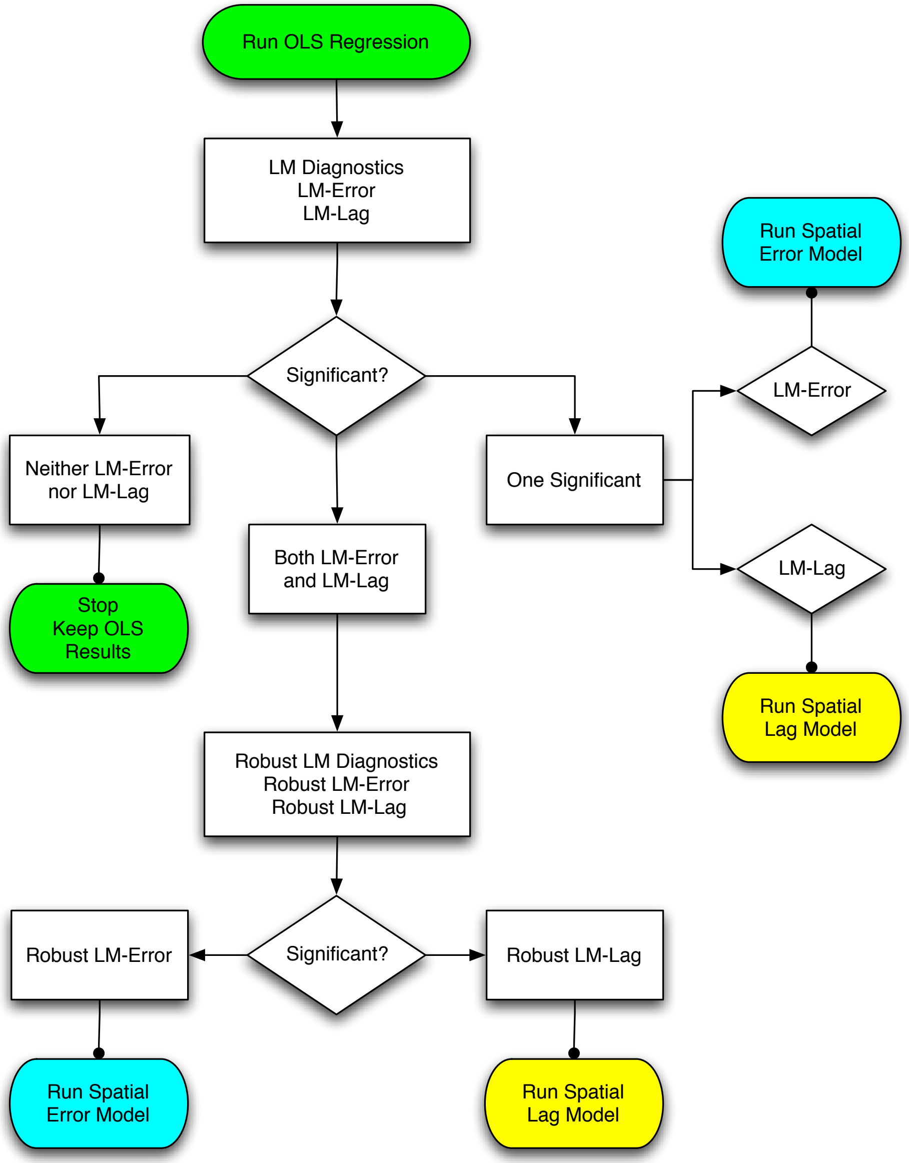


Figure 23.24: Spatial regression decision process.

will be orders of magnitude more significant than the other (e.g., $p < 0.00000$ compared to $p < 0.03$). In that case, the decision is simple: estimate the spatial regression model matching the (most) significant statistic. In the rare instance that both would be highly significant, go with the model with the largest value for the test statistic. However, in this situation, some caution is needed, since there may be other sources of misspecification. One obvious action to take is to consider the results for different spatial weights and/or to change the basic (i.e., not the spatial part) specification of the model. As shown in Figure 23.22, there are also rare instances where neither of the Robust LM test statistics are significant. In those cases, more serious misspecification problems are likely present and those should be addressed first.

23.7 Practice

Continue with the POLICE example from Section 22.6 and consider diagnostics using several spatial weights, such as rook contiguity, distance-based contiguity, etc. Compare your results to those presented in Kelejian and Robinson (1992).

Alternatively, consider the model specification used for the BALTIMORE data by Dubin (1992) and compare it to the results for the simple trend surface. Her model specification uses `PRICE` as the dependent variable and the following explanatory variables: `NROOM`, `DWELL`, `NBATH`, `PATIO`, `FIREPL`, `AC`, `BMENT`, `NSTOR`, `GAR`, `AGE`, `CITCOU`, `LOTSZ`, `SQFT` and a quadratic trend surface.[9] In addition to the rook spatial weights used in this Exercise, consider some distance-based weights as well.

[9] The variable `STIME` included in the Dubin (1992) article, is not available in the sample data set.

Exercise 24

Spatial Lag Model

24.1 Objectives

This exercise considers the estimation by means of maximum likelihood of a spatial regression model that includes a spatially lagged dependent variable.[1] Methodological aspects are reviewed in Anselin (1988) and Anselin and Bera (1998). The specific estimation algorithm used by *GeoDa* is outlined in Smirnov and Anselin (2001). Unlike the traditional approach, which uses eigenvalues of the weights matrix, this method is well suited to the estimation in situations with very large data sets.

At the end of the exercise, you should know how to:

- set up the specification for a spatial lag regression model
- interpret estimation results in the spatial lag model
- interpret measures of fit in the spatial lag model
- interpret regression diagnostics in the spatial lag model
- understand the predicted value and different notions of residuals in the spatial lag model

More detailed information on the relevant operations can be found in the *Release Notes*, pp. 53–54.

[1] Formally, this model is $y = \rho Wy + X\beta + \varepsilon$, where y is a vector of observations on the dependent variable, Wy is a spatially lagged dependent variable for weights matrix W, X is a matrix of observations on the explanatory variables, ε is a vector of i.i.d. error terms, and ρ and β are parameters.

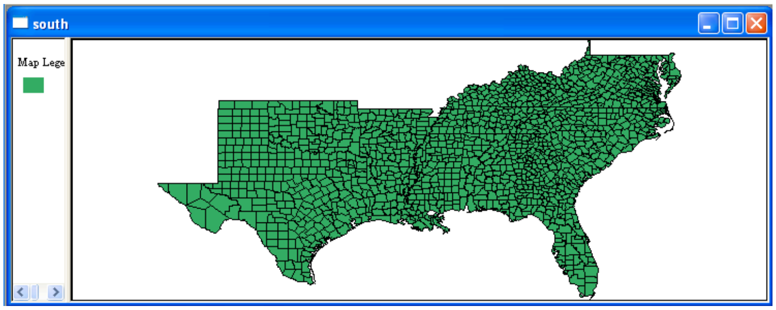


Figure 24.1: South county homicide base map.

24.2 Preliminaries

Load the sample data set with homicide rates and related variables for 1412 counties in the South of the U.S. (`south.shp` with `FIPSNO` as the `Key`). The base map should be as in Figure 24.1. Make sure to create a spatial weights matrix if you have not already done so. For this particular application, we will use a cumulative first and second order rook file. Select the `rook` criterion, change the `order of contiguity` to `2` and check the box to `Include all the lower orders` (for extensive instructions, see Section 15.2, and especially p. 114). Name the file `southrk12.GAL`.

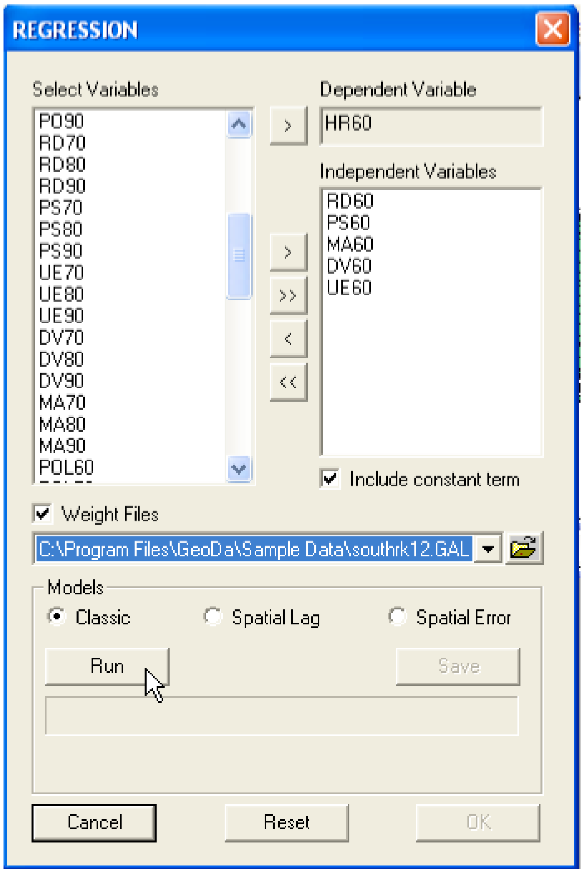
Note that the ML estimation of the spatial lag model only works for spatial weights that correspond to a symmetric contiguity relation. In other words, it works for rook and queen contiguity, as well as distance band contiguity, *but not for k-nearest neighbors.*

24.2.1 OLS with Diagnostics

As a point of reference, we will first run a `Classic` OLS regression (follow the instructions in Section 22.3) with `HR60` as the dependent variable, and `RD60`, `PS60`, `MA60`, `DV60`, and `UE60` as the explanatory variables.[2] The regression dialog should be as in Figure 24.2 on p. 203. Make sure to check the box next to `Weight Files`, and specify `southrk12.GAL` for the spatial weights.

`Run` the regression and check the results by clicking on `OK`. The OLS estimates are listed in Figure 24.3 on p. 204, with the diagnostics given in Figure 24.4 on p. 205.

[2] This replicates the analysis in Baller et al. (2001), although with slightly different spatial weights.



Dependent Variable: HR60

Independent Variables:
RD60
PS60
MA60
DV60
UE60

☑ Include constant term

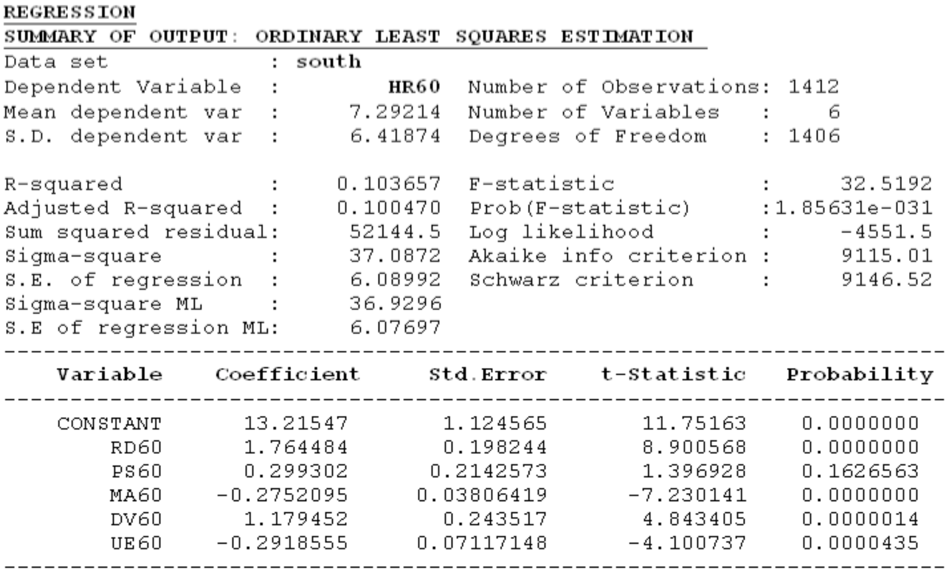
☑ Weight Files: C:\Program Files\GeoDa\Sample Data\southrk12.GAL

Models: ● Classic ○ Spatial Lag ○ Spatial Error

Figure 24.2: Homicide classic regression for 1960.

The fit of the model is not that impressive, with an adjusted R^2 of 0.10, although all but `PS60` are highly significant and with the expected sign. For comparison purposes with the spatial model, note the `Log Likelihood` of `-4551.5` and the `AIC` of `9115.01`.

The regression diagnostics reveal considerable non-normality and heteroskedasticity, as well as high spatial autocorrelation. Following the steps outlined in Section 23.6.3, we conclude that a spatial lag model is the proper alternative. Both LM-Lag and LM-Error are significant, but of the robust forms, only the Robust LM-Lag statistic is highly significant ($p < 0.00002$), while the Robust LM-Error statistic is not ($p < 0.17$). This sets the stage for the estimation of the spatial lag model.



```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set            : south
Dependent Variable  :        HR60  Number of Observations: 1412
Mean dependent var  :     7.29214  Number of Variables   :    6
S.D. dependent var  :     6.41874  Degrees of Freedom    : 1406

R-squared           :    0.103657  F-statistic           :     32.5192
Adjusted R-squared  :    0.100470  Prob(F-statistic)     :1.85631e-031
Sum squared residual:     52144.5  Log likelihood        :     -4551.5
Sigma-square        :     37.0872  Akaike info criterion :     9115.01
S.E. of regression  :     6.08992  Schwarz criterion     :     9146.52
Sigma-square ML     :     36.9296
S.E of regression ML:     6.07697
-----------------------------------------------------------------------
    Variable    Coefficient      Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT      13.21547        1.124565       11.75163     0.0000000
        RD60      1.764484        0.198244       8.900568     0.0000000
        PS60      0.299302       0.2142573       1.396928     0.1626563
        MA60    -0.2752095      0.03806419      -7.230141     0.0000000
        DV60      1.179452        0.243517       4.843405     0.0000014
        UE60    -0.2918555      0.07117148      -4.100737     0.0000435
-----------------------------------------------------------------------
```

Figure 24.3: OLS estimation results, homicide regression for 1960.

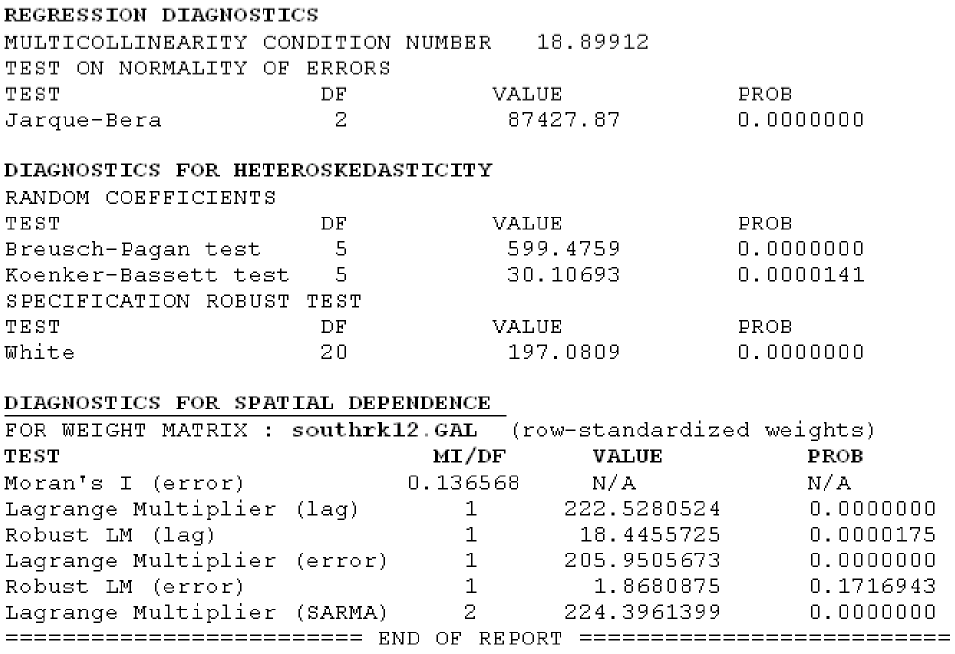
24.3 ML Estimation with Diagnostics

The ML estimation of the spatial lag model is invoked in the same manner as for the classic regression, by clicking on `Regress` on the main menu, or by choosing `Methods > Regress` before a project is loaded. The title and output file dialog is identical for all regression analyses. This allows for the specification of an output file name, such as `southlag.rtf` shown in Figure 24.5 on p. 205. Click `OK`, to bring up the familiar regression specification dialog.

24.3.1 Model Specification

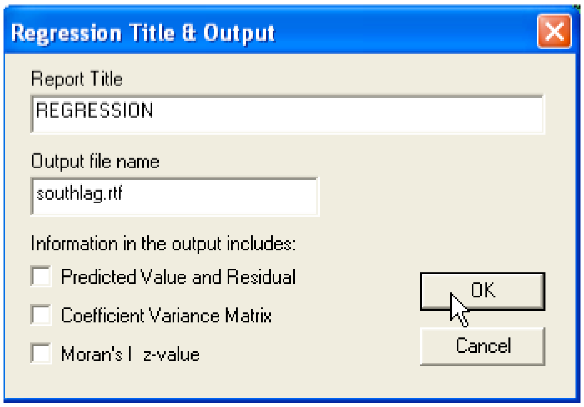
The dialog, shown in Figure 24.6 on p. 206, is the same as before. Enter the same set of dependent and explanatory variables as for the classic case and make sure to specify the spatial weights file, as illustrated in the Figure. Instead of the default `Classic`, check the radio button next to `Spatial Lag`. Invoke the estimation routine as before, by clicking on the `Run` button.

After the estimation is completed, as indicated by a progress bar and illustrated in Figure 24.7 on p. 207, make sure to click on the `Save` button *before* selecting `OK`. This will bring up the dialog to specify variable names



```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   18.89912
TEST ON NORMALITY OF ERRORS
TEST                     DF           VALUE            PROB
Jarque-Bera               2            87427.87        0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                     DF           VALUE            PROB
Breusch-Pagan test        5            599.4759        0.0000000
Koenker-Bassett test      5            30.10693        0.0000141
SPECIFICATION ROBUST TEST
TEST                     DF           VALUE            PROB
White                    20            197.0809        0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : southrk12.GAL  (row-standardized weights)
TEST                           MI/DF       VALUE            PROB
Moran's I (error)            0.136568      N/A              N/A
Lagrange Multiplier (lag)        1       222.5280524      0.0000000
Robust LM (lag)                  1        18.4455725      0.0000175
Lagrange Multiplier (error)      1       205.9505673      0.0000000
Robust LM (error)                1         1.8680875      0.1716943
Lagrange Multiplier (SARMA)      2       224.3961399      0.0000000
========================= END OF REPORT ==========================
```

Figure 24.4: OLS diagnostics, homicide regression for 1960.



Report Title

REGRESSION

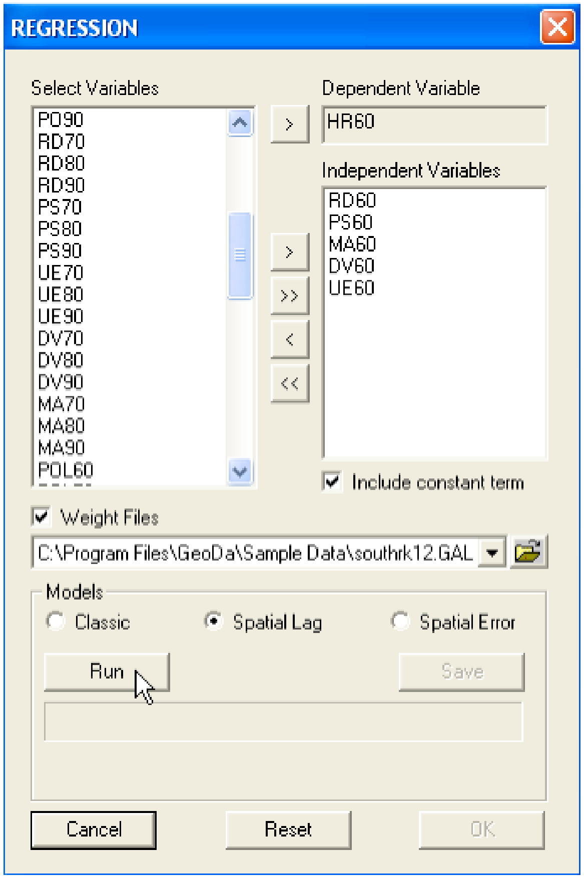
Output file name

southlag.rtf

Information in the output includes:

- ☐ Predicted Value and Residual
- ☐ Coefficient Variance Matrix
- ☐ Moran's I z-value

Figure 24.5: Title and file dialog for spatial lag regression.



Dependent Variable
HR60

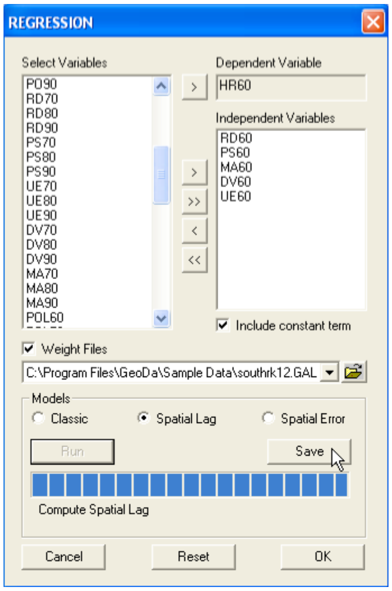
Independent Variables
RD60
PS60
MA60
DV60
UE60

☑ Include constant term

☑ Weight Files
C:\Program Files\GeoDa\Sample Data\southrk12.GAL

Models
○ Classic ● Spatial Lag ○ Spatial Error

Figure 24.6: Homicide spatial lag regression specification for 1960.



PO90
RD70
RD80
RD90
PS70
PS80
PS90
UE70
UE80
UE90
DV70
DV80
DV90
MA70
MA80
MA90
POL60

Dependent Variable: HR60

Independent Variables:
RD60
PS60
MA60
DV60
UE60

☑ Include constant term

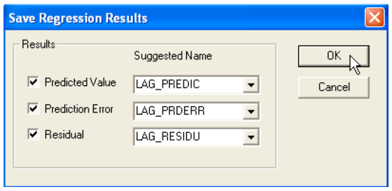
☑ Weight Files

C:\Program Files\GeoDa\Sample Data\southrk12.GAL

Models: ○ Classic ● Spatial Lag ○ Spatial Error

Compute Spatial Lag

Figure 24.7: Save residuals and predicted values dialog.



☑ Predicted Value: LAG_PREDIC

☑ Prediction Error: LAG_PRDERR

☑ Residual: LAG_RESIDU

Figure 24.8: Spatial lag predicted values and residuals variable name dialog.

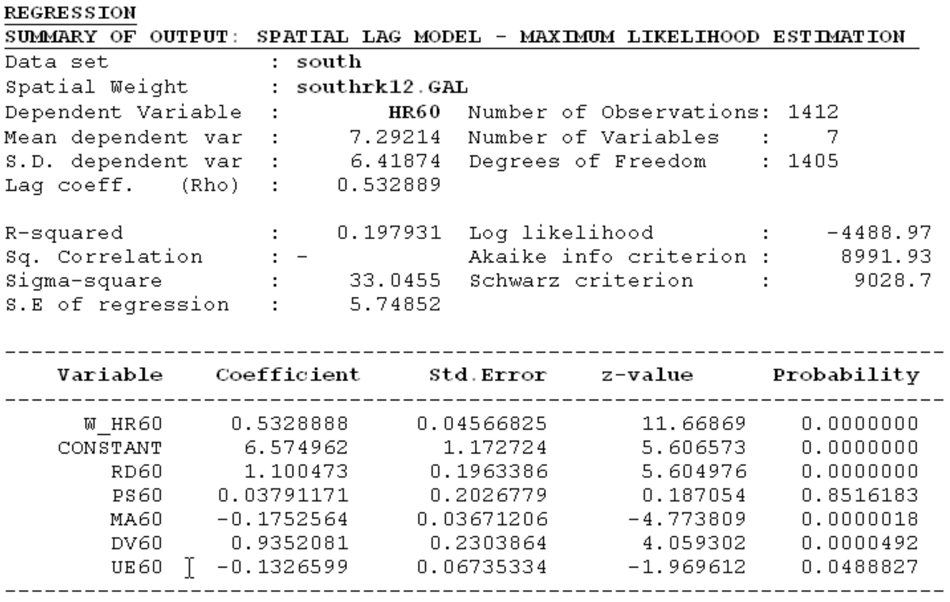
to add the residuals and predicted values to the data table.

As shown in Figure 24.8, there are three options for this in the spatial lag model. These are covered in more detail in Section 24.4. For now, select all three check boxes and keep the variable names to their defaults of `LAG_PREDIC` for the `Predicted Value`, `LAG_PRDERR` for the `Prediction Error`, and `LAG_RESIDU` for the `Residual`. Click on `OK` to get back to the regression dialog and select `OK` again to bring up the estimation results and diagnostics.

24.3.2 Estimation Results

The estimates and measures of fit are listed in Figure 24.9 on p. 208. First, a word of caution. While it is tempting to focus on traditional measures, such as the R^2, this is *not appropriate* in a spatial regression model. The value listed in the spatial lag output is not a real R^2, but a so-called pseudo-R^2, which is not directly comparable with the measure given for OLS results.

The proper measures of fit are the Log-Likelihood, AIC and SC. If we compare the values in Figure 24.9 to those for OLS in Figure 24.3 (p. 204), we notice an increase in the Log-Likelihood from `-4551.5` (for OLS) to



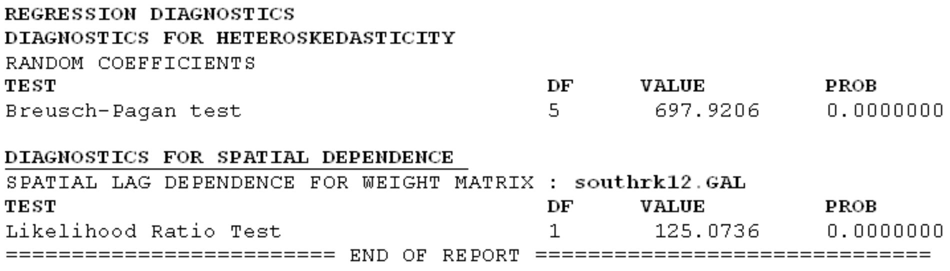
```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : south
Spatial Weight      : southrk12.GAL
Dependent Variable  :         HR60  Number of Observations: 1412
Mean dependent var  :      7.29214  Number of Variables   :    7
S.D. dependent var  :      6.41874  Degrees of Freedom    : 1405
Lag coeff.   (Rho)  :     0.532889

R-squared           :     0.197931  Log likelihood        :     -4488.97
Sq. Correlation     : -             Akaike info criterion :      8991.93
Sigma-square        :      33.0455  Schwarz criterion     :       9028.7
S.E of regression   :      5.74852

-----------------------------------------------------------------------
    Variable    Coefficient      Std.Error     z-value      Probability
-----------------------------------------------------------------------
       W_HR60     0.5328888     0.04566825        11.66869    0.0000000
     CONSTANT      6.574962       1.172724        5.606573    0.0000000
         RD60      1.100473      0.1963386        5.604976    0.0000000
         PS60    0.03791171      0.2026779        0.187054    0.8516183
         MA60    -0.1752564     0.03671206       -4.773809    0.0000018
         DV60     0.9352081      0.2303864        4.059302    0.0000492
         UE60    -0.1326599     0.06735334       -1.969612    0.0488827
-----------------------------------------------------------------------
```

Figure 24.9: ML estimation results, spatial lag model, HR60.

-4488.97. Compensating the improved fit for the added variable (the spatially lagged dependent variable), the AIC (from 9115 to 8991.9) and SC (from 9146.5 to 9028.7) both decrease relative to OLS, again suggesting an improvement of fit for the spatial lag specification.

The spatial autoregressive coefficient is estimated as 0.53, and is highly significant ($p < 0.0000000$). This is not unusual for a data set of this size (> 1000 observations) and is in part due to the asymptotic nature of the analytical expressions used for the variance.

There are some minor differences in the significance of the other regression coefficients between the spatial lag model and the classic specification: PS60 is even less significant than before ($p < 0.85$), but, more importantly, the significance of UE60 changes from $p < 0.00004$ to $p < 0.04$. The magnitude of all the estimated coefficients is also affected, all of them showing a decrease in absolute value. To some extent, the explanatory power of these variables that was attributed to their in-county value, was really due to the neighboring locations. This is picked up by the eoefficient of the spatially lagged dependent variable.



```
REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE         PROB
Breusch-Pagan test                      5        697.9206     0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : southrk12.GAL
TEST                                    DF      VALUE         PROB
Likelihood Ratio Test                   1        125.0736     0.0000000
======================== END OF REPORT ============================
```

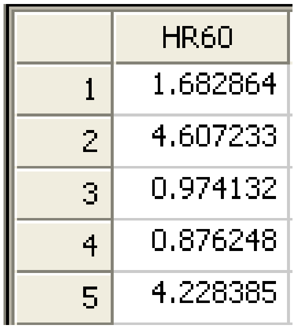
Figure 24.10: Diagnostics, spatial lag model, HR60.

24.3.3 Diagnostics

A limited number of diagnostics are provided with the ML lag estimation, as illustrated in Figure 24.10. First is a Breusch-Pagan test for heteroskedasticy in the error terms. The highly significant value of `697.9` suggests that heteroskedasticity is still a serious problem. The second test is an alternative to the asymptotic significance test on the spatial autoregressive coefficient, it is *not* a test on remaining spatial autocorrelation. The Likelihood Ratio Test is one of the three classic specification tests comparing the null model (the classic regression specification) to the alternative spatial lag model.[3] The value of `125` confirms the strong significance of the spatial autoregressive coefficient.

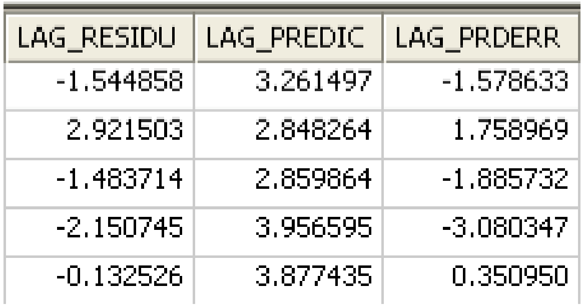
The three classic tests are asymptotically equivalent, but in finite samples should follow the ordering: $W > LR > LM$. In our example, the Wald test is $11.7^2 = 136.9$ (rounded), the LR test is 125 and the LM-Lag test was 222.5, which is not compatible with the expected order. This probably suggests that other sources of misspecification may invalidate the asymptotic properties of the ML estimates and test statistics. Given the rather poor fit of the model to begin with, the high degree of non-normality and strong heteroskedasticity, this is not surprising. Further consideration is needed of alternative specifications, either including new explanatory variables or incorporating different spatial weights.

[3] The other two classic tests are the Wald test, i.e., the square of the asymptotic t-value (or, z-value), and the LM-lag test based on OLS residuals.



	HR60
1	1.682864
2	4.607233
3	0.974132
4	0.876248
5	4.228385

Figure 24.11: Observed value, HR60.



LAG_RESIDU	LAG_PREDIC	LAG_PRDERR
-1.544858	3.261497	-1.578633
2.921503	2.848264	1.758969
-1.483714	2.859864	-1.885732
-2.150745	3.956595	-3.080347
-0.132526	3.877435	0.350950

Figure 24.12: Spatial lag predicted values and residuals HR60.

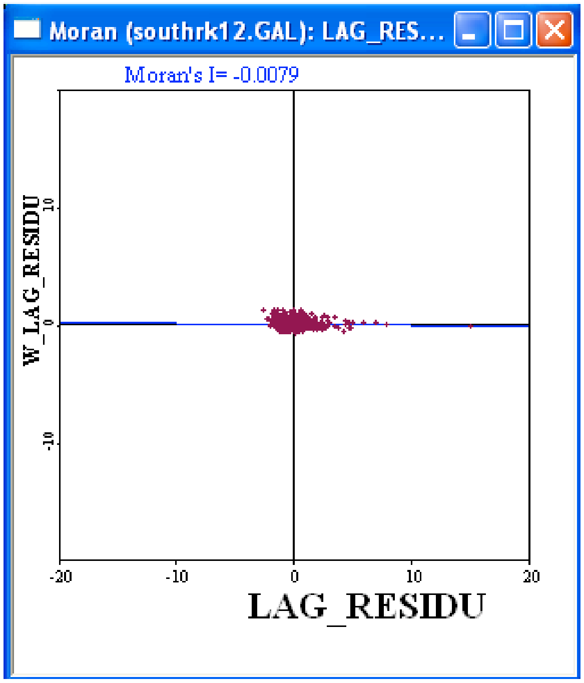


Figure 24.13: Moran scatter plot for spatial lag residuals, HR60.

24.4 Predicted Value and Residuals

In the spatial lag model, a distinction must be made between the model residuals, used in further diagnostic checks, and the prediction error. The

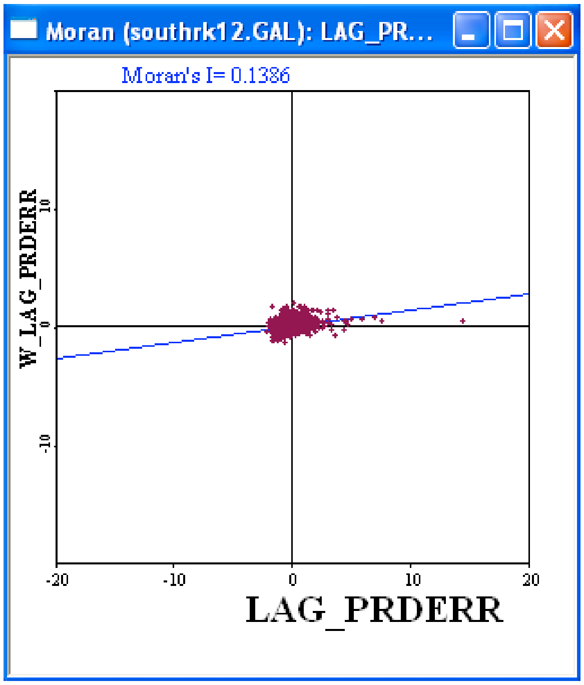


Figure 24.14: Moran scatter plot for spatial lag prediction errors, HR60.

latter is the difference between the observed and predicted values, obtained by only taking into consideration the exogenous variables.[4]

The difference between these results is illustrated in Figures 24.11 and 24.12 on p. 210. The values in the column headed by `HR60` are the observed values (y) for the first five observations. `LAG_RESIDU` contains the model residuals ($\hat{u}$), `LAG_PREDIC` the predicted values ($\hat{y}$), and `LAG_PRDERR` the prediction error ($y - \hat{y}$).

To further highlight the difference, construct a Moran scatter plot for both residuals, `LAG_RESIDU` and `LAG_PRDERR`, using `southrk12.GAL` as the weights file (select `Space > Univariate Moran` from the menu and specify the variable and weights file). The results should be as in Figures 24.13 (p. 210) and 24.14.

For `LAG_RESIDU`, the Moran's I test statistic is `-0.0079`, or essentially zero. This indicates that including the spatially lagged dependent variable

[4] Formally, the residuals are the estimates for the model error term, $(I - \hat{\rho}W)y - X\hat{\beta}$. The predicted values are $\hat{y} = (I - \hat{\rho}W)^{-1}X\hat{\beta}$, and the prediction error is $y - \hat{y}$.

term in the model has eliminated all spatial autocorrelation, as it should.[5] By contrast, the Moran's I statistic for `LAG_PRDERR` of `0.1386` is about the same as for the original OLS residuals. At first sight, this might suggest a problem, but in fact, this is as it is supposed to be. The prediction errors are an estimate for $(I - \rho W)^{-1}u$, or, the spatially transformed errors. Consequently, they are spatially correlated by construction.

24.5 Practice

Estimate the spatial lag model using ML for the same model and data set, but now using a first order rook spatial weights file. Compare the results to those obtained in this exercise. Also, check out the ranking of the W, LR and LM test statistics. What does this suggest for the specification you used?

Alternatively, for any of the OLS regression you may have run, or for the homicide model in different years, check the cases where the diagnostics suggest the Lag model as the alternative. Carry out ML estimation and compare the results of the spatial model to those in the classic regression.

[5] Note that this is not a formal hypothesis test, but only a descriptive statistic. The use of the permutation approach for residuals is *not* appropriate.

Exercise 25

Spatial Error Model

25.1 Objectives

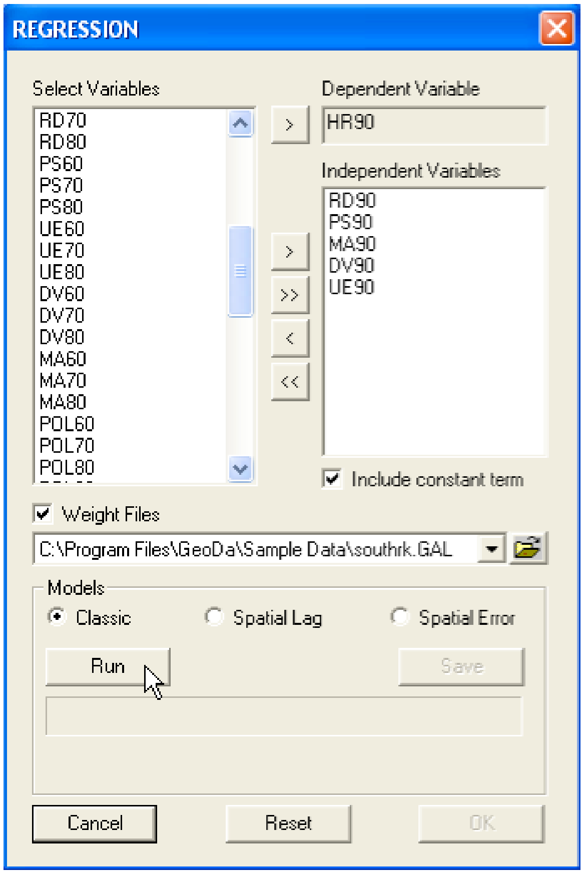
This exercise considers the estimation by means of maximum likelihood of a spatial regression model that includes a spatial autoregressive error term.[1] As for Exercise 24, methodological aspects are covered in Anselin (1988) and Anselin and Bera (1998). The same algorithm is used as in the spatial lag model, details of which can be found in Smirnov and Anselin (2001).

At the end of the exercise, you should know how to:

- set up the specification for a spatial error regression model
- interpret estimation results in the spatial error model
- interpret measures of fit in the spatial error model
- interpret regression diagnostics in the spatial error model
- understand the predicted value and different notions of residuals in the spatial error model
- compare the results of the spatial error model to those of the spatial lag model

More detailed information on the relevant operations can be found in the *Release Notes*, pp. 55–56.

[1] Formally, this model is $y = X\beta + \varepsilon$, with $\varepsilon = \lambda W\varepsilon + u$, where y is a vector of observations on the dependent variable, W is the spatial weights matrix, X is a matrix of observations on the explanatory variables, ε is a vector of spatially autocorrelated error terms, u a vector of i.i.d. errors, and λ and β are parameters.



Dependent Variable: HR90

Independent Variables: RD90, PS90, MA90, DV90, UE90

☑ Include constant term

☑ Weight Files: C:\Program Files\GeoDa\Sample Data\southrk.GAL

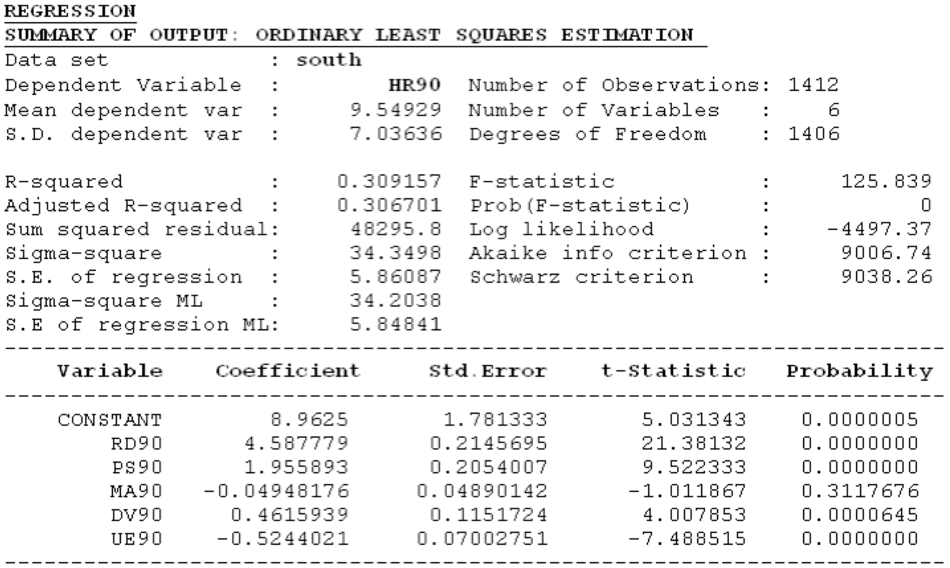
Models: ● Classic ○ Spatial Lag ○ Spatial Error

Figure 25.1: Homicide classic regression for 1990.

25.2 Preliminaries

We continue using the sample data set from the previous exercise, with homicide rates and related variables for 1412 counties in the South of the U.S (`south.shp` with `FIPSNO` as the `Key`). The base map should be as in Figure 24.1 on p. 202. If you have not already done so, make sure to create a spatial weights matrix for first order rook contiguity, as `southrk.GAL` (for specific instructions, see Section 15.2, and especially p. 114).

As for the ML estimation of the spatial lag model, this method only works for the spatial error model when the weights selected correspond to a symmetric contiguity relation. It thus works for rook and queen contiguity, as well as distance band contiguity, *but not for k-nearest neighbors.*



```
REGRESSION
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set             : south
Dependent Variable   :        HR90  Number of Observations: 1412
Mean dependent var   :     9.54929  Number of Variables   :    6
S.D. dependent var   :     7.03636  Degrees of Freedom    : 1406

R-squared            :    0.309157  F-statistic           :      125.839
Adjusted R-squared   :    0.306701  Prob(F-statistic)     :            0
Sum squared residual:      48295.8  Log likelihood        :     -4497.37
Sigma-square         :     34.3498  Akaike info criterion :      9006.74
S.E. of regression   :     5.86087  Schwarz criterion     :      9038.26
Sigma-square ML      :     34.2038
S.E of regression ML:      5.84841
-----------------------------------------------------------------------
    Variable    Coefficient      Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------
    CONSTANT         8.9625       1.781333       5.031343     0.0000005
        RD90       4.587779      0.2145695       21.38132     0.0000000
        PS90       1.955893      0.2054007       9.522333     0.0000000
        MA90    -0.04948176     0.04890142      -1.011867     0.3117676
        DV90      0.4615939      0.1151724       4.007853     0.0000645
        UE90     -0.5244021     0.07002751      -7.488515     0.0000000
-----------------------------------------------------------------------
```

Figure 25.2: OLS estimation results, homicide regression for 1990.

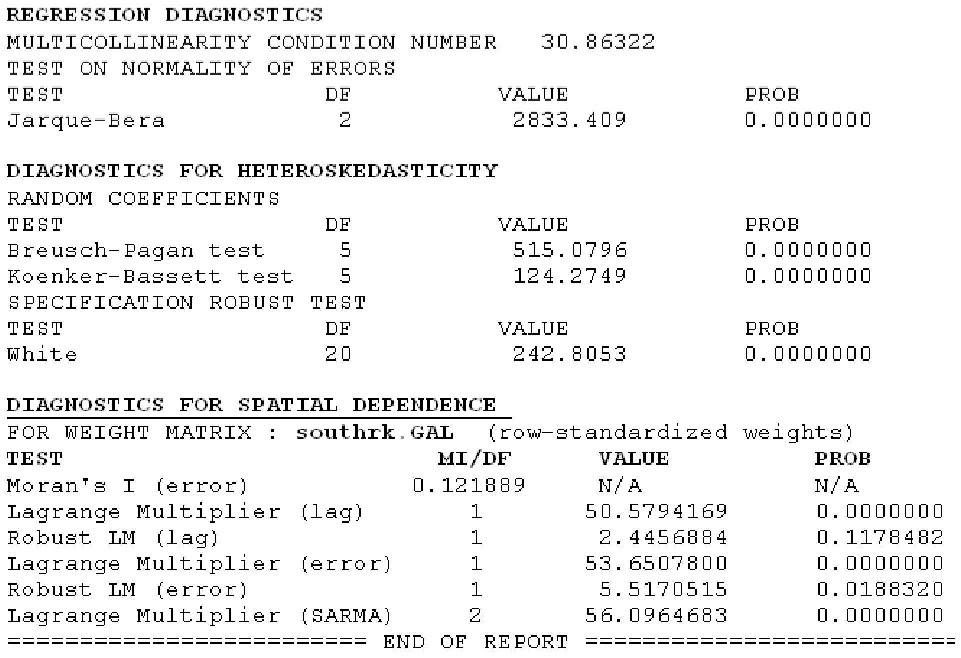
25.2.1 OLS with Diagnostics

As a point of reference, we will first run a Classic OLS regression (follow the instructions in Section 22.3), but now for the homicide rates in 1990. Specify HR90 as the dependent variable, and RD90, PS90, MA90, DV90, and UE90 as the explanatory variables. The regression dialog should be as in Figure 25.1 on p. 214. Make sure to check the box next to Weight Files, and specify southrk.GAL for the spatial weights.

Run the regression and check the results by clicking on OK. The OLS estimates are listed in Figure 25.2, with the diagnostics given in Figure 25.3 on p. 216.

The fit of the model is much better than for 1960, with an adjusted R^2 of 0.31. There is also a difference in the significance of coefficients, with PS90 now strongly significant and positive, but MA90 not significant. As before, for comparison purposes with the spatial model, note the Log Likelihood of -4497.37 and the AIC of 9006.74.

The regression diagnostics reveal considerable non-normality and heteroskedasticity, as well as high spatial autocorrelation. Following the steps outlined in Section 23.6.3, we conclude that a spatial error model is the proper alternative. Both LM-Lag and LM-Error are significant. Of the ro-



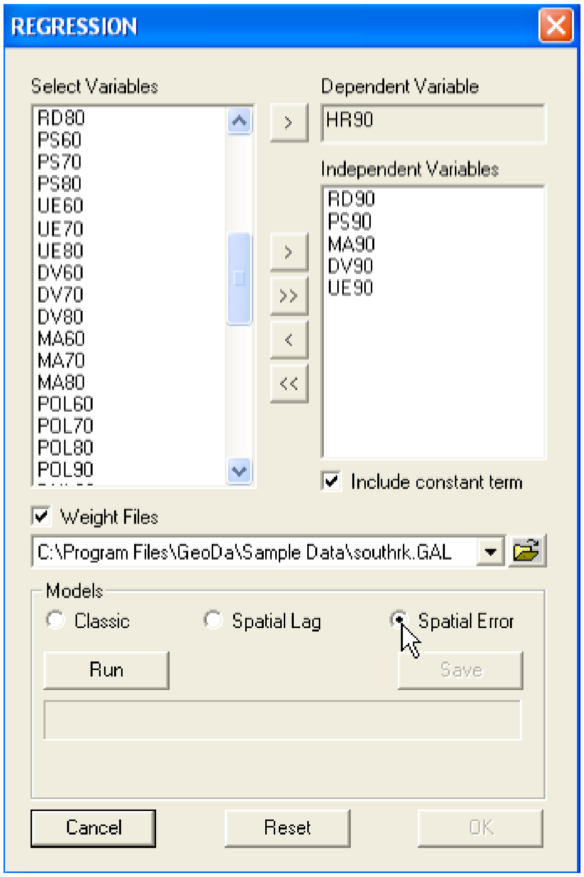
```
REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   30.86322
TEST ON NORMALITY OF ERRORS
TEST                  DF          VALUE            PROB
Jarque-Bera            2           2833.409        0.0000000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                  DF          VALUE            PROB
Breusch-Pagan test     5           515.0796        0.0000000
Koenker-Bassett test   5           124.2749        0.0000000
SPECIFICATION ROBUST TEST
TEST                  DF          VALUE            PROB
White                 20           242.8053        0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : southrk.GAL  (row-standardized weights)
TEST                          MI/DF      VALUE          PROB
Moran's I (error)           0.121889     N/A            N/A
Lagrange Multiplier (lag)       1       50.5794169      0.0000000
Robust LM (lag)                 1        2.4456884      0.1178482
Lagrange Multiplier (error)     1       53.6507800      0.0000000
Robust LM (error)               1        5.5170515      0.0188320
Lagrange Multiplier (SARMA)     2       56.0964683      0.0000000
========================= END OF REPORT ==========================
```

Figure 25.3: OLS diagnostics, homicide regression for 1990.

bust forms, the Robust LM-Error statistic is significant ($p < 0.02$), but the Robust LM-Lag statistic is clearly not ($p < 0.12$). This sets the stage for the estimation of the spatial error model.

25.3 ML Estimation with Diagnostics

The ML estimation of the spatial error model is invoked in the same manner as for the classic regression and the spatial lag model (see Section 24.3). Proceed by clicking on `Regress` on the main menu, or by choosing `Methods > Regress` before a project is loaded. The title and output file dialog is identical for all regression analyses. As before, this allows for the specification of an output file name, such as `southerr.rtf`. Click `OK`, to bring up the familiar regression specification dialog.


REGRESSION

Select Variables

RD80
PS60
PS70
PS80
UE60
UE70
UE80
DV60
DV70
DV80
MA60
MA70
MA80
POL60
POL70
POL80
POL90

Dependent Variable: HR90

Independent Variables:

RD90
PS90
MA90
DV90
UE90

☑ Include constant term

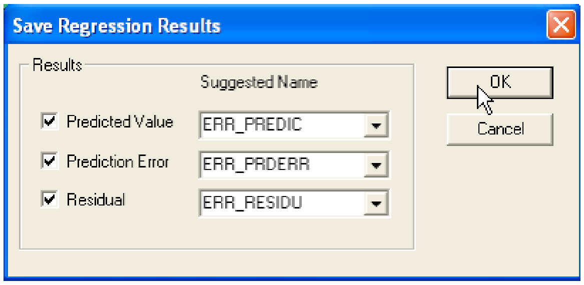
☑ Weight Files

C:\Program Files\GeoDa\Sample Data\southrk.GAL

Models

○ Classic ○ Spatial Lag ● Spatial Error

Figure 25.4: Spatial error model specification dialog.


Save Regression Results

Results

	Suggested Name
☑ Predicted Value	ERR_PREDIC
☑ Prediction Error	ERR_PRDERR
☑ Residual	ERR_RESIDU

Figure 25.5: Spatial error model residuals and predicted values dialog.

25.3.1 Model Specification

The dialog, shown in Figure 25.4 on p. 217, is the same as before. Enter the same set of dependent and explanatory variables as for the classic case and make sure to specify the spatial weights file, as illustrated in the Figure. Instead of the default `Classic`, check the radio button next to `Spatial Error`. Invoke the estimation routine as before, by clicking on the `Run` button.

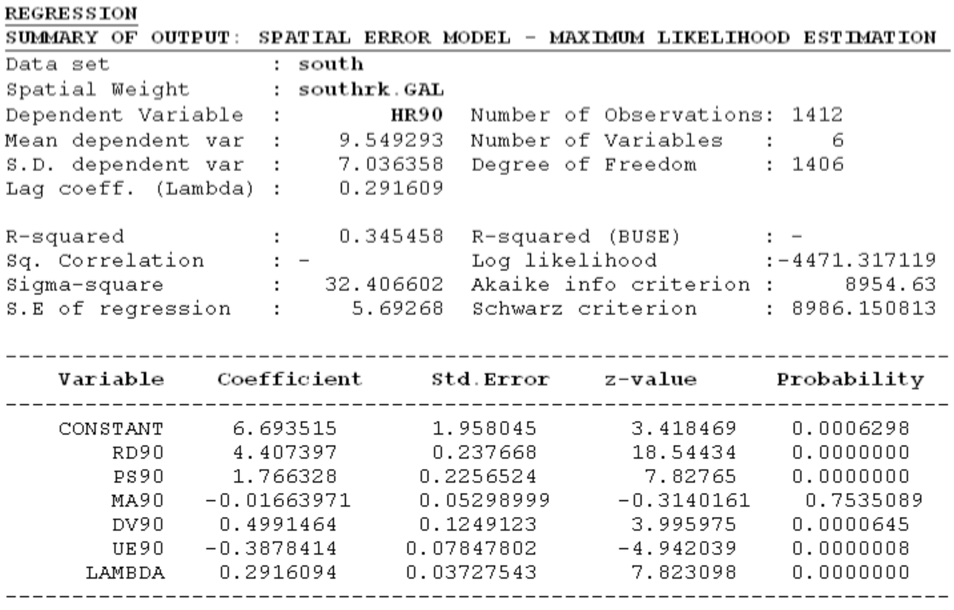
As before, after the estimation is completed, as indicated by the familiar progress bar, make sure to click on the `Save` button *before* selecting `OK`. This will bring up the dialog to specify variable names to add the residuals and predicted values to the data table.

As shown in Figure 25.5 on p. 217, there are three options for this in the spatial error model. These are covered in more detail in Section 25.4. For now, select all three check boxes and keep the variable names to their defaults of `ERR_PREDIC` for the `Predicted Value`, `ERR_PRDERR` for the `Prediction Error`, and `ERR_RESIDU` for the `Residual`. Click on `OK` to get back to the regression dialog and select `OK` again to bring up the estimation results and diagnostics.

25.3.2 Estimation Results

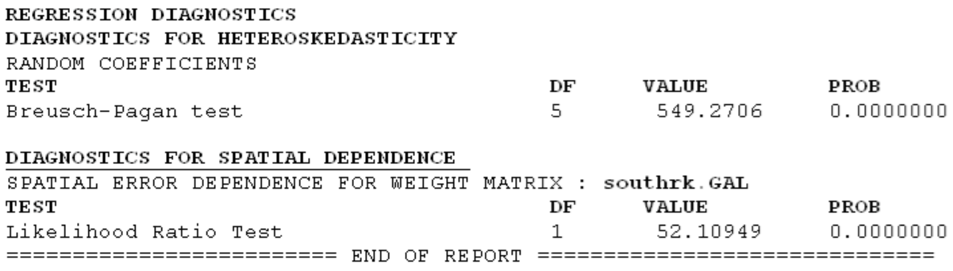
The estimates and measures of fit are given in Figure 25.6 on p. 219. Again, as in the ML Lag estimation, the R^2 listed is a so-called pseudo-R^2, which is not directly comparable with the measure given for OLS results. The proper measures of fit are the Log-Likelihood, AIC and SC. If we compare the values in Figure 25.6 to those for OLS in Figure 25.2 (p. 215), we notice an increase in the Log-Likelihood from `-4497.37` (for OLS) to `-4471.32`. Compensating the improved fit for the added variable (the spatially lagged dependent variable), the AIC (from `9006.74` to `8954.63`) and SC (from `9038.26` to `8986.15`) both decrease relative to OLS, again suggesting an improvement of fit for the spatial error specification.

The spatial autoregressive coefficient is estimated as `0.29`, and is highly significant ($p < 0.0000000$). As in the OLS case, the coefficient for `MA90` is not significant ($p < 0.75$), but all the others are. Their value is slightly less in absolute value relative to the OLS results, except for `DV90`, where the change is from `0.462` to `0.499`.



```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : south
Spatial Weight      : southrk.GAL
Dependent Variable  :        HR90  Number of Observations: 1412
Mean dependent var  :    9.549293  Number of Variables   :    6
S.D. dependent var  :    7.036358  Degree of Freedom     : 1406
Lag coeff. (Lambda) :    0.291609

R-squared           :    0.345458  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-4471.317119
Sigma-square        :   32.406602  Akaike info criterion :     8954.63
S.E of regression   :     5.69268  Schwarz criterion     : 8986.150813

-----------------------------------------------------------------------
    Variable    Coefficient      Std.Error     z-value      Probability
-----------------------------------------------------------------------
    CONSTANT      6.693515       1.958045       3.418469    0.0006298
        RD90      4.407397       0.237668       18.54434    0.0000000
        PS90      1.766328      0.2256524        7.82765    0.0000000
        MA90   -0.01663971     0.05298999     -0.3140161     0.7535089
        DV90     0.4991464      0.1249123       3.995975    0.0000645
        UE90    -0.3878414     0.07847802      -4.942039    0.0000008
      LAMBDA     0.2916094     0.03727543       7.823098    0.0000000
-----------------------------------------------------------------------
```

Figure 25.6: Spatial error model ML estimation results, HR90.

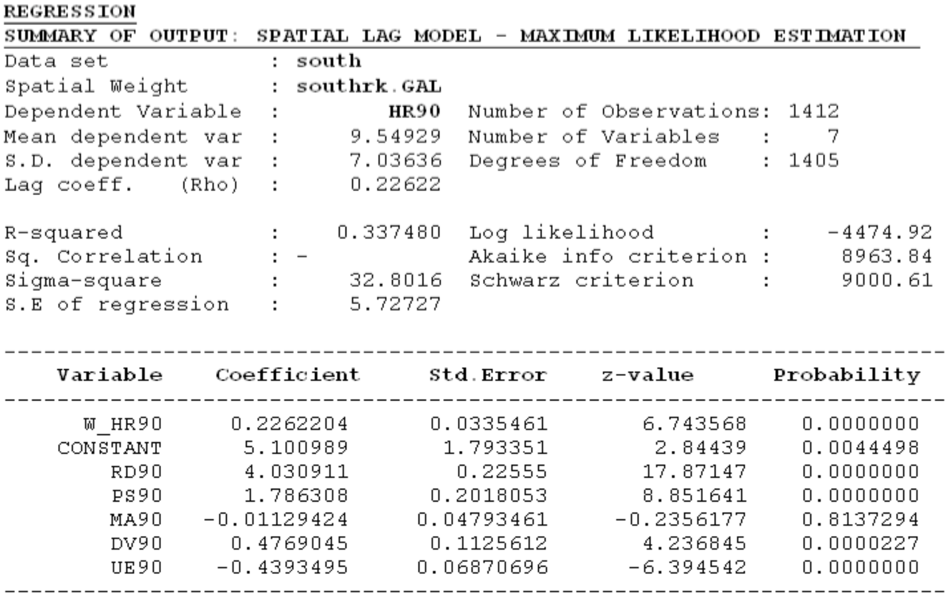


```
REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE          PROB
Breusch-Pagan test                      5        549.2706      0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : southrk.GAL
TEST                                    DF      VALUE          PROB
Likelihood Ratio Test                   1        52.10949      0.0000000
======================== END OF REPORT ==============================
```

Figure 25.7: Spatial error model ML diagnostics, HR90.

25.3.3 Diagnostics

The two model diagnostics reported for the ML error estimation are the same as for the lag specification, a Breusch-Pagan test for heteroskedasticity, and a Likelihood Ratio test on the spatial autoregressive coefficient. They are listed in Figure 25.7. Both diagnostics are highly significant, suggesting remaining specification problems in the model. Checking the order of the W, LR and LM statistics on the spatial autoregressive error coefficient, we



```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : south
Spatial Weight      : southrk.GAL
Dependent Variable  :          HR90  Number of Observations: 1412
Mean dependent var  :       9.54929  Number of Variables   :    7
S.D. dependent var  :       7.03636  Degrees of Freedom    : 1405
Lag coeff.   (Rho)  :       0.22622

R-squared           :      0.337480  Log likelihood        :     -4474.92
Sq. Correlation     : -              Akaike info criterion :      8963.84
Sigma-square        :       32.8016  Schwarz criterion     :      9000.61
S.E of regression   :       5.72727

-----------------------------------------------------------------------
    Variable    Coefficient      Std.Error     z-value      Probability
-----------------------------------------------------------------------
      W_HR90      0.2262204      0.0335461        6.743568    0.0000000
    CONSTANT       5.100989       1.793351         2.84439    0.0044498
        RD90       4.030911        0.22555        17.87147    0.0000000
        PS90       1.786308      0.2018053        8.851641    0.0000000
        MA90    -0.01129424     0.04793461      -0.2356177    0.8137294
        DV90      0.4769045      0.1125612        4.236845    0.0000227
        UE90     -0.4393495     0.06870696       -6.394542    0.0000000
-----------------------------------------------------------------------
```

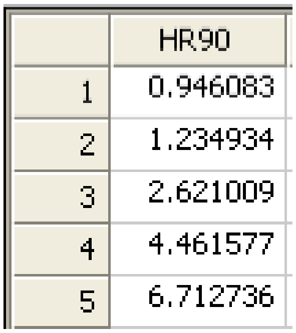
Figure 25.8: Spatial lag model ML estimation results, HR90.

find $W = 7.82^2 = 61.2$ (the square of the z-value of the asymptotic t-test in Figure 25.6), $LR = 52.1$, but $LM = 53.7$ (see Figure 25.2). As in the lag model in Exercise 24, this violates the expected order and indicates a less than satisfactory model specification at this point.

To further compare the results between an error and lag model, run the `Spatial Lag` option using the same specification as above, i.e., using `HR90` as the dependent variable (for instructions on running the spatial lag model, see Section 24.3). The estimation results are listed in Figure 25.8.

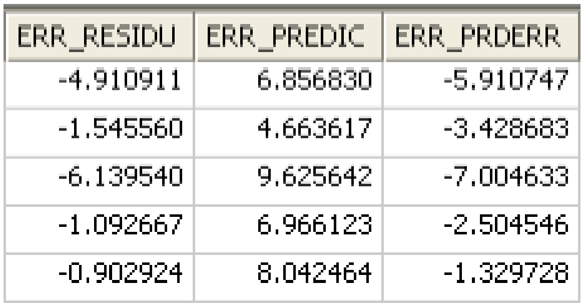
The results are very similar to those of the error model in terms of coefficient magnitude, sign and significance, further highlighting the difficulties in discriminating between the two spatial models. In terms of fit, the results confirm the indication given by our decision rule. The Log Likelihood in the error model (`-4471`) is slightly better than that in the spatial lag model (`-4474`). Similarly, the AIC is lower for the error model (`8954.63`) than for the lag model (`8963.84`).

The similarity between the results in the two models and the indication of remaining specification problems suggests that a refinement of the model may be in order.



	HR90
1	0.946083
2	1.234934
3	2.621009
4	4.461577
5	6.712736

Figure 25.9: Observed value, HR90.



ERR_RESIDU	ERR_PREDIC	ERR_PRDERR
-4.910911	6.856830	-5.910747
-1.545560	4.663617	-3.428683
-6.139540	9.625642	-7.004633
-1.092667	6.966123	-2.504546
-0.902924	8.042464	-1.329728

Figure 25.10: Spatial error predicted values and residuals HR90.

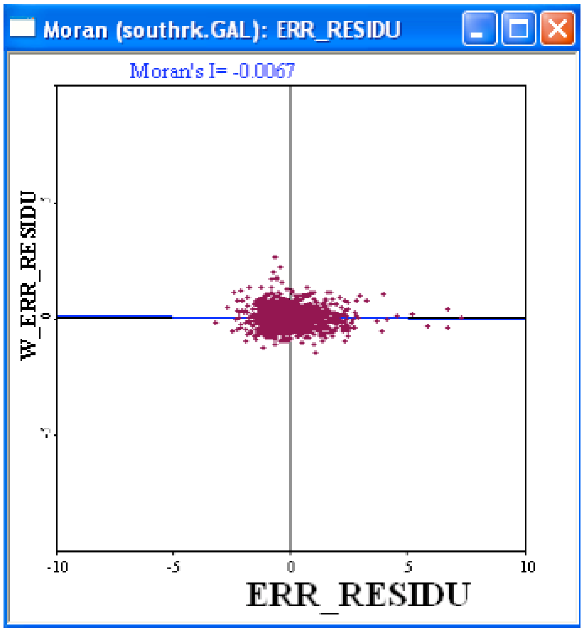


Figure 25.11: Moran scatter plot for spatial error residuals, HR90.

25.4 Predicted Value and Residuals

As in the spatial lag model, in the spatial error model a distinction must be made between the model residuals, used in further diagnostic checks, and the prediction error. The latter is the difference between the observed and

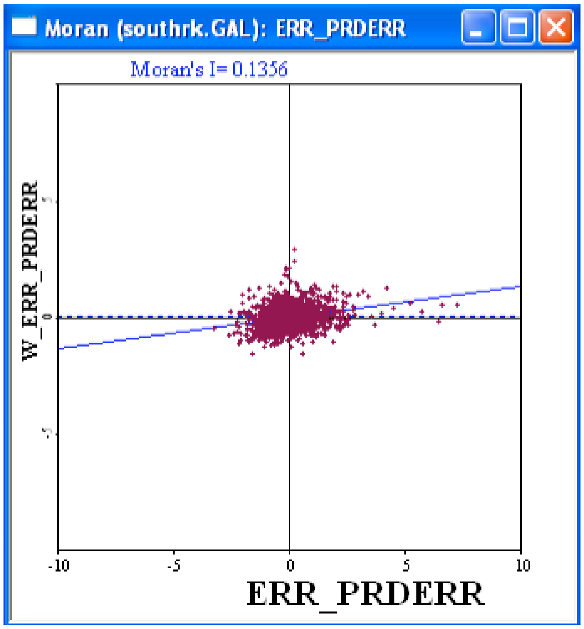


Figure 25.12: Moran scatter plot for spatial error prediction errors, HR90.

predicted values, which correspond to the conditional expectation of the y, given X.[2]

The difference between these results is illustrated in Figures 25.9 and 25.10 on p. 221. The values in the column headed by HR90 are the observed values (y) for the first five observations. ERR_RESIDU contains the model residuals ($\hat{u}$), ERR_PREDIC the predicted values ($\hat{y}$), and ERR_PRDERR the prediction error ($y - \hat{y}$).

As for the spatial lag model, construct a Moran scatter plot for both residuals, ERR_RESIDU and ERR_PRDERR, using southrk.GAL as the weights file (select Space > Univariate Moran from the menu and specify the variable and weights file). The results should be as in Figures 25.11 (p. 221) and 25.12.

For ERR_RESIDU, the Moran's I test statistic is -0.0067, or essentially zero. This indicates that including the spatially autoregressive error term

[2] Formally, the residuals are the estimates for the uncorrelated (or, spatially filtered) model error term, $\hat{u} = (I - \hat{\lambda}W)\hat{\varepsilon}$. Note that $\hat{\varepsilon} = \hat{\lambda}W\hat{\varepsilon} + \hat{u}$ is the prediction error, or $y - \hat{y}$, with $\hat{y} = X\hat{\beta}$, and $\hat{\beta}$ as the ML estimate for β.

in the model has eliminated all spatial autocorrelation, as it should.[3] By contrast, the Moran's I statistic for `ERR_PRDERR` of `0.1356` is about the same as for the original OLS residuals. At first sight, this might suggest a problem, but in fact, this is as it is supposed to be. The prediction errors are an estimate for $\varepsilon = (I - \lambda W)^{-1}u$, or, the spatially transformed idiosyncratic errors u. Consequently, they are spatially correlated by construction.

25.5 Practice

Estimate the spatial error model using ML for the same model and data set, but now using the `southrk12.GAL` weights file from Exercise 24. Compare the results to those obtained in this exercise and to the results for a spatial lag model using these weights. Also, check out the ranking of the W, LR and LM test statistics.

Alternatively, for any of the OLS regression you may have run, or for the homicide model in different years, check the cases where the diagnostics suggest the error model as the alternative. Carry out ML estimation and compare the results of the spatial model to those in the classic regression as well as to the matching lag model.

[3] Note that this is not a formal hypothesis test, but only a descriptive statistic. As in the other regression models, the use of the permutation approach for residuals is *not* appropriate.

Bibliography

Anselin, L. (1988). *Spatial Econometrics: Methods and Models.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin, L. (1993). Discrete space autoregressive models. In Goodchild, M. F., Parks, B., and Steyaert, L., editors, *GIS and Environmental Modeling*, pages 454–469. Oxford University Press, Oxford.

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, 27:93–115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Fischer, M., Scholten, H., and Unwin, D., editors, *Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences*, pages 111–125. Taylor and Francis, London.

Anselin, L. (2001). Rao's score test in spatial econometrics. *Journal of Statistical Planning and Inference*, 97:113–139.

Anselin, L. (2003a). *GeoDa 0.9 User's Guide.* Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

Anselin, L. (2003b). *An Introduction to EDA with GeoDa.* Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

Anselin, L. (2004). *GeoDa 0.95i Release Notes.* Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

Anselin, L. and Bera, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In Ullah, A. and

Giles, D. E., editors, *Handbook of Applied Economic Statistics*, pages 237–289. Marcel Dekker, New York.

Anselin, L., Bera, A., Florax, R. J., and Yoon, M. (1996). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics*, 26:77–104.

Anselin, L., Bongiovanni, R., and Lowenberg-DeBoer, J. (2004a). A spatial econometric approach to the economics of site-specific nitrogen management in corn production. *American Journal of Agricultural Economics*, 86:675–687.

Anselin, L. and Florax, R. J. (1995). Small sample properties of tests for spatial dependence in regression models: Some further results. In Anselin, L. and Florax, R. J., editors, *New Directions in Spatial Econometrics*, pages 21–74. Springer-Verlag, Berlin.

Anselin, L., Kim, Y.-W., and Syabri, I. (2004b). Web-based analytical tools for the exploration of spatial data. *Journal of Geographical Systems*, 6:197–218.

Anselin, L., Syabri, I., and Kho, Y. (2004c). GeoDa, an introduction to spatial data analysis. *Geographical Analysis*. forthcoming.

Anselin, L., Syabri, I., and Smirnov, O. (2002). Visualizing multivariate spatial correlation with dynamically linked windows. In Anselin, L. and Rey, S., editors, *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*. Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara. CD-ROM.

Assunção, R. and Reis, E. A. (1999). A new proposal to adjust Moran's I for population density. *Statistics in Medicine*, 18:2147–2161.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. John Wiley and Sons, New York, NY.

Baller, R., Anselin, L., Messner, S., Deane, G., and Hawkins, D. (2001). Structural covariates of U.S. county homicide rates: Incorporating spatial effects. *Criminology*, 39(3):561–590.

Calvo, E. and Escobar, M. (2003). The local voter: A geographically weighted approach to ecological inference. *American Journal of Political Science*, 47(1):189–204.

Cressie, N. (1993). *Statistics for Spatial Data.* Wiley, New York.

Dubin, R. (1992). Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics*, 22:433–452.

Gilley, O. and Pace, R. K. (1996). On the Harrison and Rubinfeld data. *Journal of Environmental Economics and Management*, 31:403–405.

Harrison, D. and Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5:81–102.

Kelejian, H. H. and Robinson, D. P. (1992). Spatial autocorrelation: A new computationally simple test with an application to per capita country police expenditures. *Regional Science and Urban Economics*, 22:317–333.

Lawson, A. B., Browne, W. J., and Rodeiro, C. L. V. (2003). *Disease Mapping with WinBUGS and MLwiN.* John Wiley, Chichester.

Messner, S. F. and Anselin, L. (2004). Spatial analyses of homicide with areal data. In Goodchild, M. and Janelle, D., editors, *Spatially Integrated Social Science*, pages 127–144. Oxford University Press, New York, NY.

Pace, R. K. and Gilley, O. (1997). Using the spatial configuration of the data to improve estimation. *Journal of Real Estate Finance and Economics*, 14:333–340.

Smirnov, O. and Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: A characteristic polynomial approach. *Computational Statistics and Data Analysis*, 35:301–319.

Waller, L., Carlin, B., Xia, H., and Gelfand, A. (1997). Hierarchical spatio-temporal mapping of disease rates. *Journal of the American Statistical Association*, 92:607–617.

Xia, H. and Carlin, B. P. (1998). Spatio-temporal models with errors in covariates: Mapping Ohio lung cancer mortality. *Statistics in Medicine*, 17:2025–2043.