



Center for Spatially Integrated Social Science

GeoDa™ 0.9 User's Guide

Luc Anselin

Spatial Analysis Laboratory
Department of Agricultural and Consumer Economics
University of Illinois, Urbana-Champaign
Urbana, IL 61801
<http://sal.agecon.uiuc.edu/>

and

Center for Spatially Integrated Social Science
<http://www.csiss.org/>

Revised, June 15, 2003

Copyright © 2003 Luc Anselin, All Rights Reserved

Acknowledgments

The development of the GeoDa™ software for geodata analysis and its antecedents has been supported in part by research projects funded by a variety of sources. Most recent among these are U.S. National Science Foundation grant BCS-9978058 to the Center for Spatially Integrated Social Science (CSISS). Other funding was provided by NSF grant SBR-9410612, by a grant from the National Consortium on Violence Research (NCOVR is supported under grant SBR-9513040 from NSF) and by NSF grant SES-8810917 to the National Center for Geographic Information and Analysis.

Many thanks go to the students in the Fall 2002 class of ACE 492SA, Spatial Analysis (Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign), for being such good sports in “testing” an early version of the software. Comments from users (too many to list individually) are greatly appreciated. Yanqui Ren and Widodo Baroka provided research assistance to the development of earlier versions of the software.

GeoDa™ is a trademark of Luc Anselin, All Rights Reserved.

GeoDa™ incorporates licensed libraries from ESRI’s MapObjects LT2.

ESRI, ArcView, ArcGIS and MapObjects are trademarks of Environmental Systems Research Institute.

Other companies and products mentioned herein are trademarks or registered trademarks of their respective trademark owners.

GeoDa contains code derived from publicly available and licensed sources. The Table functionality in GeoDa is derived from the MFC Grid Control 2.24 by Chris Mauder (<http://www.codeproject.com/miscctrl/gridctrl.asp>). The nearest neighbor calculations are compiled from source code contained in the ANN Code by David Mount and Sunil Arya (see Appendix B for a copy of the License Agreement). The Thiessen polygon code is based on source code by Yasuaki Oishi obtained from (<http://www.simplex.t.u-tokyo.ac.jp/~oihsi/src/voronoi/>) and described in Y. Oishi and K. Sugihara (1991). Numerically robust divide-and-conquer algorithm for constructing Voronoi diagrams (in Japanese). *Transactions of the Information Processing Society of Japan* 32 (6), 709-720.

GeoDa Project:

Project Director: Luc Anselin

Software Design and Development: Luc Anselin, Ibnu Syabri and Oleg Smirnov

Research Assistance: Younghin Kho, Yong Wook Kim

Table of Contents

Introduction	1
Getting Started	3
Project setting.....	3
Key variable	4
Menu Overview	6
General features.....	6
File menu	6
Project toolbar.....	7
Edit menu	7
Edit > New Map.....	7
Edit > Duplicate Map.....	8
Edit > Add Layer	9
Edit > Remove Layer	9
Edit > Select Variable	10
Edit > Copy to Clipboard	10
Edit toolbar	10
View menu	11
View > Toolbar.....	11
View > Status Bar	11
Tools menu.....	11
Tools > Weights.....	12
Weights toolbar.....	12
Tools > Shape	12
Tools > Data Export.....	13
Explore menu	14
Explore toolbar	15
Map menu	15
Map > Box Map.....	16
Map > Smooth	16
Map > Map Movie	17
Options menu	17
Window menu	17
Help menu.....	18
Manipulating Spatial Data.....	19

Creating a point shape file with centroids	19
Adding centroids to the data table	23
Creating a dBase file with centroids	24
Creating Thiessen polygons as shape files	26
Creating point shape files from dbf and ascii input files	28
Exporting boundary files	31
Exporting data tables	33
Mapping	35
Standard choropleth maps	35
Getting started	35
Map > Quantile	36
Map > Std Dev	38
Outlier maps	39
Map > Box Map	39
Map > Percentile	40
Map movie	40
Map options	41
Options > Selection Shape	42
Options > Zoom	42
Options > Color	43
Options > Save Image as	45
Options > Save Selected Obs	45
Options > Add Centroids to table	46
Smoothing Rate Maps	47
Map > Smooth > Raw Rate	48
Map > Smooth > Excess Risk	49
Map > Smooth > Empirical Bayes	49
Map > Smooth > Spatial Rate	50
Map > Smooth > Spatial Empirical Bayes	51
Smoothing options	51
Editing and Manipulating Tables	54
Sorting records by field	54
Editing individual table cells	55
Promoting selected records	56
Clear selection	57
Range selection	57

Save selected observations.....	59
Calculating new variables.....	59
Computing spatially lagged variables.....	61
Computing rate variables.....	62
Saving and joining tables.....	63
Undoing changes.....	64
Statistical Graphs.....	65
Explore > Histogram.....	65
Explore > Box Plot.....	66
Explore > Scatter Plot.....	68
Linking and Brushing.....	70
Selection in a map.....	70
Selection in a histogram.....	71
Selection in a box plot.....	72
Selection in a scatter plot.....	73
Selection in a table.....	74
Linking.....	75
Brushing.....	75
Creating and Manipulating Spatial Weights.....	77
Opening existing weights.....	77
Creating new weights.....	78
Spatial weights file formats.....	80
Contiguity based spatial weights.....	82
Higher order contiguity weights.....	83
Distance band spatial weights.....	83
K-nearest neighbor spatial weights.....	85
Characteristics of spatial weights.....	86
Global Spatial Autocorrelation.....	88
Univariate Moran scatter plot.....	88
Options > Exclude Selected ON.....	89
Options > Randomization.....	91
Options > Envelope Slopes ON.....	92
Options > Save Results.....	93
Other Univariate Moran options.....	94
Bivariate Moran scatter plot.....	94
Moran scatter plot matrix.....	95

Moran scatter plot for EB rates	97
Local Spatial Autocorrelation.....	99
Univariate LISA	99
Significance map.....	99
Cluster map.....	100
Box plot.....	101
Moran scatter plot.....	102
LISA options.....	102
Options > Randomization.....	102
Options > Significance filter	103
Options > Save results.....	104
Bivariate LISA	105
EB LISA	105
References	106
Appendix A – GeoDa License Agreement.....	108
Appendix B – ANN License Agreement.....	111
Appendix C – Installing GeoDa.....	112
Appendix D –New in GeoDa 0.9.3	114

List of Figures

1.	GeoDa opening screen	3
2.	GeoDa Project Setting dialog.....	4
3.	Open File dialog	4
4.	Key variable	5
5.	Opening screen with selected shape file outline	5
6.	Menu and toolbars	6
7.	File Menu for new project and active project	6
8.	Edit menu	7
9.	Multiple new maps	8
10.	Duplicate maps.....	8
11.	Point layer added to polygon layer (with choropleth map for points).....	9
12.	Variable selection	10
13.	View menu	11
14.	Tools menu.....	11
15.	Tools Weights submenu	12
16.	Tools Shape submenu.....	12
17.	Tools Data Export submenu.....	13
18.	Explore menu	14
19.	Map menu	15
20.	Map Box Map submenu.....	16
21.	Map Smooth submenu	16
22.	Map Movie submenu	17
23.	Window menu	17
24.	Help menu.....	18
25.	About GeoDa	18
26.	Shape conversion dialog.....	20
27.	Shape conversion input file selections.....	20
28.	Shape conversion input file thumbnail	20
29.	Shape conversion output save as file dialog.....	21
30.	Shape conversion create button.....	21
31.	Completed Polygon to Point conversion	21
32.	Centroids for North Carolina counties.....	22
33.	Data table for centroids point shape file.....	22
34.	Add centroids to table option in menu.....	23

35.	Add centroids to table from map.....	23
36.	Add centroids to table dialog	24
37.	Columns with centroids added to the sidcent data table.....	24
38.	Export Centroids file dialog.....	25
39.	Centroids dbf file loaded into spreadsheet.....	25
40.	Point to Polygon shape conversion dialog.....	26
41.	Completed Point to Polygon shape conversion.....	27
42.	Thiessen polygons for Baltimore point data, with point data overlaid	27
43.	Data table for Thiessen polygon shape file.....	28
44.	Baltimore point data as a comma delimited (csv) file	29
45.	Points from dbf file and variable selection dialog.....	29
46.	Baltimpoint point shape file created from dbf input	30
47.	Points from ascii file and variable selection dialog.....	30
48.	Baltipoint2 point shape file created from ascii input.....	31
49.	Boundary file format 1 (left) and format 1a (right).....	32
50.	Boundary file format 2 (left) and format 2a (right).....	32
51.	Bounding box coordinates for Columbus polygons	32
52.	Polygon shape to boundary file export dialog	33
53.	Ascii text output file from Columbus data set polygon shape file	33
54.	Exporting data dialog.....	34
55.	Data ready for export.....	34
56.	Exposing the Legend Pane.....	35
57.	Quantile Map dialog	36
58.	Columbus neighborhood crime quintile map.....	36
59.	Legend color customization dialog	37
60.	Customizing legend colors.....	37
61.	Standard deviational map for Columbus neighborhood crime	38
62.	Box Map for 1979 SIDS death rate in North Carolina counties	39
63.	Percentile map for 1979 SIDS death rates in North Carolina counties	40
64.	Map options menu	41
65.	Map options and shortcuts context menu	41
66.	Map selection rules	42
67.	Zoom options	42
68.	Percentile map after zoom in	43
69.	Map color selection options	43
70.	Changing the map color.....	44

71.	Customizing the selection hash marks color.....	44
72.	Customizing the background color for a map view	44
73.	Save map as image dialog.....	45
74.	Saved bitmap image.....	45
75.	Save selected observations indicator variable dialog	46
76.	Indicator variable added to data table.....	46
77.	Rate smoothing dialog	47
78.	Specifying spatial weights for rate smoothing.....	48
79.	Raw rate box map for 1979 SIDS death rate in North Carolina counties.....	48
80.	Excess risk map for 1979 SIDS death rates in North Carolina counties	49
81.	EB smoothed box map for 1979 SIDS death rates in North Carolina counties	50
82.	Spatial rate box map for 1979 SIDS death rates in North Carolina counties	50
83.	Spatial EB rate box map for 1979 SIDS death rates in North Carolina counties..	51
84.	Save rates option for rate maps	52
85.	Rate variable name specification.....	52
86.	Computed rate variables added to data table	52
87.	New rate variables available for analysis	53
88.	Table menu.....	54
89.	Records sorted by increasing values of POLYID	55
90.	Records sorted by decreasing values of POLYID	55
91.	Editing individual table cells.....	56
92.	Manual record selection in a table.....	56
93.	Selected records promoted to the top of the table	56
94.	Range selection interval specification	57
95.	Range selection applied to table.....	58
96.	Specifying a regime variable following a range selection.....	58
97.	Regime variable added to table	58
98.	Save selected observations in table	59
99.	Selected observations indicator variable added to data table.....	59
100.	Specifying a variable name for a new column (field)	60
101.	Field calculation, unary operations.....	60
102.	Field calculation, binary operations	61
103.	Computing a spatially lagged variable	62
104.	Rate calculation in a table	62
105.	Join tables dialog.....	63
106.	Joined variables added to data table	64

107.	Change intervals dialog for histogram.....	65
108.	Histogram for Columbus neighborhood housing values	66
109.	Histogram options	66
110.	Box plot for Columbus neighborhood housing values using 1.5 and 3 as hinge..	67
111.	Box plot options	67
112.	Scatter plot of Columbus neighborhood crime on housing values.....	68
113.	Scatter plot options	69
114.	Selected locations on the Columbus neighborhood map using line select	71
115.	Histogram with three highest categories selected	72
116.	Box plot with selected observations	73
117.	Scatter plot recomputed with selected observations excluded.....	74
118.	Linked box plot and box map.....	75
119.	Select weight dialog to open an existing spatial weights file	77
120.	Adding a spatial weights file to the project	78
121.	Using an already opened spatial weights file.....	78
122.	Creating weights dialog	79
123.	Output file specification for spatial weight.....	79
124.	Key Variable specification.....	80
125.	GAL format spatial weights file for North Carolina counties	81
126.	GWT format spatial weights file for 4 th order nearest neighbors in Columbus....	81
127.	Construction of spatial weights based on rook contiguity.....	82
128.	Completion of weights construction.....	83
129.	Specifying a higher order of contiguity	83
130.	Distance weights creation	84
131.	Specifying the threshold distance for distance band spatial weights	85
132.	Specifying the order for k-nearest neighbor spatial weights	85
133.	Selecting a weights file for the analysis of its characteristics.....	86
134.	Connectivity of first order contiguity for North Carolina counties.....	87
135.	Rook connectivity for Iredell county, NC	87
136.	Weights selection dialog.....	88
137.	Moran scatter plot for Columbus CRIME	89
138.	Univariate Moran scatter plot options	90
139.	Univariate Moran scatter plot with selected observations excluded	90
140.	Randomization option in Moran scatter plot.....	91
141.	Reference distribution for Moran's I for Columbus CRIME	92
142.	Reference distribution for Moran's I for Columbus OPEN	92

143.	Envelope slopes in the Moran scatter plot	93
144.	Save results dialog in Moran scatter plot.....	93
145.	Moran scatter plot variables added to the data table	94
146.	Bivariate Moran scatter plot.....	95
147.	Moran scatter plot matrix.....	96
148.	Non-spatial correlation matrices	96
149.	EB Moran's I variable selection dialog	97
150.	Moran scatter plot for EB standardized rates.....	98
151.	EB Moran save results variable specification dialog	98
152.	Dialog for LISA windows.....	99
153.	LISA significance map for Columbus CRIME	100
154.	LISA cluster map for Columbus CRIME	101
155.	Box plot for Local Moran statistics	101
156.	LISA map options.....	102
157.	LISA randomization option	103
158.	LISA significance filter option.....	103
159.	LISA maps after applying a significance filter	104
160.	LISA save results dialog.....	104
161.	LISA results added to data table	105

Introduction

GeoDa is the latest incarnation of a collection of software tools designed to implement techniques for exploratory spatial data analysis (ESDA) on lattice data.¹ It is intended to provide a user friendly and graphical interface to methods of descriptive spatial data analysis, such as autocorrelation statistics and indicators of spatial outliers.

The design of GeoDa consists of an interactive environment that combines maps with statistical graphics, using the technology of dynamically linked windows. Its origins trace back to initial efforts to develop a bridge between ESRI's ArcInfo GIS and the SpaceStat software package for spatial data analysis (Anselin et al. 1993). A second stage in the development of this idea consisted of a series of extensions to ESRI's ArcView 3.x GIS that implemented linked windows and brushing (see Anselin and Bao 1997, Anselin and Smirnov 1998, Anselin 2000). In contrast to these extensions, the current software is freestanding and does not require a specific GIS system. GeoDa runs under any of the Microsoft Windows flavored operating systems (Win95, 98, 2000, NT, Me and Xp). Its installation routine contains all required files.

GeoDa adheres to ESRI's shape file as the standard for storing spatial information. It uses ESRI's MapObjects LT2 technology for spatial data access, mapping and querying. The analytical functionality consists of a set of C++ classes with associated methods. A technical review of the design and architecture of the software is detailed in Anselin et al. (2001, 2002a). Extensive background on the methodology of exploratory spatial data analysis, linking and brushing, and the specific techniques included in GeoDa and its predecessors can be found in Anselin (1994, 1995, 1996, 1998, 1999b).

This document serves both as a manual and as a brief tutorial for GeoDa. It assumes some familiarity with basic GIS concepts as well as some knowledge of basic statistical principles and elementary spatial statistics. A series of specialized tutorials on the use of GeoDa for exploratory data analysis, spatial correlation analysis, etc. is available from the GeoDa web site (<http://sal.agecon.uiuc.edu/csiss/geoda.html>). Further background on methodological issues can be found in, for example, Bailey and Gatrell (1995), or

¹ Lattice data are discrete spatial units that are not a sample from an underlying continuous surface (geostatistical data) or locations of events (point patterns). GeoDa currently does not yet contain specific techniques to analyze geostatistical or point pattern data.

Fotheringham et al. (2000), as well as in the sources mentioned above.

An extended set of course notes and examples dealing with an introduction to spatial data analysis can also be found in the repository of CSISS materials and Learning Resources at http://www.csiss.org/learning_resources/content/syllabi/#gis.

GeoDa was first available as a prototype in October 2001 (as DynESDA2). Its first public release was in February 2003. The version of GeoDa covered in this user's guide is 0.9.3 (June 4, 2003). A summary of changes from the first release (Version 0.9.0) is given in Appendix D. The software is available for free for non-commercial use. Please refer to the license agreement in Appendix A before installing the software. The software is provided as is and does not come with support other than what is contained in the web pages and provided informally through the Openspace mailing list.

Some minor problems with the software can be expected. Please report anything that seems like a bug to anselin@uiuc.edu, or, alternatively, post to the Openspace mailing list: <mailto:openspace@sal.agecon.uiuc.edu>.²

Note: GeoDa **replaces** the DynESDA Extension for ArcView 3.x. This extension is now deprecated and is no longer supported.

² To subscribe to the mailing list, check <http://sal.agecon.uiuc.edu/mailman/listinfo/openspace>.

Getting Started

Once you have installed GeoDa (see Appendix C) you can launch the program by double clicking on its icon or using any of the other standard MS Windows approaches. The opening screen is as in Figure 1. You start the analysis of a data set by clicking on the **Open Project** button on the toolbar (Figure 1) or by using the menu: **File > Open Project**. Note that in order to use the data manipulation **Tools**, it is not necessary to open a project, since they operate on separate files. Since the **Project** requires you to have data in a shape file, you can use the **Tools** before the actual analysis to construct the required shape files (for example, to create a point shape file from X, Y coordinates in an ascii file).

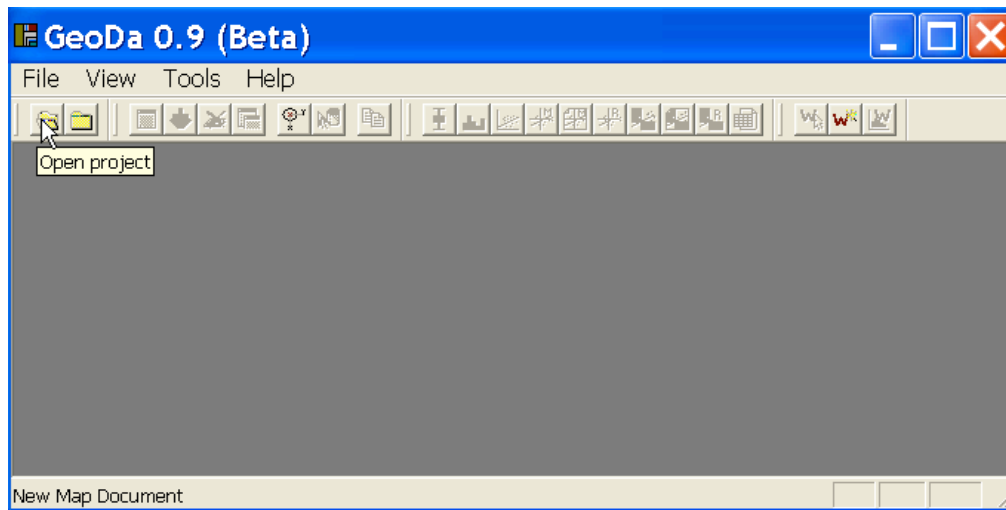


Figure 1. GeoDa opening screen.

Project Setting

The **Project Setting** dialog (Figure 2) requires you to enter the file name of a shape file that contains the data for your analysis. The shape file can be either a point or a polygon coverage. If your data is not in a shape file, you need to convert it first, either by using **Tools** (for point data) or by using ArcView, ArcGIS, or another GIS package that exports data as shape files. Think of this shape file as the spatial reference for all your analyses. You can have multiple geographical representations of the same data set, for example, as points, irregular polygons or Thiessen polygons, but they must all share a matching **Key Variable** (see below).

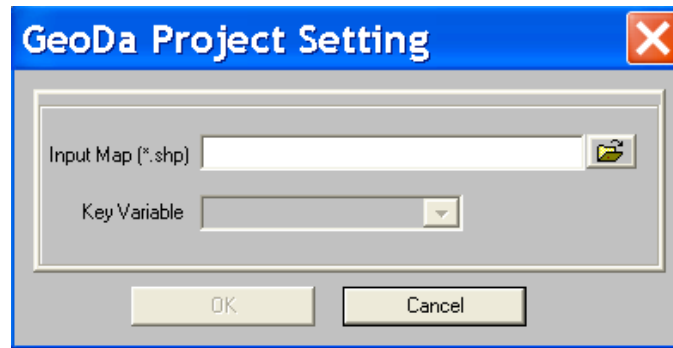


Figure 2. GeoDa Project Setting dialog

In the **Project Setting** dialog, a click on the open folder button will start the familiar Windows open file dialog (Figure 3).

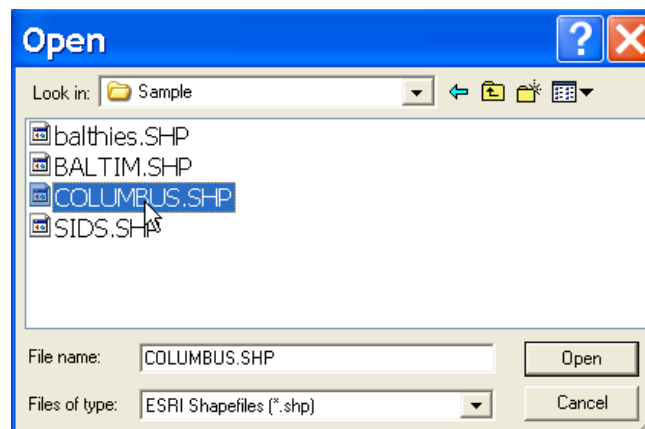


Figure 3. Open File dialog

Note that the files present in your working directory will likely differ from the ones shown in Figure 3. Using the standard Windows procedure, you should navigate to the directory that contains the shape file you want to use in the analysis. In Figure 3, that would be **COLUMBUS.SHP**. Click on **Open** to load the shape file.

Key Variable

After you select the shape file, you will need to specify a **key variable**. This variable must have a unique value for each observation. It is used to implement the linking between the maps and the different statistical graphs. GeoDa will try to guess a variable to be used as **key**. Accept the default or go down the drop down list to select the variable, as in Figure 4. If the one you picked does not have a unique value for each observation, an error message will appear. If none of the variables in your data set takes a unique

value for each observation, you must add one to the data table before you can use GeoDa. If necessary, use a spreadsheet package or ArcView to add a unique variable to the dbf table. In the Columbus example, the **Key Variable** is **POLYID**: click **OK** to accept the default, or, alternatively, click on the variable name, followed by **OK** to select it, or, double click on the variable name. After you select an appropriate **Key Variable**, a map is produced with the outline of the shape file, as in Figure 5 (for the Columbus neighborhood polygon data). At this point, the complete set of menu items and toolbar buttons becomes available as well.

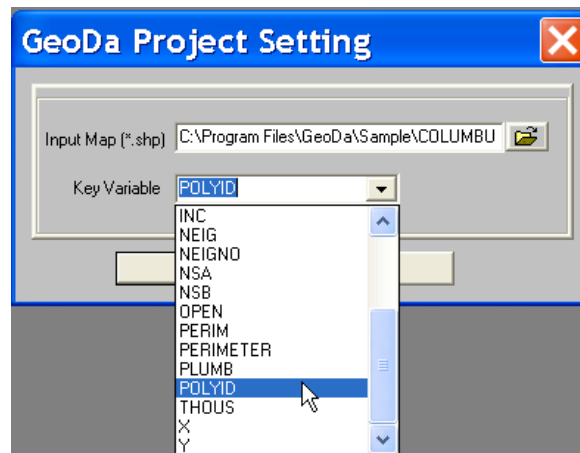


Figure 4. Key variable

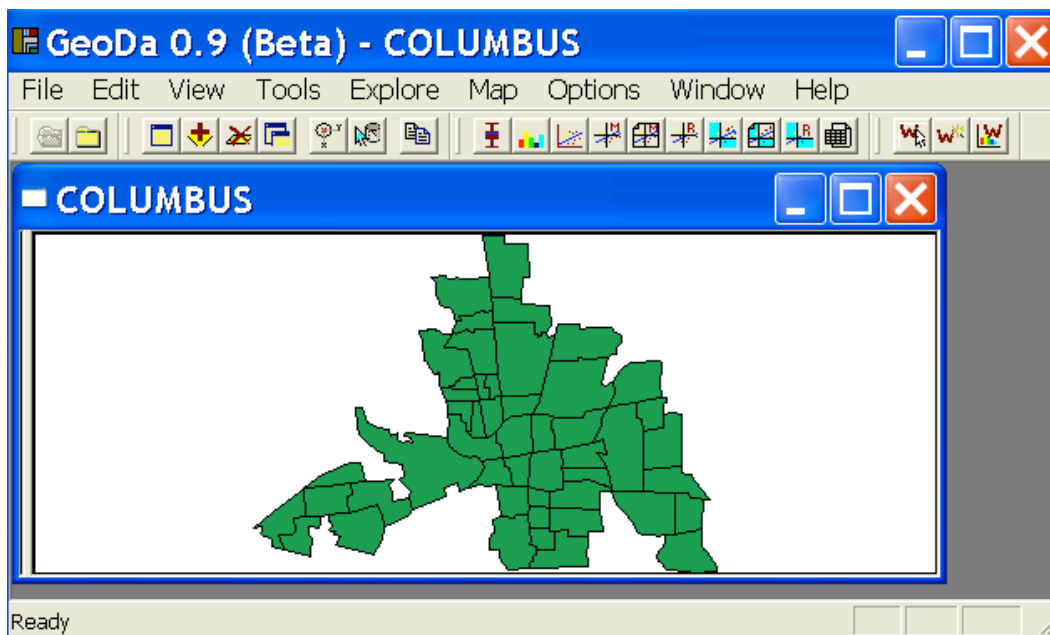


Figure 5. Opening screen with selected shape file outline.

Menu Overview

General Features

The GeoDa menu bar contains nine menu items (see Figure 6). Each of these deals with a particular functionality. The most important menu items are matched by a button on the toolbar (see Figure 6). The toolbar consists of four components that can be moved and docked independently anywhere within the program main window. To dock a toolbar, click and hold down the mouse on the raised vertical line to the left of the toolbar, and drag it to the desired location.

The four components are the **Project** toolbar (open a project, close all project windows), the **Edit** toolbar (duplicating maps, adding and removing layers, choosing variables, copying to clipboard), the **Explore** toolbar (EDA and ESDA techniques) and the **Weights** toolbar (creating and characterizing spatial weights). A detailed overview of each of the menus and their corresponding toolbar buttons follows below.



Figure 6. Menu and toolbars.

File Menu





Figure 7. File menu for new project and active project

The **File** menu contains the standard Windows functions to open a project, close the project windows and exit the program (Figure 7). Its look is different depending on whether there are active windows open. When there are no active windows in a project, the menu only contains two items. **File > Open Project** invokes the GeoDa Project

Setting Dialog (Figure 2), while **File > Exit** closes the program. When there is an active project, the menu contains four items: **close** to close the currently active window, **close All** to close all the windows in the project, **Export** to save the currently active window as a bitmap file, and **Exit** to end the program. Two buttons on the **Project** Toolbar correspond to this functionality as well: the open new project button (see also Figure 1) and the close all windows button.

Project Toolbar

-  Open project button
-  Close all windows button

Edit Menu

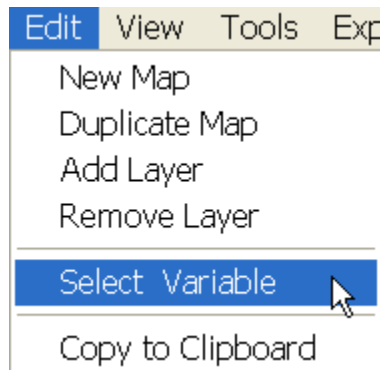


Figure 8. Edit menu.

The **Edit** menu contains three sets of functionalities, as illustrated in Figure 8. The first set deals with manipulating maps, the second set pertains to the selection of variables, with the last related to the use of the Windows clipboard.

Edit > New Map

This item opens the file dialog to select a shape file. The map will be added to a new window. The new shape file should refer to the same spatial layout as in the existing map windows. In other words, after loading a map with the Columbus neighborhoods, one could load a shape file with the Columbus Thiessen polygons that match these neighborhoods, or with the neighborhood centroids (a point file). These three shape files all must have a matching **Key Variable** (such as **POLYID**) for which the unique values must pertain to the same location in each of the shape files. GeoDa currently does not check whether this is the case, so it is possible to create nonsensical layouts! Figure 9

illustrates this functionality. **Edit > New Map** has been invoked twice, once for the **colvor** shape file (Thiessen polygons) and once for the **colpoints** shape file (centroids). Operations to create these new shape files from the original Columbus layout are illustrated below in the section on Manipulating Spatial Data. In Figure 9, the three maps are shown side by side (**Window > Tile Vertical**).

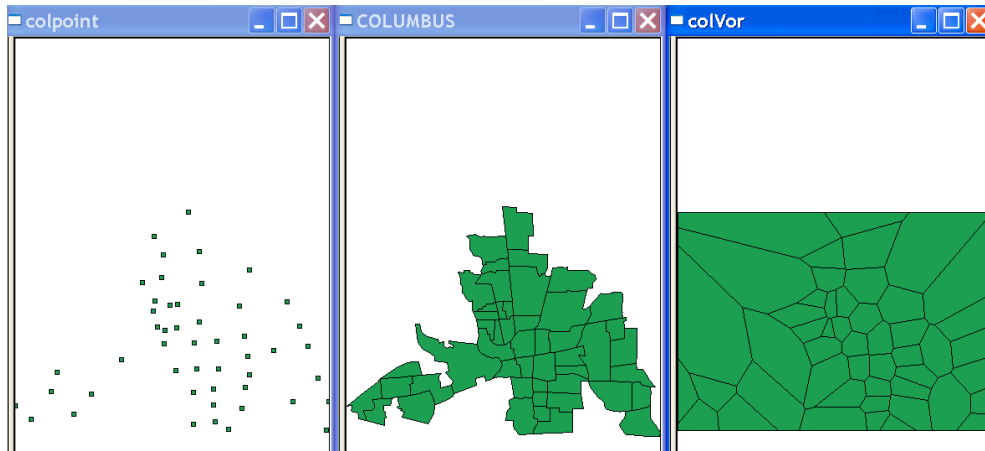


Figure 9. Multiple new maps.

Edit > Duplicate Map

Copies the layout in the active window to a new window. This creates an exact duplicate of the original in terms of the boundaries or point locations, but it does not retain the symbols or shading of the original map. This is useful when mapping different variables to compare the spatial patterns across variables. Figure 10 illustrates a duplicate map for the Columbus data, with the first layout as a choropleth map for crime. Note how the second window is only an outline (the maps are shown after **Window > Tile Vertical**).

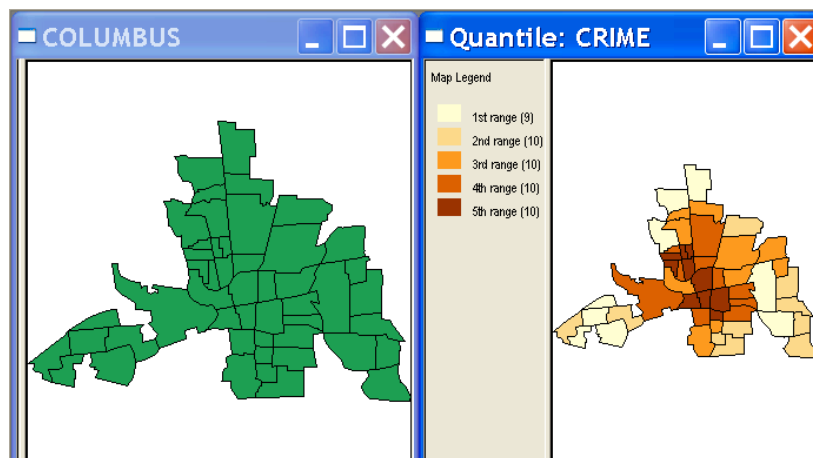


Figure 10. Duplicate maps.

Edit > Add Layer

Adds a layer to the currently active map window. The layer must conform to the same layout and the same **Key Variable** as the original map. As for **Edit > New Map**, this is the responsibility of the user, since GeoDa assumes that it is the case, but does not check for it. The last layer added becomes the active layer on the map. This is the layer to which selection, linking and brushing are applied, the lower layers are simply a backdrop.

This feature is useful when you want to add a point layer on top of a map showing areal units. For example, in Figure 11, the Columbus centroid points have been added on top of the original neighborhood layout and used to show a choropleth map for crime. Note how the neighborhoods are not colored, but only the points are. The colors match the ones for the map on the right hand side of Figure 10. The current layer functionality is limited: only the top layer is active and the order of the layers cannot be manipulated.

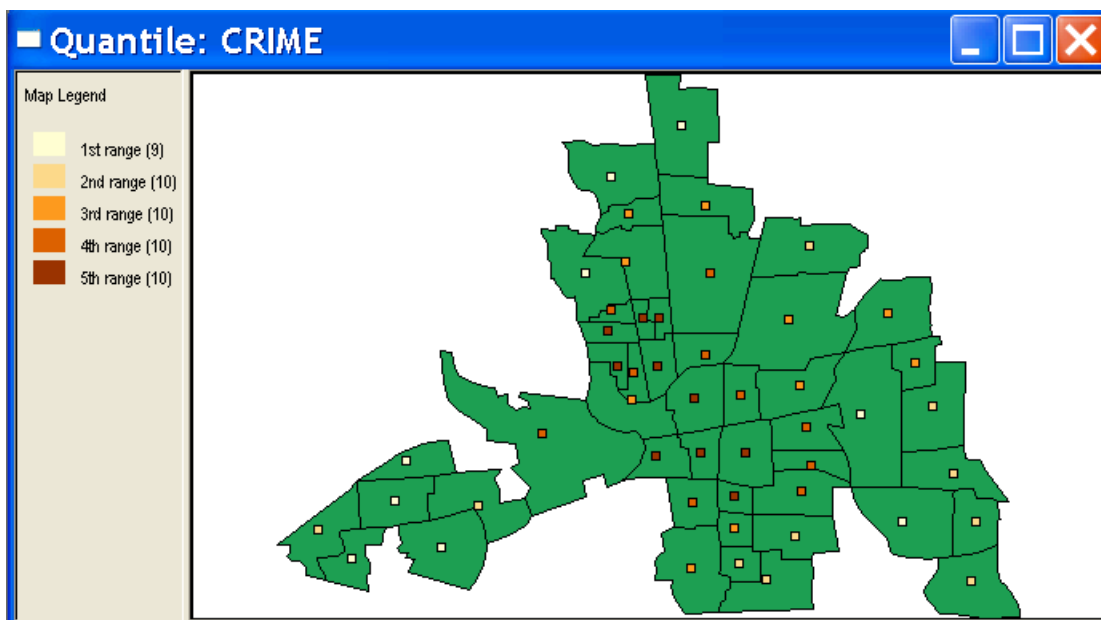


Figure 11. Point layer added to polygon layer (with choropleth map for points).

Edit > Remove Layer

This item removes the bottom layer in the active window. When the last layer is removed, the window is closed. Currently, there is a problem with this functionality, in that layers are removed bottom-up and not top-down (which is preferable).

Edit > Select Variable

Selects the variable(s) to be used for mapping and statistical analysis. This is particularly useful when setting a default variable. Once the **Set the variables as default** check box is checked, there will no longer be a dialog needed to select the variable in subsequent analyses, since all maps and statistical graphs will use the same default setting. The menu item invokes a variable selection dialog, as illustrated in Figure 12.

The dialog lists two columns, but only the left most (Y) will be used in univariate analyses. The check box is marked to set the selected variable as the default. If this check box is not marked, the variable selection dialog will open for each mapping or statistical operation.

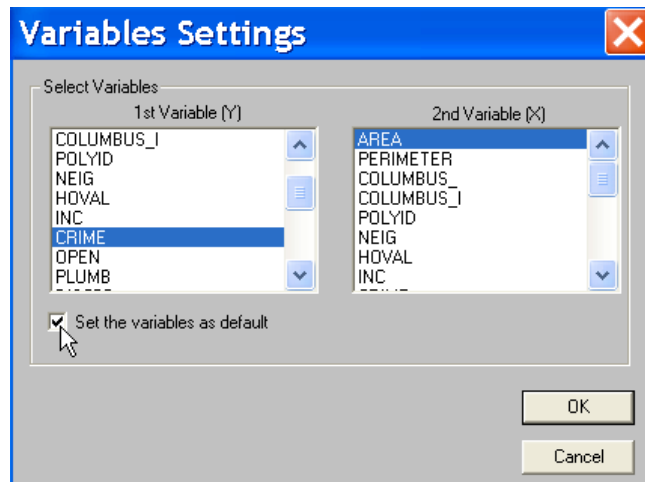





Figure 12. Variable selection.

Edit > Copy to Clipboard

Copies the contents of the current active window as a bitmap to the Windows Clipboard. The contents of the Clipboard can then be pasted into any other software package, such as a word processor or graphics package. Alternatively, any graph or map can also be saved to a bitmap file by using **File > Export** or by right clicking on the active window and selecting **Save Image As**.

Edit Toolbar

-  New map button
-  Add layer button
-  Remove layer button



- Duplicate map button
- Add centroids button (see Tools > Data Export > Centroids)
- Select variable button
- Copy to Clipboard button

View Menu

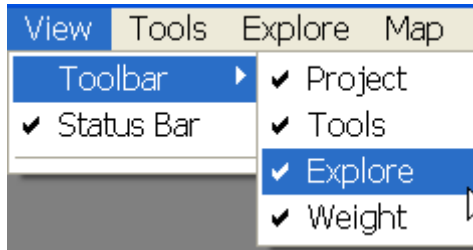


Figure 13. View menu.

The **view** menu (Figure 13) contains two options to set which items are shown in the program interface and toolbar. There are no buttons associated with these items.

View > Toolbar

Selects the toolbars that will be shown in the interface. The default is that all four toolbars will be shown: **Project** toolbar, **Tools** toolbar, **Explore** toolbar and **Weights** toolbar. Unchecking one of these items will remove them from the window toolbar.

View > Status Bar

Sets viewing of status information in the status bar on or off. The status bar is on the bottom and to the left of the program window. The default is that status messages will be shown there. Unchecking this item removes the status messages.

Tools Menu

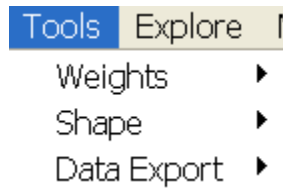


Figure 14. Tools menu

The **Tools** menu (Figure 14) contains four submenus to deal with the construction and analysis of spatial weights, the conversion and creation of point and polygon shape files, and data export.

Tools > Weights

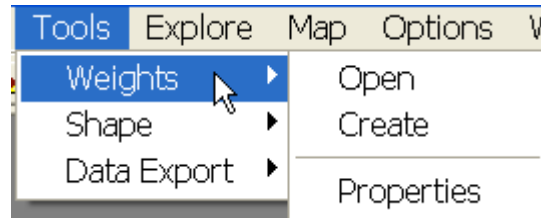


Figure 15. Tools Weights submenu.

The **Tools > Weights** submenu (Figure 15) contains three functions to **Open** a spatial weights file, **Create** a spatial weights file and analyze the **Properties** of the connectedness structure in a spatial weights files. These three functions also correspond to a button on the **Weights** toolbar. Details on their operation are given in the section on Creating and Manipulating Spatial Weights.

Weights Toolbar



Open Weights button



Create Weights button



Weights characteristics button

Tools > Shape

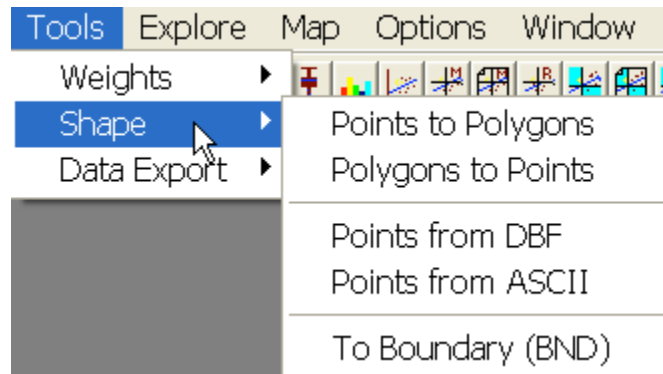


Figure 16. Tools Shape submenu.

The **Tools > Shape** submenu (Figure 16) contains the functionality to convert a point shape file to a Thiessen polygon shape file and to create a point shape file from the centroids of a polygon coverage. Both input and output files in this operation are in shape file format. In addition, this menu contains functions to convert point data from either a dbf file or from an ascii file to a point shape file. It also contains a feature to export the boundary file for a polygon shape to an ascii format (including the data on the bounding boxes for the polygons). Details are provided in the section on Manipulating Spatial Data. There are currently no toolbar buttons corresponding to this functionality.

Tools > Data Export

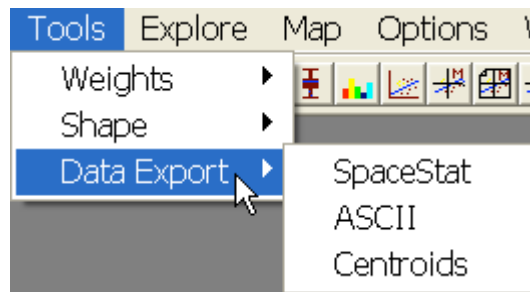


Figure 17. Tools Data Export submenu.

The **Tools > Data Export** submenu (Figure 17) includes three ways in which to export data from a shape file to other formats. The **SpaceStat** function converts selected variables from a shape file to the digital format used by SpaceStat; the **ASCII** function creates a comma delimited text file (csv) in a format useful for input in a variety of statistical software packages; and the **Centroids** function creates a **dbf** format file with the **x** and **y** coordinates of the polygon centroids. Note that this is different from the **Polygons to Points** function above, where the output file for the point coverage is in a shape file format (this includes all three files with extensions **shp**, **shx** and **dbf**, and not only the **dbf** file). Details are provided in the section on Manipulating Spatial Data.

The centroid export function has a matching toolbar button in the **Edit** toolbar.



Add centroids button

Explore Menu

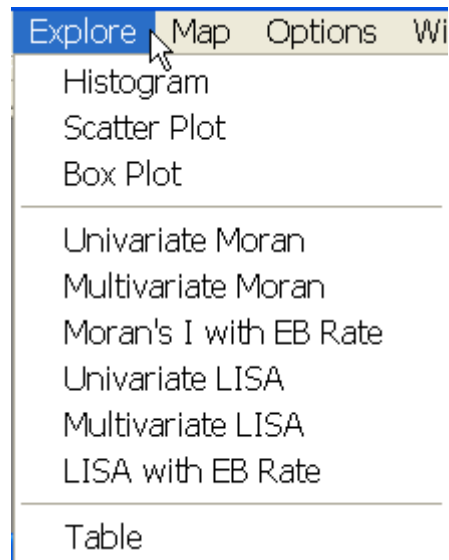


Figure 18. Explore menu











The **Explore** menu (Figure 18) contains the functionality for traditional statistical graphics used in exploratory data analysis (EDA) as well as for spatial autocorrelation analysis. In addition, the **Table** is included as an additional data “view” in the form of a simple list. All the graphs, maps and tables are linked, and all but the **Histogram** and **Table** can be brushed directly (see Linking and Brushing).

Each of these functions is invoked in the same manner. After selecting the menu item, the **Variable Settings** dialog appears, unless a variable (or two variables for the scatter plots) was earlier set as the default (**Edit > Select Variable**). The spatial statistics (**Moran** and **LISA**) also require that a spatial weights file is selected (**Tools > Weights > Open** or **Tools > Weights > Create**). After specifying these elements, a window opens with the statistical graph or map. The **Options** menu is context sensitive and contains ways to customize the graphs, such as altering the number of categories in a **Histogram** or changing the **Hinge** for a **Box Plot** (see the discussion of the Options Menu for each specific application). Specifics for the various techniques are covered in the sections on Statistical Graphs, Global Spatial Autocorrelation and Local Spatial Autocorrelation.

The **Table** item in the **Explore** menu opens a window with a table listing the variables as columns and the observations as rows. In addition to an explicit invocation by means of the **Explore > Table** menu or the **Table** toolbar button, the data table is automatically

opened for any mapping or exploratory operation. Detailed functionality is discussed in the section on Editing and Manipulating Tables.

Explore Toolbar

-  Box Plot button
-  Histogram button
-  Scatterplot button
-  Univariate Moran button
-  EB Moran button
-  Multivariate Moran button
-  Univariate LISA button
-  Multivariate LISA button
-  EB LISA button
-  Table button

Map Menu

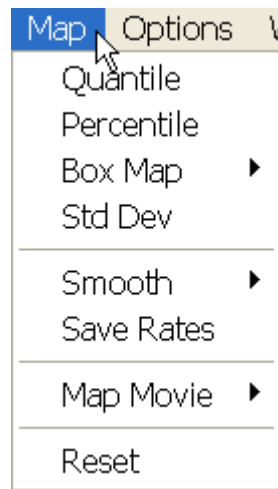


Figure 19. Map menu

The **Map** menu (Figure 19) contains the functionality to implement choropleth mapping. The first four items correspond to standard map types: **Quantile** map, **Percentile** map, **Box Map** and **Std Dev** (standard deviational) map. Each of these requires that a variable be set. Details are provided in the section on Mapping. The **Smooth** and **Map Movie** submenus implement specialized mapping functions. The **Smooth** group also includes an option to save the rates computed in the smoothing procedure (**Save Rates**). The **Reset**

item clears the map and returns the window to a simple outline. There are no toolbar buttons matching the entries in the **Map** menu.

Map > Box Map



Figure 20. Map Box Map Submenu

The **Box Map** item on the **Map** menu contains two options for the fences (**Hinge**) used to identify outliers (Figure 20). These need to be specified before the box map is computed. See the section on Mapping.

Map > Smooth

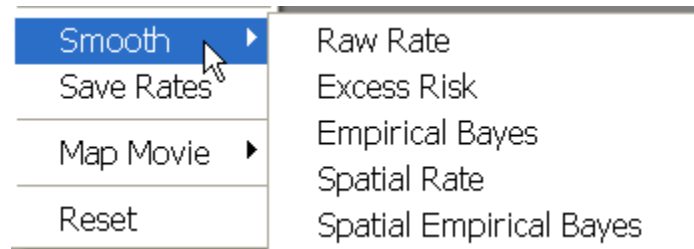


Figure 21. Map Smooth submenu.

The **Map > Smooth** submenu (Figure 21) contains five methods to produce choropleth maps for variables that are expressed as rates. Technical details are provided in the section on Smoothing Rate Maps.

Map > Map Movie

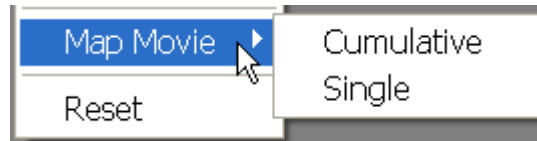


Figure 22. Map Movie submenu.

The **Map > Map Movie** submenu (Figure 22) allows a choice between two ways in which the locations are highlighted in the “movie.” **Cumulative** gradually fills out the complete map, starting with the location with the lowest value for the selected variable and adding new highlighted locations as the values increase. At the end of the movie, the complete map is filled out with the highlight. In contrast, the **single** option flashes each location in turn, going from lowest to highest. The speed by which this happens depends on machine hardware and is not predictable in the current version. It is therefore recommended to use the **Cumulative** option. More details on the **Map Movie** are provided in the section on Mapping.

Options Menu

The **Options** menu controls several settings for specific statistical graphs and statistical analyses. Only those options are shown in the menu list that are relevant for the graphs present in the program window at the time. The look of the menu is therefore slightly differently in each case. The **Options** menu items will be discussed with the techniques to which they refer. The **Options** can also be invoked by right clicking on the active graph or map.

Window Menu

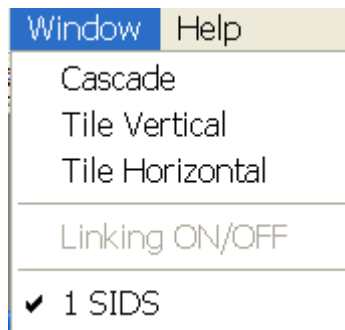


Figure 23. Window menu

The **Window** menu (Figure 23) lists all the open windows (in Figure 23, this would only be the **SIDS** map) and provides some means to rearrange the windows: **Cascade**, **Tile Vertical** or **Tile Horizontal**. **Cascade** stacks the windows, **Tile Vertical** puts them side by side and **Tile Horizontal** arranges them vertically (Note that in the current version of GeoDa, this is the opposite of the usual convention). The **Linking on/off** switch is currently always on and cannot be changed (it is grayed out in the menu). There are no toolbar buttons corresponding to the **Window** menu.

Help Menu



Figure 24. Help menu.

The **Help** menu (Figure 24) is currently not active and only contains the usual **About** item with copyright notices and credits. Clicking on this yields an information box which lists the current version of GeoDa (Figure 25). There is no built in help system in the current release of GeoDa.

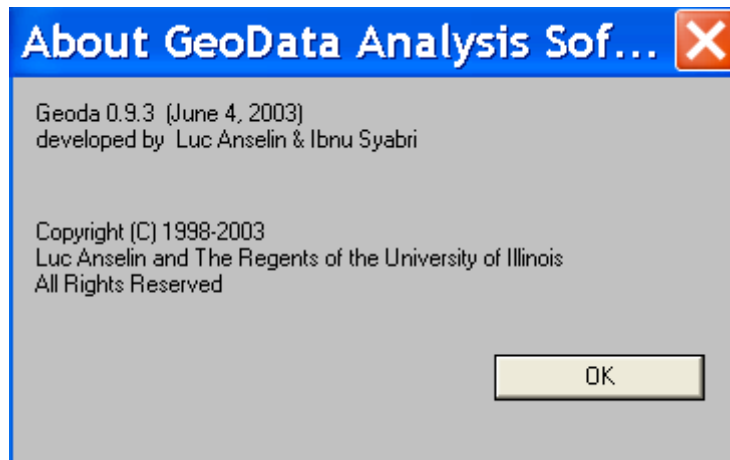


Figure 25. About GeoDa

Manipulating Spatial Data

GeoDa is geared to the analysis of lattice data. Observations are represented as spatial objects, either as points (with X, Y coordinates) or as polygons (with the X, Y coordinates of the boundary outline). Which representation is most useful depends on the analysis context. For example, for the calculation of spatial weights based on a common boundary, a polygon representation is necessary. In contrast, when the spatial weights are derived from a distance criterion, a point representation is needed. The **Tools > Shape** submenu contains the functionality to switch between point and polygon representations for the same data set, and to import point coordinate information from ascii and dBase format data sets. In addition, it includes a function to export the boundary file information contained in a polygon shape file to several common ascii formats. The **Tools > Data Export** submenu allows selected variables from the current project to be exported in formats suitable for use as input into other statistical software.

Creating a Point Shape File with Centroids

This function converts the contents of a polygon shape file to a matching point shape file, where the points are the centroids of the polygons. You have to specify the input shape file and a file name for the output shape file (which will be a point shape file). Specifically, after selecting **Tools > Shape > Polygons to Points**, a dialog opens in which the file name for the **Input file** (polygon shape file) and the **Output file** (point shape file) must be specified. The shape conversion dialog looks as in Figure 26.

Click on the open folder icon to select the **Input file** (simply typing in the name of the file will *not* work). Select the input shape file (for example, **SIDS**) in the open file dialog and click **Open** (Figure 27). The file name will now appear in the **Input file** text box and a thumbnail of the polygon shape file will appear in the space below (Figure 28). You select the output shape file using a similar procedure. You need to click on the file save icon in the dialog (simply typing in a file name will *not* work), and enter the file name in the file name text box of the **Save As** file dialog. Click **Save** to confirm the choice (Figure 29). The file name of the output file will now appear in the text box of the **Shape Conversion** dialog. Click on **Create** (now active) to carry out the conversion (Figure 30). A thumbnail of the point shape file will appear in the space below the output file text box. Once the progress bar at the bottom of the dialog (blue line in Figure 31) is completely filled out, click **Done** to leave the procedure.

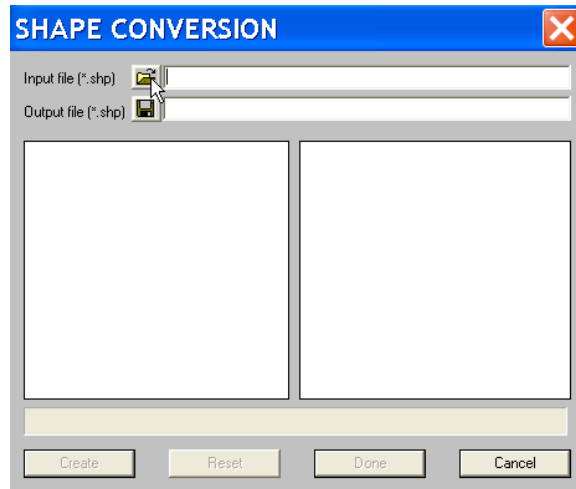


Figure 26. Shape conversion dialog

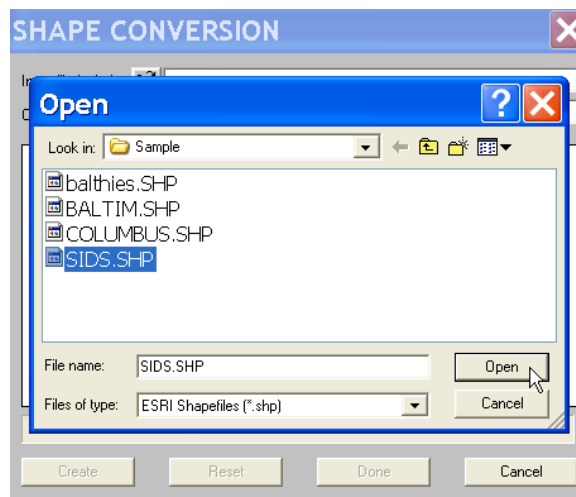


Figure 27. Shape conversion input file selection.

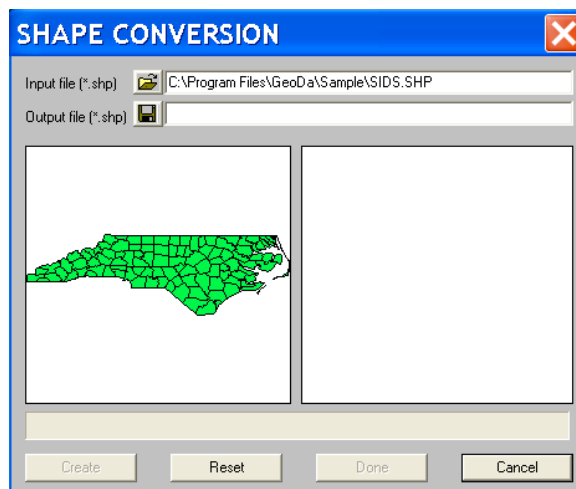


Figure 28. Shape conversion input file thumbnail.

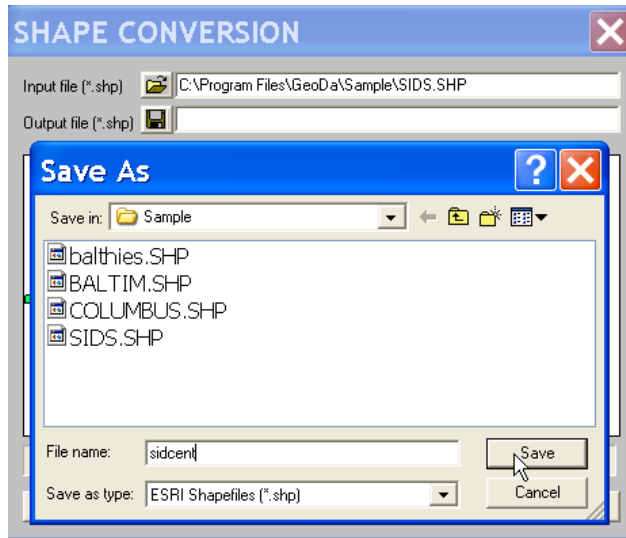


Figure 29. Shape conversion output save as file dialog.

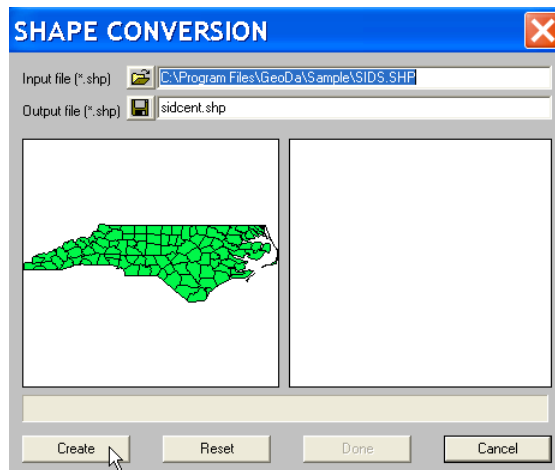


Figure 30. Shape conversion create button.

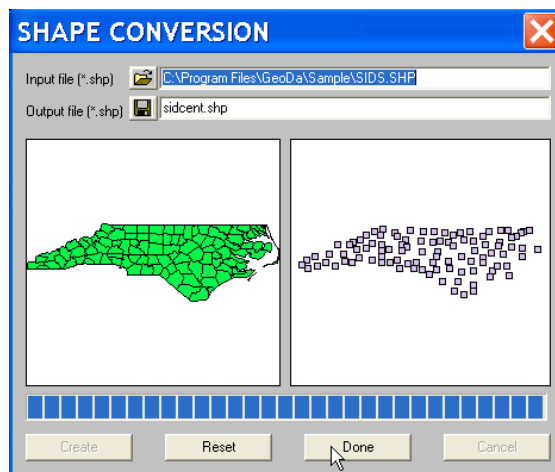


Figure 31. Completed Polygon to Point conversion.

If you check the contents of the working directory, you will notice three new files: **sidcent.shp**, **sidcent.shx** and **sidcent.dbf**. You can now add the point map to your project, for example, using **Edit > New Map**, or by clicking on the **New Map** button on the toolbar. This is illustrated in Figure 32 for the North Carolina county **SIDS** data.

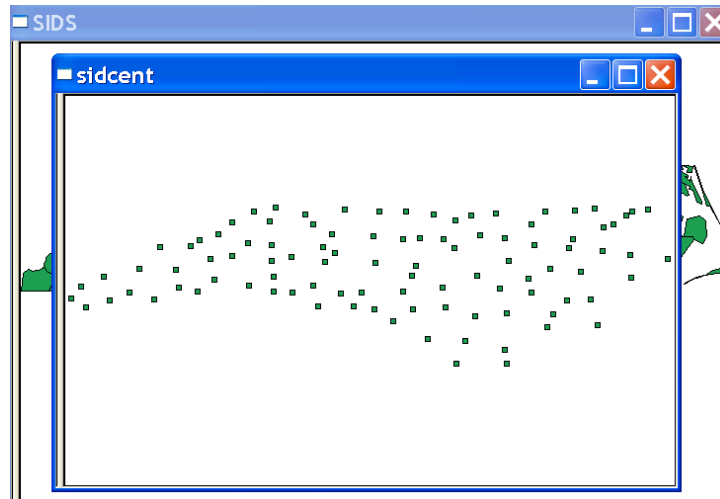


Figure 32. Centroids for North Carolina counties.

The data table for the new point shape file contains all the variables from the original polygon shape file, including **AREA** and **PERIMETER** (which of course don't make sense for points). You make the table visible by clicking on the **Table** toolbar button or using **Explore > Table** from the menu. The result is as shown in Figure 33. You can delete the two variables using the **Table** edit functions (right click and **Delete Column**). Note also how the table does not show an explicit field for the **x** and **y** coordinates for the centroids. Since this is a point *shape file*, the point coordinates are invisible (just as the boundary coordinates are invisible in a polygon shape file). However, you can add the points to the table separately (see next section on Adding Centroids to the Data Table).

Table : sidcent					
	AREA	PERIMETER	CNTY_	CNTY_ID	NAME
1	0.114000	1.442000	1825	1825	Ashe
2	0.061000	1.231000	1827	1827	Alleghany
3	0.143000	1.630000	1828	1828	Surry
4	0.070000	2.968000	1831	1831	Currituck

Figure 33. Data table for centroids point shape file.

Adding Centroids to the Data Table

In many instances, it is not necessary to create a new shape file for the centroid points, but the main interest is in adding x and y coordinates to the current data table. These coordinates can then be used in distance computations to construct spatial weights. In GeoDa, this is accomplished as an **option** in any map view (whether for polygons or for points). When a map view is active, the context sensitive **options** menu contains **Options > Add Centroids to Table** (Figure 34). Alternatively, right clicking on the map invokes an **options** menu with the same items. For example, in Figure 35, this is shown for the **sidcent** point shape file. Similar functionality is present for maps constructed from a polygon shape file.

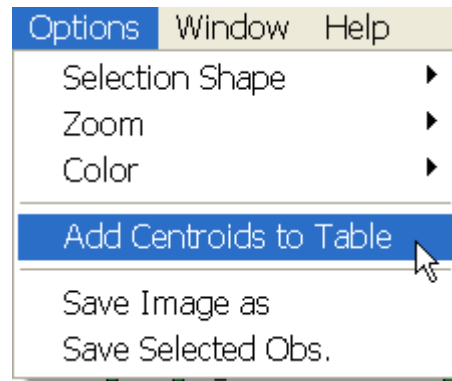


Figure 34. Add Centroids to Table option in Menu.

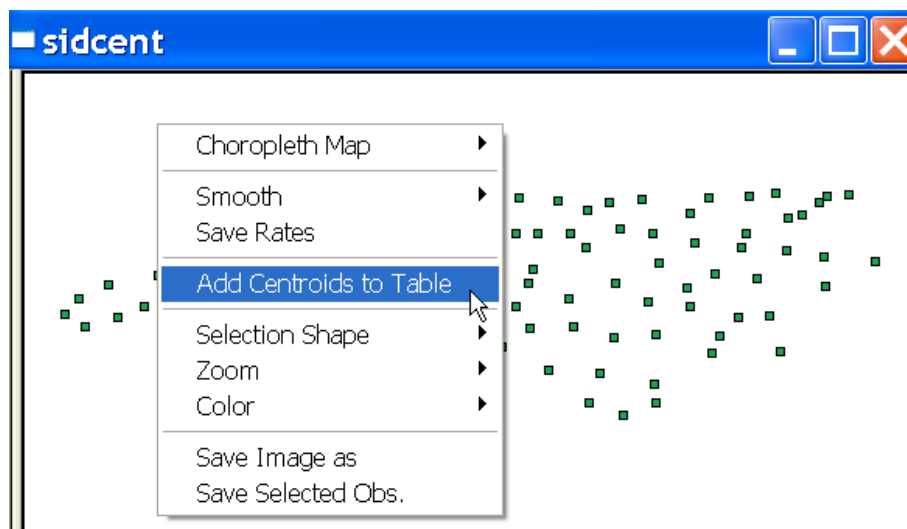


Figure 35. Add Centroids to Table from Map.

Invoking this item brings up a dialog to specify the variable names for the coordinates (Figure 36). Click on the check boxes and either choose the default or enter your own choice for variable names for the coordinates. Click **OK** to add the new variables to the table. In Figure 37, the last two columns of the new `sidcent` table are shown (the original table is illustrated in Figure 33), with the coordinates added as new variables. Note that the centroid coordinates only become permanently included in the data table after the table is *saved* to a shape file (see the section on Editing and Manipulating Tables).

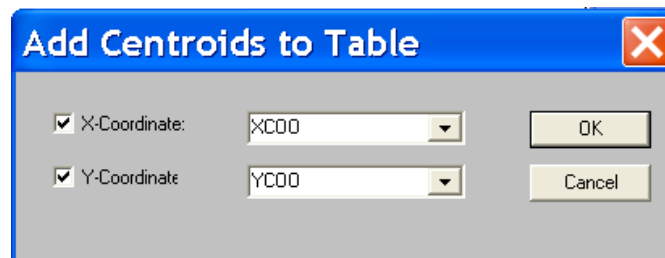


Figure 36. Add Centroids to Table dialog.

	XCOO	YCOO
1	-81.428840	36.372571
2	-81.123626	36.434344
3	-80.713185	36.337643
4	-75.984813	36.401059

Figure 37. Columns with centroids added to the `sidcent` data table.

Creating a dBase File with Centroids

A slightly different perspective is taken when the **Add Centroids** toolbar button is invoked or by using **Tools > Data Export > Centroids**. This does not create a new shape file, but only a data base file (dBase format) that contains a **key** variable and the **x** and **y** coordinates, but no other attributes. This is useful when the coordinates need to be joined with a data base file in a statistical package, and a complete shape file is not required. You can always load this dbf file back into GeoDa by converting it first to a point shape file using **Tools > Shape > Points from DBF**. With the table functionality, you can then use the **key** variable to join the point shape file (which will

not have any other attributes) to another dBase format file with attributes. However, this will usually not be the way you want to proceed (creating a centroid point shape file directly avoids the extra join operation and is more straightforward).

After clicking on the **Add Centroids** toolbar button, or selecting **Tools > Data Export > Centroids**, a file dialog opens as in Figure 38. You need to click on the folder buttons and enter the file names for both input shape file and output (dbf) files. You also must select a **Key Variable**. This will be included as a field in addition to the **x** and **y** coordinates, so that the resulting dbf file can be easily joined with other data sets for the same observations. An additional field contains a simple sequence number (**RECNUM**), which is used as a key in some statistical packages. For the **SIDS** data, the appropriate **Key Variable** is **FIPSNO**. Clicking the **Create** button will then compute the centroids and write the output file.

You will note the new **sidcent_2.dbf** file in your working directory. If you open it with a spreadsheet package like Excel, you will see the four data columns, as in Figure 39.

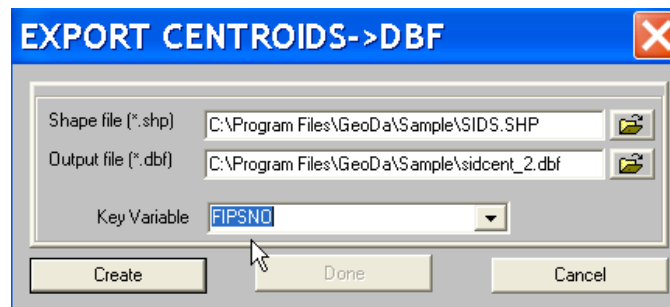


Figure 38. Export Centroids file dialog.

	A	B	C	D	E
1	RECNUM	FIPSNO	X_COORD	Y_COORD	
2	1	37009	-81.428800	36.372600	
3	2	37005	-81.123600	36.434300	
4	3	37171	-80.713200	36.337600	
5	4	37053	-75.984800	36.401100	
6	5	37131	-77.410000	36.372700	
7	6	37091	-76.993300	36.391300	
8	7	37029	-76.205000	36.379400	

Figure 39. Centroids dbf file loaded into spreadsheet.

Creating Thiessen Polygons as Shape Files

Thiessen polygons are created as a polygon shape file derived from a point shape file. Each Thiessen polygon encloses the original points in such a way that all the points in a polygon are closer to the enclosed point than to any other point. This corresponds to the notion of geographic market area from economic geography. Thiessen polygons are most useful for the visualization of variables when the point patterns are hard to distinguish. In addition, they allow the computation of contiguity based spatial weights for point data, using the boundaries of the polygons to establish contiguity.

The Thiessen polygon procedure is started by **Tools > Shape > Points to Polygons**. The interface is similar to that for the **Tools > Shape > Polygons to Points** procedure (see Figures 26-31). A **Shape Conversion** dialog is opened, and you need to specify the **Input** file (for example, **BALTIM.SHP**) and the **Output** file (for example, **balthi.shp**). The file names are specified by clicking on the folder button for the input file and entering the file name in the **Open file** dialog (followed by clicking **Open**). Similarly, you need to click on the save file icon to select the output file, followed by entering the file name in the **Save As** file dialog (finish by clicking **Save**). At this point, the **Shape Conversion** dialog will show the name of the **Input file**, a thumbnail for the point pattern associated with that file name, and the name of the **Output file**, as in Figure 40.

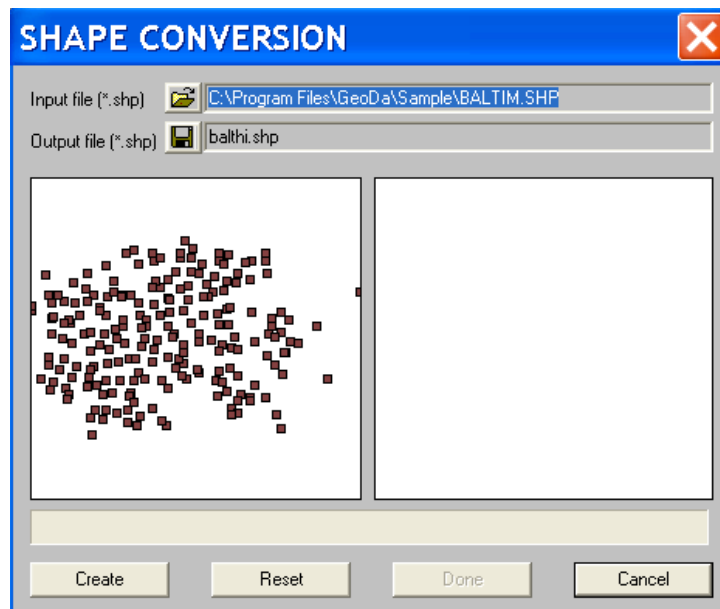


Figure 40. Point to Polygon shape conversion dialog.

Clicking on **Create** will start the computation process. When the blue progress bar at the bottom of the dialog has reached the end, a thumbnail for the Thiessen polygons will appear below the **Output file** text box (Figure 41). Click **Done** to end the procedure. There will now be three new files in the working directory: **balthi.shp**, **balthi.shx** and **balthi.dbf**. The polygon shape file can be added to the current project by using **Edit > New Map**. Overlay the original points with **Edit > Add Layer**, as in Figure 42.

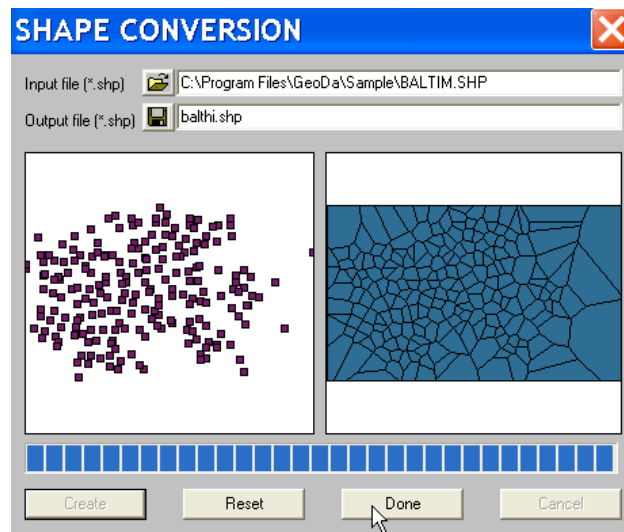


Figure 41. Completed Point to Polygon shape conversion.

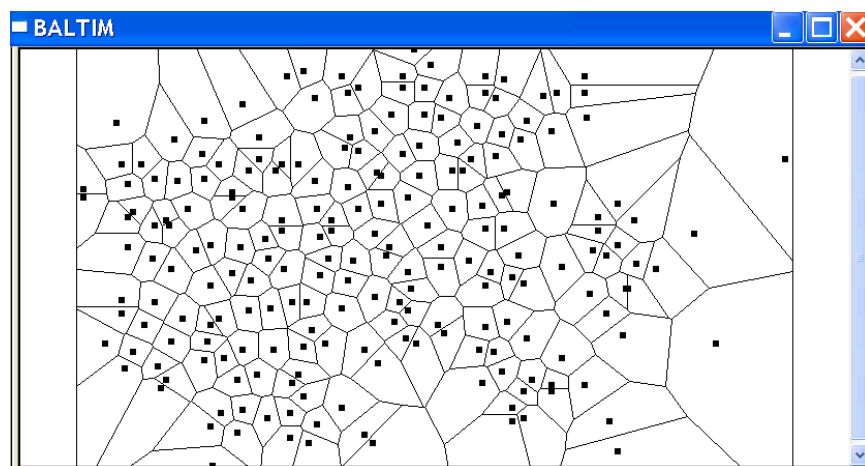


Figure 42. Thiessen polygons for Baltimore point data, with point data overlaid.

The data table for the Thiessen polygon shape file contains **AREA** and **PERIMETER** variables computed for the new polygons, as well as all the attributes contained in the original point shape file, as illustrated in Figure 43. Note that the **AREA** and **PERIMETER**

calculations are currently only supported for *projected coordinates* (Euclidean distance). For point shape files in unprojected Latitude and Longitude, the results will *not* be correct. Also, the coordinates of the points themselves will not be part of this table by default, but only if they had been added to the original point shape file explicitly (by means of the **Add Centroids to Table** option).

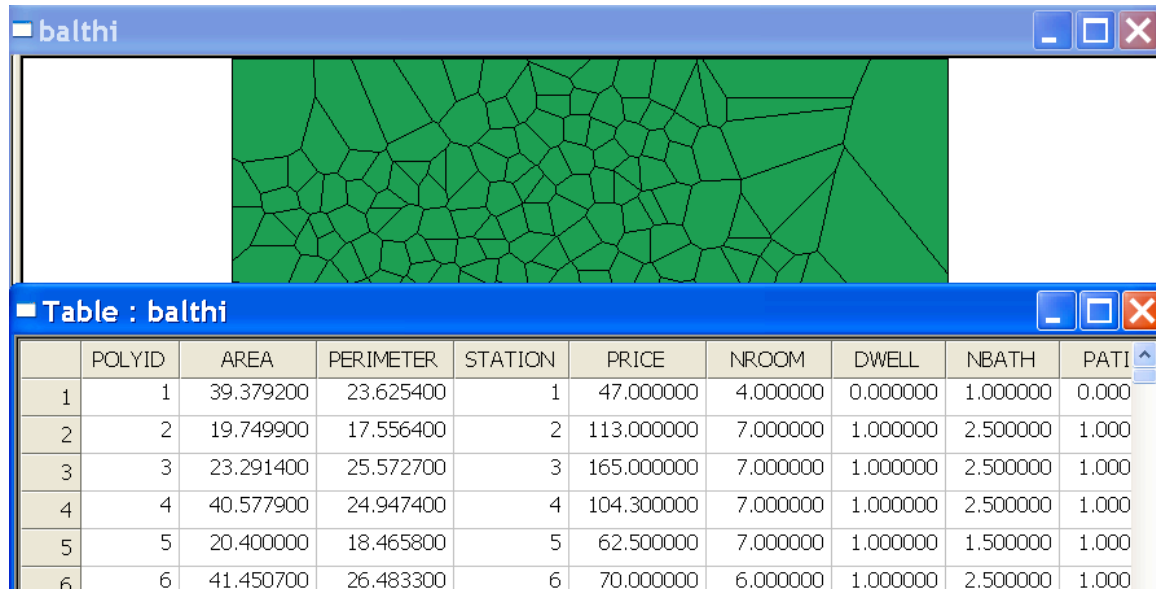


Figure 43. Data table for Thiessen polygon shape file.

Creating Point Shape Files from DBF and ASCII Input Files

While GeoDa requires project input files to be in the shape file format, the **Tools** functions contain means to construct point shape files from input files that contain the **x** and **y** coordinates as fields. **Tools > Shape > Points from DBF** takes a dBase file as input, while **Tools > Shape > Points from ASCII** reads a comma delimited ascii file. The latter follows a specific format, starting with a header line with the number of observations and the number of variables. The second line contains the variable names, separated by a comma. Next follow the actual data, with each row corresponding to an observation, and the values separated by a comma. Figure 44 illustrates the layout of the comma delimited **baltim.csv** file corresponding to the **BALTIM.shp** sample point data set. This file was constructed by exporting the dbf file as **csv** in the Excel spreadsheet and adding the header line in any text editor. Note how the file in Figure 44 contains the **x** and **y** coordinates as variables.

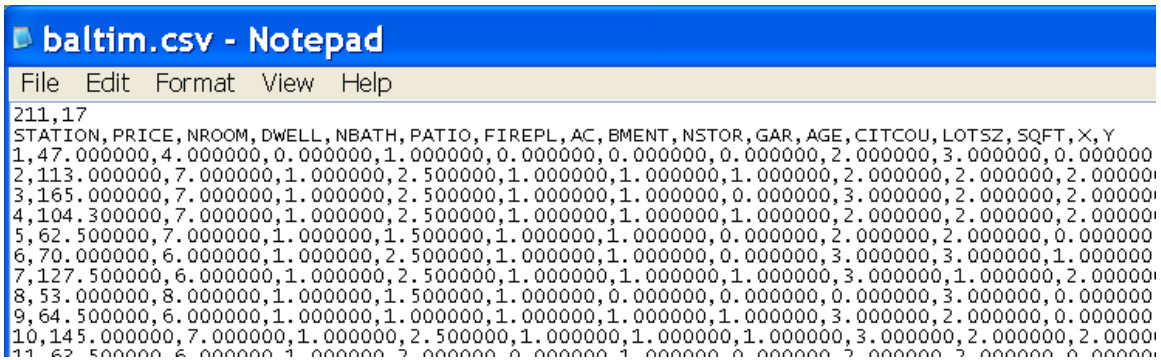


Figure 44. Baltimore point data as a comma delimited (csv) file.

After invoking **Tools > Shape > Points from DBF**, a dialog appears in which the input file name and output file name need to be entered. For example, in Figure 45, the **BALTIM.dbf** file (from the Baltimore shape file trio) is used and **baltimpoint.shp** is specified as the output file. You must click on the icon and enter the file name for both input and output files. The **x-coord** and **y-coord** drop down lists show the variables contained in the input file. Select the variable names for the coordinates and click the **create** button. The shape file will be saved to the working directory. You can next open it in GeoDa. As shown in Figure 46, the points and data table are identical to the ones for the Baltimore sample point shape file.

The procedure for **Tools > Shape > Points from ASCII** is similar. Again, you need to specify the input and output file name in a dialog (Figure 47), as well as the variable names for the **x** and **y** coordinates. If the input ascii file does not correspond to the comma delimited format or does not have the proper header line, an error message appears (you can use any text editor to make sure the header line is in the proper format).

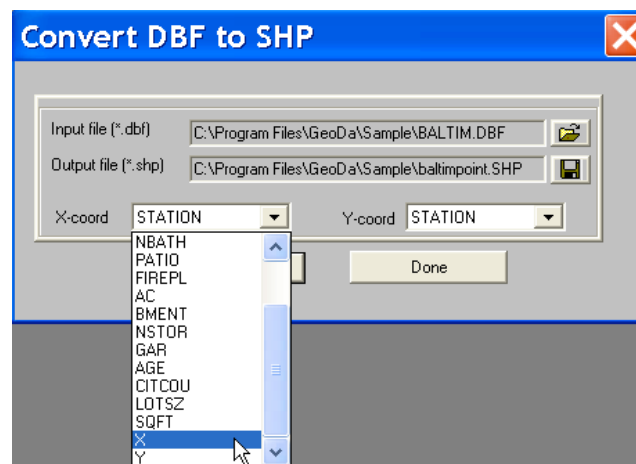


Figure 45. Points from DBF file and variable selection dialog.

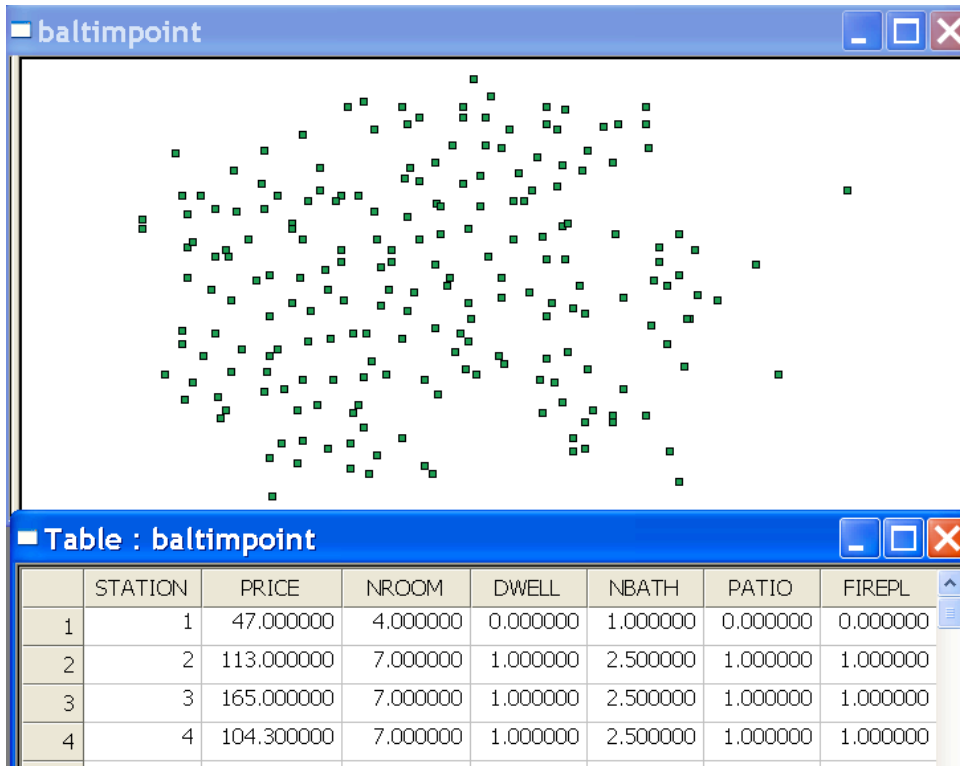


Figure 46. Baltimpoint point shape file created from dbf input.

In Figure 47, the input file is `baltim.csv` and the output file is set as `baltimpoint2.shp`. Again, `x` and `y` are the coordinates. Click on the `Create` button to save the point shape file to the working directory. As before, the point shape file can now be opened in GeoDa. Figure 48 illustrates how the point pattern and the data table are identical to the other incarnations of the Baltimore data.

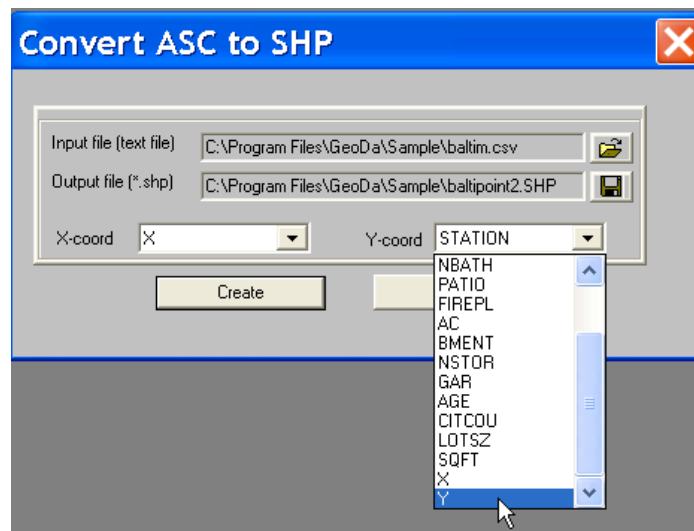


Figure 47. Points from ASCII file and variable selection dialog.

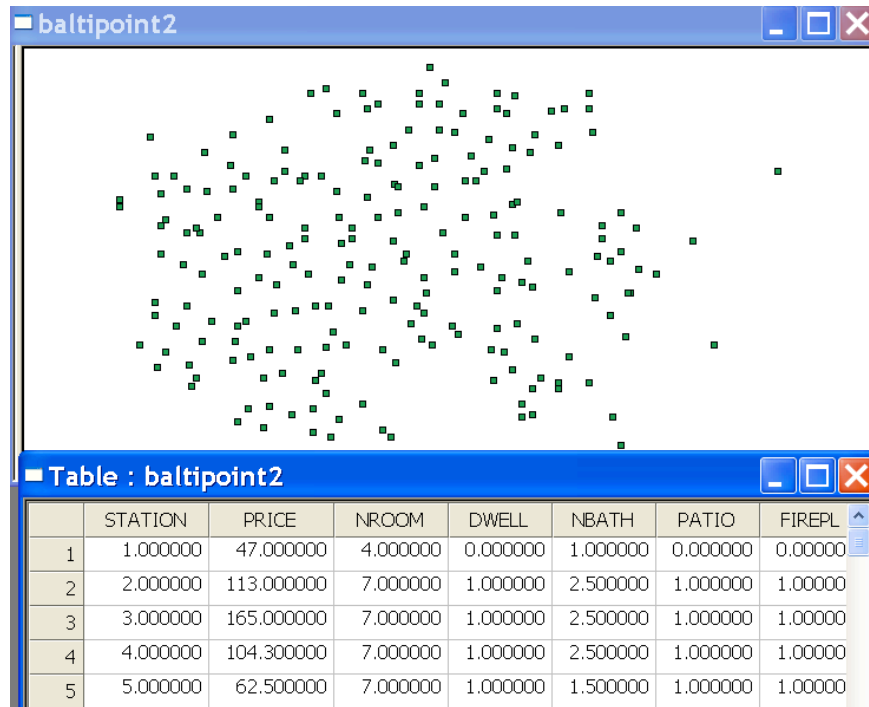


Figure 48. Baltipoint2 point shape file created from ascii input.

Exporting Boundary Files

The **Tools > Shape > To Boundary (BND)** command allows you to export the coordinates of the polygon boundaries in a shape file to an ascii output file. Four commonly used formats are supported, allowing the input of these boundary files in a variety of mapping and graphics software packages. Type 1 lists the polygon **ID**, the number of points in the polygon boundary, followed by a list of **x, y** coordinates, as in the left panel of Figure 49. Type 1a (right hand panel of Figure 49) is identical to Type 1, except that a header line is included with the number of polygons and the name of the variable used as **key**. Type 2 uses a common convention to cycle the coordinates in the list, in the sense that for each polygon the last set of **x, y** values is identical to the first. Only the polygon **ID** is given (not the number of points in the polygon boundary). For example, in the left panel of Figure 50, the values pair **8.62413, 14.237** is listed twice for polygon 1. Type 2a is the same as Type 2, except for the inclusion of a header line with the number of polygons and the **key** variable. In addition to the boundary, it is also possible to create an ascii file with the coordinates of the bounding box for each polygon. This is used to derive contiguity information in the **R spdep** package. The file contains an **ID** for each polygon, and the **x, y** coordinates for the lower left and upper right corners, as in Figure 51.

```

columbus1.bnd - Notepad
File Edit Format View Help
1,14
8.62413,14.237
8.5597,14.7424
8.80945,14.7344
8.80841,14.6365
8.9193,14.6385
9.08714,14.6305
9.09997,14.2448
9.01505,14.2418
9.00895,13.9951
8.81814,14.0021
8.65331,14.0081
8.6429,14.0897
8.63259,14.1706
8.62583,14.2237
2,46
8.25279,14.2369
8.28276,14.2299
8.33071,14.2299
8.38366,14.2289
8.4446,14.2289
8.5445,14.2349
8.62413,14.237
8.62583,14.2237
8.63259,14.1706
8.6429,14.0897
8.65331,14.0081

```

```

columbus1a.bnd - Notepad
File Edit Format View Help
49,POLYID
1,14
8.62413,14.237
8.5597,14.7424
8.80945,14.7344
8.80841,14.6365
8.9193,14.6385
9.08714,14.6305
9.09997,14.2448
9.01505,14.2418
9.00895,13.9951
8.81814,14.0021
8.65331,14.0081
8.6429,14.0897
8.63259,14.1706
8.62583,14.2237
2,46
8.25279,14.2369
8.28276,14.2299
8.33071,14.2299
8.38366,14.2289
8.4446,14.2289
8.5445,14.2349
8.62413,14.237
8.62583,14.2237
8.63259,14.1706
8.6429,14.0897

```

Figure 49. Boundary file format 1 (left) and format 1a (right).

```

columbus2.bnd - Notepad
File Edit Format View Help
1
8.62413,14.237
8.5597,14.7424
8.80945,14.7344
8.80841,14.6365
8.9193,14.6385
9.08714,14.6305
9.09997,14.2448
9.01505,14.2418
9.00895,13.9951
8.81814,14.0021
8.65331,14.0081
8.6429,14.0897
8.63259,14.1706
8.62583,14.2237
8.62413,14.237
2
8.25279,14.2369
8.28276,14.2299
8.33071,14.2299
8.38366,14.2289
8.4446,14.2289
8.5445,14.2349
8.62413,14.237
8.62583,14.2237
8.63259,14.1706

```

```

columbus2a.bnd - Notepad
File Edit Format View Help
49,POLYID
1
8.62413,14.237
8.5597,14.7424
8.80945,14.7344
8.80841,14.6365
8.9193,14.6385
9.08714,14.6305
9.09997,14.2448
9.01505,14.2418
9.00895,13.9951
8.81814,14.0021
8.65331,14.0081
8.6429,14.0897
8.63259,14.1706
8.62583,14.2237
8.62413,14.237
2
8.25279,14.2369
8.28276,14.2299
8.33071,14.2299
8.38366,14.2289
8.4446,14.2289
8.5445,14.2349
8.62413,14.237
8.62583,14.2237
8.63259,14.1706
8.6429,14.0897
8.65331,14.0081

```

Figure 50. Boundary file format 2 (left) and format 2a (right).

```

columbus1_r.bnd - Notepad
File Edit Format View Help
1,8.5597,13.9951,9.09997,14.7424
2,7.95009,13.7274,8.66655,14.2639
3,8.65331,13.5444,9.35149,14.0081
4,8.1986,13.5865,8.68527,13.8617
5,8.67758,12.8611,9.40138,13.7222
6,9.3333,13.2724,10.1806,13.6982
7,7.80197,12.942,8.45657,13.6445
8,8.10498,13.1041,8.73397,13.6444
9,9.12428,12.5952,10.0954,13.2985
10,10.0154,12.724,10.6497,13.2725

```

Figure 51. Bounding box coordinates for Columbus polygons.

After selecting the **Tools > Shape > To Boundary (BND)** command, a dialog appears in which you need to specify the file names for the input (a polygon shape file) and output files (an ascii text file), as in Figure 52. In addition, you must select one of the four output formats by checking the corresponding radio button. If you also want the bounding box file, you must check that option. The bounding box file is created as **cover_r.bnd** (where **cover** is the name of the polygon shape file). The boundary file is created in the working directory and appears as in Figures 49-50.

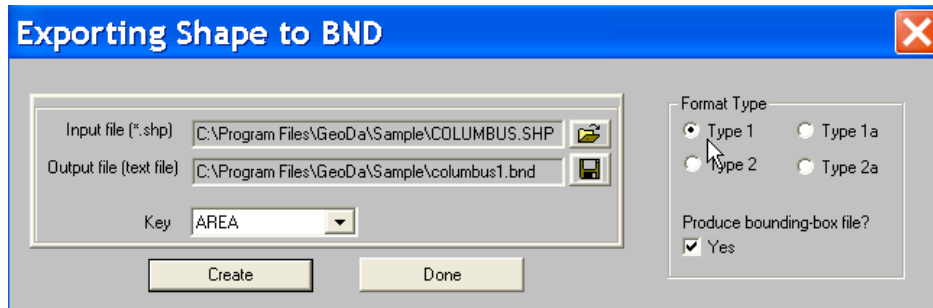


Figure 52. Polygon shape to boundary file export dialog.

Exporting Data Tables

The **Tools > Data Export** submenu supports conversion of data tables between a shape file and output files in a format suitable for analysis by other software packages. Specifically, the **SpaceStat** option creates a binary file in the format used by SpaceStat. The **ASCII** option creates a comma delimited text file (csv format) in the format suitable for input into many software packages. It contains a header line that specifies the number of observations and the number of variables, followed by a list of the variable names and a listing of the data by observation, as shown in Figure 53 for a subset of variables from the Columbus data set. The header line may have to be removed or edited to accommodate the requirements of specific software.

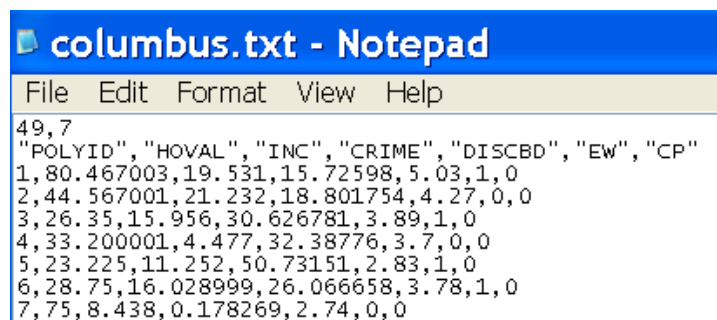


Figure 53. Ascii text output file from Columbus data set polygon shape file.

Both `SpaceStat` and `ASCII` options work in the same way. After the menu item is invoked, an `Exporting Data` dialog appears in which the `Input file` (shape file) and `Output file` name must be specified. This operates in the same manner as in the `Shape Conversion` dialogs. You need to click on the folder button to specify the `Input file` (click `Open` to confirm the choice) and on the save button to specify the `Output file` (click `Save As` to confirm the choice). The `Exporting Data` dialog then lists all the variables in the `Input file` that are available for export. You select variables by clicking the double arrow `>>` to select all, or individually by highlighting variables and clicking on the single arrow `>`, as in Figure 54. After the list of variables is complete, click on `Export` to create the new file, as in Figure 55. After the exported file is written to disk, the `Done` button becomes active. Click this button to close the dialog. The new file will appear in the working directory.



Figure 54. Exporting data dialog.



Figure 55. Data ready for export.

Mapping

GeoDa contains rudimentary functionality for mapping and geovisualization. This is limited to choropleth maps, with some specialized flavors that aid in highlighting extreme values or outliers. All maps are invoked from the **Map** menu.

Standard Choropleth Maps

Getting Started

The two most commonly used types of choropleth maps in GeoDa are the **Quantile** map and the **Std Dev** (standard deviational) map. Before you can make a choropleth map, you need to load the shape file with the data using the **Edit > New Map**, **Edit > Duplicate Map** or similar commands. In the window containing the layout for the shape file, you can expose a legend pane by dragging the bar on the left hand side of the window slightly to the right, as in Figure 56.

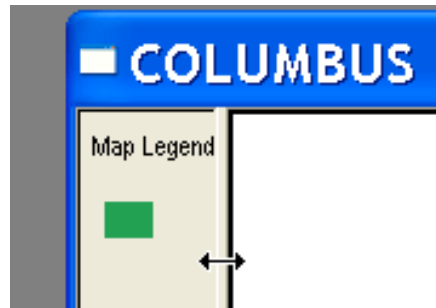


Figure 56. Exposing the Legend Pane.

You also need to specify a variable to be mapped. If you earlier used the **Select Variables** toolbar button or **Edit > Select Variable**, you may have set a variable to be the default, as in Figure 12. If that is the case, you will not be queried for a variable name. If you don't like the current default, you will need to reset the selected variable by invoking **Edit > Select Variable** and changing your choice. If you have not specified a default variable, the **Variables Settings** dialog will appear, as in Figure 12. If you had previously specified two default variables (bivariate variable selection), the first one (**x**) will be used for mapping. Note that the first time a data set is analyzed, the data table is created and it may need to be moved out of the way to access the map functionality (minimizing the table window will get it out of the way).

Map > Quantile

In a quantile map, the data are sorted and grouped in categories with equal numbers of observations, or quantiles. The **Map > Quantile** command invokes a simple dialog to specify the number of quantiles or categories (assuming a variable has been specified). The default number of categories is 4 for a *quartile* map. In the example in Figure 57, the number of categories was changed to 5, for a *quintile* map. Clicking **OK** generates the map and associated legend, as in Figure 58, for Columbus neighborhood crime. Note how the legend pane indicates the type of map and variable name, as well as, in parentheses, the number of observations in each category.

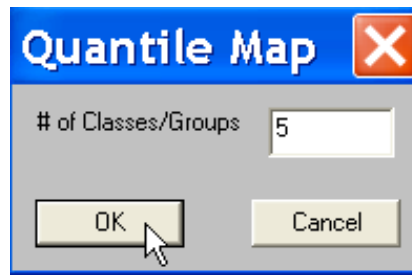


Figure 57. Quantile Map dialog.

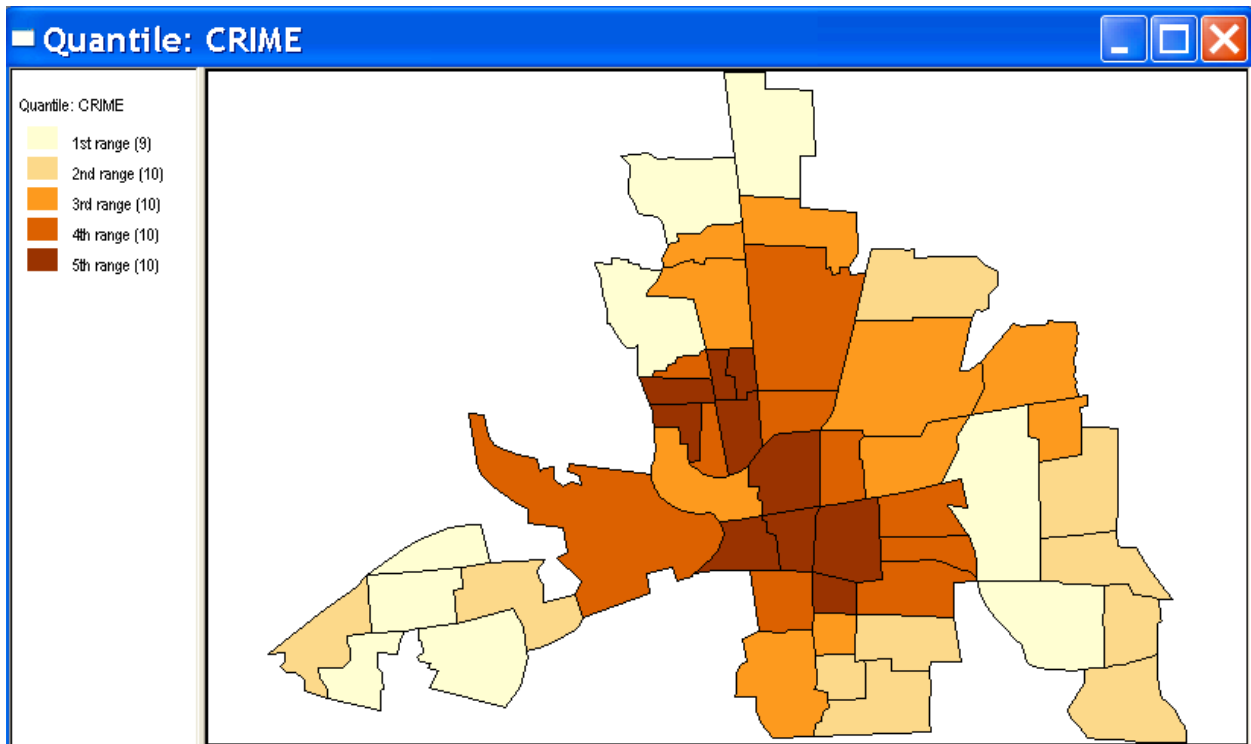


Figure 58. Columbus neighborhood crime quintile map.

The legend colors are pre-coded using the color schemes suggested in Cynthia Brewer's ColorBrewer color picker.³ They can also be individually specified for each category. Double click on the square box in the legend and the standard MS Windows color dialog appears, as in Figure 59. Select a new color from this palette by clicking on a square. For example, selecting a blue for the upper quintile will change the matching locations on the map, as in Figure 60.

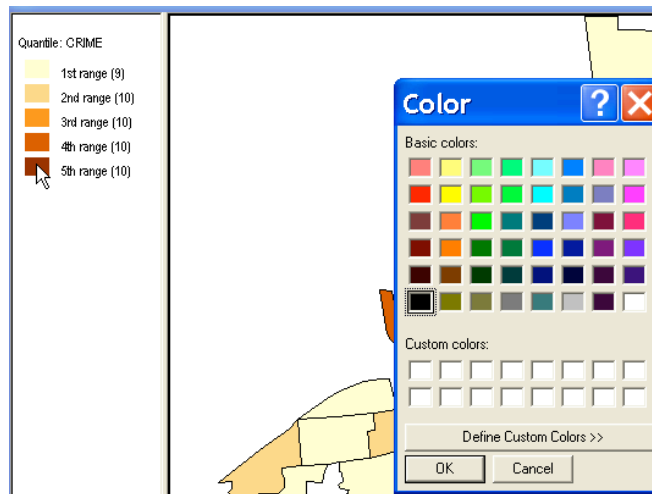


Figure 59. Legend color customization dialog.

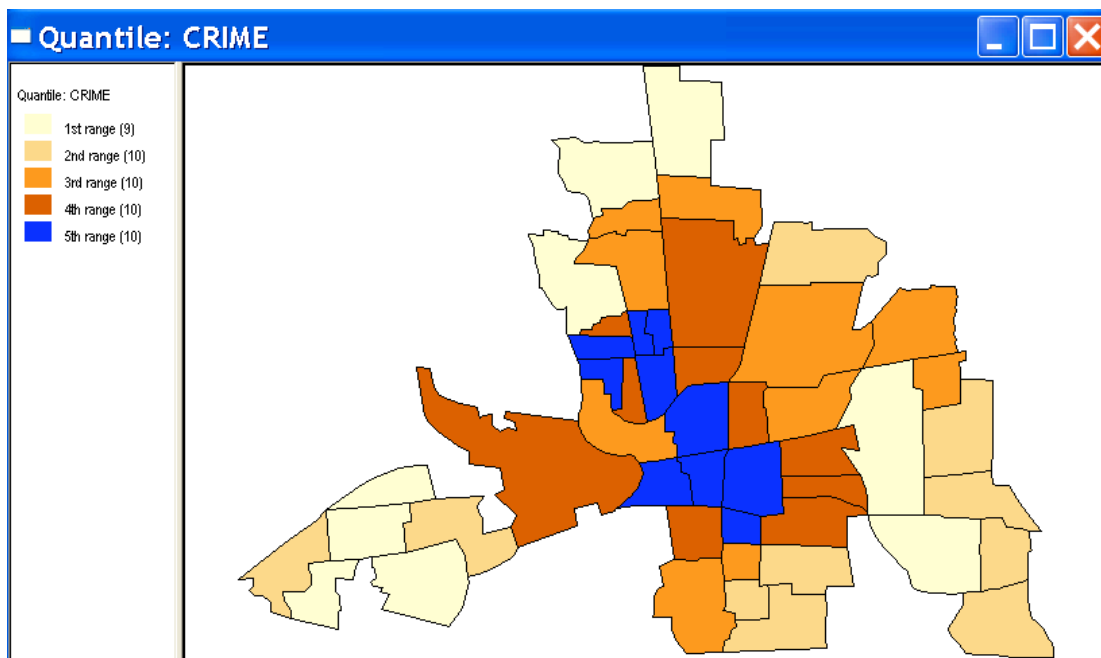


Figure 60. Customized legend colors.

³ ColorBrewer can be found at <http://www.personal.psu.edu/faculty/c/a/cab38/ColorBrewerBeta.html>.

A quantile map may yield surprising results when the data contain a lot of observations with similar values (e.g., a value of zero for a map of rare events). In that case, some quantiles may be undefined, since it is not possible to resolve ties and allocate observations with the same value to different groupings. In such an instance, all observations are given the color of the highest quantile for which there is no problem with ties. For example, if there were more than 10 observations with zero crime rate in the Columbus case (there are 49 observations), all 19 observations in quintile #1 and #2 would receive the color of quintile #2.

Map > Std Dev

A standard deviational map groups observations according to where their values fall on a standardized range, expressed as standard deviational units away from the mean.⁴ The **Map > Std Dev** command creates a choropleth map with the categories corresponding to multiples of standard deviational units. No additional dialog is required. For example, a standard deviational map for the Columbus neighborhood crime data is illustrated in Figure 61. The map legend shows the type of map, variable, its mean value, the break points for each category, and, in parentheses, how many observations fall in that category. As for the quantile map, the legend colors can be customized for each individual category.

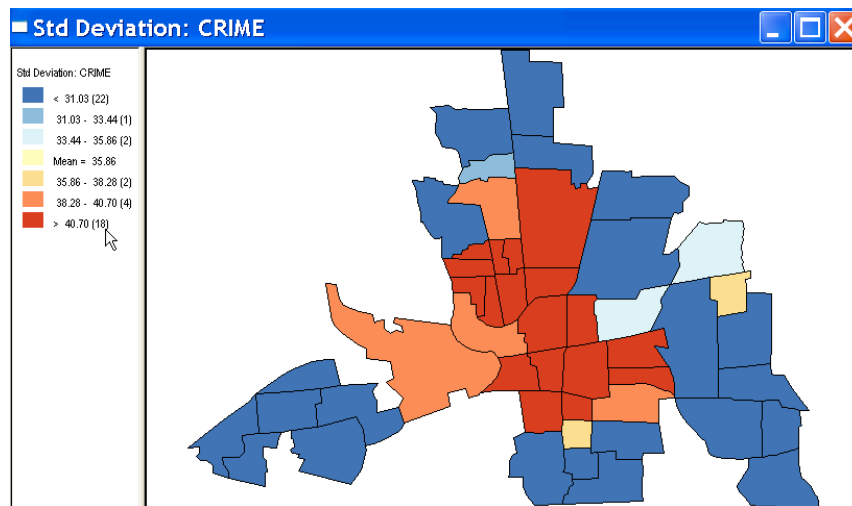


Figure 61. Standard deviational map for Columbus neighborhood crime.

⁴ A standardized variable has a mean of zero and a standard deviation of 1, by construction. Hence a standardized value can be interpreted as multiples of standard deviational units.

Outlier Maps

Outlier maps highlight locations with extreme values (both high as well as low). GeoDa contains two types of outlier maps, a **Box Map** and a **Percentile Map**. These are choropleth maps and as such they require that a shape file has been loaded into the project. In addition, a variable must have been specified. If no variable was set as default, a **Variables Settings** dialog (Figure 12) will appear to request a variable name.

Map > Box Map

A **Box Map** (Anselin 1994, 1999b) is a special case of a quartile map where the outliers (if present) are shaded differently. As a result, there are six legend categories: four base categories (one for each quartile), one for outliers in the first quartile (extremely low values) and one for outliers in the fourth quartile (extremely high values). The legend shows for each of the categories in parentheses the number of observations that fall in this category. For the second and third quartile, this is always $\frac{1}{2}$ of the number of observations. For the first and fourth quartile, this number will vary, depending on how many outliers there are. For example, a **Box Map** of 1979 SIDS death rates in North Carolina counties reveals four upper outliers, but no lower outliers, as shown in Figure 62. This is invoked by the **Map > Box Map > Hinge = 1.5** (the default). To use a stricter definition of outlier, select **Map > Box Map > Hinge = 3.0**.

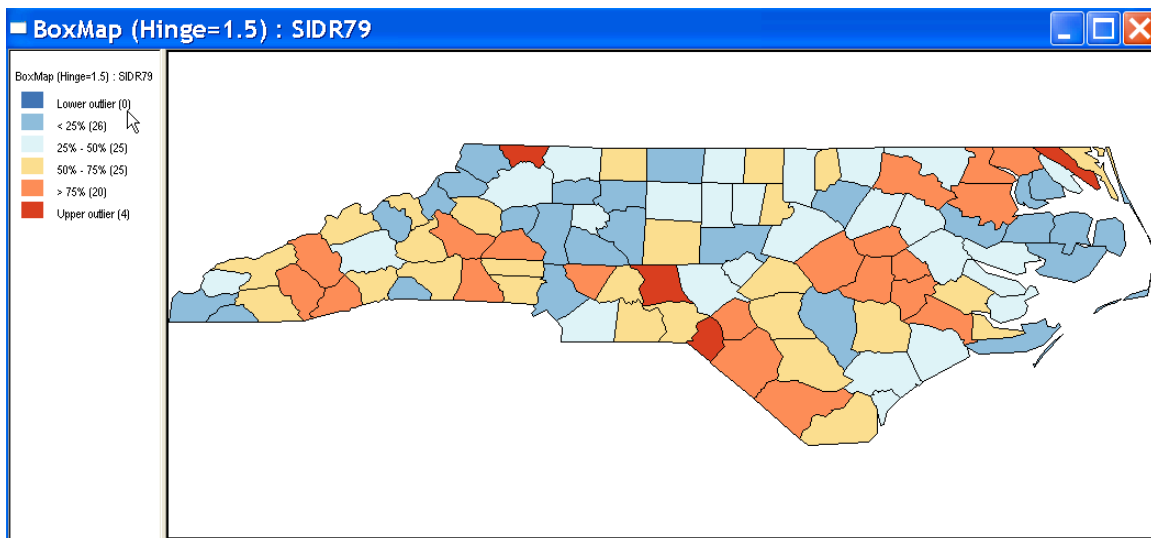


Figure 62. Box Map for 1979 SIDS death rate in North Carolina counties.

Map > Percentile

A **Percentile** map is also a special case of a quantile map. In this case, no additional categories are created, but the categories are grouped to accentuate the extreme values. Specifically, six legend categories are created, corresponding to < 1%, 1% to <10%, 10% to <50%, 50% to <90%, 90% to <99% and >99%. This is illustrated for the 1979 SIDS death rates in North Carolina counties in Figure 63. This is invoked by the **Map > Percentile** command without additional dialogs.

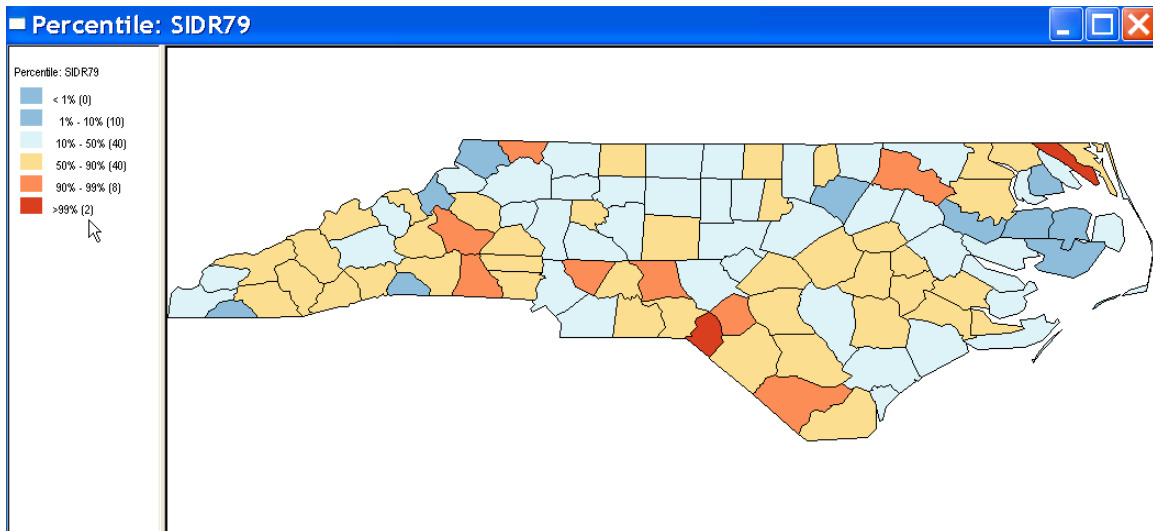


Figure 63. Percentile map for 1979 SIDS death rates in North Carolina counties.

Note some particular features in Figure 63. The parentheses next to the first category show no observations in this category, while one would expect 1 (NC has 100 counties). In fact, due to ties, that first observation cannot be assigned to a separate percentile and is grouped with the 1 to <10% category. Similarly, at the high end, there are two observations, one for #99 and one for #100, due to the decision rule used to assign observations to categories.

Map Movie

A **Map Movie** is an animated sequence highlighting locations on the map in increasing order of magnitude for a specified variable. It may suggest spatial structure in the data, in the form of spatial regimes. For example, a **Map > Map Movie > Cumulative** command applied to the Columbus neighborhood crime data will show the outlying neighborhoods

first, followed by a converging pattern to the center city. This suggests a pattern of spatial heterogeneity where the lower values are found in the periphery and the higher values in the core. A **Map Movie** requires that a variable is specified but no other dialog is used. The **single** option for the **Map Movie** is currently not very reliable and too dependent on specific machine hardware.

Map Options

When a map is active, the **Options Menu** contains several ways to customize the map, as illustrated in Figure 64. Similarly, right clicking on any active map invokes a menu that includes several customization options. In addition, the right click also provides a short cut to the choropleth mapping functions discussed earlier in this section, as well as to the rate maps (see the section on Smoothing Rate Maps), as shown in Figure 65.

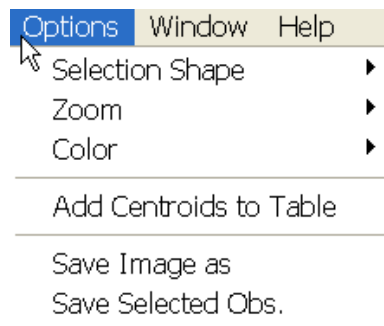


Figure 64. Map Options menu.

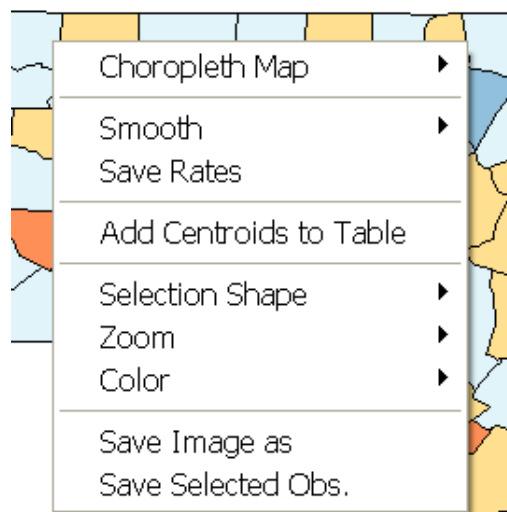


Figure 65. Map Options and shortcuts context menu.

Options > Selection Shape

The **Selection Shape** is the graphical mechanism by which observations are selected on a map. GeoDa supports five different shapes, as illustrated in Figure 66: point, rectangle, polygon, line and circle. Selection by forming a **Rectangle** through clicking and dragging the mouse is the default. The observations are selected following a “spatial select” rule, which selects those locations whose centroid falls within the rectangle. The other selection rules are self explanatory. The selection rules can also be set by right clicking on a map. A selection rule stays active until a different one is chosen. Note that a selection rule is specific to each window, so different windows (different maps) may have different selection rules (such as **Rectangle** in one and **Point** in the other).

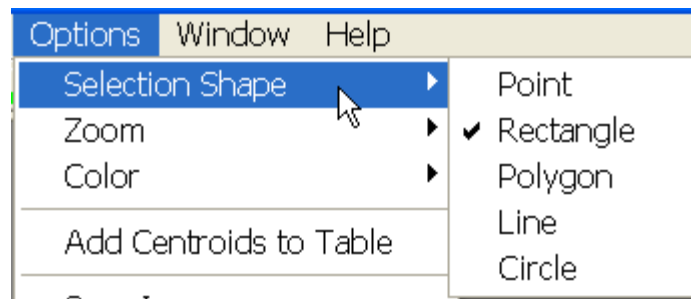


Figure 66. Map selection rules.

Options > Zoom

The map view is not static, but supports the usual zoom in and zoom out functions. You invoke this as **Options > Zoom** and select one of the three choices (Figure 67). Alternatively, you can right click on the map. **Zoom In** requires you to draw a rectangle with the pointer to rescale the picture (Figure 68). To **Zoom Out** you need to click on the map (no rectangle). **Zoom > Full Extent** restores the original map view.

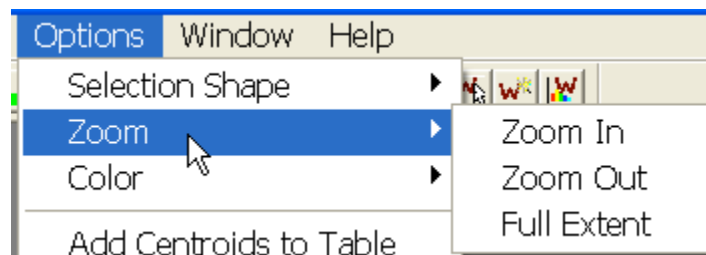


Figure 67. Zoom options.

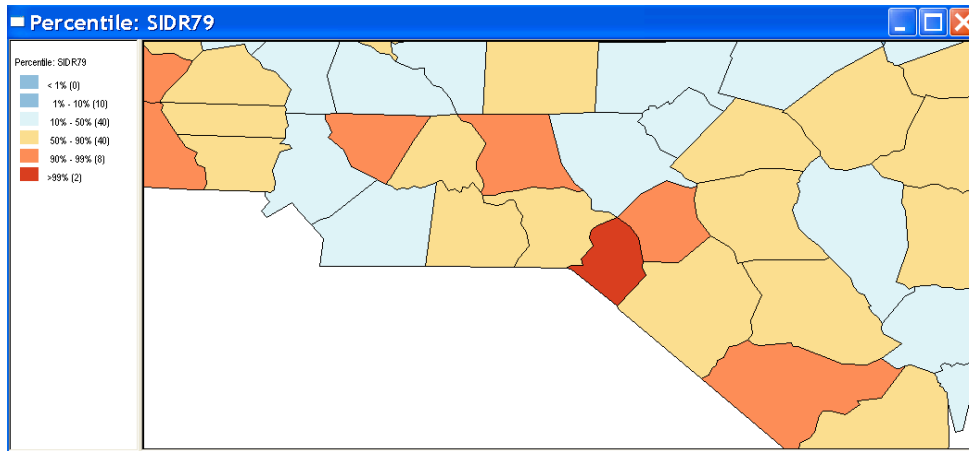


Figure 68. Percentile map after zoom in.

Options > Color

The color options allow four different aspects of the map view to be customized: the **Map**, **Shading**, **Movie** and **Background** (Figure 69). Each of these operates in the same fashion. Selecting an option brings up the Windows **Color** dialog (as shown in Figure 59). Clicking on one of the color boxes changes the current setting for a map aspect to the new color. The **Color > Map** option alters the base color for the unclassified map, i.e., the map that appears when selecting **Edit > Duplicate Map** or clicking on the **Duplicate Map** toolbar button. For example, in Figure 70, this color has been changed from the default green (as in Figure 5) to a red-brown base. The **Color > Shading** option alters the color used for the hatch marks that identify selected observations (see the section on Linking and Brushing). For example, in Figure 71, this has been changed to red from the default yellow. The **Color > Movie** option changes the color of the animated locations in the **Map Movie**. Finally, **Color > Background**, changes the color for the background in the map view. For example, in Figure 72, this is set to gray. Note that the legend background color can also be changed, after right clicking in its window pane. You also use this short cut to save the legend pane to the clipboard (see Figure 72).

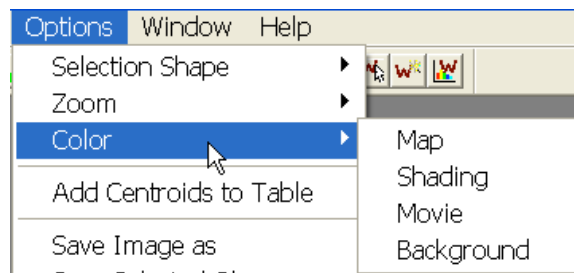


Figure 69. Map color selection options.

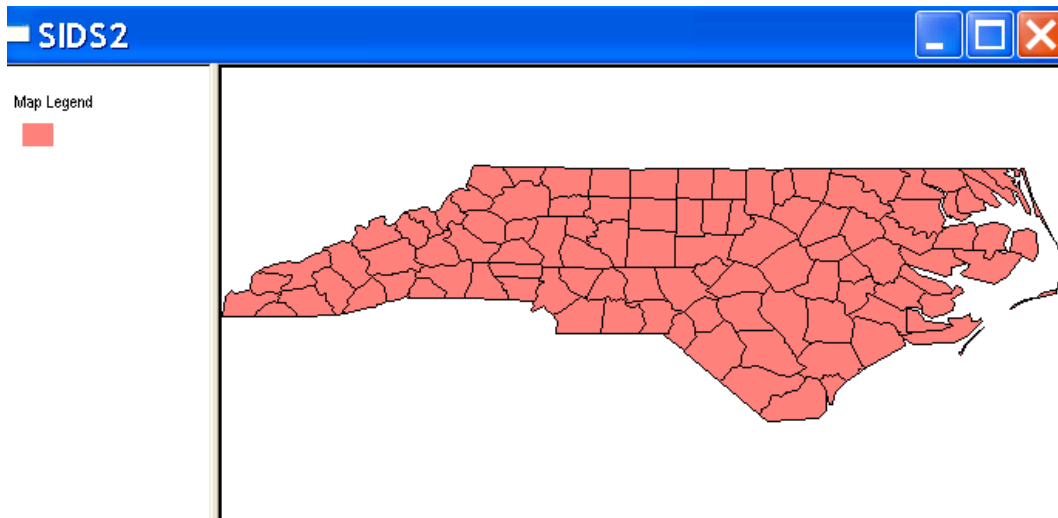


Figure 70. Changing the Map color.

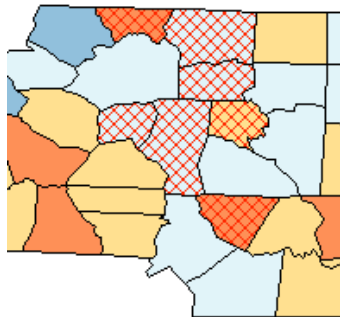


Figure 71. Customizing the Selection hash marks color.

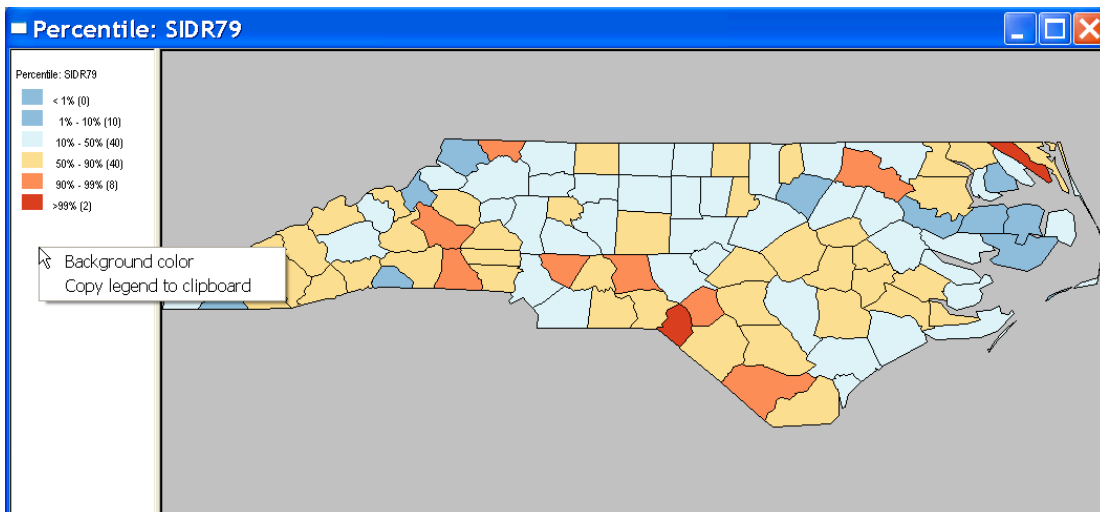


Figure 72. Customizing the Background color for a map view.

Options > Save Image as

Besides capturing the currently active window to the clipboard (**Edit > Copy to clipboard**), a map view can be saved to a bitmap file through **Options > Save Image as**. This invokes a file **Save as** dialog in which the name for the **bmp** file must be specified in the usual fashion (Figure 73). The saved file can then be incorporated into other documents. For example, Figure 74 shows the bitmap file that was saved by applying the **Save Image as** operation to the map in Figure 72.

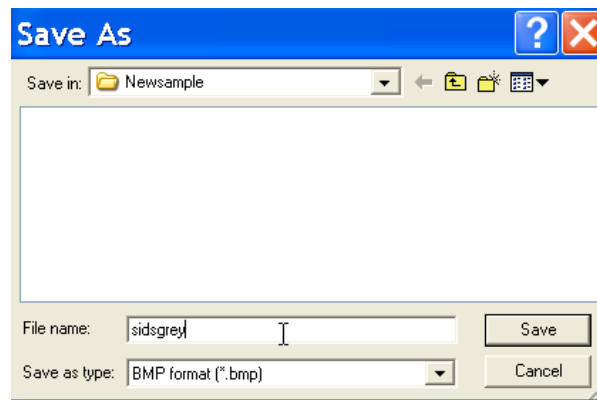


Figure 73. Save map as image dialog.

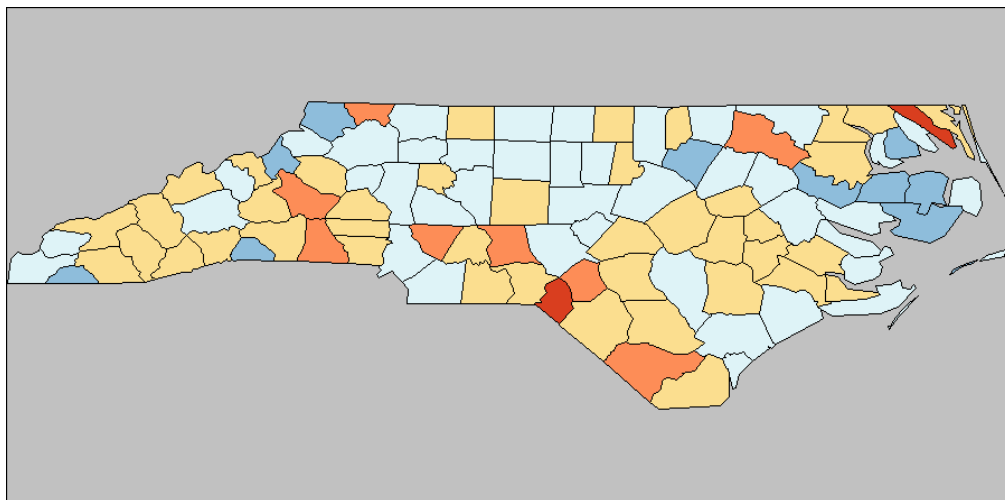


Figure 74. Saved bitmap image.

Options > Save Selected Obs.

Whenever observations are selected, whether explicitly on the map or through linking and brushing (see Linking and Brushing), a dummy variable can be saved to the data

table that contains values of 1 for the selected records and 0 for the others. Choosing **Options > Save Selected Obs.** brings up a dialog in which the variable name for the indicator must be specified, as in Figure 75 (the default variable is **SELECT_1**). Clicking **OK** adds a new column to the data table, as in Figure 76. Note that this addition is not permanent until the table has been saved (see the section on Editing and Manipulating Tables).

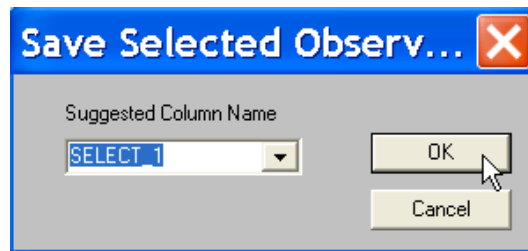


Figure 75. Save selected observations indicator variable dialog.

NWR.74	NWR.79	SELECT_1
9.165903	13.929619	0
20.533881	22.140221	0
65.244668	71.902655	0
242.125984	174.698795	1
750.175932	745.330012	1
657.024793	673.014146	1
402.097902	397.142857	1
604.761905	624.579125	1
772.727273	709.243697	1
99.255583	86.359176	0
531.400966	476.456504	0

Figure 76. Indicator variable added to data table.

Options > Add Centroids to Table

See the section on Manipulating Spatial Data.

Smoothing Rate Maps

Using choropleth maps to represent the spatial distribution of rates or proportions represents a number of challenges. These are primarily due to the inherent *variance instability* (unequal precision) of rates as estimates for an underlying “risk.” A number of procedures have been suggested to correct for this variance instability by “smoothing” the risk estimate. GeoDa contains five specialized mapping routines to deal with the visualization of rates in the **Map > Smooth** submenu (see Figure 21): **Raw Rate**, **Excess Risk**, **Empirical Bayes**, **Spatial Rate** and **Spatial Empirical Bayes**. Technical background and further illustration of these procedures is given in Anselin et al. (2002b).

All five procedures require that a shape file be loaded and that two variables be specified: one is the **Event Variable**, the other the **Base Variable**. The rates themselves need not be specified, they are computed internally from the events (counts of deaths, disease incidence, homicides, etc.) and the “population at risk” (the population to which the event pertains, such as total births for the SIDS death rate). The **Event** and **Base Variable** are specified in a **Rate Smoothing** dialog, as in Figure 77. After selecting the variables, click **OK** to generate the map.

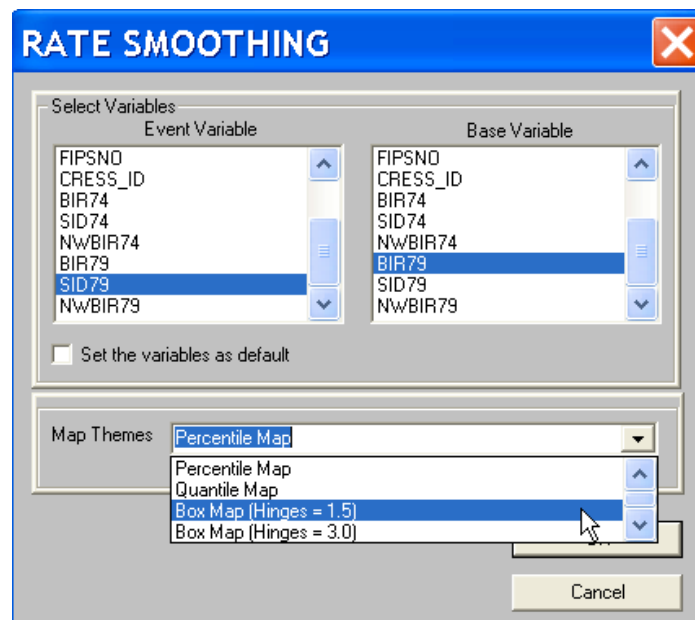


Figure 77. Rate smoothing dialog.

Figure 77 shows four of the five **Map Themes** available for mapping rates in a drop down list: **Percentile Map**, **Quantile Map**, two **Box Maps** (hinges 1.5 or 3.0) and a **Std**

Deviational map. These five types can be used for all but the **Excess Rate** map, for which there is a customized legend.

The **Spatial Rate** and **Spatial Empirical Bayes** smoothing procedures also require that a spatial weights file be specified. If there is no current default spatial weights file in the project (that is, if it had not been previously set using a **Tools > Weights > Open** command and set as the default), a warning will appear to set the weight. You must first **Open** or **Create** the weights file. You might also want to set it as default (see Figure 78).

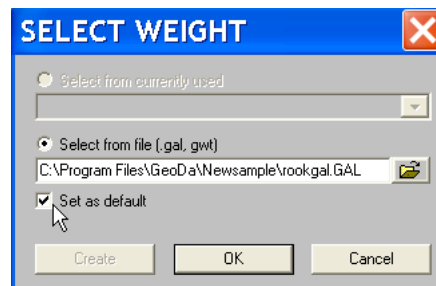


Figure 78. Specifying spatial weights for rate smoothing.

Map > Smooth > Raw Rate

This computes the “raw rate” as a simple ratio of the event count to the base population at risk. For example, using the 1979 SIDS death rates for North Carolina counties, this would require **SID79** as the **Event Variable** and **BIR79** as the **Base Variable**, as in Figure 77. With the map type set to **Box Map** (Hinges = 1.5), the same map results as generated earlier in Figure 62 (compare to Figure 79). The only difference is the map heading, which specifies **Raw Rate** and the variables from which the rate is computed.

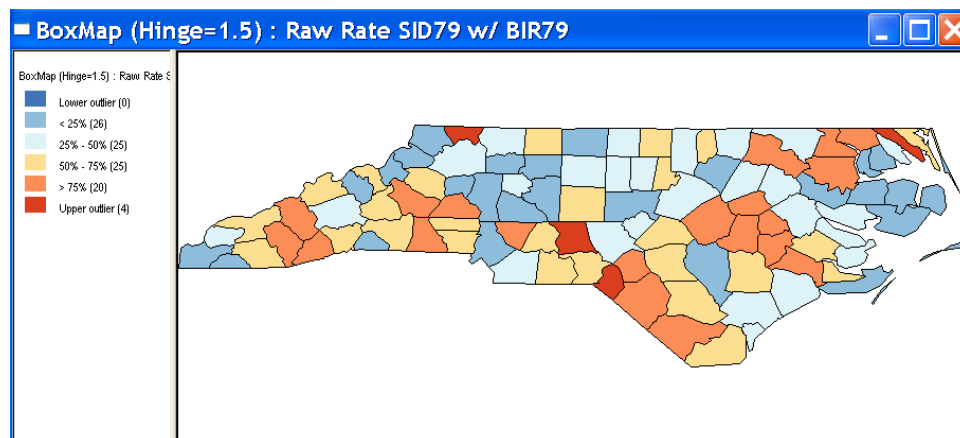


Figure 79. Raw rate box map for 1979 SIDS death rate in North Carolina Counties

Map > Smooth > Excess Risk

This computes a map showing the *relative risk* or excess risk as a ratio of the observed number of events over the expected number of events. The expected number is computed by applying the average risk (total number of events in all the locations over total population) to the population at risk in each location. Values of the **Excess Risk** less than one show locations with fewer than expected events, values greater than one show locations where the number of events exceeds the expectation. Note that the **Excess Risk** measure is a non-spatial measure in that it ignores any effect of spatial autocorrelation. The **Excess Risk** map uses a specialized legend that classifies locations by the extent to which they vary around 1 (see Figure 80 for the 1979 SIDS data). The map header lists the type of rate map as well as the events and base population used in the calculation.

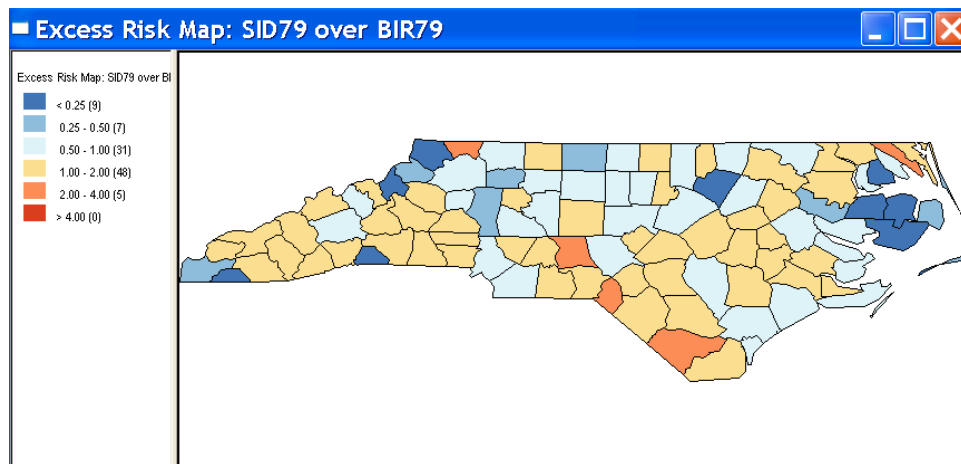


Figure 80. Excess risk map for 1979 SIDS death rates in North Carolina counties.

Map > Smooth > Empirical Bayes

This creates a choropleth map for rates that are “shrunk” towards the overall mean using the classic method of moments estimator in the **Empirical Bayes** procedure (Marshall 1991). This smoothing procedure particularly affects the value for locations with small populations at risk (the “small area” problem). It will also typically remove the problem associated with many ties (especially zero values) in quantile maps. This map is illustrated in Figure 81 for a **Box Map** of the 1979 SIDS death rates. Note the contrast with Figure 79: there is now one outlier county at the low end of the scale, and five outliers at the high end.

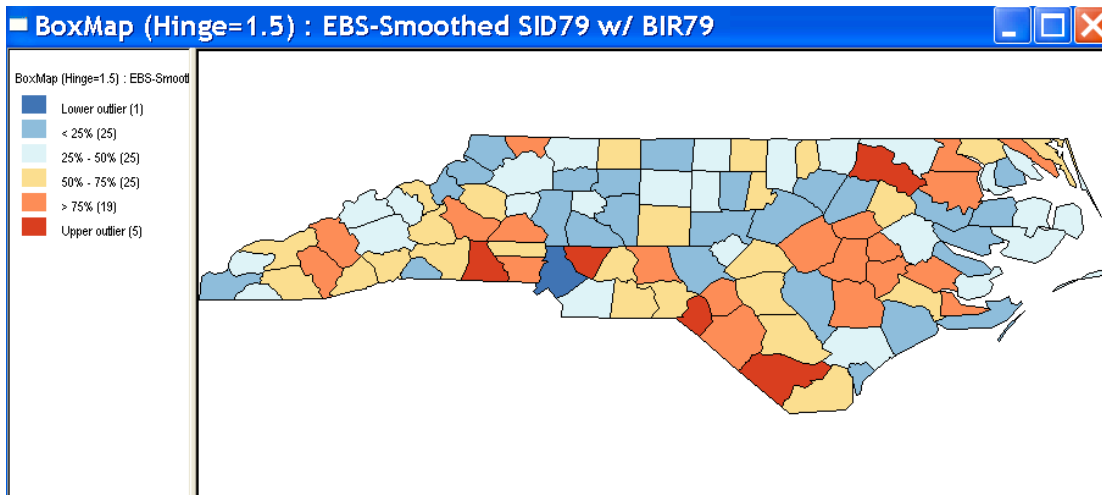


Figure 81. EB smoothed box map for 1979 SIDS death rates in North Carolina counties

Map > Smooth > Spatial Rate

Creates a smoothed map using a spatial window average. The smoothed rate is computed from the ratio of the total number of events in a spatial “window” to the total population at risk in that window. The window is specified using a spatial weights file and includes both the neighbors as well as the location itself. This is illustrated in Figure 82 with a **Box Map** for the 1979 SIDS data, using a first order contiguity (rook) spatial weights file to define the window. The **Spatial Rate** smoother brings out broad spatial trends in the data and typically greatly reduces the number of outliers. The map header lists the type of smoothing and the variables used in the calculation.

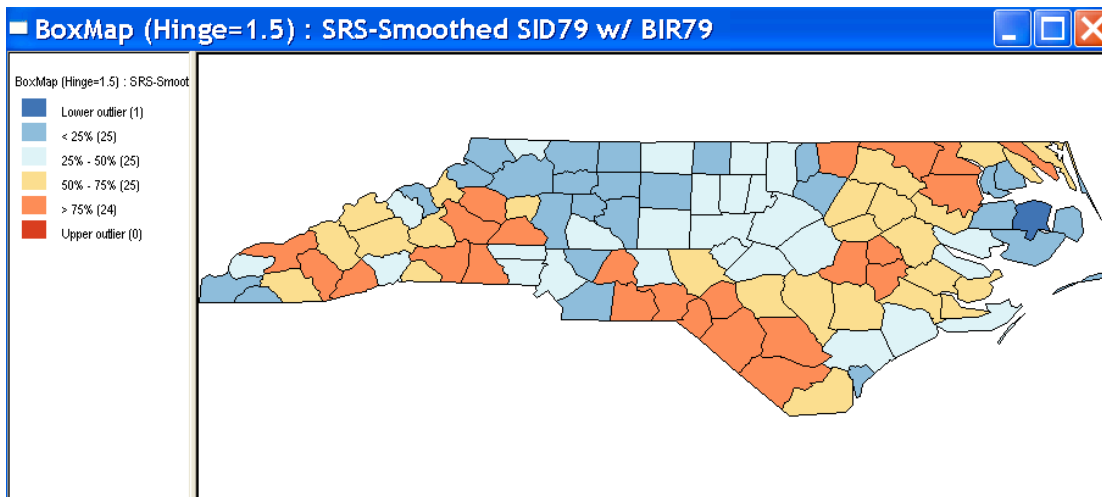


Figure 82. Spatial rate box map for 1979 SIDS death rates in North Carolina counties.

Map > Smooth > Spatial Empirical Bayes

Creates a smoothed map using the **Empirical Bayes** procedure, but with the window average used as the reference for adjustment, rather than the overall mean. This results in some degree of spatial smoothing, but less so than for the **Spatial Rate** smoother. In Figure 83, this is shown for the 1979 SIDS data, using the first order contiguity (rook) spatial weights to define the spatial window. The map header lists the type of smoothing and the variables used in the calculation.

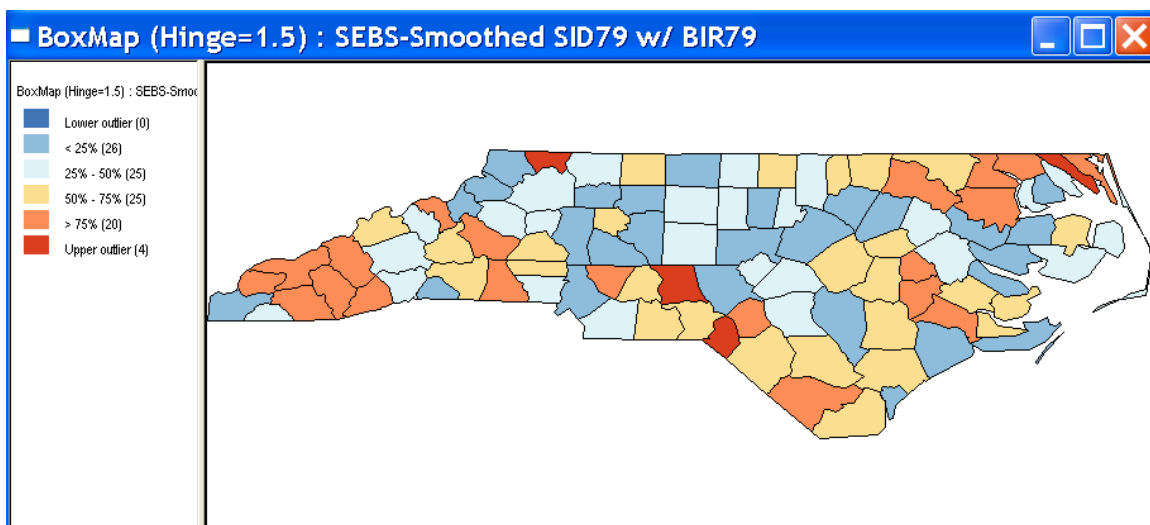


Figure 83. Spatial EB rate box map for 1979 SIDS death rates in North Carolina counties.

Smoothing Options

The **Options** menu for the smoothed rate maps is the same as for the other choropleth maps. Specific to the rate calculations is the item to save the computed rates to the data table, invoked by right clicking on the map and selecting **Save Rates** (Figure 84). A dialog appears requesting a variable name for the rates. Depending on the context, a different default is suggested, respectively, **R_RAWRATE** for raw (unsmoothed) rates, **R_EXCESS** for the relative risk, **R_EBS** for the empirical Bayes smoothing, **R_SPATRATE** for the spatial smoothing, and **R_SPATEBS** for spatial empirical Bayes (see Figure 85 for the raw rate example). You don't have to accept the default, but can specify any variable name in the text box. Click **OK** to add the variable to the data table, as illustrated in Figure 86 for various smoothed rates in the **SIDS** example. The new rates are immediately available for analysis. For example, Figure 87 shows the variable selection dialog for a

box map of the raw rates without using the smoothed rates function.

As with all table manipulations, the addition of rate variables is not permanent until after the table has been saved.

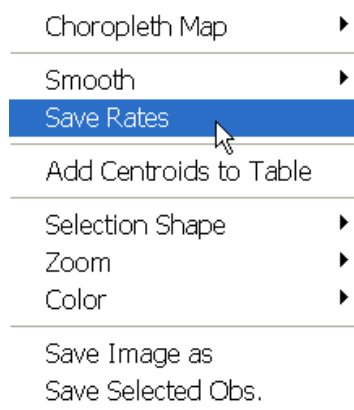


Figure 84. Save rates option for rate maps.

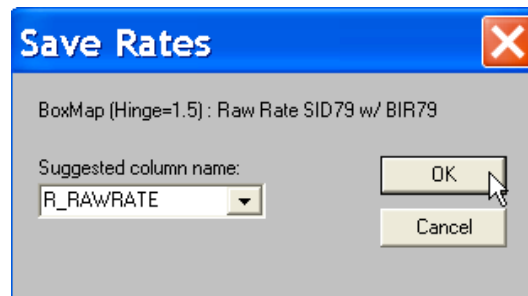


Figure 85. Rate variable name specification.

Table : SIDS						
	BIR79	SID79	NWBIR79	R_RAWRATE	R_EBS	R_SPATRATE
1	1364.000000	0.000000	19.000000	0.000000	0.001625	0.001485
2	542.000000	3.000000	12.000000	0.005535	0.002263	0.001730
3	3616.000000	6.000000	260.000000	0.001659	0.001862	0.001471
4	830.000000	2.000000	145.000000	0.002410	0.002030	0.002233
5	1606.000000	3.000000	1197.000000	0.001868	0.001956	0.002987
6	1838.000000	5.000000	1237.000000	0.002720	0.002148	0.002653
7	350.000000	2.000000	139.000000	0.005714	0.002177	0.002470
o	594.000000	2.000000	371.000000	0.003367	0.002100	0.002111

Figure 86. Computed rate variables added to data table.

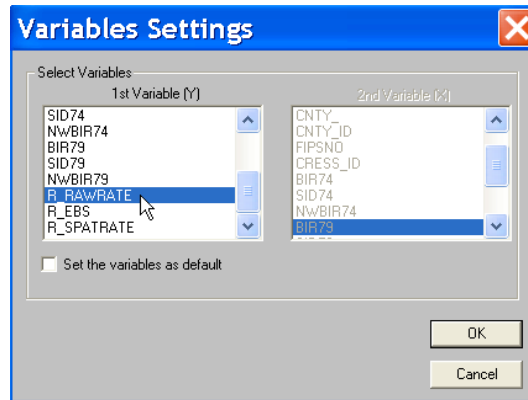


Figure 87. New rate variables available for analysis.

Editing and Manipulating Tables

The **Table** is included as one of the views on the data in the exploratory toolbox implemented in GeoDa. The first time a request is made to analyze or map a variable (in a map function or to construct a statistical graph) the data table is loaded in memory and becomes available for editing and querying. It can also be loaded explicitly using the **Explore > Table** command or by clicking on the **Table** toolbar button. Using the standard convention, observations or records are shown as rows and variables or fields are shown as columns.

The **Table** functionality is invoked by right clicking anywhere in the active table. This brings up a menu, as shown in Figure 88. There are three main sets of functions. The first set deals with selection of records, the second with editing and data transformations and the third with saving and joining. In addition, sorting of records and editing of individual cells are invoked by means of mouse clicks.



Figure 88. Table menu.

Sorting Records by Field

As initially listed, the records in the table are given in the order in which they appear in the shape file that is stored on disk, which may not be the most useful. Sorting of records according to the values taken by a given variable is done directly by double clicking on the field (variable) name at the top of the table, as shown in Figure 89 for the **POLYID** variable in the Columbus data set. The result is a table sorted by increasing values of

POLYID. Note the small triangle pointing up next to the variable name. Double clicking again toggles the order and sorts the records by decreasing value of the variable. This is indicated by the triangle pointing down, as in Figure 90. The sorted records can be restored to the original order by double clicking on the header for the left most column. This column contains sequence numbers for the original order. The sorting mechanism is applied to it in the same way as to any other variable, hence the triangle appears and the order can be reversed (double click to toggle the sort order). Note that the sorted order of records is not permanent, nor does it affect the way the data are stored on disk.

Table : COLUMBUS					
	AREA	PERIMETER	COLUMBUS_	COLUMBUS_I	POLYID ▲
1	0.309441	2.440629	2	5	1
2	0.259329	2.236939	3	1	2
3	0.192468	2.187547	4	6	3
4	0.083841	1.427635	5	2	4
5	0.488888	2.997133	6	7	5

Figure 89. Records sorted by increasing values of **POLYID**.

Table : COLUMBUS					
	AREA	PERIMETER	COLUMBUS_	COLUMBUS_I	POLYID ▼
49	0.205964	2.199169	50	26	49
48	0.069762	1.102032	49	27	48
47	0.245249	2.079986	48	15	47
46	0.124728	1.841029	47	48	46
45	0.256431	2.193039	46	28	45

Figure 90. Records sorted by decreasing value of **POLYID**.

Editing Individual Table Cells

Individual entries for table cells can be edited by double clicking on the cell. The cell is highlighted and ready for editing, as shown in Figure 91. Typing in a new value, followed by **Enter** will change the entry in the selected cell. The new value is used in all subsequent analyses, but is not permanent until the table is **saved as** a different file. Only then will the changes be written to disk. Changes can be undone by selecting **Refresh Data**. However, this will undo all changes since the last save.

Table : COLUMBUS				
	AREA	PERIMETER	COLUMBUS_	COLUMBUS_I
1	0.309441	2.440629	2	5
2	0.259329	2.236939	3	1
3	0.192468	2.187547	4	6
4	0.083841	1.427635	5	2

Figure 91. Editing individual table cells.

Promoting Selected Records

Records are selected by clicking on the left most column in the table, as shown in Figure 92. Also, clicking in the left most column and dragging down will select a set of consecutive records (rows). Records are also selected indirectly, through linking and brushing of maps and statistical graphs (see the section on Linking and Brushing). The selected records can be moved up to the top of the table by choosing **Promotion** in the table menu. The result for the selection in Figure 92 would be as shown in Figure 93.

Table : COLUMBUS				
	PERIMETER	COLUMBUS_	COLUMBUS_I	POLYID
1	2.440629	2	5	1
2	2.236939	3	1	2
3	2.187547	4	6	3
4	1.427635	5	2	4
5	2.997133	6	7	5

Figure 92. Manual record selection in a table.

Table : COLUMBUS				
	PERIMETER	COLUMBUS_	COLUMBUS_I	POLYID
3	2.187547	4	6	3
1	2.440629	2	5	1
2	2.236939	3	1	2
4	1.427635	5	2	4
5	2.997133	6	7	5

Figure 93. Selected records promoted to the top of the table.

Clear Selection

Any selection can be undone by choosing **Clear Selection** for the table menu. The removal of the selection affects all statistical graphs and maps in the project (see the section on Linking and Brushing).

Range Selection

The table contains some limited functionality to carry out queries, where records are selected as the result of a logical operation comparing the value of a variable (field) to a numerical range. This is invoked by **Range Selection** in the table menu, which brings up a dialog to specify the parameters for the range interval, as in Figure 94.

The drop down list and set of text boxes under the heading **Range Selection** pertains to the interval itself. Note how in the current version the range is inclusive of the lower bound (\leq) and exclusive of the upper bound ($<$). For example, using the Columbus sample data set, the records with a **CRIME** rate over 30 can be selected by specifying 30 as the lower bound and a large number (larger than any in the data set) as the upper bound, as in Figure 94. In order to implement an equality condition, you can trick the upper bound inequality by adding a very small amount to the equality value (e.g., to set **CRIME = 30** one could use $30 \leq \text{CRIME} < 30.001$). Click **Apply** to generate the selection, as shown in Figure 95. The selection is applied to all current graphs and maps in the project (see the section on Linking and Brushing).

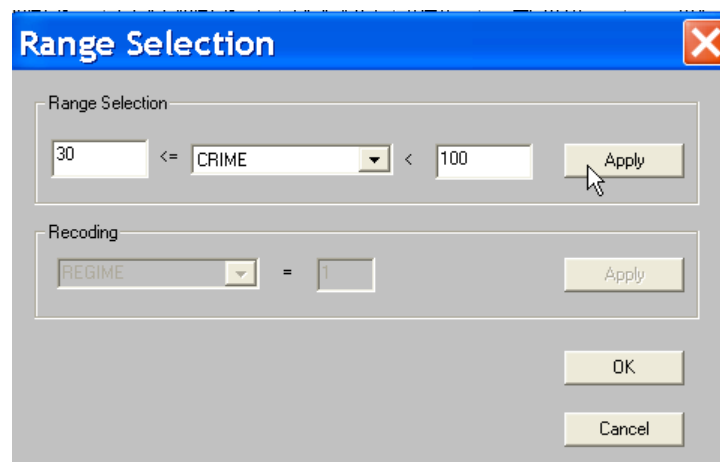


Figure 94. Range selection interval specification.

Table : COLUMBUS					
	POLYID	NEIG	HOVAL	INC	CRIME
1	1	5	80.467003	19.531000	15.725980
2	2	1	44.567001	21.232000	18.801754
3	3	6	26.350000	15.956000	30.626781
4	4	2	33.200001	4.477000	32.387760
5	5	7	23.225000	11.252000	50.731510
6	6	8	28.750000	16.028999	26.066658

Figure 95. Range selection applied to table.

After the **Apply** button is clicked, the second row of options becomes active, under the heading **Recoding** (Figure 96). This allows you to create an indicator variable that takes on the value of 1 for the selected records and 0 elsewhere. The drop down list gives **REGIME** as the default variable name, but this can be changed by typing in a different name. The **REGIME** variable is added to the table, as shown in Figure 97. The variable is available for analysis, but the addition is not permanent until the table has been saved as a different file. You can also skip this stage by clicking on **OK**, which closes the dialog.

The dialog box titled "Range Selection" has a close button (X) in the top right corner. It contains two main sections: "Range Selection" and "Recoding".

In the "Range Selection" section, there is a text input field containing "30", followed by a dropdown menu showing "<=", then a dropdown menu showing "CRIME", followed by a text input field containing "100", and an "Apply" button.

In the "Recoding" section, there is a dropdown menu showing "REGIME", followed by an equals sign "=", then a text input field containing "1", and an "Apply" button. A mouse cursor is pointing at this "Apply" button.

At the bottom of the dialog box, there are three buttons: "OK", "Cancel", and a button that is partially obscured by the "Apply" button in the "Recoding" section.

Figure 96. Specifying a regime variable following a range selection.

NEIGNO	PERIM	REGIME
1005.000000	2.440629	0
1001.000000	2.236939	0
1006.000000	2.187547	1
1002.000000	1.427635	1
1007.000000	2.997133	1
1008.000000	2.335634	0

Figure 97. Regime variable added to table.

Save Selected Observations

A new indicator variable with a value of 1 for the selected observations and 0 for the others can be added to the table by right clicking and choosing **Save Selected Obs.**. This brings up a dialog to select the variable name for the indicator variable, as in Figure 98. The default is **SELECT_1**, but any other name can be specified. Clicking **OK** will add a new column to the data table, as illustrated in Figure 99. The new variable is immediately available for analysis, but will not be added permanently until the table is saved as a new file.

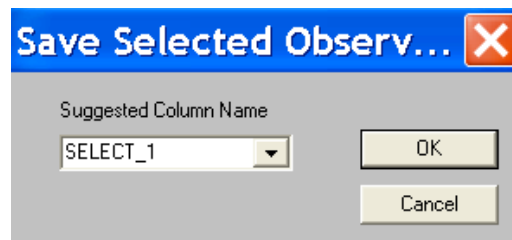


Figure 98. Save selected observations in table.

NEIGNO	PERIM	REGIME	SELECT_1
1005.000000	2.440629	0	0
1001.000000	2.236939	0	0
1006.000000	2.187547	1	0
1002.000000	1.427635	1	1
1007.000000	2.997133	1	1
1008.000000	2.335634	0	1
1004.000000	2.554577	0	1
1003.000000	2.139524	1	1
1018.000000	3.169707	1	1
1010.000000	2.087235	1	0

Figure 99. Selected observations indicator variable added to data table.

Calculating New Variables

Adding new variables to the data table is a two step process. First, a new column must be added and a variable name specified for the new field. This is invoked by right clicking on the table and choosing **Add Column**. This brings up a dialog to select the variable name, as in Figure 100. Similarly, **Delete Column** removes a field from the table.

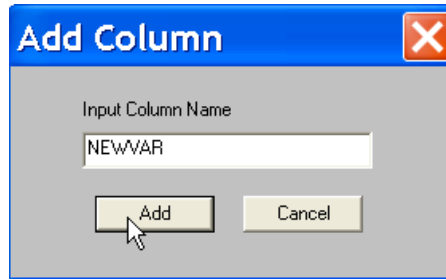


Figure 100. Specifying a variable name for a new column (field).

Once a variable (field) name has been created by *adding a column*, its values can be computed using the **Field Calculation** command. This brings up a dialog, as in Figure 101, with four tabs, each pertaining to a class of calculations: **Unary Operations**, **Binary Operations**, **Lag Operations** and **Rate Operations**. Each tab brings up a slightly different dialog, although the operation of the dialog is similar: in each case, you specify the target variable name (**Result**), the type of operation and any parameters needed for that operation (such as a spatial weights matrix for spatial lag computations), and the required variables (**Variables**). The result is inserted into the new field. As before, this variable is immediately available for analysis, but not permanent until the table has been saved as a new file.

In Figure 101, the dialog is shown for **Unary Operations**, i.e., transformations of existing variables. Five such operations are currently supported: equality (**Assign**), reverse sign (**Negate**), 1/value (**Invert**), **Square Root** and **Log**. In Figure 101, the new variable **NEWVAR** becomes the square root of the variable **CRIME** (in the Columbus data set). Note that you can enter any numeric value instead of a variable name in the **Variables** text box. Click **OK** to carry out the calculation.

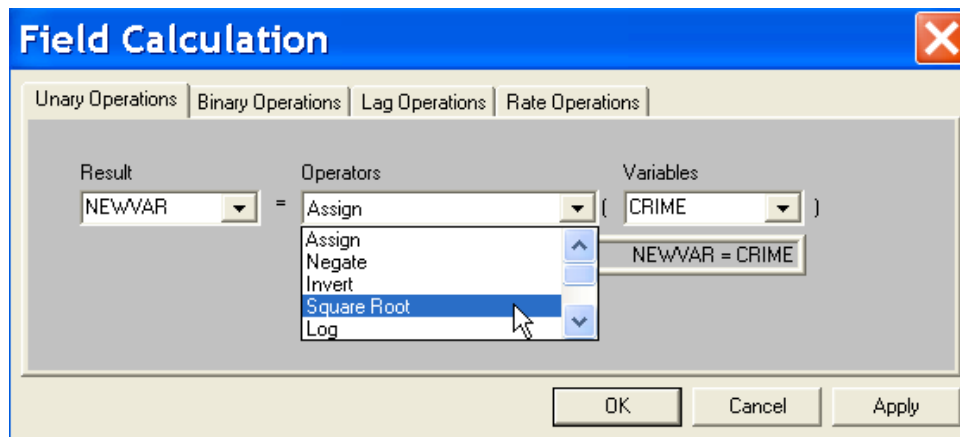


Figure 101. Field calculation, unary operations.

Binary operations are mathematical operations on two variables (or numerical values). You invoke its interface by clicking on the second tab in the **Field Calculation** dialog. As shown in Figure 102, five operations are currently supported: **Add**, **Subtract**, **Multiply**, **Divide** and **Power**. You need to specify the **Result** variable, and two existing variables (**Variable-1** and **Variable-2**), as well as the **Operator**. For example, in Figure 102, the variable **NEWVAR** becomes **HOVAL/AREA** (in the Columbus data set). You could also replace either of the variables by a numeric value, such as **1000** instead of **AREA** (this would divide the house values by 1000). Click **OK** to carry out the computation.

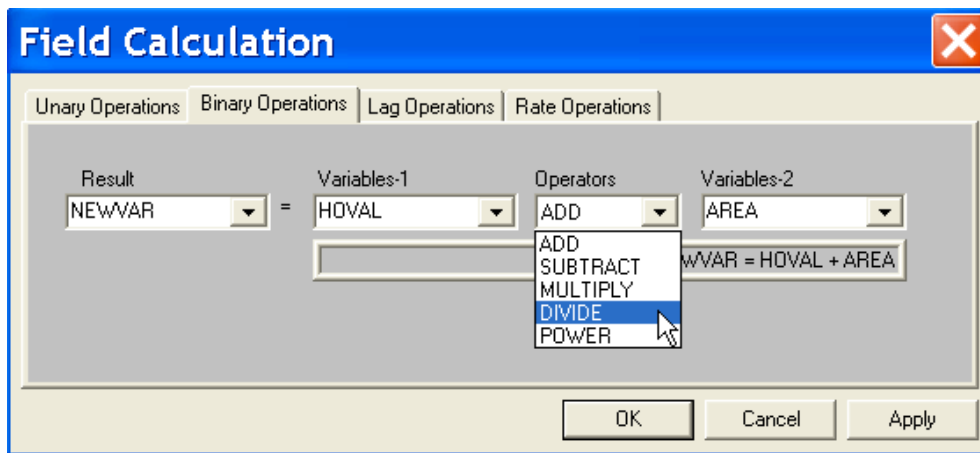


Figure 102. Field calculation, binary operations.

Computing Spatially Lagged Variables

Spatially lagged variables are weighted averages of the values for neighboring locations, as specified by a spatial weights matrix. This is implemented in the third tab of the **Field Calculation** dialog. In addition to the target (**Result**) and variable, you also need to specify a spatial weights file. This file must first be opened or created with **Tools > Weights > Open** or **Tools > Weights > Create** (see the section on Creating and Manipulating Spatial Weights). All currently opened spatial weights files will appear in the drop down list under **Weights files** (if the list is empty, you must open a weights file). The spatial lag is implemented for row-standardized spatial weights.

For example, in Figure 103, a spatially lagged variable is constructed for **CRIME** in the Columbus data set, using a rook-based contiguity spatial weights matrix. After the table is saved as a different file, the spatial variables are available for use in other statistical software (for example, to include in spatial regression models in the **R** `spdep` package).

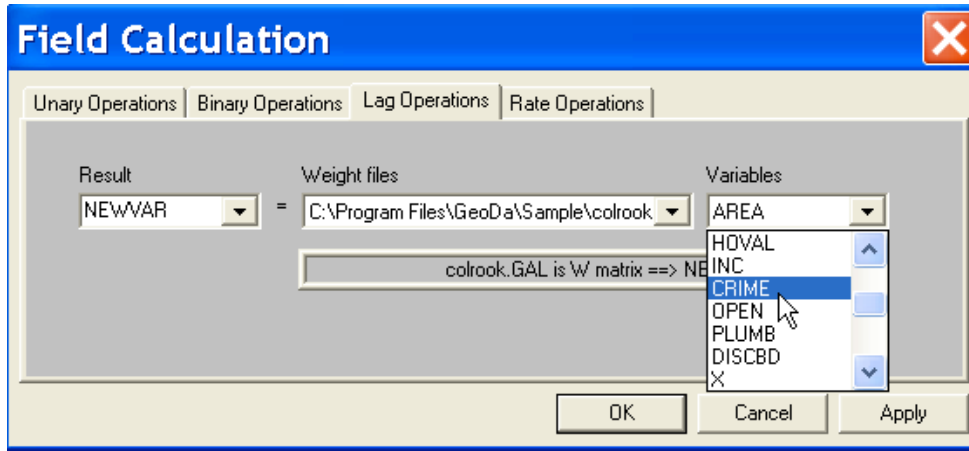


Figure 103. Computing a spatially lagged variable.

Computing Rate Variables

Rates or proportions can be added to a table by using the **Save Rates** option in a rate map. In addition, they can also be calculated directly, from the fourth tab in the **Field Calculation** dialog. In addition to the five rate types implemented for the maps (**Raw Rate**, **Excess Rate**, **Empirical Bayes**, **Spatial Rate**, and **Spatial Empirical Bayes**), there is also the **EB Standardized Rate** which is used in the **EB Moran Scatter Plot** and **LISA Maps** (see the sections on Global and Local Spatial Autocorrelation). The operation is similar to the other dialogs: specify the target variable name, the method (and a spatial weights file for the spatial rates), and the **Event** and **Base** variables to compute the rates. For example, in Figure 104, the rate calculation is illustrated using the **SID74** and **BIR74** variables from the NC **SIDS** data set.

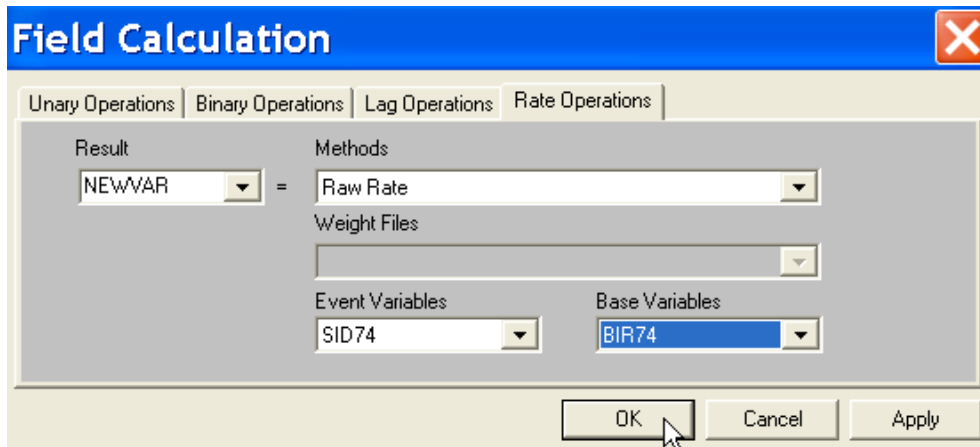


Figure 104. Rate calculation in a table.

Saving and Joining Tables

The changes and additions made to a table only reside in memory and are not permanent. In order to make them permanent, the table must be saved to a new file. This is implemented in the **Save to Shape File As ...** command. It invokes the usual **Save As** file dialog. You must specify a file name, which should be different from the current shape file in the project. This results in three files to be saved, with file extensions **.shp**, **.shx** and **.dbf**. The **.dbf** file is the new table, the other two files are copies of the shape file geography that was associated with the table. If you only want to use the table as a data set in non-GIS statistical software, you can ignore or remove the **.shp** and **.shx** files.

Other tables can be joined with the current data table in a project, provided that they have an exact match in terms of the number of records and the **key** variable. Right click on the table and choose **Join Tables** from the menu. This brings up a **Join Tables** dialog. In the example in Figure 105, the **Input file (*.dbf)** is a file created by the **Tools > Data Export > Centroids** command and contains the variable **POLYID** as **Key**, as well as a sequence number, and the **X_COORD** and **Y_COORD** centroid coordinates for the Columbus data set. In the dialog, you need to specify the **key** variable in the drop down list and move the variables over that you wish to join with the current data table. You use the same **>>** and **>** symbols as in the **Data Export** interface (e.g., Figure 54). Click on **Join** to carry out the procedure. Three new columns will be added to the data table in your project, as shown in Figure 106. Note that the **key** variable is not duplicated. The joined variables are immediately available for use in any of the analyses. However, as always, you must save the table as a new file to make the join permanent.

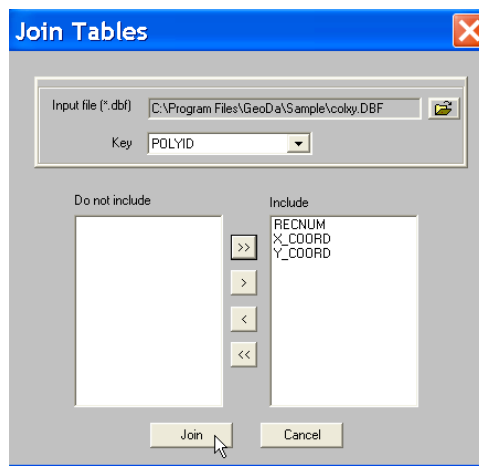


Figure 105. Join Tables dialog.

NEIGNO	PERIM	RECNUM	X_COORD	Y_COORD
1005.000000	2.440629	1	8.795270	14.322100
1001.000000	2.236939	2	8.335270	13.986100
1006.000000	2.187547	3	9.027840	13.763000
1002.000000	1.427635	4	8.434140	13.714200
1007.000000	2.997133	5	9.084150	13.420700
1008.000000	2.335634	6	9.865510	13.482700
1004.000000	2.554577	7	8.075840	13.327400
1003.000000	2.139524	8	8.374980	13.486000
1018.000000	3.169707	9	9.615140	12.879900

Figure 106. Joined variables added to data table.

Undoing Changes

Any changes made to the data table is temporary, in that it only resides in memory. In order to make the changes permanent, the table must be saved as a new file. To undo all changes, select **Refresh Data** from the menu. This will restore the table to its original state when it was first loaded from disk. There is currently no more subtle undo facility. All the changes are removed, even those you may wish to keep. The safest procedure to follow in this regard is to save the table each time you have added some new variables or made some edits that you are likely to use in the future.

Statistical Graphs

GeoDa contains standard EDA functionality in the form of three statistical graphs. The graphs are dynamically linked, allowing for full brushing of the data (see the section on Linking and Brushing). They are invoked from the **Explore** menu (see Figure 18), or by clicking on the corresponding **Explore** toolbar button.

The three statistical graphs included in GeoDa are the **Histogram**, **Scatter Plot** and **Box Plot**. They require that a shape file is loaded into the project and that a variable be specified using a **Variables Settings** dialog (see Figure 12). The **Histogram** and **Box Plot** require one variable to be selected, the **Scatter Plot** two. If these variables were set earlier as default, there will be no **Variable Settings** dialog when the statistical graphs are invoked. To change the default variable, it is necessary to explicitly reselect it using **Edit > Select Variable** or by clicking the **Select Variable** toolbar button.

Explore > Histogram

Creates a **Histogram** for the specified variable. The default number of classifications is set to 7. This can be changed by invoking **Options > Intervals**, which will bring up a dialog where a new value can be specified (for example, 12 in Figure 107). Clicking on the **OK** button will create a new **Histogram** with the specified number of intervals. For example, using the Columbus neighborhood shape file (**COLUMBUS.SHP**), a histogram of housing values (**HOVAL**) is given in Figure 108. The category break points are listed on the right hand side and the number of observations in each of the groups is given on top of the histogram bar. In the current version, the colors of the histogram bars are randomly assigned and cannot be changed, nor can the break points be manipulated.

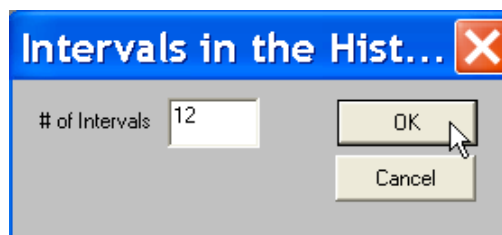


Figure 107. Change intervals dialog for histogram.

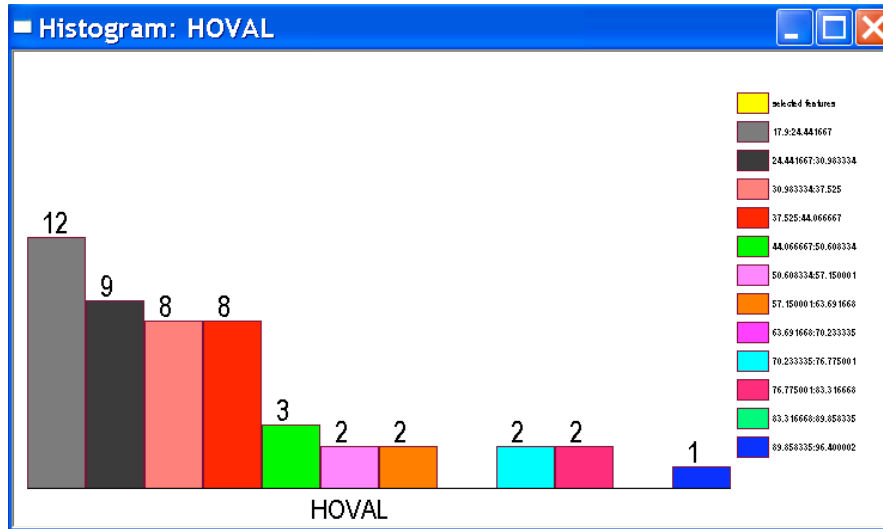


Figure 108. Histogram for Columbus neighborhood housing values.

Besides allowing a choice of the number of intervals, the **Histogram** also has three other settings that can be customized. These are invoked from the **Options** menu or by right clicking on the histogram window, as shown in Figure 109.

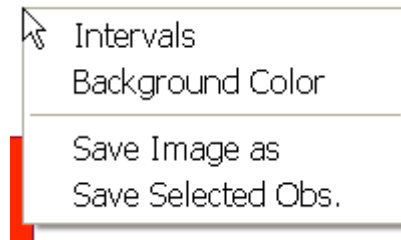


Figure 109. Histogram options.

The **Background Color**, **Save Image as** and **Save Selected Obs.** options work in the same manner as for the **Map** window (see, respectively, Figure 72, Figures 73-74 and Figures 75-76). They allow the background color to be changed, the window to be captured as a bitmap file and the selected observations (see Linking and Brushing) to be added to the data table as an indicator variable (see the discussion in the section on Mapping for further details).

Explore > Box Plot

Creates a **Box Plot** for the specified variable. The plot consists of a box going from the lowest value (bottom) to the highest value (top). The interquartile range is indicated by a

purple box in the middle of the graph: its lowest edge is the first quartile (25%), the upper edge the third quartile (75%). A blue dot in this box corresponds to the median value. The fence is indicated by a purple horizontal bar at the end of the “T”. If the fence is below the lowest value or above the highest value, the fence is drawn on the edge of the box. In Figure 110, two **Box Plots** are shown for the Columbus **HOVAL** variable, with the hinge set to 1.5 (default) on the left (five outliers) and to 3.0 on the right (no outliers). All box plots are scaled to have the same size.

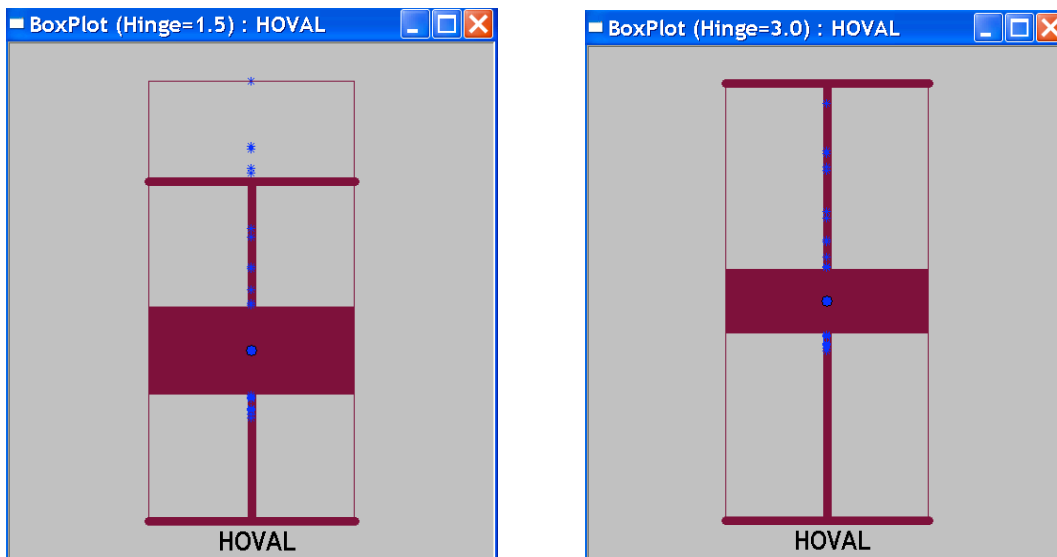


Figure 110. Box Plot for Columbus housing values using 1.5 and 3 as hinge.

The **Hinge** value for the **Box Plot** can be changed by means of the **Options > Hinge** command. The **Options** menu can also be invoked by right clicking on the plot window, as in Figure 111. Besides the **Hinge** option, which is unique to the **Box Plot** functionality, the **Background Color**, **Save Selected Obs.** and **Save Image as** options work the same as for the other graphs and maps (note how in Figure 110, the background color was changed to gray from the default white).

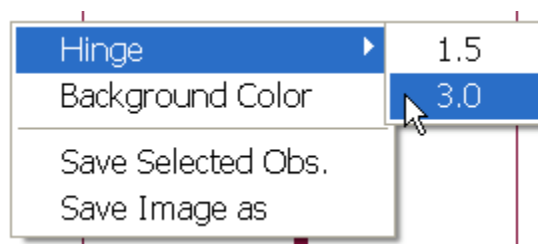


Figure 111. Box Plot options.

Explore > Scatter Plot

This function creates a bivariate **Scatter Plot** with the first specified variable (Y) on the vertical axis and the second variable (X) on the horizontal axis. A least squares linear regression fit is superimposed on the scatter and its slope is listed at the top of the graph. The window header lists the explanatory variable first (X), the dependent variable second (Y), as in the left pane of Figure 112 for the Columbus neighborhood variables **CRIME** and **HOVAL**.

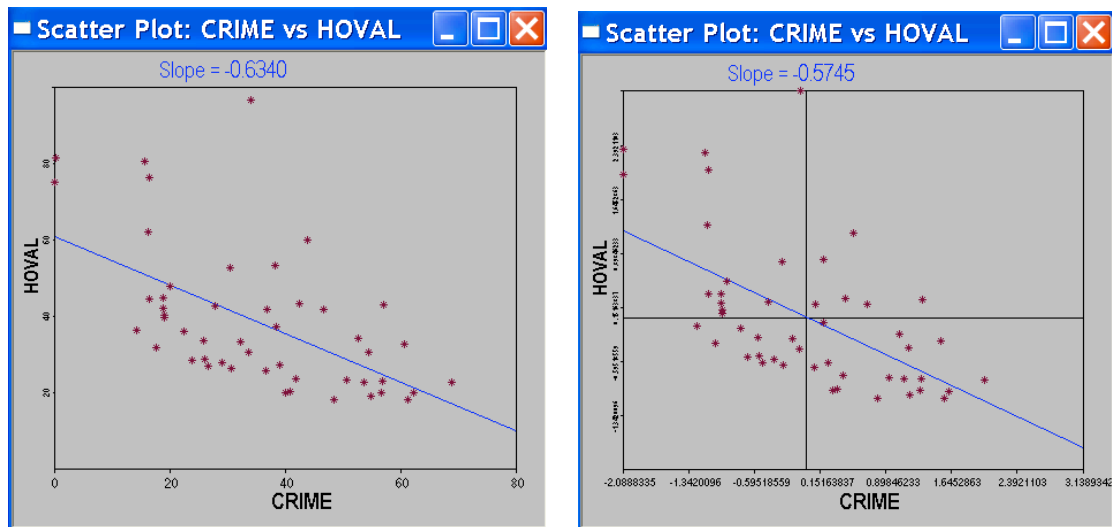


Figure 112. Scatter plot of Columbus neighborhood crime on housing values.

The default is to compute the scatter plot for the data as they appear in the shape file. Using the **Options > Standardized data** command (see Figure 113), the scatter plot can be shown for standardized values, such that the slope of the regression line corresponds to the bivariate correlation coefficient. GeoDa also shows the four quadrants of the scatter plot (similar to the **Moran Scatterplot**) so that it is straightforward to identify locations where above mean (or below mean) values on both variables coincide, or, alternatively, locations where above mean (or below mean) values in one variable coincide with below mean (or above mean) values for the other. This is illustrated in right hand pane of Figure 112.

As is the case in all the graphs, the **Options** menu for the **Scatter Plot** can be invoked by right clicking on the graph, as illustrated in Figure 113. Besides the **Standardized data** feature, an important option is **Exclude selected**. Choosing this activates the dynamic recalculation of the regression slope as observations are selected. The new slope

excludes the selected observations. Typically, this is used when brushing a scatter plot (or map), as discussed in more detail in the section on Linking and Brushing. The **Exclude selected** option is off by default and must be activated explicitly. The other options, **Background Color**, **Save Selected Obs.** and **Save Image as** work the same as for the other graphs and maps (note how in Figure 112, the background color was changed to gray from the default white).

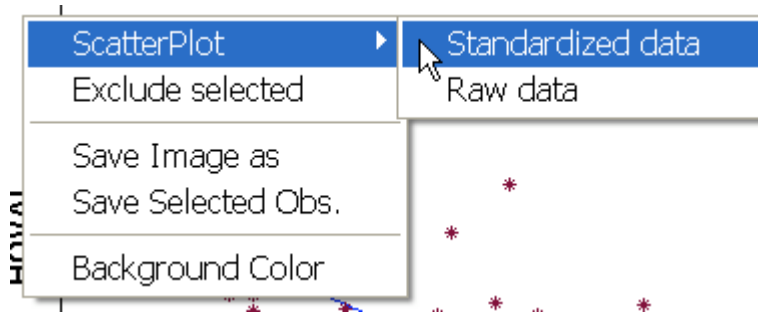


Figure 113. Scatter plot options.

Linking and Brushing

A distinguishing feature of GeoDa (and its precursor DynESDA) is the implementation of full two-way dynamic linking. This approach consists of treating each window (statistical graph, map, table) as a “view” of the data and connecting all these views together. The connection is such that any time an observation is selected in any view, it is simultaneously selected (highlighted) in *all* the other views. This allows the user extensive interaction with the data and is a fundamental tool in EDA. GeoDa implements the selection of individual observations or groups of observations for all statistical graphs and maps, as well as for the table. Any time more than one view is open, selection automatically leads to linking. The linking is dynamic in the sense that any change in the selection in any of the views directly leads to an update of the selection in all the other views.

Brushing is a slightly more involved form of dynamic linking in that a “brush” is moved over the observations in one of the views. In GeoDa, the brush is a rectangle specified by the user. As the brush moves over the data, the selection changes and is updated in all the other views in the project. When the **Exclude selected** option is set in the **Scatter Plot**, brushing results in an instantaneous recalculation of statistics in a **Scatter Plot**, such as the slope of the regression line and Moran’s I spatial autocorrelation statistic.

In each of the maps and graphs, an indicator variable can be created and added to the **Table** using the **Save Selected Obs.** function in the **Options** (use the **Options** menu or right click on the graph or map). This indicator variable will take on the value of one for the selected observations and zero for the others. Several such indicator variables may be constructed, using different selection criteria.

Selection in a Map

Selection of an individual observation or a group of observations is implemented slightly differently for each type of graph. In a **Map**, selection is based on the option set by the **Options > Selection shape** command (Figure 66), or by using the pop up menu (right click) in the map window itself. You select a location by clicking on it (**Point** selection), by clicking and dragging to create the desired shape (**Rectangle** or **Circle**), or by consecutive clicking to form a **Line** or **Polygon**. For the latter two, the selection must be finalized by a **Double-Click**. The other selection mechanisms do not require this.

The selected locations (following a “spatial select” rule) are highlighted using a cross-hatched pattern (see Figure 114). The default color of the selection is yellow, but this can be changed using the `Options` command (for example, as in Figure 71). A click outside the solid part of the map clears the selections. A `Double-Click` outside the solid part of the map selects all locations.

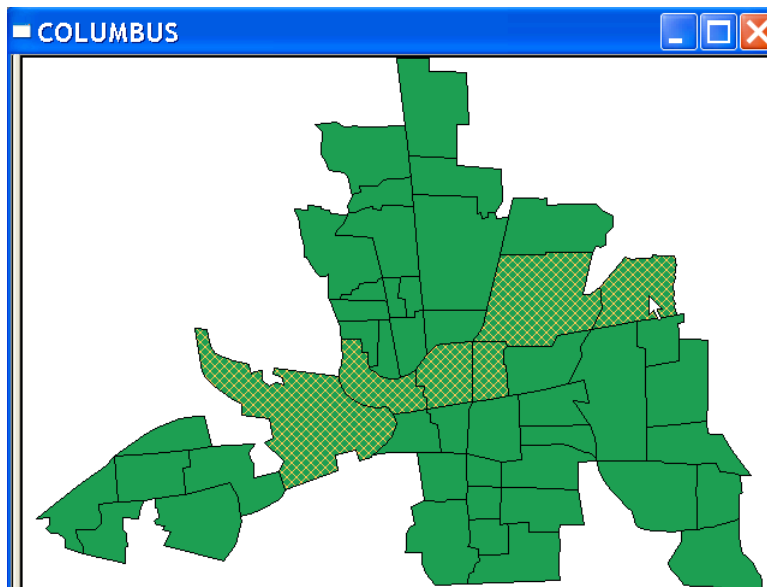


Figure 114. Selected locations on the Columbus neighborhood map using Line Select.

For a `Point` selection, additional items may be added to the selection using `Shift-Click`. More generally, `Shift-Click` will change the selection status of individual locations (turning them on or off). A `Double-Click` on an individual location will select all other locations (i.e., the full map except that location). This may have unexpected effects on individual graphs when the `Exclude Selected` option is on, in the sense that a regression will be computed for a single observation.

Selection in a Histogram

In a `Histogram`, it is not possible to select individual observations, only groups of observations contained in the same category. A category is selected by clicking on its bar, or by clicking on the small square shown next to the breakpoints on the right hand side of the window. Additional categories may be selected by using `Shift-Click`. In general, the selection status of a category can be turned on or off using `Shift-Click`. Selected

categories do not have to be adjoining in the **Histogram**. A **Double-Click** anywhere in the **Histogram** will reverse the selection to its complement (the selected will be unselected and vice versa). The selected category or categories will be highlighted in yellow, as shown in Figure 115 for the **CRIME** variable in the Columbus neighborhood data set (**COLUMBUS.SHP**).

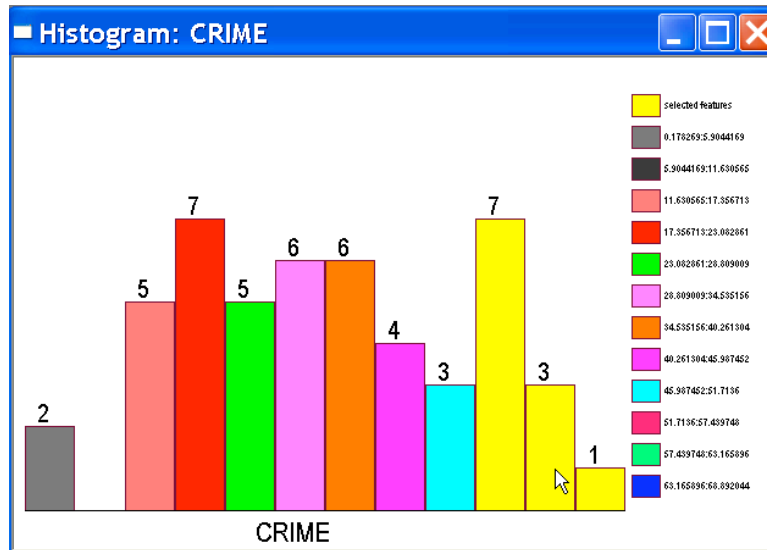


Figure 115. Histogram with three highest categories selected.

When the selection is carried out in the **Histogram**, only whole categories can be highlighted. However, subsets of categories in a histogram may be highlighted as a result of selections in linked windows.

Note that in the current version of GeoDa, you can only undo the selection in a **Histogram** by clicking outside the solid part in a **Map**. A slightly more convoluted way to obtain the same result is to select all categories, followed by **Double-Click** to select the complement (no categories).

Selection in a Box Plot

In a **Box Plot**, individual observations can be selected by clicking on the blue dots that correspond to them or by dragging a rectangle around them. **Shift-Click** will add more selected observations to the current selection. In general, **Shift-Click** will change the selection status of individual observations. The rectangle selection mechanism also allows you to select multiple observations.

The selected observations are highlighted in yellow. For values outside the interquartile range (shown as dots below and above the main purple box in the **Box Plot**), the selection is shown as a yellow dot, for values inside the interquartile range, the selection is a yellow line (see Figure 116 for **HOVAL**). A **Double-Click** selects the complement.

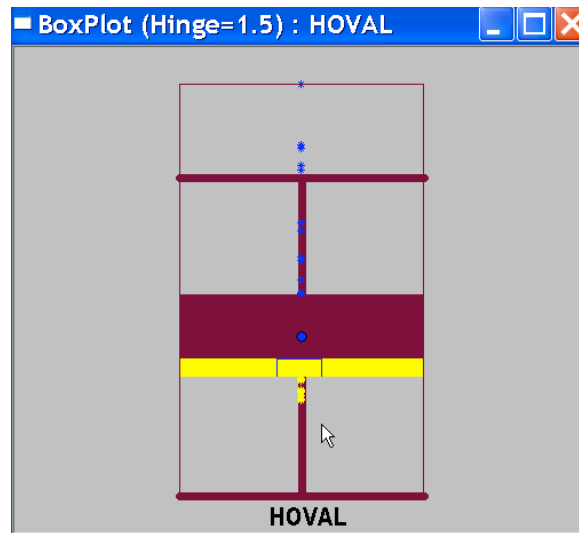


Figure 116. Box plot with selected observations.

Note how in the current version of GeoDa, you can only undo the selection in a **Box Plot** by clicking outside the solid part in a **Map**. A slightly more convoluted way to obtain the same result is to select all observations (drag a rectangle around the whole **Box Plot**), followed by **Double-Click** to select the complement.

Selection in a Scatter Plot

Selection in a **Scatter Plot** operates in much the same way as for the **Box Plot**. Individual observations are selected by clicking on the matching points in the plot. Groups of observations are selected by dragging a rectangle around them. The selected points are highlighted in yellow (default selection color).

Observations can be added to an existing selection by using **Shift-Click** for individual points or by holding down the **Shift** key while dragging a rectangle. In general, the selection status of any single observation or group of observations (using a rectangle) can be switched by means of the **Shift** key.

With the **Exclude Selected** option on, any selection on a **Scatter Plot** results in the regression fit to be recalculated. A new line is drawn in the graph (in brown, the original fit remains blue) and the slope in the new line appears on the right hand side above the graph (Figure 117). The new fit is computed for a data set *without the selected observations*. To get the result for the selection only, you need to **Double-Click** to switch the selection to its complement. The **Exclude Selection** option allows for the visual inspection of the effect of outliers and leverage points on a regression slope.

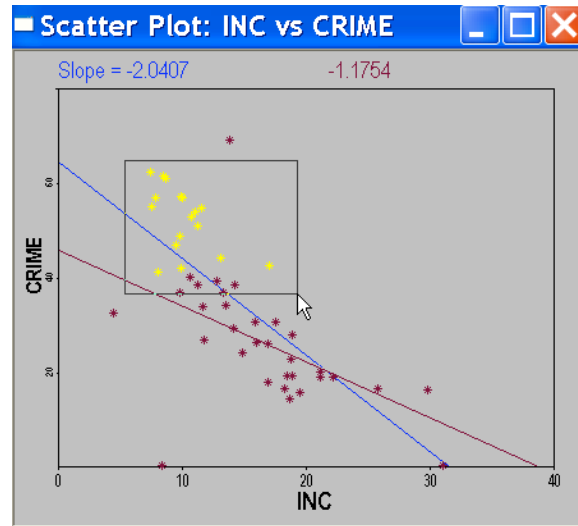


Figure 117. Scatter Plot recomputed with selected observations excluded.

Note that in the current version of GeoDa, you can only undo the selection in a **Scatter Plot** by clicking outside the solid part in a **Map**. A slightly more convoluted way to obtain the same result is to select all observations (drag a rectangle around the whole **Scatter Plot**), followed by **Double-Click** to select the complement.

Selection in a Table

Selection in a **Table** is carried out by clicking on the left side tab in the row corresponding to the selected observation. Multiple rows can be selected by **shift-click** on individual rows or by **click and drag** over the left side column. Selected rows can be promoted as well as obtained as the result of range queries. Details on these options are provided in the section of Manipulating and Editing Tables.

Linking

Any time more than one window is open in a project, selection of observations in any of the windows automatically triggers the same selection in the other windows. This is referred to as *linking*. In Figure 118, a **Box Plot** for **HOVAL** is linked to a **Box Map**, to illustrate how the five outlying observations are connected between the two views.

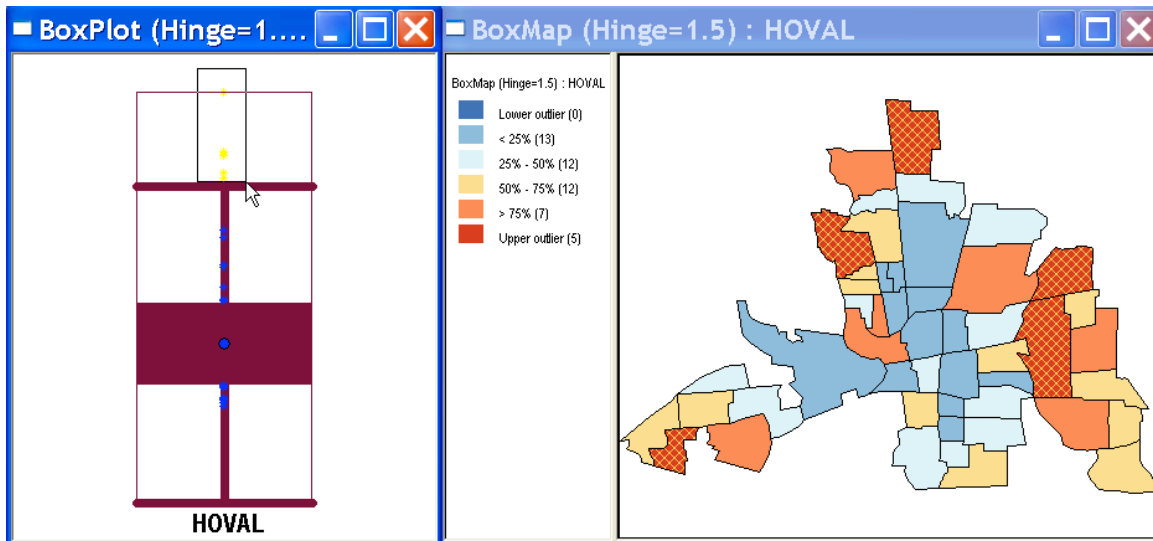


Figure 118. Linked box plot and box map.

Linking is most effective when multiple views are connected that pertain to different variables, allowing a visualization and exploration of the multivariate association between variable. In practice, the number of views open at the same time is limited by screen size. More importantly, the simultaneous consideration of more than 9-11 graphs on a screen becomes a challenge.

Brushing

Brushing is a dynamic implementation of the linking concept. The selection in a given graph or map is continuously updated by moving a “brush” over the observations in a view. In GeoDa, the brush is a rectangle. It is constructed by clicking and dragging the shape while holding down the **CTRL** key. After a few seconds, the rectangle will start to blink, indicating that brushing is now active. The selection is updated when the brush is moved. Brushing is turned off by a single **click** anywhere in the view.

Active brushing works for all graphs (except the histogram) as well as for multiple maps. It is most effective between **Scatter Plots** and **Maps**, especially when the **Exclude Selected** option is set. Moving the brush over a **Scatter Plot** will immediately show how the slope is affected, both in the form of a newly drawn regression line as well as by the changing slope value above the graph. Brushing also affects the selected rows in a **Table**.

Creating and Manipulating Spatial Weights

Spatial weights are essential for the computation of spatial autocorrelation statistics. In GeoDa, they are also used to implement **Spatial Rate** smoothing. Weights can be constructed based on contiguity from polygon boundary files (the original layout or a set of Thiessen polygons), or calculated from the distance between points (points in a point shape file or any **x, y** coordinates in a file).

Opening Existing Weights

Existing spatial weights are added to a project by invoking **Tools > Weights > Open** or by clicking on the **Open weights matrix** toolbar button. This creates a **Select Weight** dialog as shown in Figure 119. When no weights have been opened for the current project, the second radio button is checked by default. This forces you to select an existing weights file in either **GAL** or **GWT** format (see next section). Click on the **Open File** icon to obtain the usual dialog with the contents of the current working directory (your contents will likely differ from those shown in Figure 119). Select the weights file and click on **Open** to add it to the **Select Weight** dialog. Finally, click **OK** to add it to the project, as in Figure 120.

Once a weights file has been opened, it becomes available for any spatial analysis. Invoking the **Tools > Weights > Open** at this point brings up the **Select Weight** dialog with the first radio button checked and the available weights given in a drop down list, as shown in Figure 121. At this point, you can also still check the other radio button to open a new weights file.

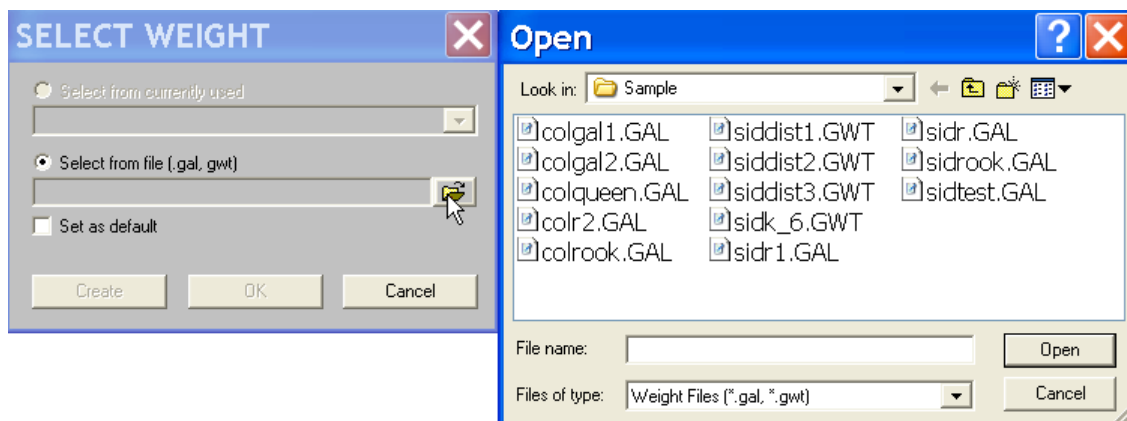


Figure 119. Select weight dialog to open an existing spatial weights file.

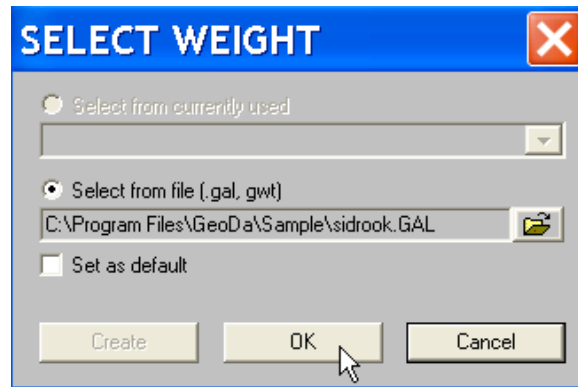


Figure 120. Adding a spatial weights file to the project.

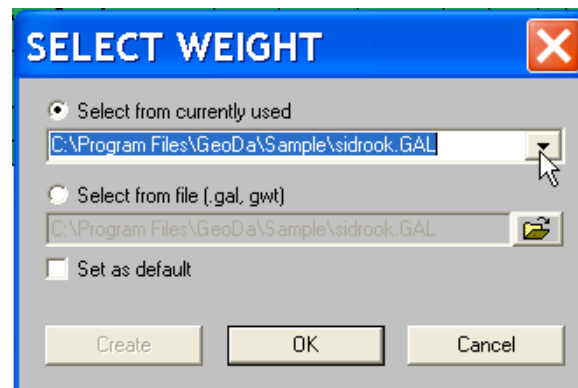


Figure 121. Using an already opened spatial weights file.

Creating New Weights

New spatial weights are constructed by invoking **Tools > Weights > Create** or by clicking on the **Create weights** toolbar button. This starts the **Creating Weights** dialog, as in Figure 122. Before any weights can be constructed, an **Input File** and **output File** must be specified, as well as a **Key Variable**. The former is selected by clicking on the open file icon in the dialog and selecting a shape file (select the file in the **open File** dialog and click the **open** button). The output file must be specified by clicking on the save icon and entering the file name in the text box of the **save As** dialog, followed by clicking on the **save** button (as shown in Figure 123). Once the two files are specified, their path names will be listed in the text box of the **Creating Weights** dialog. In addition, all the options in the **Creating Weights** dialog become active, that is, they change from being grayed out to being visible.

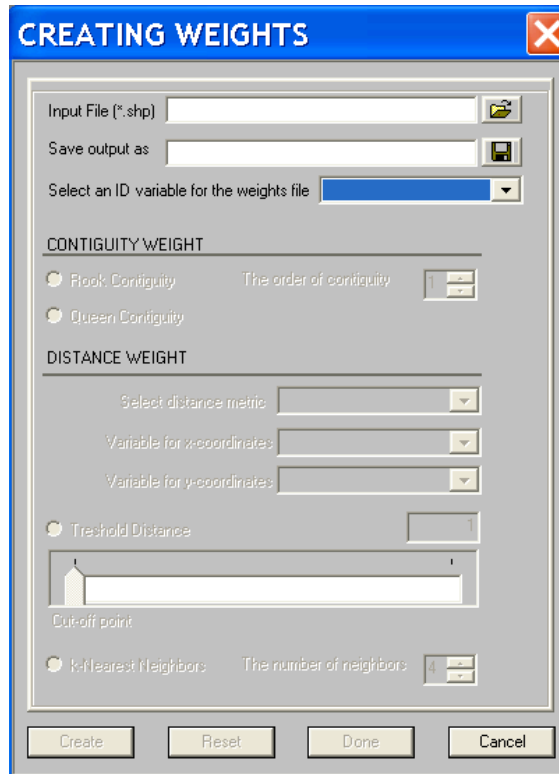


Figure 122. Creating weights dialog

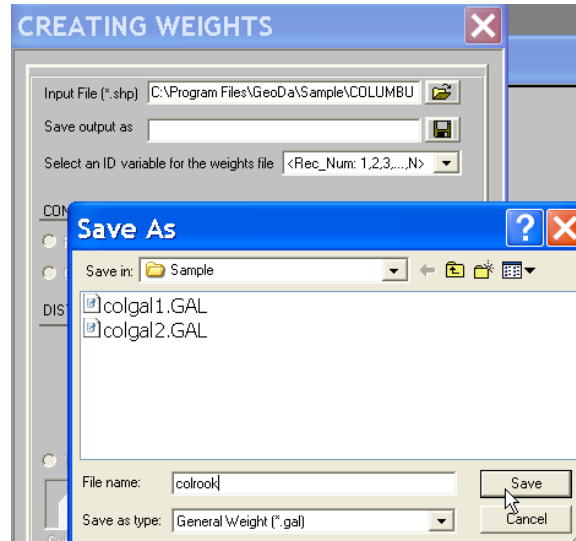


Figure 123. Output file specification for spatial weight

In addition to the input and output files, a **key variable** must be specified that corresponds to the **ID** values used in the weights files (Figure 124). This can be the same **key variable** as used in the project, but this is not required, as long as the variable takes on a unique value for each observation. If no **key variable** is specified, the default is

the sequence number of the observation. Using this default is usually not a good idea, since the order of the observations may be different between different file formats (e.g., shape file vs. ascii output file).

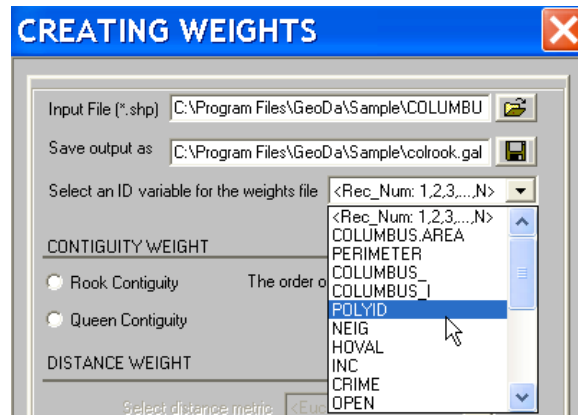


Figure 124. Key Variable specification

Spatial Weights File Formats

The spatial weights are saved as text files with a **GAL** file extension (for contiguity) or a **GWT** file extension (for distance). They can be readily used by a number of other software packages as well. The **GAL** file format is illustrated in Figure 125 for the “rook” contiguity of North Carolina counties. On the left, the first part of a file is shown that results when the **Key Variable** is set to **FIPSNO**, while on the right the file is illustrated for the case where *no Key Variable* was specified. The main difference between the two files is in the first line, the so-called header line. When a **Key Variable** is specified, that line contains four values: 0 (reserved for future use), the number of observations (100), the name of the shape file (**SIDS**) and the variable name for the **Key Variable** (**FIPSNO**). When sequence numbers are used to label the observations, the header line only contains the number of observations, as in the right hand panel in Figure 125.

After the first line, the structure of the two **GAL** files is identical. For each observation, there are two lines with information. The first contains the observation **ID** and the number of neighbors. The second line contains the **ID** values for the neighboring locations. Note how in the left hand panel, the FIPS codes are used, while in the right panel simple sequence numbers appear. **GAL** files can be edited with any text editor. The numeric values are separated by a single space, which can be replaced by another separator through a global **Find-Replace** command in a text editor.

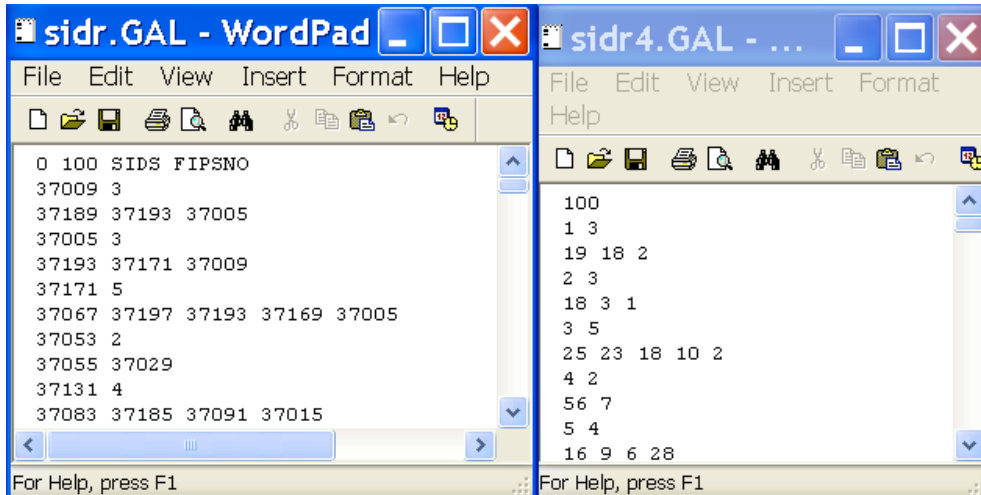


Figure 125. GAL format spatial weights file for North Carolina counties.

The **GWT** format for spatial weights is used when the “neighbor” information is derived from the distance between points. Again, the result is a text file with a header line followed by the neighbor information. When a **Key Variable** has been specified, the header line is as in Figure 126, for k-nearest neighbors of order 4 in the Columbus data set. It contains four items: 0 (for future use), the number of observations (**49**), the name of the shape file (**COLUMBUS**) and the **Key Variable** (**POLYID**). When no **Key Variable** is specified, but sequence numbers are used, the header line consists only of the number of observations. The remainder of the file contains, on each line, the “origin” **ID** value (i.e., the **ID** for the observation), the “destination” **ID** (i.e., the **ID** for the neighbor), and the distance on which the contiguity definition was based.

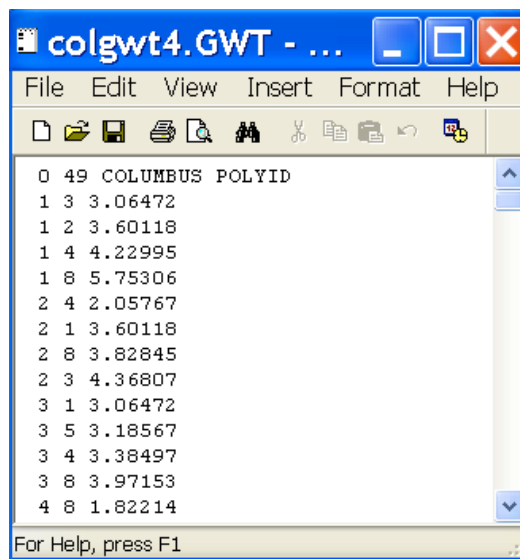


Figure 126. GWT format spatial weights file for 4th order nearest neighbors in Columbus.

Contiguity Based Spatial Weights

Contiguity based spatial weights can be created when the input file is specified as a polygon shape file. After both input and output files are specified, both weights options in the Creating Weights dialog become active. For the **CONTIGUITY WEIGHT** option, a choice is available between **Rook Contiguity** and **Queen Contiguity**. One of these radio buttons selects the type of contiguity criterion (see Figure 127 for the use of the **Rook Contiguity**). **Rook Contiguity** uses only common boundaries to define neighbors, while **Queen Contiguity** includes all common points (boundaries and vertices) in the definition. Spatial weights based on **Queen Contiguity** therefore always have a denser connectedness structure (more neighbors).

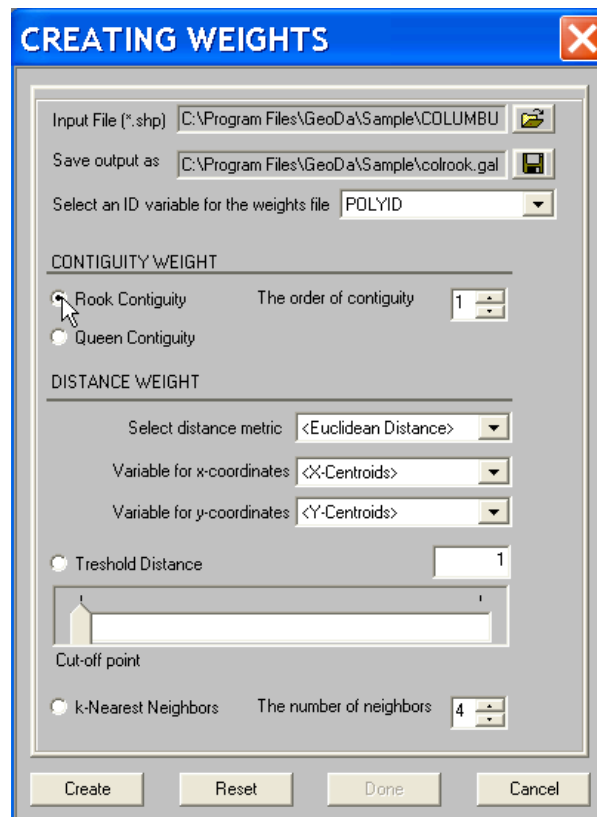


Figure 127. Construction of spatial weights based on rook contiguity.

After selecting one of the two contiguity types, click on **create** to construct the weights. A message will indicate successful completion (Figure 128). Click on **done** to remove the dialog.

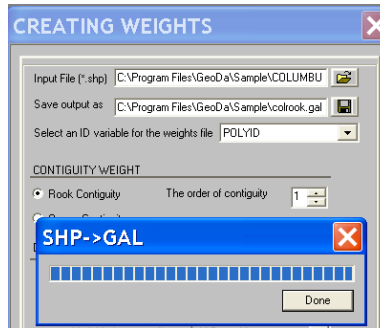


Figure 128. Completion of weights construction

Higher Order Contiguity Weights

Spatial weights need not be limited to first order contiguity, but higher order weights can be constructed as well. In the **Creating Weights** dialog, change the **order** default value of 1 to any higher order (Figure 129). The remainder of the procedure is identical to that for first order contiguity. The higher order contiguity weights are based on the algorithm by Anselin and Smirnov (1996) that removes redundancies and circularities in the weights construction.

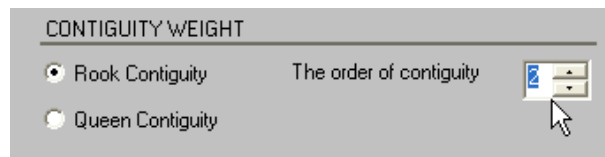


Figure 129. Specifying a higher order of contiguity.

Distance Band Spatial Weights

When a shape file is specified as the **Input File**, or when **x, y** coordinates are available as fields in a data table, the spatial weights can be derived from the distance between these points. This is carried out in the second panel of the **Creating Weights** dialog (see Figure 130). The distance option requires input of the type of **distance metric** (**Euclidean** or **Arc Distance**), the variable holding the **x-coordinate** and the variable holding the **y-coordinate**. Note that these need not be actual “coordinates” but can be any two variables in the data table. For a shape file, the coordinates need not be made explicit. The default is to use the **x-y** coordinates in a point shape file and the polygon centroids in a polygon shape file. It is important to use **Euclidean Distance** only for projected maps and **Arc Distance** for coordinates in unprojected latitude-longitude (note that the arc distance computed by GeoDa is approximate).

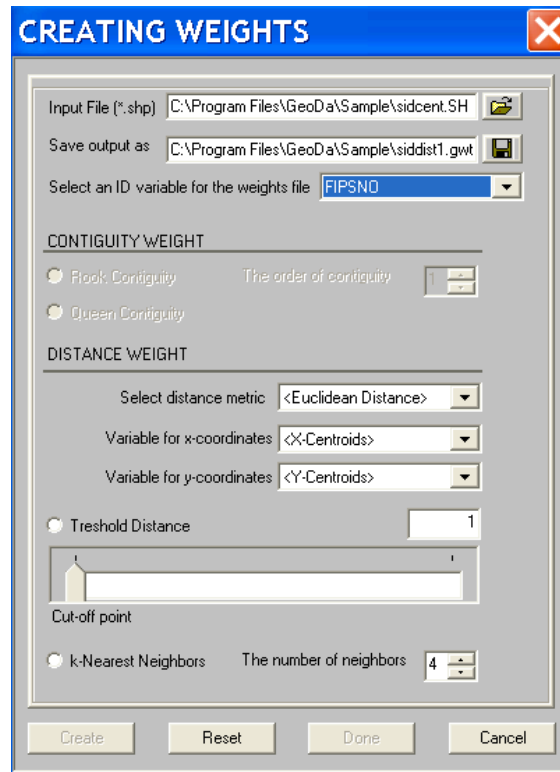


Figure 130. Distance weights creation.

In order to construct a spatial weights file based on a distance band, the **Threshold Distance** button must be checked and a critical distance specified. GeoDa internally computes the minimum distance required to assure that each observation has at least one neighbor. This distance appears in the text box when the slider is clicked. This is shown in Figure 131 for the SIDS centroids shape file (**sidcent.shp**) and with **Arc Distance** selected as the method. Moving the slider to the right increases the cut off distance. The slider must be activated before a weights file can be created.

Click on **create** to build the file and a progress bar appears indicating successful completion, similar to Figure 128. Click on **Done** to close the dialog. Note that the scale of the distance shown depends on the scale and projection for the coordinates used in the input shape file and may not be in any meaningful units.

Using the distance band criterion, all points are selected that are within the specified distance from the observation under consideration.

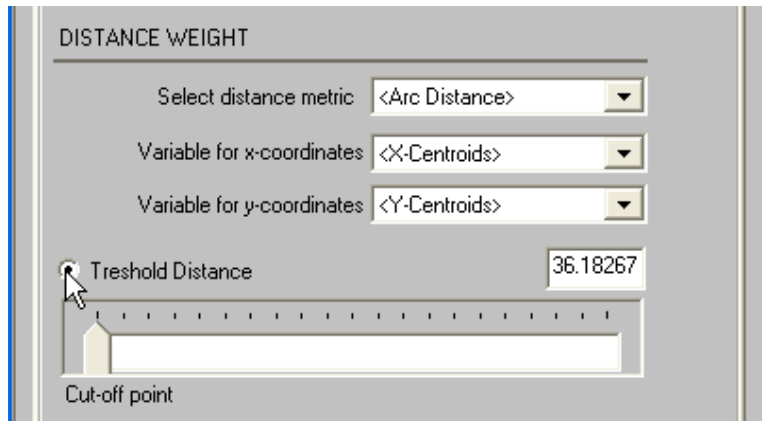


Figure 131. Specifying the threshold distance for distance band spatial weights.

K-Nearest Neighbor Spatial Weights

Spatial weights based on a simple **Distance Threshold** criterion often lead to a very unbalanced connectedness structure. For example, this is the case when the spatial units have very different areas, resulting in the smaller units having many neighbors, while the larger ones may have very few (or none, yielding “islands”). A commonly used alternative consists of considering the **k-Nearest Neighbors**. This is the second option in the **DISTANCE WEIGHT** section of the **Creating Weights** dialog.

K-Nearest Neighbor weights are constructed by checking the appropriate radio button and specifying the order. The default order is 4, but alternatives are readily specified in the dialog, as shown in Figure 132. Again, click on **create** to start building the weights file and a progress bar will indicate successful completion. Click on **Done** to close the dialog.

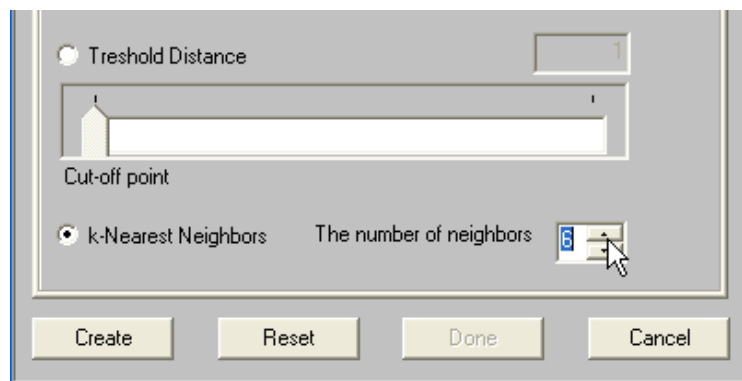


Figure 132. Specifying the order for k-nearest neighbors spatial weights.

Characteristics of Spatial Weights

The **Tools > Weights > Properties** command or the **Weights Characteristics** toolbar button creates a histogram of the frequency distribution of the number of neighbors in a spatial weights file. The command brings up a **Weights Characteristics** dialog in which the weights file must be specified. Clicking on the open file button brings up the usual **Open File** dialog. Spatial weights with either **GAL** or **GWT** file extensions may be specified, as illustrated in Figure 133. Select the file and click on **open** to complete the dialog, then click on the **OK** button in the **Weights Characteristics** dialog to create the histogram.

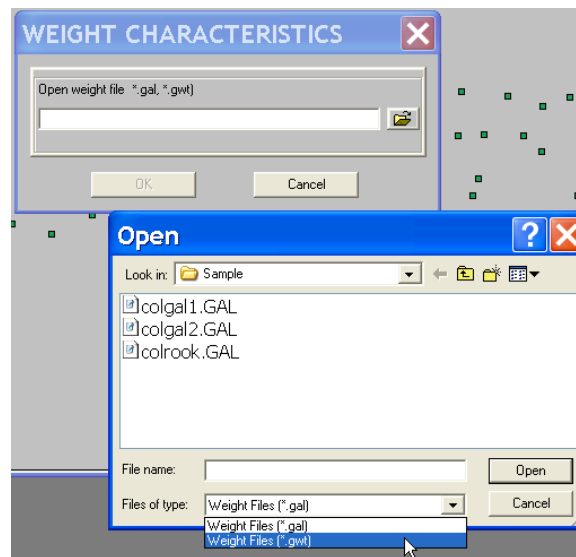


Figure 133. Selecting a weights file for the analysis of its characteristics.

The **Connectivity** histogram shows the number of observations (on top of each bar) by number of neighbors (the numbers in the **Histogram** legend on the right). You may need to change the default number of categories in the histogram to suit the distribution of contiguities (use **Options > Intervals**).

The **Connectivity Histogram** has all the properties of a standard **Histogram** and can be linked to the other views in a project. For example, in Figure 134, the North Carolina counties that have two neighbors are shown by selecting the right most histogram category. Conversely, by selecting one or more counties in the map, as in Figure 135, one can find how many neighbors they have as defined in a given weights file.

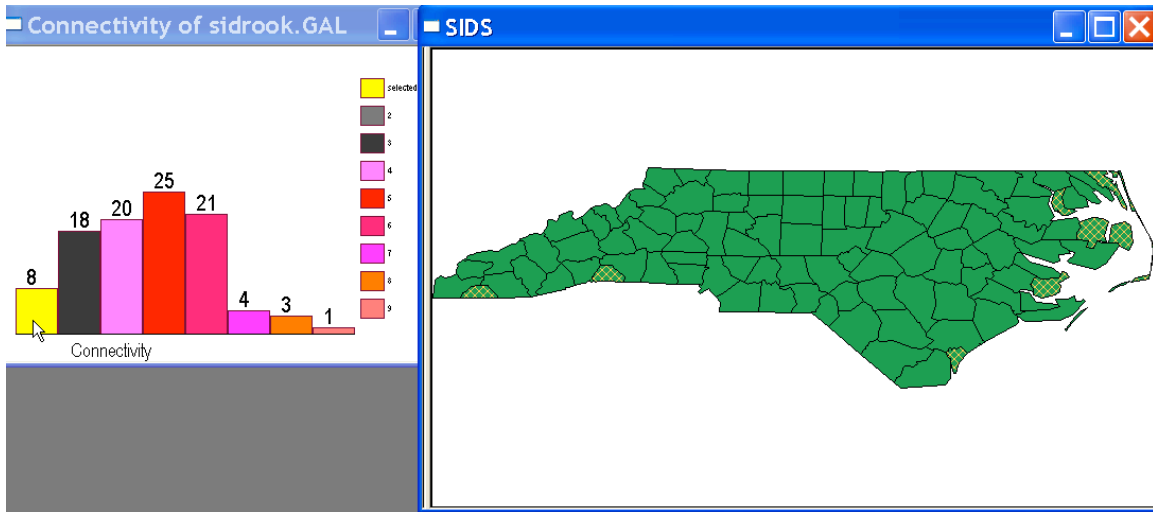


Figure 134. Connectivity of first order contiguity for North Carolina counties.

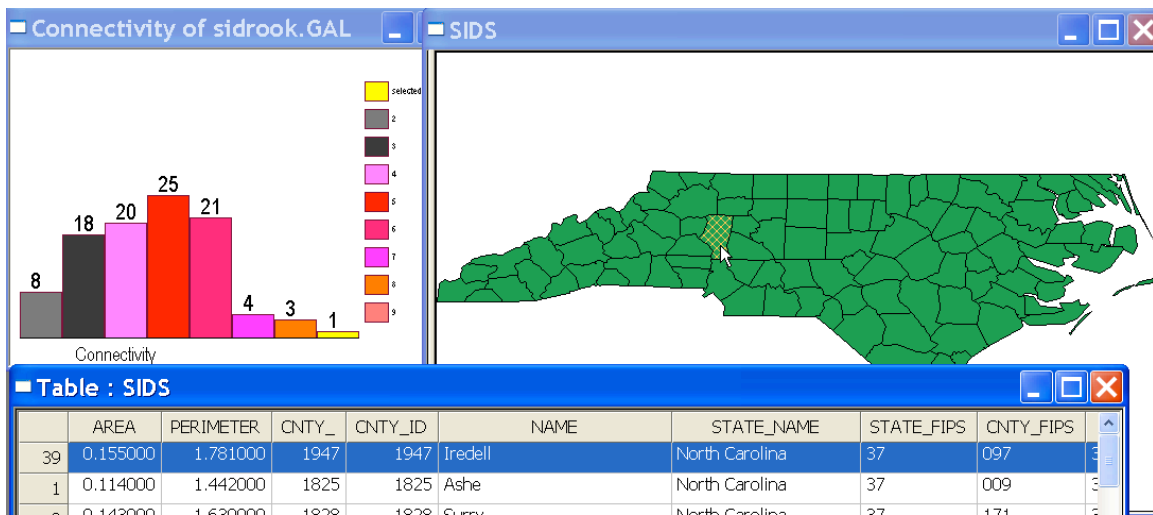


Figure 135. Rook connectivity for Iredell county, NC.

Global Spatial Autocorrelation

Global spatial autocorrelation analysis is handled in GeoDa by means of Moran's I spatial autocorrelation statistic and its visualization in the form of a Moran Scatter Plot (Anselin 1995, 1996). The Moran Scatter Plot is a special case of a **scatter plot** and as such has the same basic options. It is linked to all the graphs and maps in the project, allowing full brushing.

Global spatial autocorrelation analysis requires that both a variable be specified (using the **Variables Settings** dialog) as well as a spatial weights file. The latter is specified through a dialog that asks to select either a new file from disk, or one from a list of currently in use spatial weights (as the result of an earlier **Tools > Weights > Open** command), as in Figure 136.

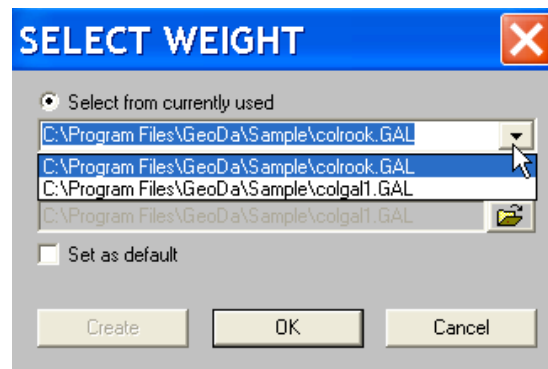


Figure 136. Weights selection dialog

Spatial autocorrelation analysis is implemented in its traditional univariate form, as well as in a bivariate form (Anselin et al. 2002a).

Univariate Moran Scatter Plot

The **Univariate Moran Scatter Plot** is invoked by **Explore > Univariate Moran**, or by clicking the **Univariate Moran** button on the **Explore** toolbar. After the variable of interest and a spatial weights file are specified (or using the default), a window is created with a scatter plot that shows the spatial lag of the variable on the vertical axis and the original variable on the horizontal axis, as shown in Figure 137 for the Columbus **CRIME**

data using a rook-based contiguity file.

The variables are standardized so that the units in the graph correspond to standard deviations. The four quadrants in the graph provide a classification of four types of spatial autocorrelation: *high-high* (upper right), *low-low* (lower left), for positive spatial autocorrelation; *high-low* (lower right) and *low-high* (upper left), for negative spatial autocorrelation. The slope of the regression line is Moran's I, listed at the top of the graph (in blue). The weights file used to compute the statistic is listed in parentheses (`colrook.GAL` in Figure 137).

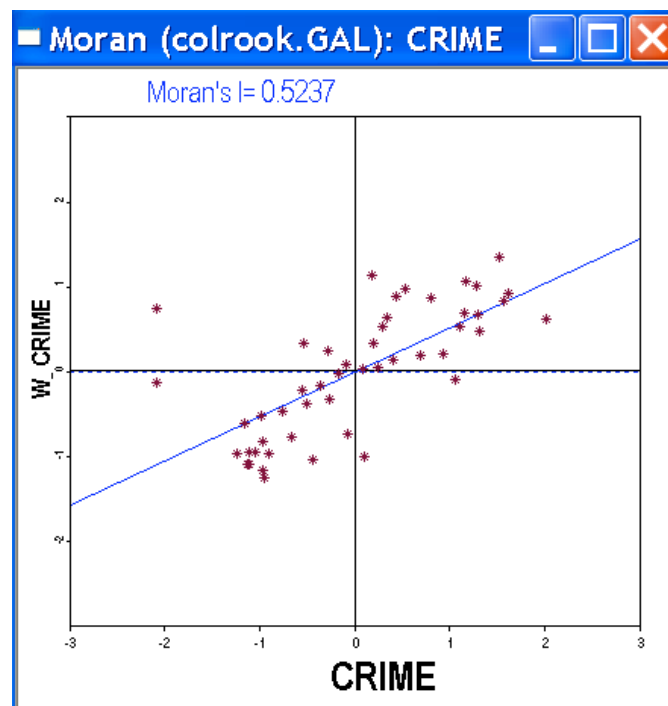


Figure 137. Moran Scatter Plot for Columbus **CRIME**.

Options > Exclude Selected ON

The **Univariate Moran** has several options, invoked by using the **Options** menu or by right clicking on the graph, as illustrated in Figure 138. The **Exclude Selected** option works in the same fashion as for the standard Scatter Plot. When the **Exclude Selected** option is active (**ON**), the selection of observations in the graph (or, through linking, the selection in any other graph) will result in the recalculation of Moran's I for a layout *without* the selected observations. The new regression line is shown in brown and the corresponding Moran's I is listed above the graph on the right hand side. For example, in

Figure 139, the two observations highlighted in yellow (on the left hand side of the graph) have been selected. The value of Moran's I on the right, corresponding to the slope of the brown line is that for the dataset without the selected observations.⁵ The slope of the blue line is for the complete data set.

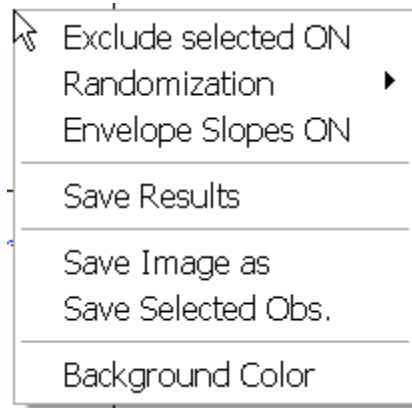


Figure 138. Univariate Moran scatter plot options.

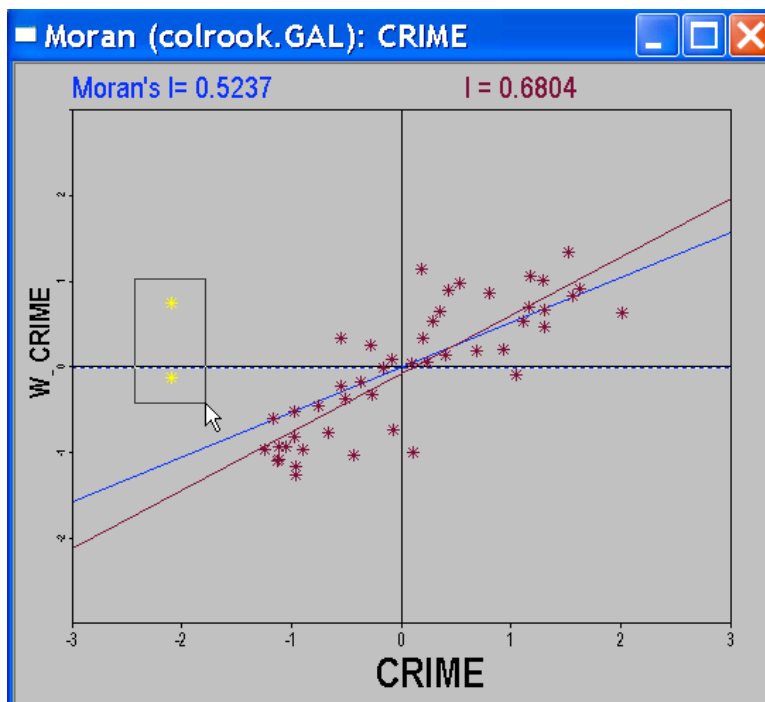


Figure 139. Univariate Moran scatter plot with selected observations excluded.

⁵ While the observations are excluded from the calculation, the spatial weights are not reconstructed, but use a subset from the weights for the complete data set. In Figure 139, the background color has been changed to gray.

Options > Randomization

Inference for Moran's I is based on a permutation approach, in which a reference distribution is calculated for spatially random layouts with the same data (values) as observed. The randomization uses an algorithm to generate spatially random simulated data sets outlined in Anselin (1986). It is invoked by **Options > Randomization > xxx**, or by right clicking anywhere in the graph window. The **xxx** is the number of random permutations used in constructing the reference distribution (such as 99, or 999), as shown in Figure 140.

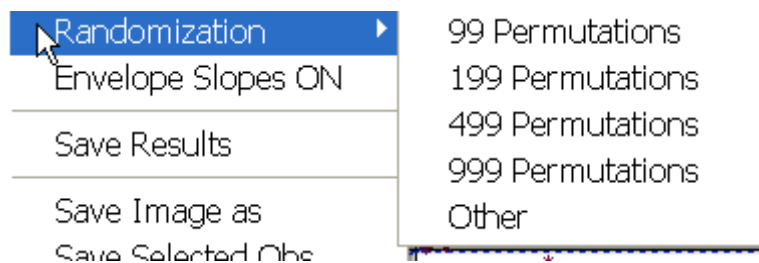


Figure 140. Randomization option in Moran scatter plot.

The result is a window depicting a histogram for the reference distribution, with the observed Moran's I shown as a yellow bar, as in Figure 141 for the Columbus variable **CRIME**. In addition to the histogram, a small number of summary statistics are listed as well. In the top left corner, the number of permutations used (999) and the pseudo-significance level (0.001) are given. The latter is computed as the ratio of the number of statistics for the randomly generated data sets that are equal to or exceed the observed statistic + 1, over the number of permutations used + 1. Hence, the value of 0.001 in Figure 141 indicates that none of the simulated values were larger than the observed 0.52. In the bottom left of the graph, the value for Moran's τ is listed, its theoretical mean ($E[\tau]$), the average of the reference distributions (**Mean**) and the standard deviation for the reference distribution (**sd**). You can generate another set of simulated values by clicking on the **Run** button. Clicking **close** removes the **Randomization** window.

An example of a reference distribution where the statistic does not turn out to be significant is shown in Figure 142, for the Columbus variable **OPEN**: Moran's I is -0.067, with pseudo-significance of 0.31. Note how the mean computed for the reference distribution (-0.0181) is slightly different from its theoretical value (-0.0208).

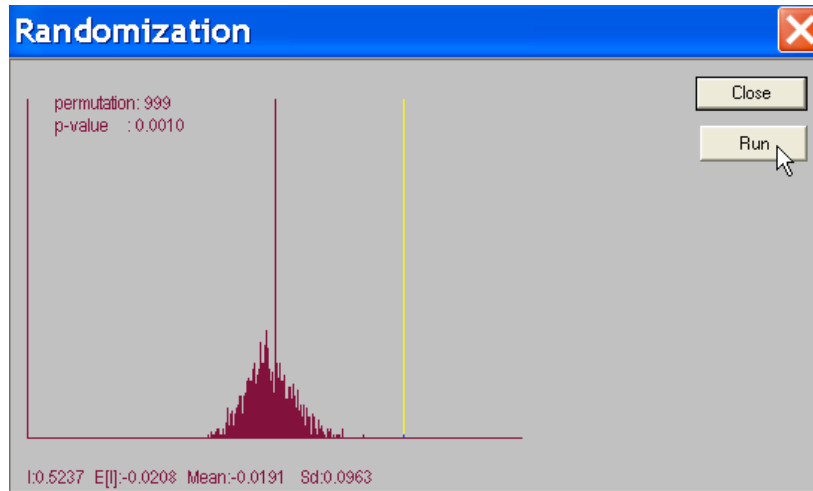


Figure 141. Reference distribution for Moran's I for Columbus **CRIME**.

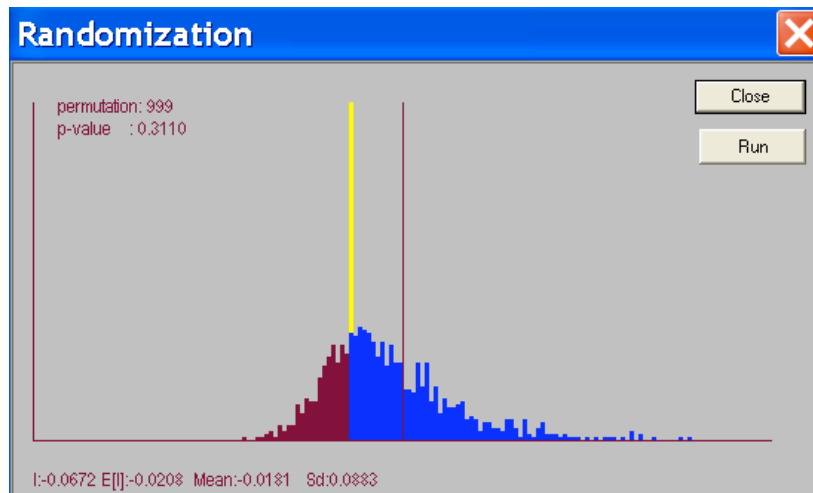


Figure 142. Reference distribution for Moran's I for Columbus **OPEN**.

Options > Envelope Slopes ON

A different way to visualize the range of autocorrelation statistics that can be obtained in spatially random simulated data sets is shown with **Options > Envelope Slopes ON**. As for the other options, it is invoked by using the **options** menu or by right clicking anywhere in the scatter plot window. This plots the lower 5th percentile and upper 95th percentile of the reference distribution as the slope of dashed lines in the Moran Scatter Plot. In Figure 143, this is shown for the Columbus **CRIME** data, illustrating the degree of extremeness of the observed statistic.

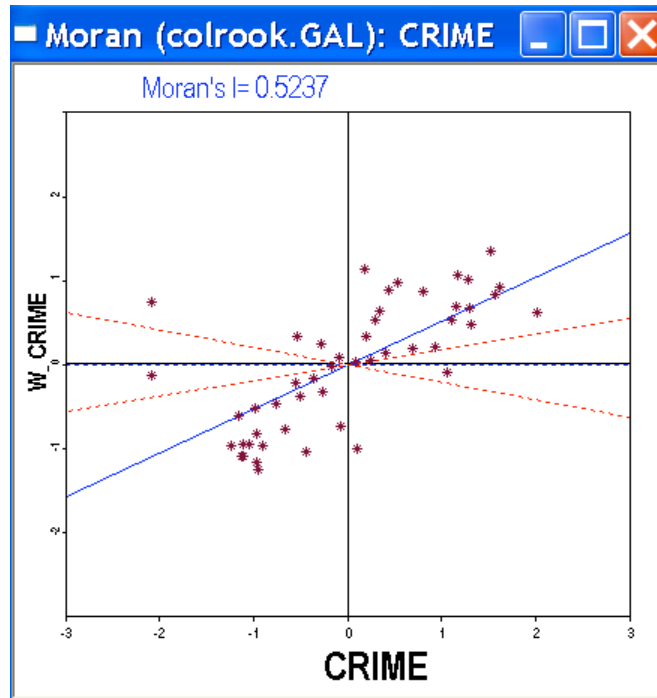


Figure 143. Envelope slopes in the Moran Scatter plot

Options > Save Results

The **Save Results** option is invoked from the menu as **Options > Save Results**, or by right clicking in the Moran scatter plot window. It allows you to add the standardized variable and its spatial lag to the data table for use in other analyses. Selecting this option brings up a dialog (Figure 144) to choose the variable to be added to the table by clicking on the corresponding check box and selecting a variable name. The default variable names are **STD_VAR** and **LAG_VAR**, where **VAR** is the variable name used in the analysis. Click **OK** to add the variables to the table, as shown in Figure 145. As before, the addition is not permanent until the table has been **Saved As** a different file.

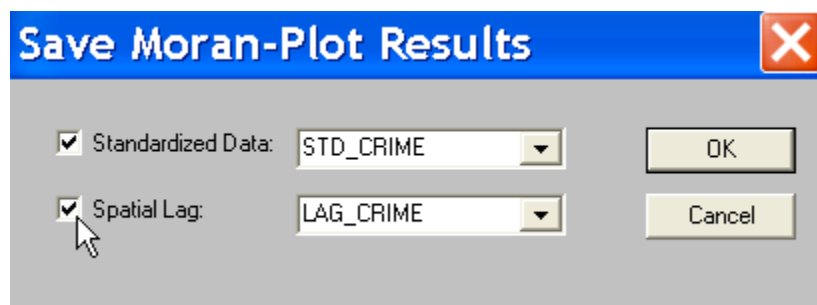


Figure 144. Save results dialog in Moran scatter plot.

STD_CRIME	LAG_CRIME
-1.159619	-0.622430
-0.975794	-0.530835
-0.269066	-0.341683
-0.163821	-0.028828
0.932501	0.196220
-0.541604	0.328404

Figure 145. Moran scatter plot variables added to the data table.

Other Univariate Moran Options

As shown in Figure 138, the Moran Scatter plot also has three other common options: **Save Image as**, **Save Selected Obs.** and **Background Color**. These allow the scatter plot to be saved to a bitmap file, to add an indicator variable for the selected observations to the data table and to change the background color for the graph. For example, in Figure 139, the background color was changed to gray. These options work in the same way as described before.

Bivariate Moran Scatter Plot

A bivariate Moran's I is invoked by **Explore > Multivariate Moran** or by clicking on the **Multivariate Moran** toolbar button. This creates a scatter plot with the spatial lag of the first variable on the vertical axis and the second variable on the horizontal axis. Both variables are standardized internally (such that their mean is zero and variance one), and the spatial lag operation is applied to the standardized variables.

The slope of the regression line shows the degree of linear association between the variable on the horizontal axis and the values for the variable on the vertical axis at *its neighboring locations* (as defined by a spatial weights file). For example, in Figure 146, the spatial lag of **CRIME**, **W_CRIME**, is on the vertical axis and the variable **INC** on the horizontal axis. Inference is carried out by invoking **Options > Randomization**. As for the **Univariate Moran**, this computes a histogram for the reference distribution. The Options are the same as for the **Univariate Moran** (see Figure 138 and the discussion immediately afterwards).

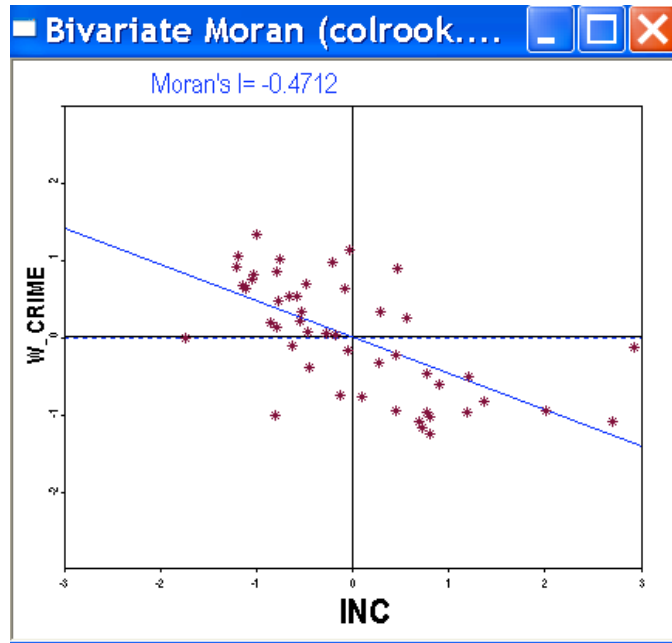


Figure 146. Bivariate Moran Scatter plot.

Moran Scatter Plot Matrix

Univariate and bivariate Moran scatter plots may be computed for a number of variables and arranged as a scatter plot matrix, as in Figure 147 for the Columbus variables **CRIME** and **INC**. On the bottom axis are the variables under consideration (all in standardized units), on the vertical axis the spatially lagged variables (with the spatial lags applied to the standardized variables). This provides an overview of the spatial patterning of each variable with itself as well as with the spatial lags of the other variables. For example, on the left hand side of Figure 147, the correlation between **CRIME** and neighboring **CRIME** is positive (top panel), whereas the correlation between **CRIME** and neighboring **INC** is negative.

Other combinations are insightful as well. For example, the relationship between **INC** and **w_CRIME** (top right panel in Figure 147) can be compared to the usual correlation between **INC** and **CRIME** (left panel in Figure 148), as well as to the correlation between **w_CRIME** and **INC** (right hand panel in Figure 148). The latter is constructed as a regular scatter plot *after* the variables **STD_CRIME** and **LAG_INC** were added to the data table. In contrast to the **Univariate Moran**, it is legitimate to use the spatially lagged variable on the x-axis in a **Bivariate Moran**, since ordinary least squares remains an unbiased estimator for the slope of the regression line (this is *not* the case in a regression of **CRIME** on **w_CRIME**).

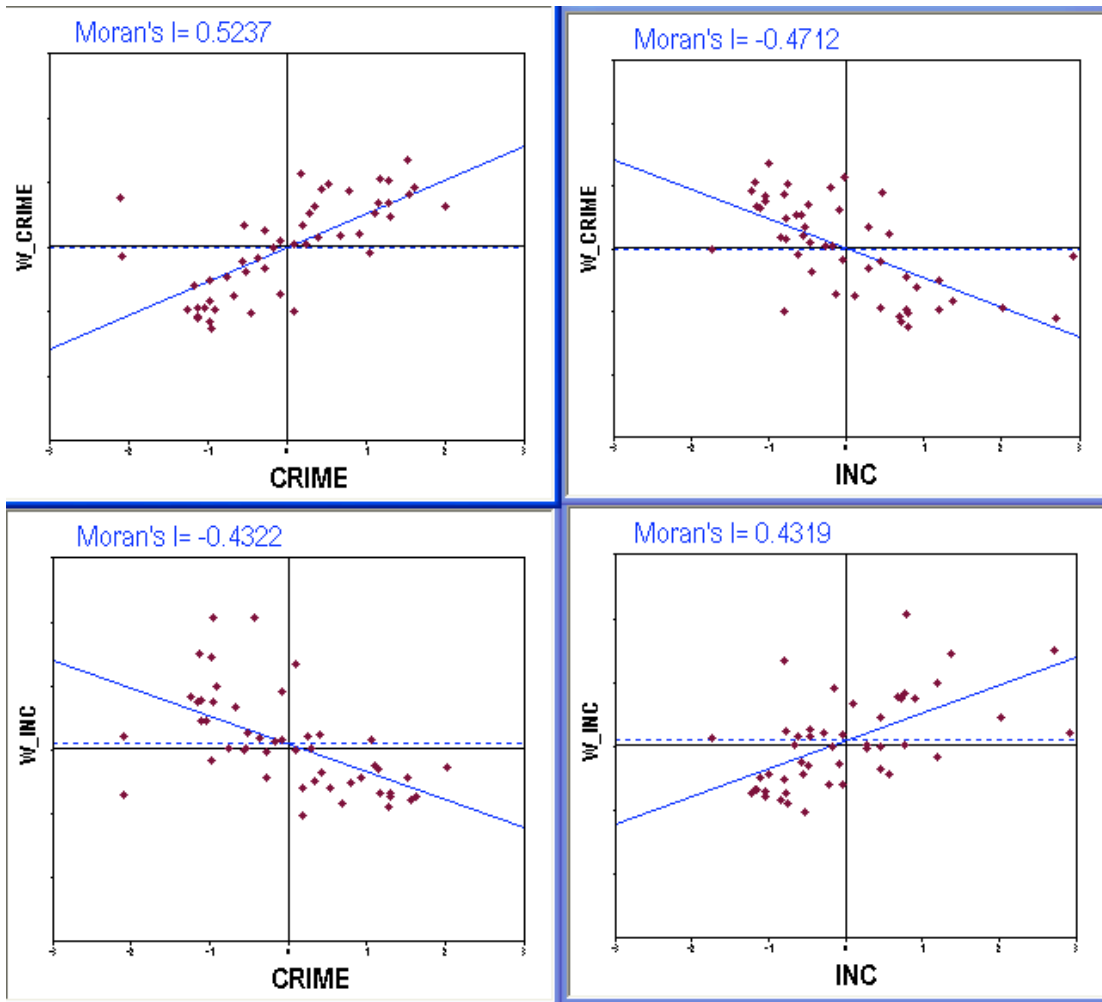


Figure 147. Moran Scatter plot matrix

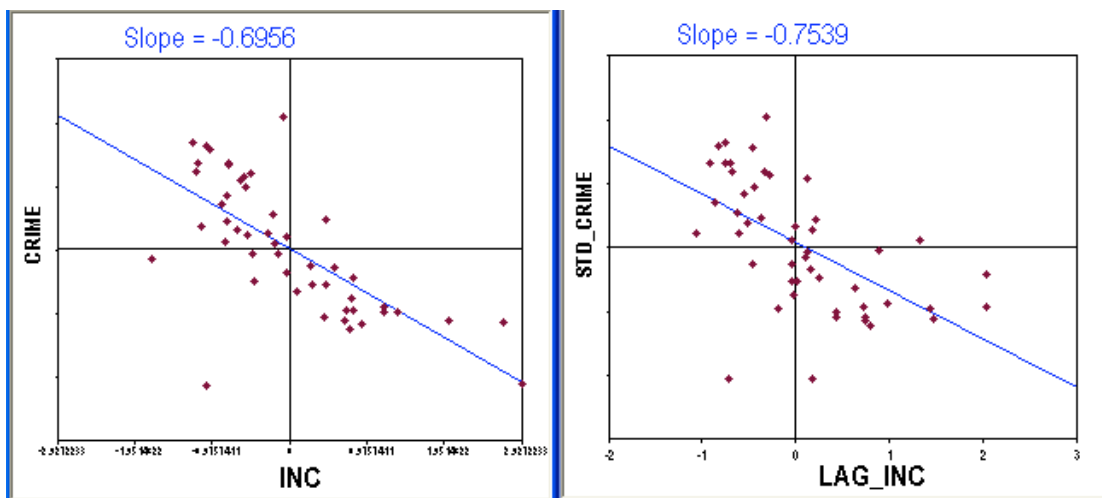


Figure 148. Non-spatial correlation matrices

Moran Scatter Plot for EB Rates

The problem with variance instability for rates or proportions, which served as the motivation for applying smoothing techniques to maps may also affect the inference for Moran's I test for spatial autocorrelation. GeoDa implements the adjustment procedure of Assuncao and Reis (1999) which uses a variable transformation based on the Empirical Bayes principle. This yields a new variable that has been adjusted for the potentially biasing effects of variance instability due to differences in the size of the underlying population at risk.

The EB standardized autocorrelation statistics is invoked as **Explore > Moran's I with EB rate** or by clicking on the matching toolbar button. This brings up a dialog, similar in appearance to the one used in map smoothing: the **Event Variable** and **Base Variable** must be specified, as shown in Figure 149 for the **SIDS** data set.

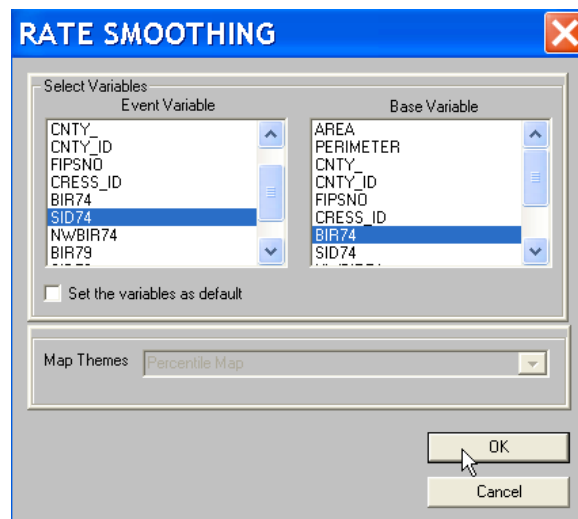


Figure 149. EB Moran's I variable selection dialog.

As for the standard Moran scatter plot, you next need to specify a spatial weights file. The EB Moran Scatter Plot appears as in Figure 150. Note that the actual variable names are not listed on the x and y axes, but instead a generic **RATES** and **w_RATES** is used. The variables names for the **Event** and **Base** are listed at the top of the window. Note that the EB standardized rates in this analysis are not the same as the EB smoothed rates used in the map smoothing (for technical details, compare the Marshall and Assuncao and Reis references). The EB Moran Scatter Plot has the same options as the standard Moran Scatter Plot. They are invoked from the menu or by right clicking in the window.

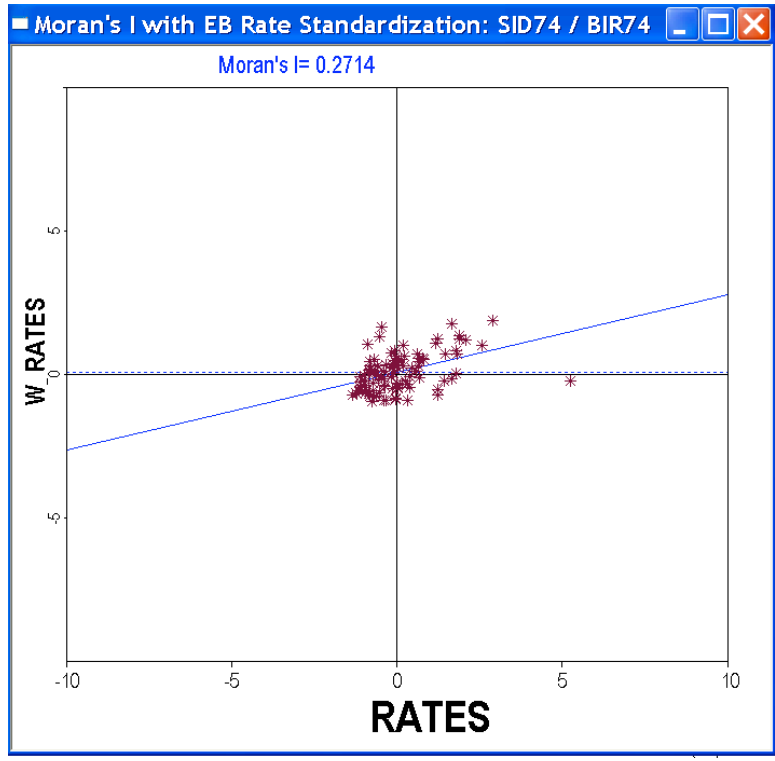


Figure 150. Moran scatter plot for EB standardized rates.

The **Save Results** option allows you to add the standardized rate and its spatial lag to the data table. The dialog is similar to that for the standard Moran scatter plot, except that the default variable names are different. As shown in Figure 151, the default for the **Standardized Data** is **STD_RATES**, whereas the default for the **Spatial Lag** is **LAG_RATES**. As always, the addition of the new variables to the data table is not permanent until the table has been saved as a new file.

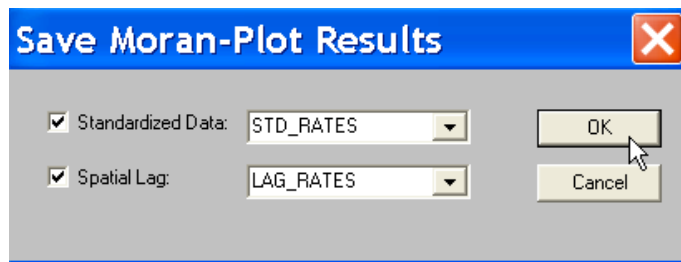


Figure 151. EB Moran save results variable specification dialog.

Local Spatial Autocorrelation

Local spatial autocorrelation analysis is based on the Local Moran **LISA** statistics (Anselin, 1995). This yields a measure of spatial autocorrelation for each individual location. Both **Univariate LISA** as well as **Multivariate LISA** are included in GeoDa. The latter is based on the same principle as the **Bivariate Moran's I**, but is localized. In addition, the **LISA** can be computed for **EB Standardized Rates**. The input needed for the **LISA** statistics is the same as for the global spatial autocorrelation statistics. First, one or two variable names (or **Event** and **Base** for rates) must be selected. Next, a spatial weights file must be specified (for details, see Global Spatial Autocorrelation).

Univariate LISA

Univariate LISA statistics are invoked by **Explore > Univariate LISA** or by clicking on the **Univariate LISA** button on the **Explore** toolbar. This brings up a dialog that lets you specify which of the four output options you want to generate, as shown in Figure 152. The most relevant of these options are **The Significance Map** and **The Cluster Map**, which are unique to the **LISA** functionality. **The Box Plot** and **The Moran Scatter Plot** are identical to their standard counterparts, with the only difference that the **Box Plot** pertains to the distribution of the Local Moran statistics. **Box Plot** and **Moran Scatter Plot** also have the same options as their standard counterparts.

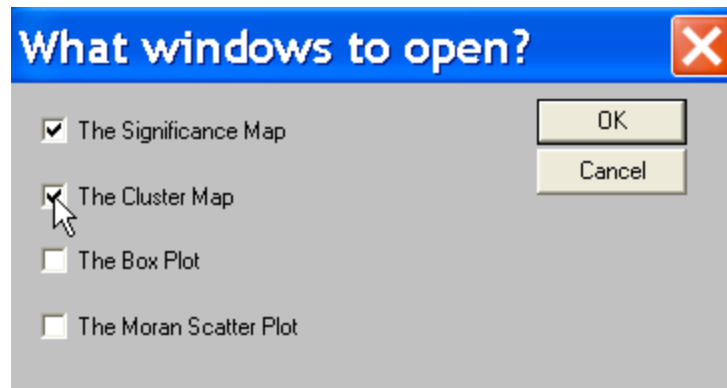


Figure 152. Dialog for LISA windows.

Significance Map

The **Significance Map** is selected by clicking on the matching check box in the dialog

of Figure 152. The result is a special choropleth map showing those locations with a significant Local Moran statistic as different shades of green, depending on the significance level. For example, in Figure 153, the **Significance Map** is shown for the **CRIME** variable in the Columbus data set, using rook contiguity. The map you obtain may be slightly different, since the results are derived from a randomization procedure that may yield slightly different significance levels, depending on the number of replications used in the randomization procedure (see also below). Four significance levels are shown, $p < 0.05$, $p < 0.01$, $p < 0.001$, $p < 0.0001$. As always in a randomization procedure, the “highest” level of significance that can be obtained depends on the number of replications. For example, with only 99 replications, the two more extreme significance levels will never appear. The **Significance Map** can be linked and brushed in the same way as any other map in GeoDa.

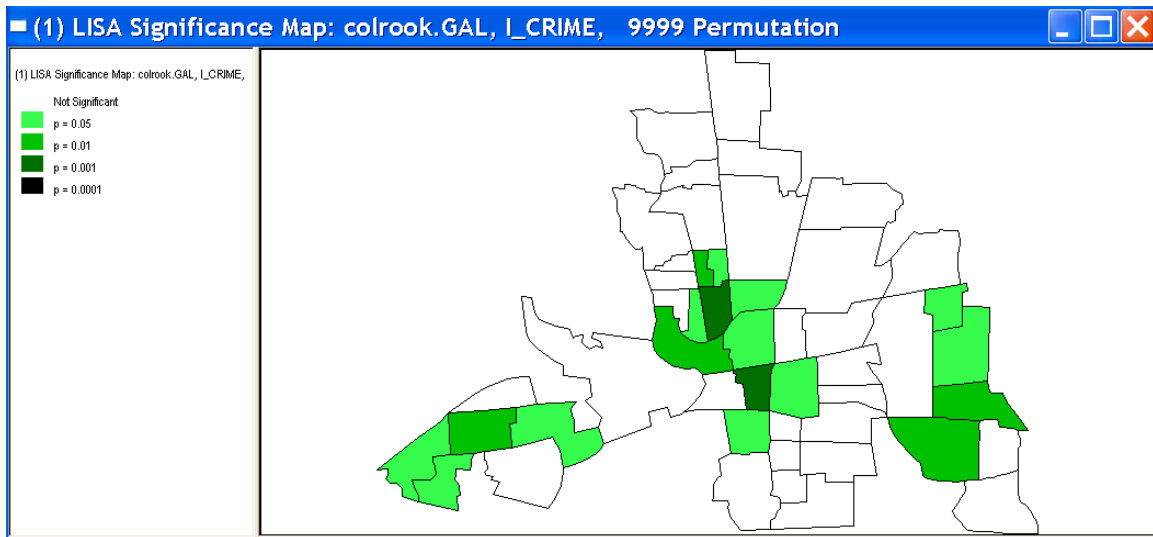


Figure 153. LISA significance map for Columbus **CRIME**.

Cluster Map

The **cluster Map** is selected by clicking on the matching check box in the dialog of Figure 152. The result is a special choropleth map showing those locations with a significant Local Moran statistic classified by type of spatial correlation: bright red for the high-high association, bright blue for low-low, light blue for low-high, and light red for high-low (Figure 154). The high-high and low-low locations suggest clustering of similar values, whereas the high-low and low-high locations indicate spatial outliers. The **cluster Map** can be linked and brushed in the same way as any other map in GeoDa.

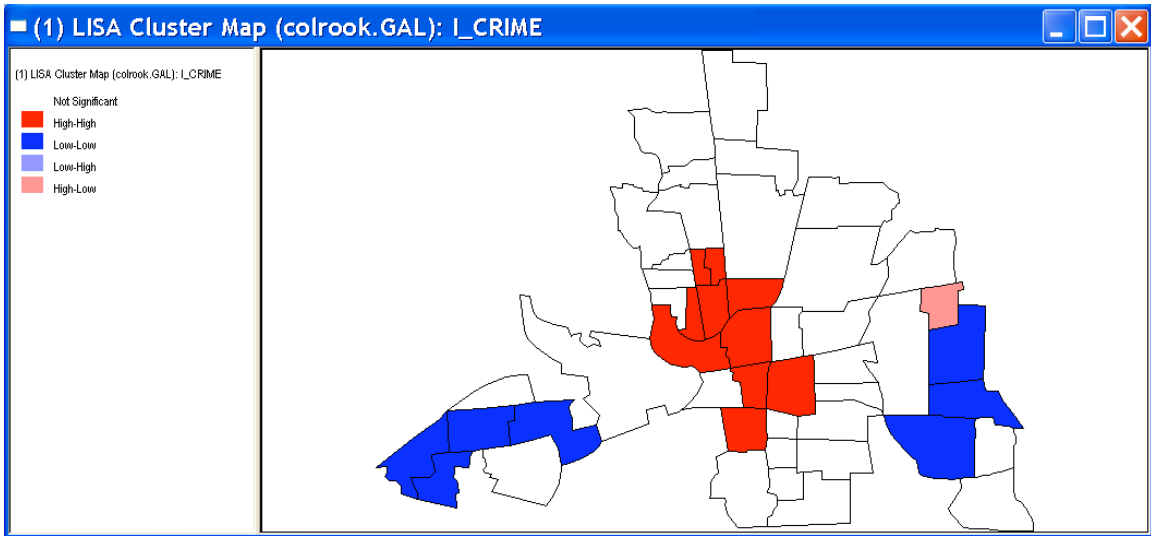


Figure 154. LISA cluster map for Columbus **CRIME**.

Box Plot

The **Box Plot** is selected by clicking on the matching check box in the dialog of Figure 152. The result is a slightly customized box plot of the distribution of the Local Moran statistics, as in Figure 155. This plot supports linking and brushing in the same way as the standard **Box Plot**. Outliers identified in this graph pertain to the Local Moran statistics and not to the variable itself. They are most useful as an informal diagnostic to identify “pockets of local non-stationarity” (Cressie, 1993).

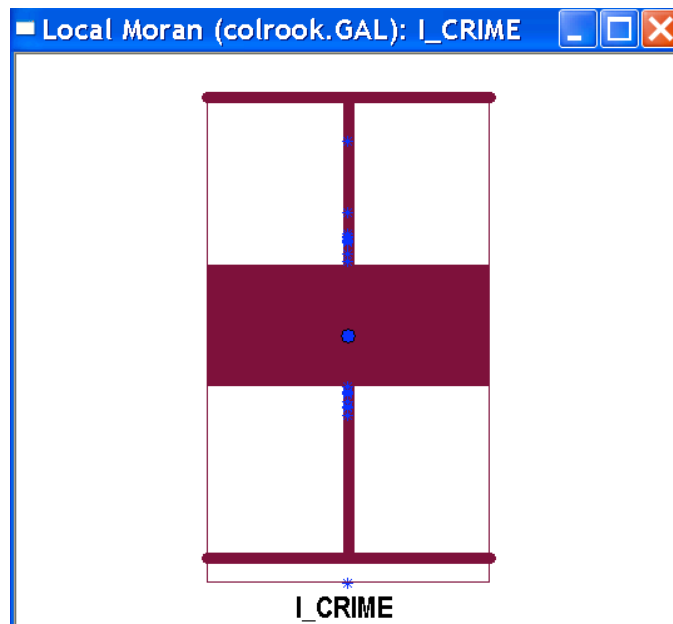


Figure 155. Box plot for Local Moran statistics.

Moran Scatter Plot

The **Moran Scatter Plot** is selected by clicking on the matching check box in the dialog of Figure 152. The result is identical to the graph obtained with **Explore > Univariate Moran** (Figure 137). The scatter plot supports linking and brushing and all the same options as the standard Moran Scatter Plot. If such a graph has already been created prior to the LISA analysis, there is not much point in checking this option in the dialog.

LISA Options

Each of the graphs and map windows created as part of a LISA analysis supports the same options as their standard counterpart. They are invoked by using the **Options Menu** or by right clicking in the graph. In addition, for the **Significance Map** and **Cluster Map**, three more options appear in the menu, as shown in Figure 156: **Randomization**, **Significance Filter** and **Save Results**. The other map options are the familiar ones to change the look of the window, save the map as an image, and to create an indicator variable for selected observations.

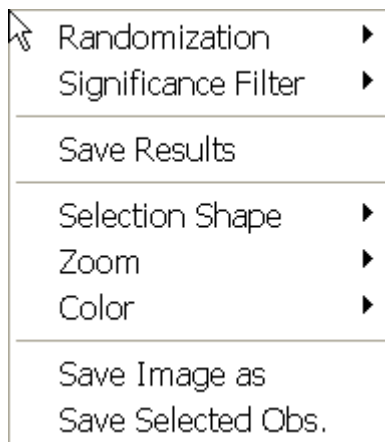


Figure 156. LISA map options.

Options > Randomization

The **Randomization** option, shown in Figure 157, allows the number of permutations to be specified and a new run of randomizations to be computed. The number of permutations is applied to each observation in turn, such that the total number of

computations equals the number of observations times the selected option. For large data sets (several thousand observations) this can be time consuming. Each new set of permutations yields a **Significance Map** and matching **Cluster Map** that take on a slightly different look. This is a way to assess how sensitive the indication of “significance” is to the choice of the number of permutations.

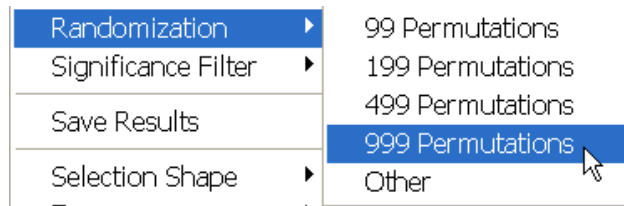


Figure 157. LISA randomization option.

Options > Significance Filter

The **Significance Filter** option, shown in Figure 158, changes the threshold at which locations with significant Local Moran statistics are displayed on the **Significance Map** and the **Cluster Map**. The default is $p < 0.05$, which tends to yield more significant locations that would be warranted if an adjustment were made for the multiple comparisons involved in inference for local statistics of spatial correlation. The **Significance Filter** allows you to set the threshold at a higher level of significance, such as $p < 0.01$ in Figure 158. As a result, fewer locations will be portrayed as significant, as illustrated in the LISA maps for the Columbus **CRIME** variable in Figure 159.

It is strongly recommended that sensitivity analysis be carried out before interpreting the results of LISA maps as “significant” clusters or outliers. The **Randomization** option provides a way to address numerical stability of the results. The **Significance Filter** is designed to assess how conclusions depend on the chosen significance level, and thus provides an informal mechanism to deal with multiple comparisons.

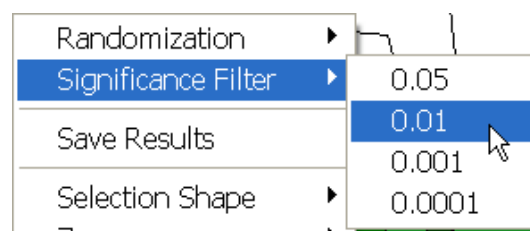


Figure 158. LISA significance filter option.

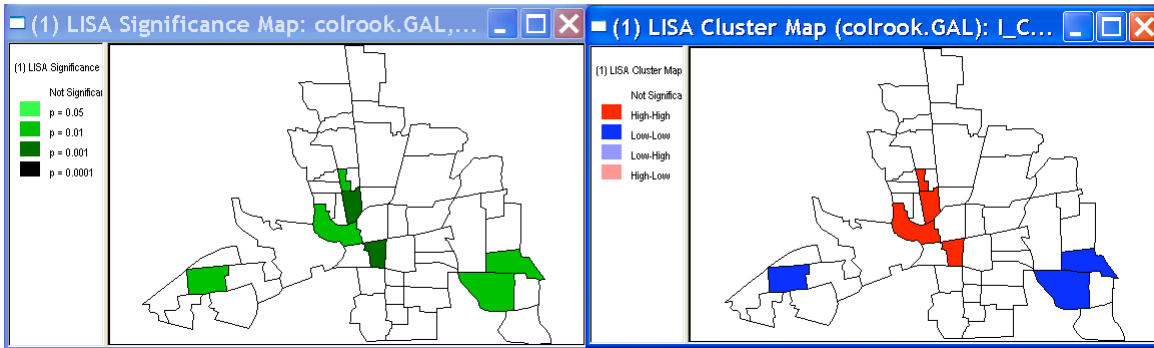


Figure 159. LISA maps after applying a significance filter.

Options > Save Results

The LISA Local Moran statistics for each location, the association p-value and the classification of significant (at $p < 0.05$) locations by type of spatial correlation can be saved to the data table by means of the **Options > Save Results** command. For both the Local Moran and the significance level the actual values are stored. For the spatial correlation type an indicator value is stored, which takes on the value of 1 for *high-high*, 2 for *low-low*, 3 for *low-high* and 4 for *high-low*. A dialog appears to select the variables to be added to the table and to specify the variable names, as in Figure 160. The default names are **I_VAR** for the Local Moran, **CL_VAR** for the spatial correlation category, and **PVAL_VAR** for the p-value (where **VAR** is the name of the variable).

Clicking **OK** in the dialog adds the specified variables to the data table, as shown in Figure 161. As always, this addition is not permanent until the table has been saved under a new file name. Note that in Figure 161, the observations have been sorted in increasing order of their p-value, with the observation with the most significant Local Moran listed at the top of the table. Using selection, sorting and promotion in the table allows you to investigate other significance ranges than the categories used by default.

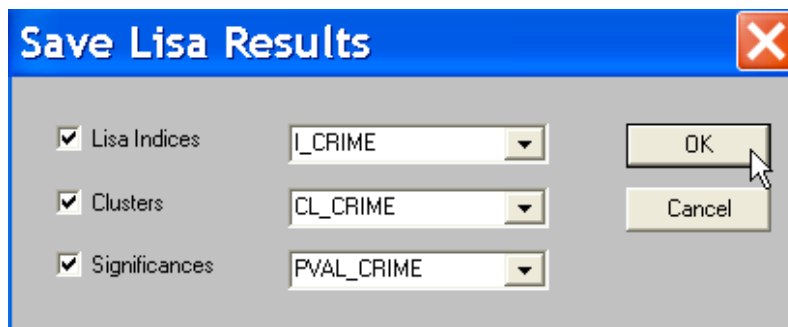


Figure 160. LISA save results dialog.

I_CRIME	CL_CRIME	PVAL_CR... △
2.037207	1.000000	0.000700
1.229924	1.000000	0.000800
0.212836	1.000000	0.001600
1.214552	2.000000	0.001800
1.289821	1.000000	0.005600
1.255139	2.000000	0.005900
1.217870	2.000000	0.007200
0.383639	1.000000	0.016900

Figure 161. LISA results added to data table.

Bivariate LISA

The **LISA** principle can be applied to a bivariate measure of local spatial autocorrelation in a straightforward way. It is invoked by **Explore > Multivariate LISA** or by clicking on the **Multivariate LISA** button in the **Explore** toolbar. The same four graphs can be generated as for the **Univariate LISA**, except that they pertain to a bivariate measure of local spatial autocorrelation. All options are the same as for the **Univariate LISA**.

EB LISA

The **LISA** principle can also be applied to an EB standardized rate variable. This operates the same as for the standard univariate measure of local spatial autocorrelation, except that the variable specification dialog asks for both **Event** and **Base** variables (as in Figure 149). It is invoked by **Explore > LISA with EB Rate** or by clicking on the matching button on the **Explore** toolbar. The same four graphs can be generated as for the **Univariate LISA**, except that they pertain to a measure of local spatial autocorrelation computed for EB rates. All options are the same as for the **Univariate LISA**.

References

- Anselin, L. (1986). *MicroQAP: a Microcomputer Implementation of Generalized Measures of Spatial Association*, Working Paper, Department of Geography, University of California, Santa Barbara.
- Anselin, L. (1994). Exploratory Spatial Data Analysis and Geographic Information Systems. In M. Painho (ed.), *New Tools for Spatial Analysis*, Eurostat, Luxembourg, 1994, pp. 45–54.
- 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 M. Fischer, H. Scholten, and D. Unwin (eds.), *Spatial Analytical Perspectives on GIS*. London: Taylor and Francis, pp. 111–125.
- Anselin, L. (1998). Exploratory Spatial Data Analysis in a Geocomputational Environment. In P. Longley, S. Brooks, B. Macmillan and R. McDonnell (eds.), *GeoComputation, a Primer*. New York: Wiley, pp. 77–94.
- Anselin, L. (1999). Interactive Techniques and Exploratory Spatial Data Analysis. In P. Longley, M. Goodchild, D. Maguire and D. Rhind (eds.), *Geographical Information Systems: Principles, Techniques, Management and Applications*. New York: Wiley, pp. 251–264.
- Anselin, L. (2000). Computing Environments for Spatial Data Analysis, *Journal of Geographical Systems* 2: 201–225.
- Anselin, L. and S. Bao (1997). Exploratory Spatial Data Analysis Linking SpaceStat and ArcView. In M. Fischer and A. Getis (eds.), *Recent Developments in Spatial Analysis*. Berlin: Springer-Verlag, pp. 35–59.
- Anselin, L., R.F. Dodson and S. Hudak (1993). Linking GIS and Spatial Data Analysis in Practice, *Geographical Systems* 1: 3–23.
- Anselin, L. and O. Smirnov (1996). Efficient Algorithms for Constructing Proper Higher Order Spatial Lag Operators, *Journal of Regional Science* 36: 67–89.
- Anselin, L. and O. Smirnov (1998). The DynESDA Extension for ArcView. Bruton Center, University of Texas at Dallas, Richardson, TX.
- Anselin, L., I. Syabri, O. Smirnov, Y. Ren (2001). Visualizing Spatial Autocorrelation with Dynamically Linked Windows. In *Computing Science and Statistics 33*, Proceedings of Interface 01, Orange County, CA, June 13-16, 2001 (CD-ROM).
- Anselin, L., I. Syabri and O. Smirnov (2002a). Visualizing Multivariate Spatial Correlation with Dynamically Linked Windows. In L. Anselin and S. Rey (Eds.), *New Tools for Spatial Data Analysis: Proceedings of a Workshop*. Center for Spatially Integrated Social Science, University of California, Santa Barbara, May 2002 (CD-ROM).
- Anselin, L., Y-W Kim and I. Syabri (2002b). *Web-Based Spatial Analysis Tools for the Exploration of Spatial Outliers*. GIScience 2002. The Second International Conference on Geographic Information Science. Boulder, CO, Sept. 25-28.
- Assuncao, R. and E.A. Reis (1999). A new proposal to adjust Moran's I for population density. *Statistics in Medicine* 18, 2147-2161.
- Bailey, T. and A. Gatrell (1995). *Interactive Spatial Data Analysis*. London: Longman
- Cressie, N. (1993). *Statistics for Spatial Data*. New York: Wiley.

- Fotheringham, A.S., C. Brunsdon and M. Charlton (2000). *Quantitative Geography, Perspectives on Spatial Data Analysis*. London: Sage Publications.
- Marshall, R.J. (1991). Mapping disease and mortality rates using Empirical Bayes estimators. *Applied Statistics* 40, 283-294.

Appendix A – GeoDa License Agreement

Copyright © 1998-2003 Luc Anselin and The Regents of the University of Illinois, All Rights Reserved. This software is subject to the License Agreement set forth in the License. Please read and agree to all terms before using this software.

BY INSTALLING THE GEODA SOFTWARE (THE SOFTWARE), YOU ARE CONSENTING TO BE BOUND BY THIS AGREEMENT. IF YOU DO NOT AGREE TO ALL OF THE TERMS OF THIS AGREEMENT, THEN DO NOT USE THE SOFTWARE.

End User License Agreement

The Regents of the University of Illinois grant you a non-exclusive License to use the Software free of charge if a) you are a student, faculty member or staff member of an educational institution (K-12, junior college, college or university); b) you are a United States federal, state or local government employee; or c) your use of the Software is exclusively at home for non-business purposes. Government contractors are not considered government employees for the purposes of this Agreement. If you do not meet the requirements for free use of the Software, you must contact the Spatial Analysis Laboratory at the Department of Agricultural and Consumer Economics of the University of Illinois, Urbana-Champaign (the DISTRIBUTOR), to obtain express written permission prior to any such use. Commercial users must contact the Distributor to negotiate terms of use. If you are using the Software free of charge under the terms of this Agreement, you are not entitled to hard-copy documentation, support or telephone assistance.

DISCLAIMER OF WARRANTY. Software is provided on an "AS IS" basis, without warranty of any kind, including without limitation the warranties of merchantability, fitness for a particular purpose and non-infringement. The entire risk as to the quality and performance of the Software is borne by you. Should the Software prove defective, you and not the copyright holder assume the entire cost of any service and repair. In addition, the security mechanisms implemented in the Software have inherent limitations, and you must determine that the Software sufficiently meets your requirements. This disclaimer of warranty constitutes an essential part of the agreement.

SOME JURISDICTIONS DO NOT ALLOW EXCLUSIONS OF AN IMPLIED WARRANTY, SO THIS DISCLAIMER MAY NOT APPLY TO YOU AND YOU MAY HAVE OTHER LEGAL RIGHTS THAT VARY BY JURISDICTION.

SCOPE OF GRANT.

You may: use the Software on one or more computers; use the Software on a network, provided that each person accessing the Software through the network must have a copy licensed to that person; copy the Software for archival purposes, provided any copy must contain all of the original Software's proprietary notices.

You may not: redistribute the Software in any form; permit other individuals to use the Software except under the terms listed above; modify, translate, reverse engineer, decompile, disassemble (except to the extent applicable laws specifically prohibit such restriction), or create derivative works based on the Software; rent, lease, grant a security interest in, or otherwise transfer rights to the Software; or

remove any proprietary notices or labels on the Software.

TITLE. Title, ownership rights, and intellectual property rights in the Software shall remain in Luc Anselin and The Regents of the University of Illinois. The Software is protected by copyright laws and treaties. Title and related rights in the content accessed through the Software is the property of the applicable content owner and may be protected by applicable law. This License gives you no rights to such content.

TERMINATION. The License will terminate automatically if you fail to comply with the limitations described herein. On termination, you must destroy all copies of the Software.

EXPORT CONTROLS. None of the Software or underlying information or technology may be downloaded or otherwise exported or re-exported (i) into (or to a national or resident of) Cuba, Iraq, Libya, North Korea, Iran, Syria or any other country to which the U.S. has embargoed goods; or (ii) to anyone on the U.S. Treasury Department's list of Specially Designated Nationals or the U.S. Commerce Department's Table of Denial Orders. By downloading or using the Software, you are agreeing to the foregoing and you are representing and warranting that you are not located in, under the control of, or a national or resident of any such country or on any such list.

LIMITATION OF LIABILITY. UNDER NO CIRCUMSTANCES AND UNDER NO LEGAL THEORY, TORT, CONTRACT, OR OTHERWISE, SHALL LUC ANSELIN, THE REGENTS OF THE UNIVERSITY OF ILLINOIS OR THE DISTRIBUTOR BE LIABLE TO YOU OR ANY OTHER PERSON FOR ANY INDIRECT, SPECIAL, INCIDENTAL, OR CONSEQUENTIAL DAMAGES OF ANY CHARACTER INCLUDING, WITHOUT LIMITATION, DAMAGES FOR LOSS OF GOODWILL, WORK STOPPAGE, COMPUTER FAILURE OR MALFUNCTION, OR ANY AND ALL OTHER COMMERCIAL DAMAGES OR LOSSES. IN NO EVENT WILL THE DISTRIBUTOR BE LIABLE FOR ANY DAMAGES IN EXCESS OF THE AMOUNT THE DISTRIBUTOR RECEIVED FROM YOU FOR A LICENSE TO THE SOFTWARE, EVEN IF THE DISTRIBUTOR SHALL HAVE BEEN INFORMED OF THE POSSIBILITY OF SUCH DAMAGES, OR FOR ANY CLAIM BY ANY OTHER PARTY. THIS LIMITATION OF LIABILITY SHALL NOT APPLY TO LIABILITY FOR DEATH OR PERSONAL INJURY TO THE EXTENT APPLICABLE LAW PROHIBITS SUCH LIMITATION. FURTHERMORE, SOME JURISDICTIONS DO NOT ALLOW THE EXCLUSION OR LIMITATION OF INCIDENTAL OR CONSEQUENTIAL DAMAGES, SO THIS LIMITATION AND EXCLUSION MAY NOT APPLY TO YOU.

HIGH RISK ACTIVITIES. The Software is not fault-tolerant and is not designed, manufactured or intended for use or resale as on-line control equipment in hazardous environments requiring fail-safe performance, such as in the operation of nuclear facilities, aircraft navigation or communication systems, air traffic control, direct life support machines, or weapons systems, in which the failure of the Software could lead directly to death, personal injury, or severe physical or environmental damage ("High Risk Activities"). The Copyright Holders specifically disclaim any express or implied warranty of fitness for High Risk Activities.

MISCELLANEOUS. This Agreement represents the complete agreement concerning this License and may be amended only by a writing executed by both parties. If any provision of this Agreement is held to be unenforceable, such provision shall be reformed only to the extent

necessary to make it enforceable. This Agreement shall be governed by the laws of the State of Illinois. The application of the United Nations Convention of Contracts for the International Sale of Goods is expressly excluded.

Appendix B – ANN License Agreement

ANN: Approximate Nearest Neighbors

Version: 0.1 (Beta release)

Copyright © 1997-1998 University of Maryland and Sunil Arya and David Mount
All Rights Reserved.

This software and related documentation is part of the
Approximate Nearest Neighbor Library (ANN).

Permission to use, copy, and distribute this software and its
documentation is hereby granted free of charge, provided that
(1) it is not a component of a commercial product, and
(2) this notice appears in all copies of the software and
related documentation.

ANN may be distributed by any means, provided that the original
files remain intact, and no charge is made other than for reasonable
distribution costs.

Disclaimer

The University of Maryland and the authors make no representations
about the suitability or fitness of this software for any purpose.
It is provided "as is" without express or implied warranty.

Version 0.1 03/04/98

Preliminary release

Version 0.2 06/24/98

Changes for SGI compilation.

Authors

David Mount
Dept of Computer Science
University of Maryland,
College Park, MD 20742 USA
mount@cs.umd.edu

Sunil Arya
Department of Computer Science
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong
arya@cs.ust.hk

Appendix C – Installing GeoDa

GeoDa comes with a simple installation script. Locate the `Setup.exe` file and double click (or use `Run` from the `Start` button and type the path to `Setup.exe`). The Installshield script will open a welcome window. Type in the required information and the program will copy all the necessary files. A typical installation will move `GeoDa.exe` to the `C:\Program Files\GeoDa` directory.⁶ There are two additional directories, `C:\Program Files\GeoDa\Samples`, which contains the sample data sets and `C:\Program Files\GeoDa\Docs`, which contains this document. A shortcut will be created as well and appear as an icon on your desktop. At the end of the installation, you can check the box to start the program before clicking on the `Finish` button.

Installation Issues

There have been some problems with incompatibilities between various MS Windows systems libraries on Win98 and Win2000 platforms. Before reporting these problems, make sure the files listed below are present in two important directories. If some files are missing, they may have been copied to the wrong directory by the install program. Try to locate them on your system and if found, copy them manually into the appropriate directory (this is sometimes the case with misplaced ESRI MapObjects LT libraries).

Required Files in the Directory `c:\Program Files\Common Files\ESRI`

(files in bold must be registered on the target computer, the installation program should take care of this, but you can always do it manually as well; refer to the Windows Registry instructions for your system)

04/05/2000	01:15 PM	520,192	<code>AFLT20.dll</code>
04/05/2000	01:21 PM	380,928	<code>MOLT20.ocx</code>
04/05/2000	12:51 PM	622,592	<code>Pe.dll</code>
04/05/2000	12:51 PM	248,320	<code>Sg.dll</code>
04/05/2000	01:20 PM	77,824	<code>ShapeLT20.dll</code>

Required System Files

The following files should be in the systems directory for your flavor of the Windows operating system: `c:\WINNT\SYSTEM32`, `C:\WINDOWS\SYSTEM32` or `C:\WINDOWS\SYSTEM`

⁶ Until you have access to a stable release of GeoDa, it is highly recommended that you stick with the default location for the files.

(files in bold must be registered on the target computer, the installation program should take care of this, but you can always do it manually as well; refer to the Windows Registry instructions for your system).

mfc42.dll
Msvcirt.dll
Msvcp60.dll
msvcrt.dll
Msvcrt40.dll
oleaut32.dll
olepro32.dll
Stdole2.tlb

Installation Error Messages

Sometimes, MS Windows reports an error during installation, such as

Msvcirt.dll is linked to missing export msvcrt.dll

The likely cause is that GeoDa's installation routine cannot find the correct version of the msvcirt.dll file. Another program may have installed a different version of the file, or the file may have been damaged.

It is important to note that there are two files with very similar file names that can easily be confused: **Msvcirt.dll** and **msvcrt.dll**. The latter is used by the MS operating system and should **never** be replaced or renamed.

To fix the problem:

- Search the computer to locate all copies of msvcirt.dll
- Rename each of these files, e.g., to msvcirt.old
- Reinstall GeoDa
- Reboot the computer before running GeoDa

Check the Openspace mailing list (<http://sal.agecon.uiuc.edu/mailman/listinfo/openspace>) for any other messages pertaining to installation problems.

Appendix D – New in GeoDa 0.9.3

Fixes

- Removed the 8.3 file name limitation
- Fixed some issues with randomization procedures in LISA
- Streamlined Menu items and Options
- Removed extra spaces in GAL file format (eliminates a problem reading them into the R spdep package)
- Both GAL and GWT files with either sequence number ID or Key Variable

New General Features

- New Table functionality
 - Calculator: new variables, variable transformations, standardization, spatial lags, rates, rate smoothers
 - Queries, sort, select, promote
 - Save edited table
 - Join table with other dbf tables (with common key)
- Save selection dummy variable in all graphs, maps and table (new variable added to table)
- Save results of operations to data table: rates, smoothed rates, spatial lags, local Moran (statistic, classification, p-value)
- Window/map customization for all graphs and maps: background, selection color, etc.
- Newly created shape files (centroid points, Thiessen polygons) include all variables from data file

Statistical Functions

- EB Moran and EB LISA correction for variance instability for rates
- LISA maps and significance maps have user-specified number of permutations and significance filter for sensitivity analysis

Input-Output

- Read point data from dbf, ascii files (not only from point shape files)
- Data export as comma delimited files (csv)
- Capture maps and graphs as bitmap files
- Export polygon boundary files as ascii (with bounding box)

Data Manipulation

- Tools (weights creation, data conversion) available without opening project
- More flexible spatial weights construction for polygon shape files: include distance for arbitrary x, y coordinates, centroid x, y coordinates as default for distance computation
- Thiessen polygon shape files include area and perimeter as variables