Center for Spatially Integrated Social Science

GeoDa™ 0.9.5-i Release Notes

Luc Anselin

Spatial Analysis Laboratory
Department of Agricultural and Consumer Economics
University of Illinois, Urbana-Champaign
Urbana, IL 61801

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

Center for Spatially Integrated Social Science

http://www.csiss.org/

Revised, January 20, 2004

Contents

List of Figures

iv

Preface

These release notes pertain to the third official release of the $GeoDa^{TM}$ *software for geodata analysis*, an upgrade to Version 0.9.5-i, released on January 23, 2004. The first release dates back to February 5, 2003. The notes complement the $GeoDa^{TM}$ *0.9 User's Guide* (Anselin 2003) that accompanied the second official release of the software, Version 0.9.3, released on June 4, 2003. In the remainder of the release notes, that document will be referred to as the *User's Guide*.

Many important aspects of the use of the software are not repeated here. All the basic functions, background information on the software and the full text of all relevant licenses are included in the *User's Guide*. The current release notes only document additions and changes to the software, and should be used together with the Version 0.9 *User's Guide*.

The development behind this release of $GeoDa^{TM}$ has been facilitated by the continued research support through the U.S. National Science Foundation grant BCS-9978058 to the Center for Spatially Integrated Social Science (CSISS), and by grant RO1 CA 95949-01 from the National Cancer Institute. Funding sources for earlier version of the software and its antecedents can be found in the *User's Guide*.

Many thanks go to the students in the Fall 2003 classes of ACE 492SA, Spatial Analysis, and ACE 492SE, Spatial Econometrics, offered through the Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, for being such good sports in serving as guinea pigs for various iterations of what became called "version 095i." $GeoDa$'s growing user community contributed considerably as well, with bug reports, requests for features and other useful feedback from too many users to be listed individually. Their continued interest is greatly appreciated.

Trademarks and Licenses

- $GeoDa^{TM}$ is a trademark of Luc Anselin, All Rights Reserved.

- *GeoDa* incorporates licensed libraries from ESRIs MapObjects LT2; ESRI, ArcView, ArcGIS and MapObjects are trademarks of Environmental Systems Research Institute, Redlands, CA

- *GeoDa* incorporates code derived from publicly available sources under various generous licenses (the detailed licenses are listed in the Appendix of the *User's Guide*):

    - the MFC Grid Control 2.24 by Chris Mauder
    - the ANN Code by David Mount and Sunil Arya
    - the Thiessen polygon algorithm of Yasuaki Oishi

- other companies and products mentioned herein are trademarks or registered trademarks of their respective trademark owners

The GeoDa Team

- Project Director: Luc Anselin

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

- Technical Documentation and Training Materials: Luc Anselin and Julia Koschinsky

What's New in GeoDa 0.9.5-i

GeoDa 0.9.5-i contains several minor improvements and bug fixes to the previous version, as well as totally new functionality for mapping (cartogram), exploratory data analysis (parallel coordinate plot, 3D scatter plot and conditional plots) and spatial regression. A brief outline of the major changes and innovations is given next.[1] A more complete discussion of the features, methodological background and relevant user interface is given in the remaining sections.

New Features

Data Manipulation

- the creation of polygon shape files for regular lattices or grids from basic user input on the structure of the lattice

- the construction of polygon shape files from boundary information contained in an ascii input file

Mapping

- a circular cartogram, implementing Dorling's cellular automata algorithm (Dorling 1996), fully linked and brushable

- conditional maps (see conditional plots)

Exploratory Data Analysis

- parallel coordinate plot (PCP) for multivariate data exploration, with linking and brushing

---

[1]To highlight the new items, they are given blue section headings in the remainder of the release notes.

- three-dimensional scatter plot, with linking and brushing

- conditional plots, using two conditioning variables to explore the distribution of a third variable

  - conditional map, box plot, histogram and scatter plot

Spatial Regression

- Ordinary Least Squares regression with full diagnostics for spatial effects (Moran's I, Lagrange Multiplier statistics), as well as the usual tests against heteroskedasticity and non-normality

- Maximum Likelihood estimation of the spatial lag and the spatial error models, with asymptotic inference

Refinements and Improvements of Existing Features

User Interface

- the standard menu structure has been slightly reorganized

  - new items for `Table`, `Space`, and `Regress`
  - the spatial autocorrelation analysis was moved from the earlier `Explore` menu to `Space`

- an item for `Methods` has been added to the opening menu to allow regression analysis without starting a project (i.e., without loading a shape file into the project)

- the toolbar buttons have been slightly reorganized

  - new toolbar buttons have been added for the `Cartogram`, `PCP`, `Conditional Plot`, and the `3-D Scatter Plot`
  - a new dockable toolbar is included with buttons to activate the various map functions (quantile map, box map, standard deviational map, percentile map, cartogram and map movie)
  - the spatial autocorrelation toolbar buttons were separated from the EDA toolbar

- the toolbar buttons for opening a new map and duplicating a map have a new look

4

- the default for the map window now shows the legend; in earlier versions the user needed to explicitly open the legend pane by dragging the left side of the window to the right

Spatial Data Manipulation

- a custom bounding box can be specified in the creation of Thiessen polygons; previously, only the bounding box for the points themselves was used

- higher order contiguity calculation now contains an option for the inclusion of all lower order neighbors (the previous default) or the computation of a "pure" higher order contiguity

- distance weights files and k-nearest neighbor weights now contain the "correct" distance between the points as the third column of the GWT file; previously this value was rescaled and not useful for interpretation

- the "treshold" typo has been fixed

- the weights characteristics histogram has a new look and more flexible classifications, with islands properly included as having zero neighbors

Mapping

- the default selection tool for point shape files is now the circle (previously, it was a rectangle)

- the `Map Movie` was thoroughly revised, with a new interface, allowing interactive starting and stopping (pause) as well as rewind and step through

Exploratory Data Analysis

- the `Histogram` sports a new look and uses a continuous color ramp

- the `Box Plot` has been slightly redesigned and shows the median more distinctly

Table

- `Table` functions can now be invoked from the main menu; previously, this was only by right-clicking in the table

- tables can be saved after the deletion of columns (variables)

Bug Fixes

- the classification for the percentile map is now correct; in the earlier versions, an extra observation may have been included in the top percentile

- the computation of the standard deviation is fixed; this could affect many functions, including the classifications in the standard deviational map and the computations in the Moran scatter plot and the Local Moran

- the coordination between LISA maps for different variables is fixed; there were situations where all the maps became identical when a second variable was analyzed

- problems with selection in the map when using different selection tools are fixed; when switching between selection tools there was some strange behavior

- various minor bugs in table and rate calculations were fixed

- several issues related to "weird" *out of memory* errors when loading tables or shape files were fixed; the out of memory error had in fact nothing to do with memory, but indicated problems with file formats, all known such problems have been fixed (including the one where the file name could not start with a "T")

Menu Structure and Toolbar Buttons

The menu structure has been slightly reorganized and new toolbars have been added to facilitate map construction and spatial autocorrelation analysis. Below, an overview of the main structure is given, followed by a detailed look at the changes in and additions to individual menus and toolbars.

Overview

As in Version 0.9.3, the window that appears after the program has been launched contains a simplified menu that allows access to `Tools`, such as spatial weights construction and spatial data transformations, without having to explicitly start a project (and load a shape file). A `Methods` item has been added to this menu, to invoke the spatial regression functionality directly. This is especially useful when analyzing larger data sets, since it avoids the need to update all linked windows, including potentially a very large data table. The initial menu is shown in Figure 1. As before, only two items on the toolbar are active, the first of which is used to launch a project, as illustrated in the figure.
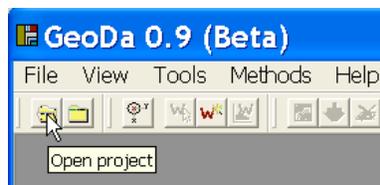
Figure 1: The initial menu and toolbar

After opening the project, the usual dialog requests the file name of the

shape file and the `Key` variable. After clicking on the `OK` button, a map window is opened, showing the base map for the analyses, as in Figure 2. The main difference with earlier versions is that the default window shows (part of) the legend pane on the left hand size. As before, this can be resized by dragging the separator between the two panes (the legend pane and the map pane) to the right or left.
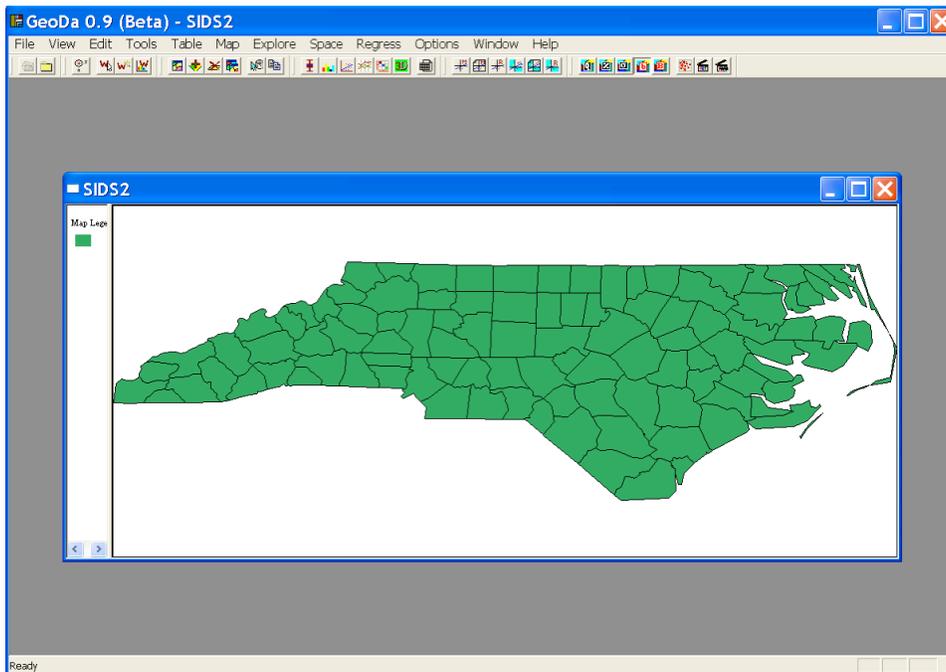
Figure 2: Opening window after loading the SIDS2 sample data set

With a shape file loaded, the complete menu and all toolbars are active, as in Figure 3. The menu bar contains three new items: `Table`, `Space` and `Regress`. The toolbar has two new dockable sets of buttons, `Space` and `Map`, and a slightly reorganized `Explore` toolbar. The `Weights` toolbar has been moved to the left. Two icons on the `Edit` toolbar sport a new look.

Figure 3: The complete menu and toolbar buttons

Menu Items

Tools Menu

The `Tools` menu is available both with and without a loaded shape file and is identical in both cases. As show in Figure 4, there are two new items. `Tools > Shape > Polygons from Grid` constructs a polygon shape file for a regular lattice or grid, based on simple user input, such as the coordinates of the lower-left and upper-right corners, the number of rows and the number of columns (see p. 13). `Tools > Shape > Polygons from BND` creates a polygon shape file based on the boundary coordinates contained in an ascii input file (see p. 15).
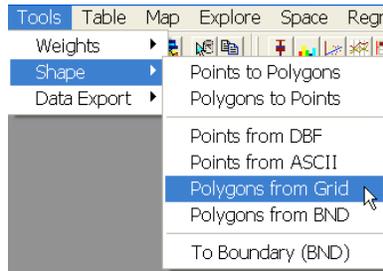
Figure 4: The tools menu item

Methods Menu

The `Methods` menu is only available when no shape file has been loaded into the project. Its only use is to invoke the spatial regression functionality, as shown in Figure 5. The interface is identical to that used in the `Regress` menu inside a project (see p. 45).

Figure 5: The methods menu item

There is a major difference between the use of the regression functionality through the menu shown in Figure 5 and through the `Regress` item in the main menu in Figure 3. When invoked without a shape file with the `Methods` menu, the regression analysis reads the data directly from the dBase file, without showing the table contents in the window. This avoids the overhead required for the linking and brushing and is typically the only practical way to analyze large data sets (10,000 observations and more).

Map Menu

The `Map` menu (Figure 6) contains one new item, the `Cartogram` (see p. 23). The `Map` toolbar, shown in Figure 7, is new. It contains buttons to invoke the familiar choropleth map types (from left to right, quantile, percentile, standard deviational, and box map –with two fences), the cartogram and the map movie (both single and cumulative, see p. 26).
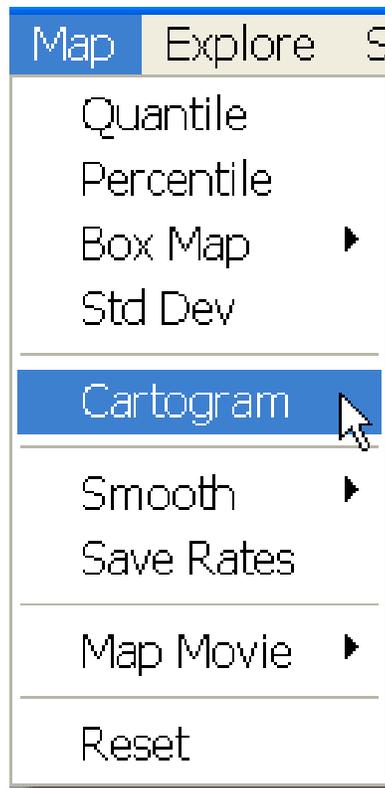
Figure 7: Map toolbar

Figure 6: Map menu

Explore Menu

The Explore menu includes three new items: the Parallel Coordinate Plot (see p. 29), the 3D Scatter Plot (see p. 33), and the Conditional Plot (see p. 38), as shown in Figure 8. The spatial autocorrelation analysis items and the table have been moved to their own menu (see p. 12). The Explore toolbar, Figure 9, contains the icons for the six EDA functions, as well as a button to activate the data Table.

Figure 9: Explore toolbar
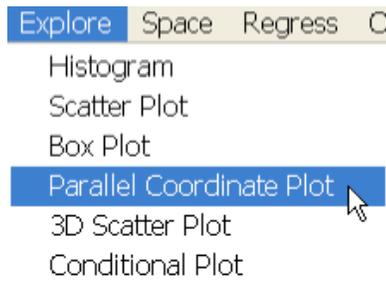
Figure 8: Explore menu

Table Menu

The Table menu (Figure 10) contains all the operations on table elements.
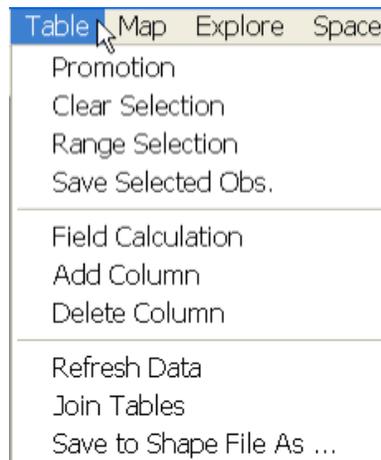
Figure 10: The table menu item

11

Note that the items in the `Table` menu are identical to what is obtained when right clicking in an active table.

Space Menu

The `Space` menu is new and groups the functions to carry out spatial auto-correlation analysis, as illustrated in Figure 11. In previous versions, these were included in the `Explore` menu. The matching toolbar buttons (Figure 12) are combined together in a separately dockable toolbar.
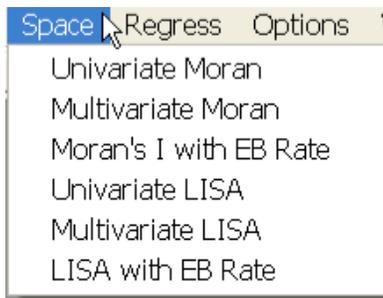
Figure 11: Space menu

Figure 12: Space toolbar

Manipulating Spatial Data

Two new spatial data input functions have been added to the `Tools` menu. There are also minor changes in the spatial weights calculations, and a new option was added to the construction of Thiessen polygons.

Creating Grid Polygon Shape Files

`Tools > Shape > Polygons from Grid` gives the ability to construct a polygon shape file for a regular lattice or grid from simple user input. The regular grid is either square or rectangular and has the observation numbers starting in the upper left corner and increasing to the right, and then down, row by row. Figure 13 illustrates the main dialog.
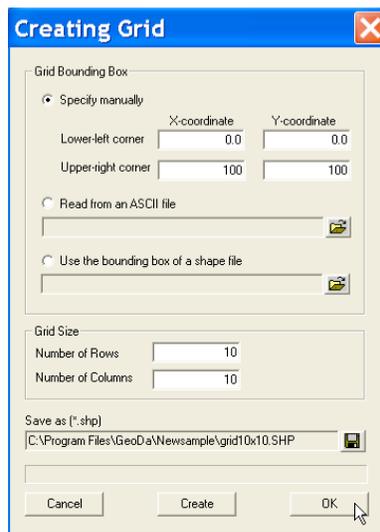
Figure 13: The creating grid dialog

The simplest approach is to enter the coordinates for the lower left and upper right corner and to specify the number of rows and columns, as shown in Figure 13. For the example given there, the result is a 10 by 10 regular lattice, as in Figure 14. The shape file contains three fields: an identifier (POLYID), the area of the grid cell (AREA), and the perimeter (PERIMETER). Any other data need to be added by means of the table join functionality.
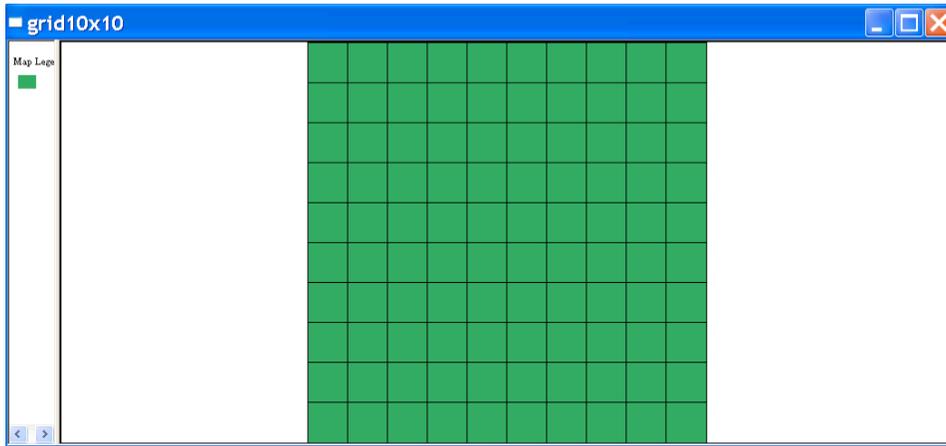
Figure 14: A 10 by 10 regular lattice

A second approach is to read the bounding box information from a text file. This file has a very simple format, as shown in Figure 15. It contains the number of rows, the number of columns, the X,Y coordinates for the lower left corner, and the X,Y coordinates for the upper right corner. These items can be on the same line, separated by white space (space or tab), or on consecutive lines; a comma- separated file does *not* work.
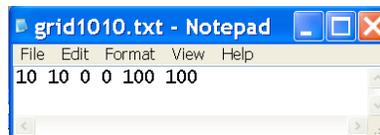
Figure 15: Format of bounding box text input file

Yet a third approach bases the grids on the bounding box associated with a shape file. Note that this approach is only correct for projected shapes. When the coordinates are unprojected lat-lon, there may be distortions for

larger extents. The corner coordinates of the bounding box (as read from the shape file) determine the extent of the lattice. The size of the individual grid cells follows from the number of rows and columns specified, as shown in Figure 16, using the extent of the SIDS shape file as the bounding box.
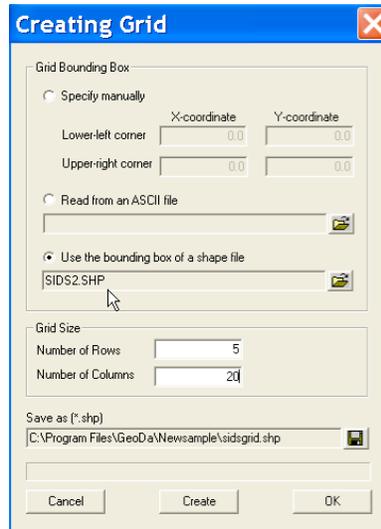
Figure 16: Create a grid from bounding box in a shape file

The result is illustrated in Figure 17. To illustrate the effect of the bounding box choice, the original outline of the North Carolina counties is superimposed on the 5 by 20 lattice, after applying an `Edit > Add Layer` command. Note the slight distortion in the grid cells, due to the fact that the SIDS shape file is unprojected.

Creating Polygon Shape Files from BND Input

`Tools > Shape > Polygon from BND` creates a polygon shape file from the boundary information contained in a text input file. The dialog, as shown in Figure 18, requires the name of the output shape file and the input text file. The input file must follow a very specific format, similar to the formats used in the shape to BND function. The only format supported so far is the "1a" BND format, as specified in *GeoDa*'s shape output function. The format is spelled out when the help feature is invoked, by clicking the question mark in the dialog shown in Figure 18, yielding Figure 19.

15
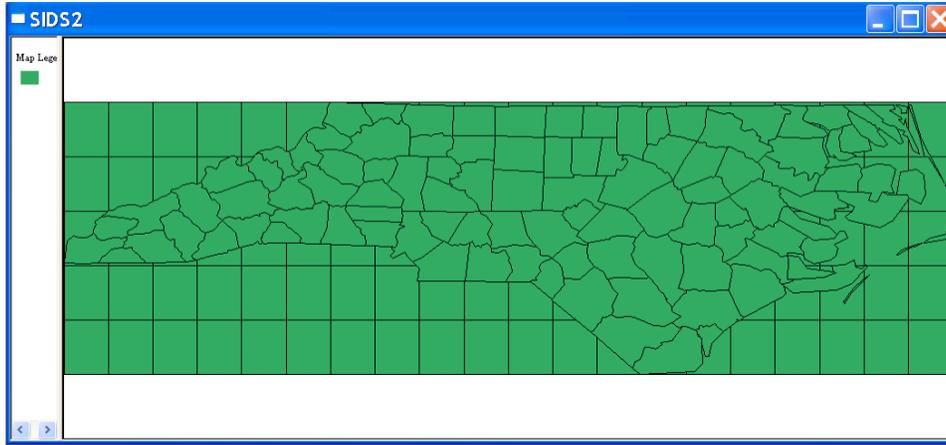
Figure 17: North Carolina counties with matching 5 by 20 regular lattice
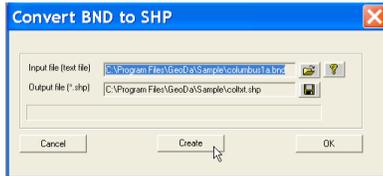
Figure 18: Shape file from a boundary text file
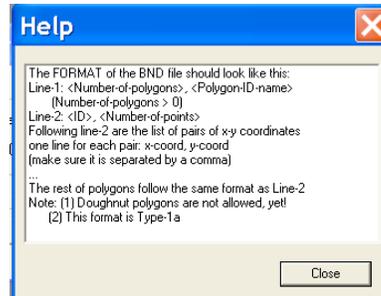
Figure 19: Boundary file format

The supported file format for the text file consists of a header line, containing the number of observations and the variable name for the Key variable, separated by a comma. Next, for each observation follows a line with the ID and the number of vertices that define the polygon, again comma-separated. Then, the X,Y coordinates are given, comma-separated and on a separate line for each point. This is repeated for each polygon in the data set. For example, the contents of the input file for the Columbus data would be:

```
49,POLYID
1,14
8.62413,14.237
```

```
8.5597,14.7424
8.80945,14.7344
...
8.6429,14.0897
8.63259,14.1706
8.62583,14.2237
2,46
...
```

The resulting shape file can be loaded into *GeoDa* in the usual way. For example, using the input text file for Columbus yields the shape file shown in Figure 20. The data table, also illustrated in the figure, contains three fields: the original identifier (POLYID), AREA, PERIMETER, and a simple sequential identifier (RECORD_ID).
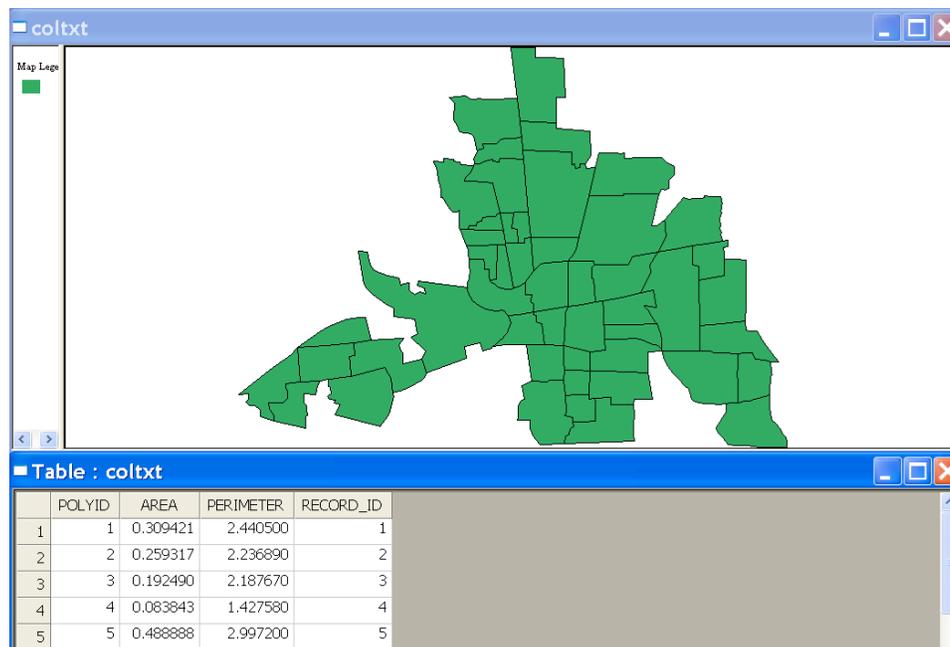
Figure 20: Columbus shape and table from text boundary file

Creating Spatial Weights

The `Weights` functionality in the `Tools` menu has been revised slightly. This affects higher order contiguity computation, distance-based weights and the weights characteristics.

Higher Order Contiguity

`Tools > Weights > Create` invokes the usual dialog. There is a new check box below the selection of the order of contiguity, as shown in Figure 21. Selecting this option includes all the lower order neighbors up to the order specified. The default (check box left unchecked) only computes "pure" higher order contiguity, which does *not* include the lower order neighbors.
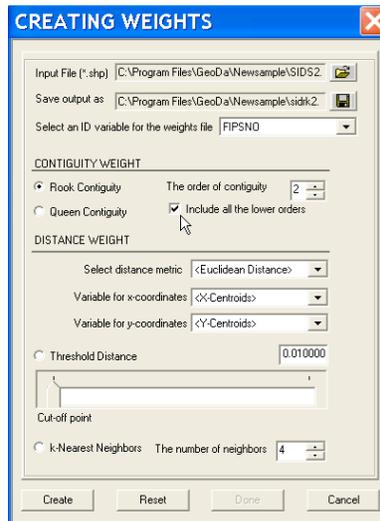
Figure 21: Options for higher order contiguity

Creating Distance Weights

The distance weights calculation now uses the correct distance metric, both in the user interface as well as in the resulting weights file. The distance units depend on the units for the coordinates of the base map. When those points are stored as unprojected lat-lon decimal degrees, the resulting distance will be in *miles*. Previously, the distance shown was rescaled and did not have a meaningful interpretation.

In Figure 22, the cut off distance shown in the interface using the North Carolina counties is (approximately) 29.9 miles. The distances calculated are included as the third column in the GWT file, both for distance-based contiguity as well as for k-nearest neighbors. For example, in Figure 23, the distances are listed (in miles) for the 4 nearest neighbors in the North Carolina example.

Note that in the current version of *GeoDa*, the distances themselves are not used, but only the resulting contiguity information is taken into account.
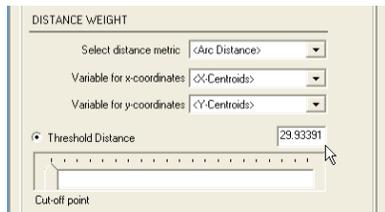
Figure 22: Distance cutoff in miles

Figure 23: Distance in k nearest neighbor weights files

Weights Characteristics

The design of the histogram used to depict the connectivity structure in spatial weights has been revised. The classification into discrete categories has been made more flexible and allows the adjustment to the necessary number through the `Options > Intervals` command (in the `Options` menu or by right clicking on the histogram). Also, a continuous color ramp is used for the histogram bars (see also p. 43). Islands are properly identified and shown as polygons with 0 contiguities.

In Figure 24, this is illustrated for distance-based contiguity using a cut off distance of 28 miles for the North Carolina counties. As shown in Figure 22 this is less than the necessary distance to ensure connectivity for all counties. As a result, two counties are identified as islands. Their location is shown by linking with the base map.
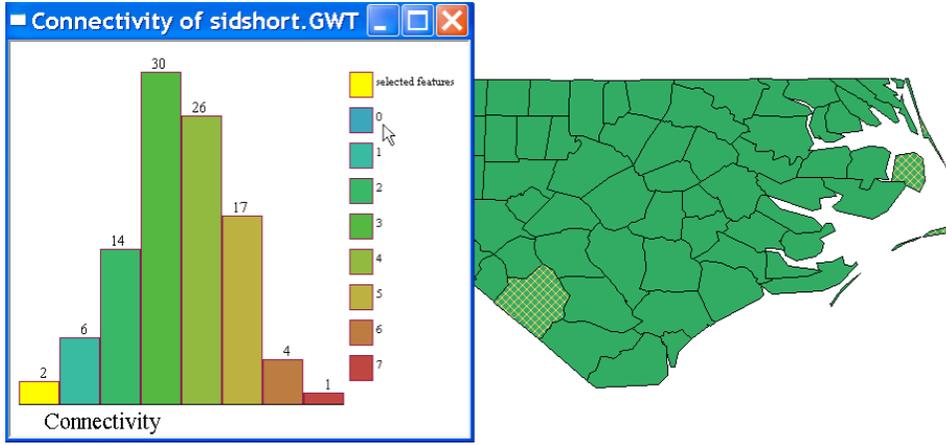
Figure 24: Weights characteristics with islands

Thiessen Polygons

`Tools > Shape > Points to Polygons` brings up a dialog to specify the options for the creation of Thiessen polygons from a point shape file. A new option has been included, which allows the use of an external bounding box to determine the extent of the enclosing "rectangle" for the polygons. In the interface, a check box selects this option, which requires a shape file to be specified, as in Figure 25.
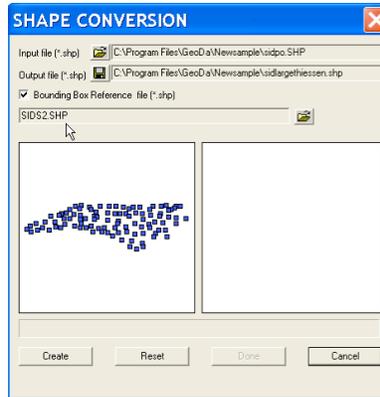
Figure 25: Bounding box option for Thiessen polygons

The difference between the default and the use of this option is illustrated in Figures 26 and 27, using the centroids of the North Carolina counties as the input point shape file. The default (Figure 26) uses the bounding box for the point file, which has the extreme points on the boundary. Typically, the resulting rectangle will be smaller than the extent of the original counties. In Figure 27, the county polygon shape file was specified as the bounding box. Note that there is now some space between the centroids and the outer boundary of the rectangle. The latter is identical to the bounding box of the county shape file, facilitating overlay in a GIS.

While this option provides a degree of flexibility in setting the bounding box, it does not allow for an external shape bounding box that would be internal to the default for the point shape. In other words, the bounding box will never *exclude* points from the Thiessen polygons.
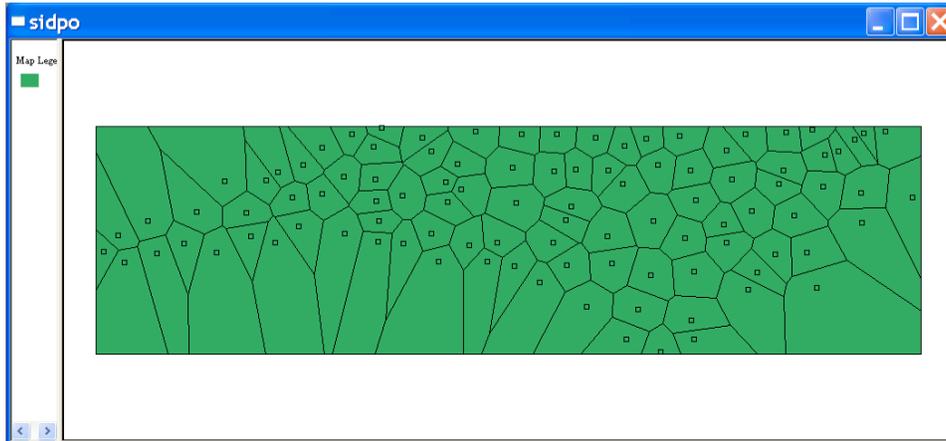
Figure 26: Default bounding box for Thiessen polygons
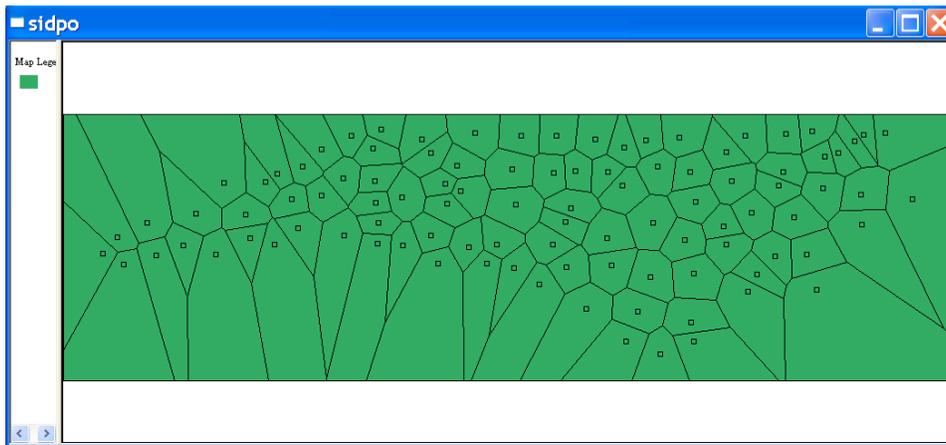
Figure 27: Polygon-based bounding box for Thiessen polygons

Mapping

A `Cartogram` has been added as a new type of map and the `Map Movie` functionality has been fine tuned considerably.

Cartogram

A cartogram is a map where the original layout of the areal units is replaced by a layout in which the size of the area is proportional to a given variable. *GeoDa* implements a so-called circular cartogram, in which the original irregular polygons are replaced by circles. The placement of the circles is such that the original pattern is mimicked as much as possible, both in terms of absolute location as in terms of relative location (neighbors, or *topology*). This is based on a non-linear cellular automata algorithm due to Dorling (1996). The size (area) of the circles is proportional to the value of the selected variable.

The cartogram is invoked by selecting `Map > Cartogram` from the menu or by clicking on the cartogram toolbar button. In the usual fashion, the variable selection dialog appears. After selecting the variable and clicking on the `OK` button, the cartogram is drawn.

For example, in Figure 28 a cartogram is shown for the 1974 Sids rates (`SIDR74`) for North Carolina counties. The cartogram uses a color code to provide additional information about specific values, such as negative values, zero and outliers. The default color is green. Negative values are shown as black and zeros as transparent (white in the default background). Upper outliers are red and lower outliers are blue. The default hinge used to identify outliers is 1.5, which results in four such observations in Figure 28.

The default for the outlier criterion can be changed: in the `Options` menu; by right clicking in the cartogram; or by clicking on the matching `Box Map` toolbar button. This dialog is illustrated in Figure 29. Selecting 3 as the value results in a cartogram with only one outlier, as in Figure 30.
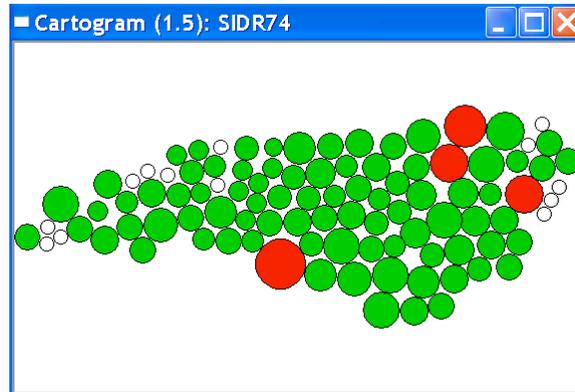
Figure 28: Circular cartogram for North Carolina Sids rates (`SIDR74`)
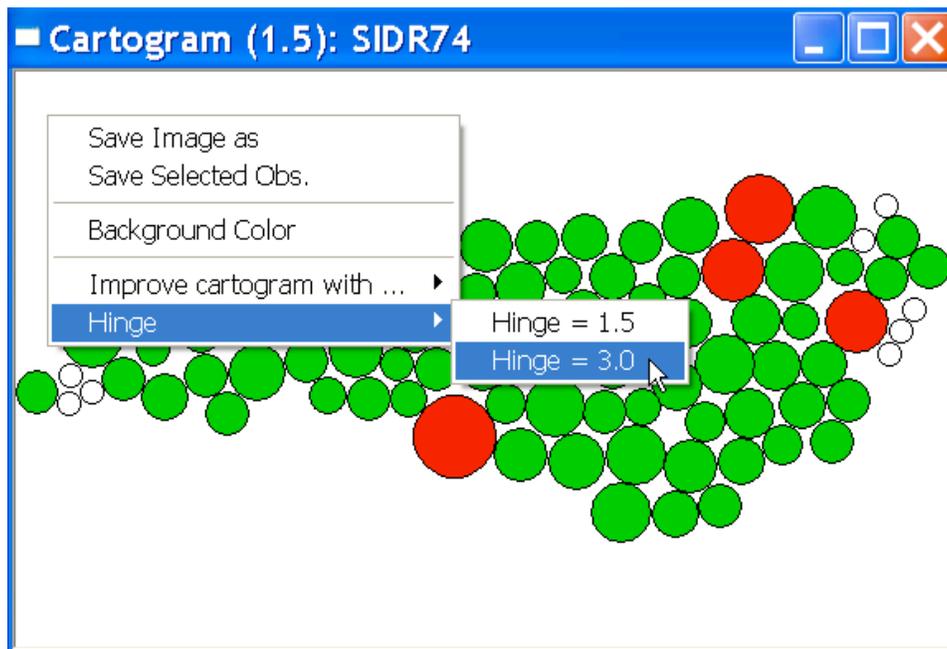
Figure 29: Selection of outlier hinge in cartogram
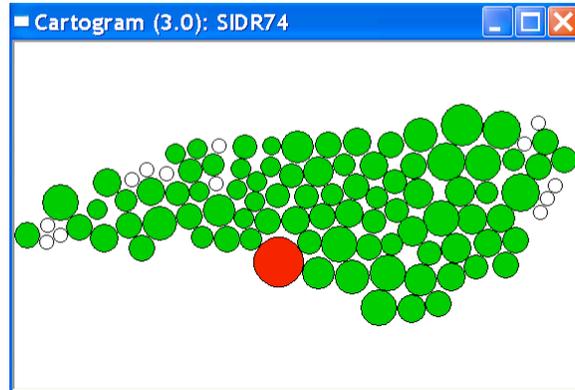
Figure 30: Outliers in Sids rate cartogram using a hinge of 3
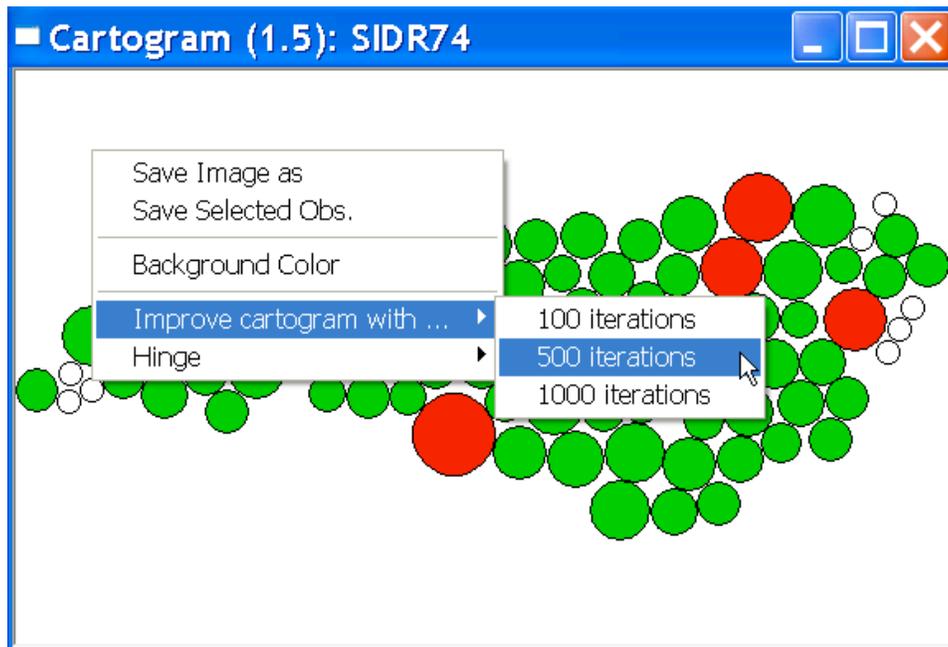
Figure 31: Improving the layout of the cartogram

The cartogram uses a nonlinear algorithm to position and size the circles, which does not necessarily converge to an acceptable solution after the default number of iterations. An option is provided to compute an additional 100, 500 or 1000 iterations and improve upon the current solution, as illustrated in Figure 31.

The cartogram is treated in the same way as other windows when it comes to brushing and linking. Any selection in another window will also be highlighted in the cartogram, and vice versa. For example, Figure 32 shows the outliers in the cartogram linked to their actual locations in the North Carolina county map.
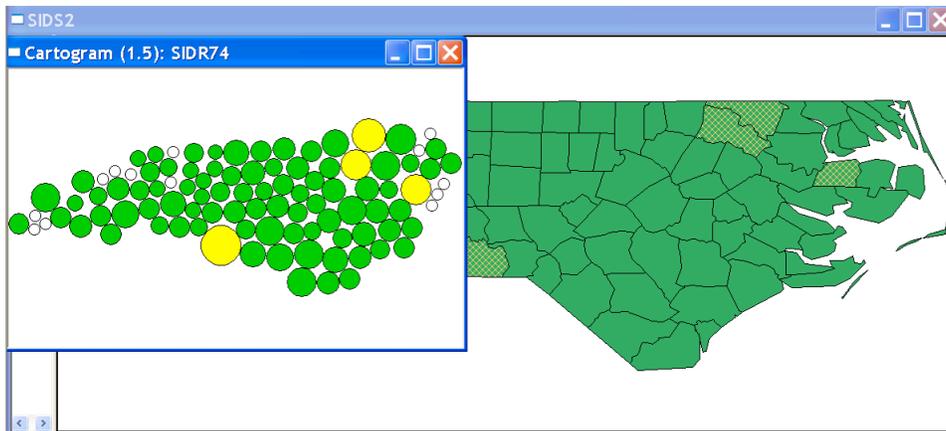
Figure 32: Outliers in Sids rate cartogram linked to base map

Map Movie

The Map Movie is an attempt at providing a simple form of map animation in *GeoDa*. This is accomplished by highlighting locations according to their order for a given variable, from low to high. This gives the same effect as when a box plot would be brushed from the bottom to the top, one observation at a time. The Map Movie is implemented either in a Cumulative form or in a Single form. In the Cumulative version, the observations are added to a cumulative selection set, which ultimately covers the whole map. In contrast, in the Single form, only one location is shown at any time.

The Map Movie is invoked from the main menu by selecting Map > Map Movie > Cumulative , or Map > Map Movie > Single, or by clicking the

toolbar button. Once a variable is chosen in the usual dialog, the map movie window opens, as in Figure 33 for the Columbus neighborhoods. This consists of some controls at the top and the usual areal outline.
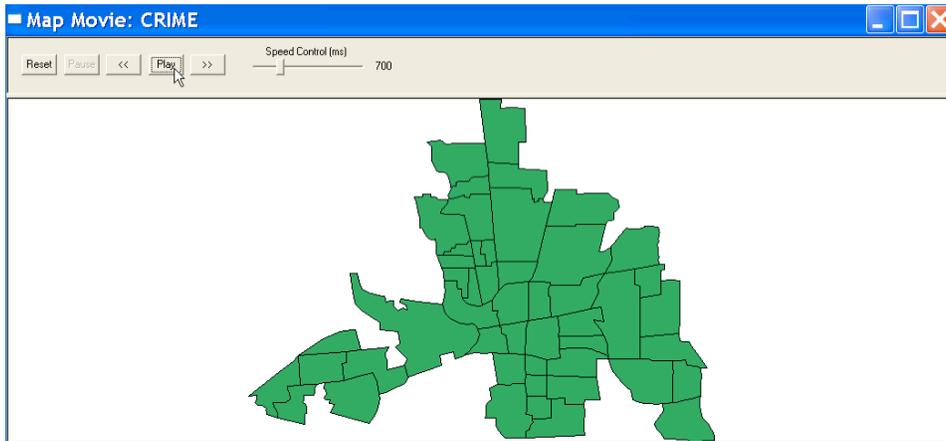
Figure 33: Starting a cumulative map movie

There are five main controls and one slider bar. The `Play` button starts (or re-starts) the operation of the movie. The speed by which locations are shown on the map depends on the setting for the slider bar. This is a function of the machine clock speed and is hardware dependent. Moving the button on the slider bar to the left speeds things up, moving it to the right slows the movie down. The `Pause` button stops the movie, as in Figure 34, and `Reset` clears the map. After the movie has been paused (or at the start), the arrow buttons, `>>` and `<<`, step through the movie one observation at a time, either forward (`>>`) or backward (`<<`). At the end of a cumulative map movie, all locations are selected, as in Figure 35.

The map movie is linked to all the other windows in the current project. However, this linking is only one-way, in the sense that the selected locations that appear in the map movie will also be highlighted in the other windows. However, selections in other windows do not affect the map movie (this would defeat the purpose of the animation). More importantly, any change to the selections in other windows during the operation of a map movie will break the linking mechanism. For the same reason, there is no brushing in the map movie. Finally, while it is possible to do so, it is not a good idea to run two map movies at the same time. This is fine in terms of the movies themselves, but the linking mechanism will be inconsistent.
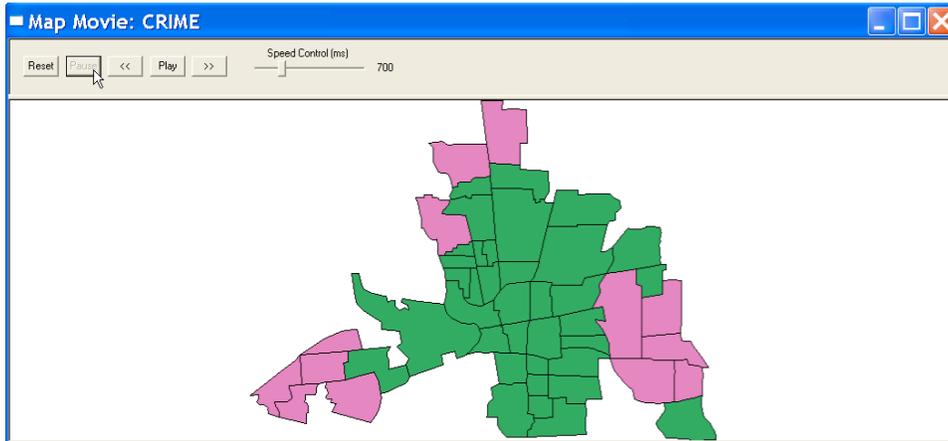
Figure 34: Pausing a cumulative map movie
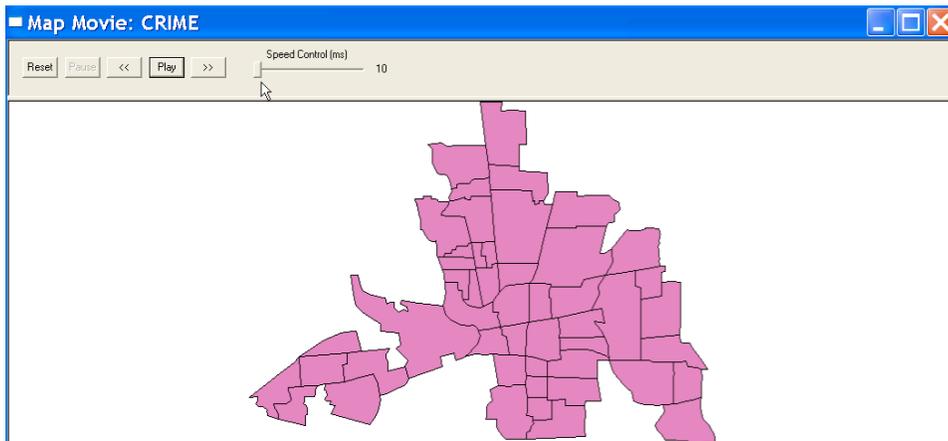
Figure 35: A completed cumulative map movie

Exploratory Data Analysis

GeoDa's functionality for exploratory data analysis has been extended with three new types of dynamically linked graphs: the parallel coordinate plot, the three dimensional scatter plot, and four conditional plots (conditional map, box plot, histogram and scatter plot). In addition, the histogram and box plot graphs were redesigned slightly.

Parallel Coordinate Plot

The `Parallel Coordinate Plot` (PCP) is a method to explore multivariate relationships. Each variable under consideration is drawn as a parallel line on which the (coordinates of the) observations are recorded as points. The matching points for each observation are connected and form a line. As a result there are as many lines as observations in the PCP. Background on the fundamental ideas and methodological issues can be found in, among others, Inselberg (1985) and Wegman (1990).

The PCP can be used to discover "clusters" among observations when their *lines* show similar patterns (i.e., group together in a distinct way in the graph). In addition, a common pattern in the slopes of the lines connecting coordinates on different variable axes indicates the nature of the "correlation" between those variables (positive or negative, or no patterning). The PCP is linked to all the other graphs and maps and can be brushed.

The `Parallel Coordinate Plot` is launched by selecting it from the main menu, using `Explore > Parallel Coordinate Plot`, or by clicking on the PCP toolbar button. This opens up the PCP variable selection dialog, as in Figure 36. Variables are included by selecting them in the left hand side panel and using the `>` arrow button. Alternatively, `>>` selects all variables, but this is usually not advised for a PCP. The selection can be edited by means of the reverse button. Click on `OK` (Figure 37) to launch the plot, which yields the PCP as shown in Figure 38.
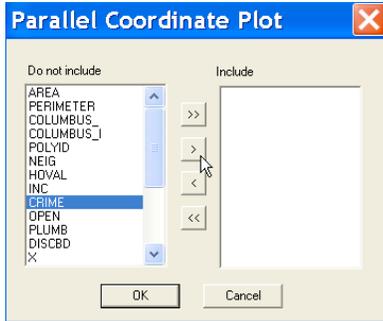
Figure 36: Variable selection for PCP
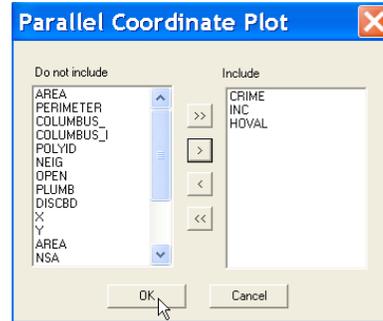
Figure 37: PCP variables selected
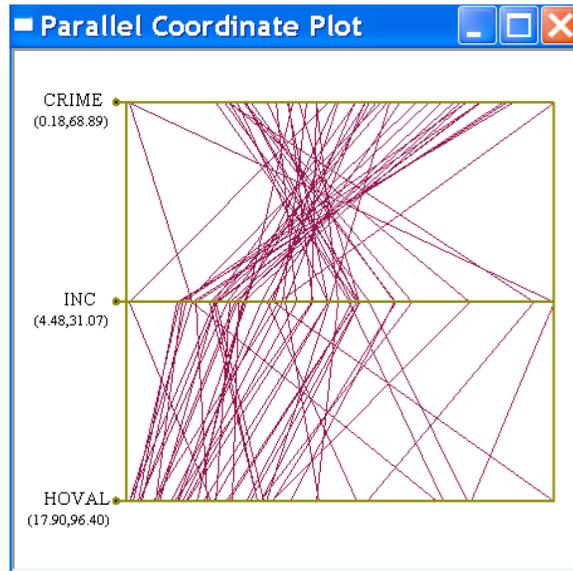
Figure 38: PCP for Columbus variables

A closer look at the graph shows the range for each variable listed in parentheses next to the variable name (on the left hand size). The order of the axes (variables) can be changed by clicking on the small dot next to the variable name (as in Figure 39) and dragging it to "drop" it on top of another variable. As a result, the two axes switch places in the plot. Rearranging the order of variables in this manner can sometimes facilitate the discovery of clusters and patterns.

The PCP implemented in *GeoDa* has a limited number of options, which are invoked by right clicking on the graph, illustrated in Figure 40. The first three of these are standard options for any graph: saving the image as a bitmap file, adding the selected observations as a dummy variable to the table, and changing the `Background Color`. The latter is often useful for better visibility of selected observations, since the default selection color of yellow is not easy to see on the default white background in the plot.

The last two of the five options PCP options are non-standard. They pertain to the scale used for the horizontal axes. The default is to keep the variables in their original scales (this is not necessarily a good idea when the scales are very different). The alternative is to convert the variables to standard deviational units, which is obtained with the `Standardize Data Set` option. This is a toggle switch, so one of the two is always selected. Figure 41 illustrates the standardization on a dark grey background.

The PCP can be brushed like any other graph. A rectangular selection can be moved over the lines as in Figure 42. This selects the matching observations in all the other open graphs and maps.

Figure 39: PCP change variable order
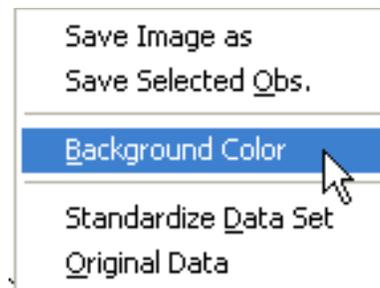
Figure 40: PCP options
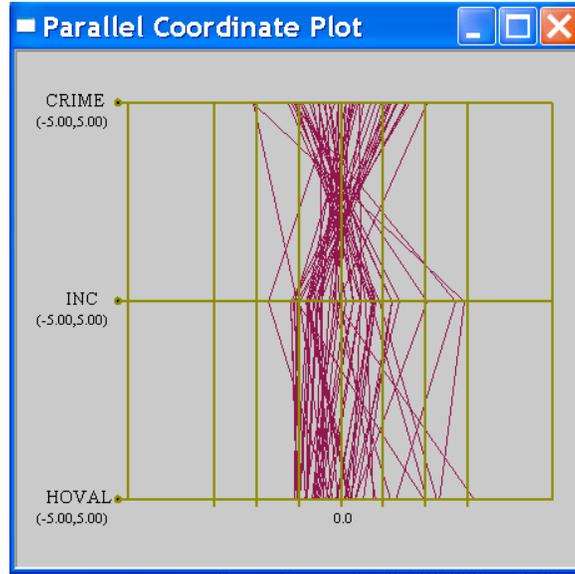
Figure 41: PCP using standardized variables
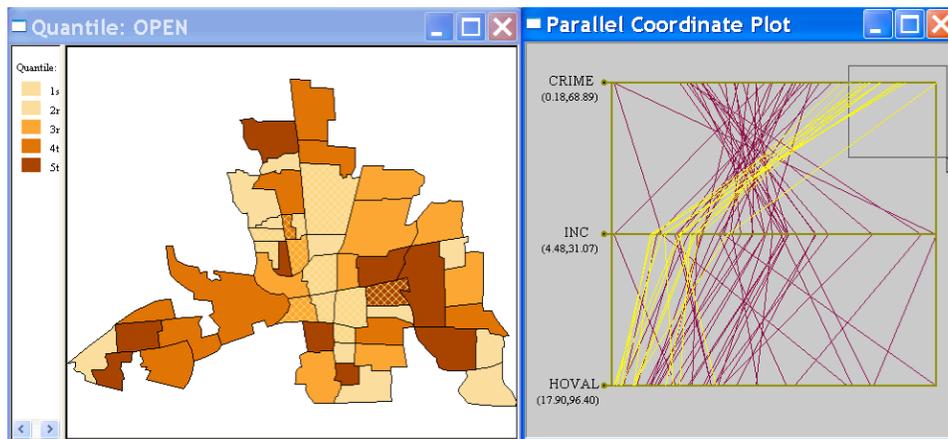
Figure 42: Brushing the PCP

3D Scatter Plot

Multivariate data exploration in *GeoDa* is further facilitated by the inclusion of a three-dimensional scatter plot. This feature is still somewhat experimental and may not be totally stable at this point. It implements the usual 3-D point manipulations, such as rotating, zooming and translation of the graph, as well as linking and brushing.

The `3D Scatter Plot` is started as `Explore > 3D Scatter Plot` from the menu, or by clicking the matching toolbar button. This brings up the `Axis Selection` dialog, as shown in Figure 43. For each of the axes in
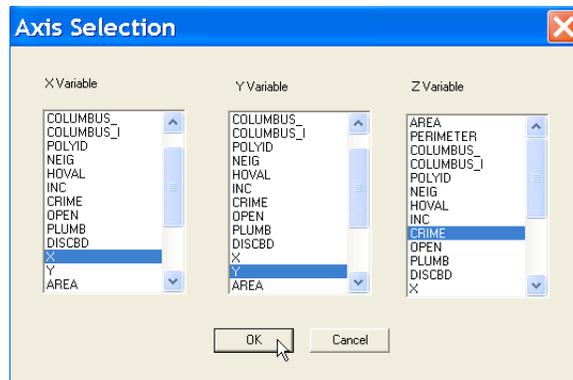
Figure 43: Variable selection for 3D scatter plot

the plot, the variable is selected from the drop down list in the usual way. Clicking `OK` generates the initial view of the 3-D plot, as in Figure 44. Note the position of the axes, with the z-axis coming out towards the viewer (the axes are color coded to facilitate keeping track of them during rotation and translation).

The plot is manipulated by means of the mouse buttons. The left button is used to *rotate* the plot, the right button to *zoom* in or out (by moving the mouse up or down), and both buttons to *translate* the plot (move it up or down, or sideways). Figure 45 shows a rotation, where the z-axis is made vertical (the highest crime locations are the most vertical) and the x-y axes form the horizontal plane. Figure 45 also illustrates the *projection* of the points onto one of the side planes. In the left hand side of the interface, the check box next to `Project x-y` is checked, which yields the points on the horizontal plane. In the illustration, since the X and Y axes are the coordinates, these are the locations of the Columbus neighborhood centroids.
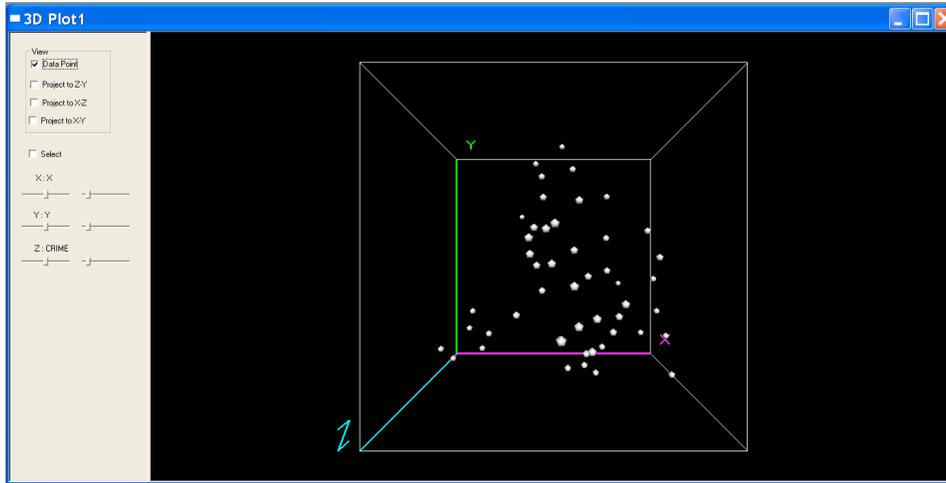
Figure 44: 3D scatter plot initial view
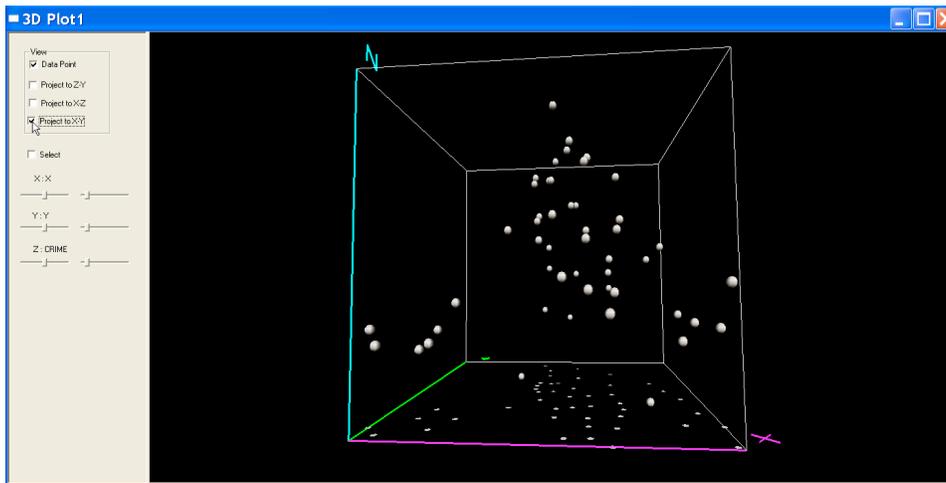
Figure 45: Rotated 3D scatter plot

The selection of observations in the 3D scatter plot is implemented by means of a three-dimensional selection box or volume. Checking the `Select` box in the left hand pane generates the default volume. This can be resized by moving the sliders on the *right hand side* for each of the dimensions, as shown in Figure 46. The selected points (spheres) are highlighted in yellow.

34

Figure 46: Selection box in 3D scatter plot

The selection can be changed (brushing) in two different ways. In one, the sliders on the *left hand side* in the pane next to each of the dimensions can be moved to change the position of the selection box along this axis. For example, moving the slider for the X-axis, as shown in Figure 47, will change the position of the box along the X dimension, but will keep its position along the two other dimensions fixed. Alternatively, CTRL-left mouse button allows free movement of the selection box in all dimensions (Figure 48).

The selected points in the 3D Scatter Plot are *linked* to all the other graphs and maps. This is slightly different from the standard approach, in the sense that the direction of selection matters. When the Select check box is activated in the 3D plot, the points selected are highlighted in the other plots. However, this is not continuous (as in other brushing), but the selection is refreshed each time the brush stops, i.e., each time the red box on the plot stops moving. This is illustrated in Figure 49. Alternatively, when brushing is carried out in a different map or graph, this invalidates the Select check box in the 3D plot. The selection from the other graphs is highlighted as yellow in the 3D plot, but without the red selection box, as shown in Figure 50.
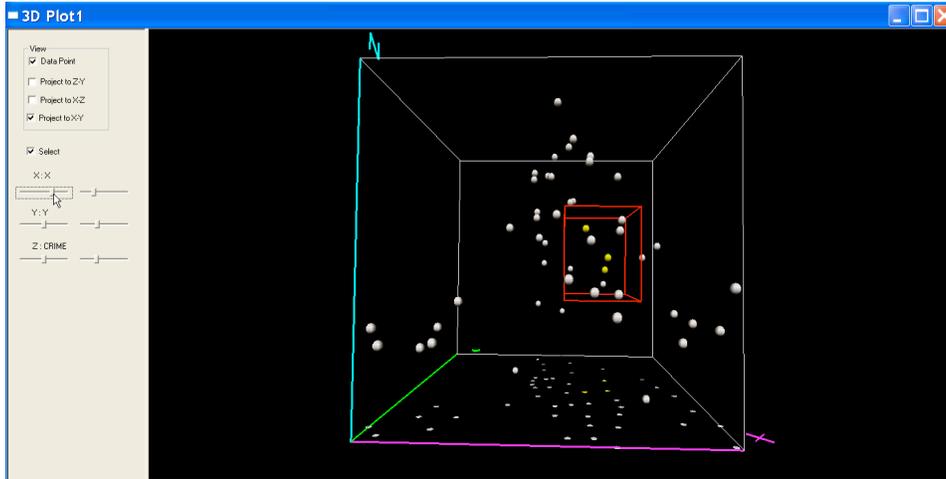
Figure 47: Brushing the 3D scatter plot using the slider
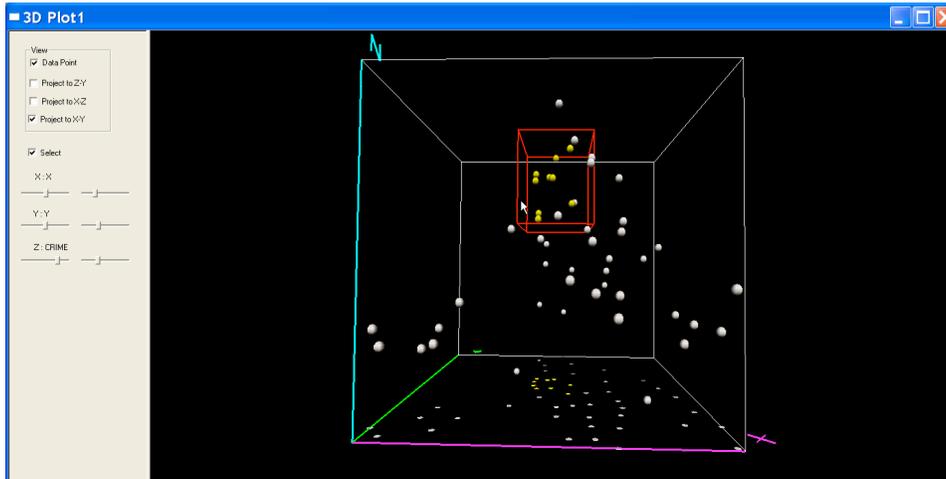
Figure 48: Free form brushing of the 3D scatter plot
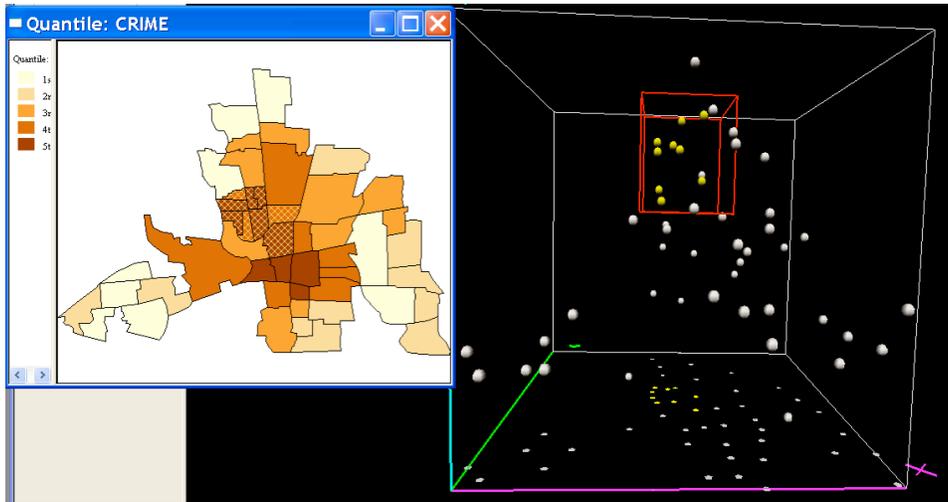
Figure 49: Linking between 3D scatter plot and other windows
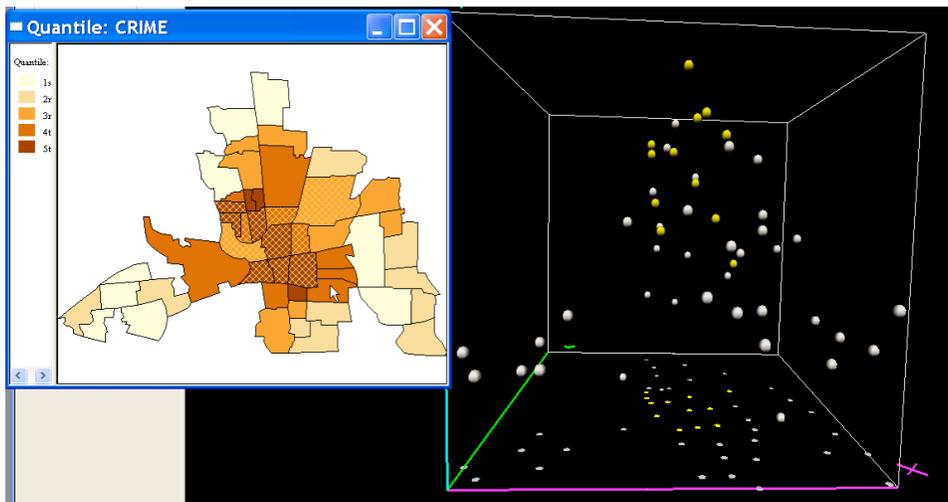
Figure 50: Linking between map and 3D scatter plot

37

Conditional Plot

The conditional plots are yet another way to carry out multivariate data exploration. The main principle behind these plots is to use two *conditioning variables* to subset the data sample into distinct categories. The observations in each of these categories fall into a specific range for the conditioning variables. A separate graph or map is drawn for a third variable in each of the subsets. The fundamental ideas behind this approach are outlined in Becker et al. (1996) and Carr et al. (2002), among others.

In *GeoDa*, each of the conditioning variables can have three subsets, yielding a total of nine subgraphs. Four types of conditional plots are supported: a conditional map, conditional box plots, conditional histogram and conditional scatter plots.

The conditional plots are invoked as `Explore > Conditional Plot` from the menu, or by clicking the matching toolbar button. This brings up a simple dialog to select the type of graph, as in Figure 51. With the radio button checked next to the desired plot type, clicking `OK` brings up the variable selection dialog.

Variables are moved to the respective axes by selecting them from the drop down list and clicking on the matching `>` button, as shown in Figure 52. After the variables are entered for all axes, `OK` (Figure 53) will start the selected graph. Since the map, box plot and histogram are univariate plots, only three axes are required. For the conditional scatter plot, a fourth axis is needed (the third is for the dependent variable, or vertical axis in the scatter plot, the fourth the explanatory variable, or horizontal axis in the scatter plot).
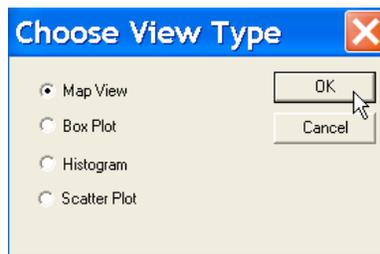
Figure 51: Types of conditional plots
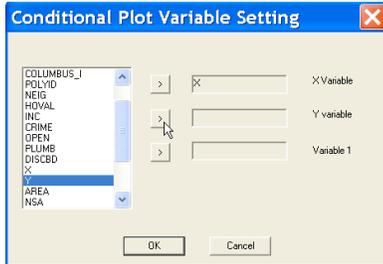
Figure 52: Conditional plot variable selection
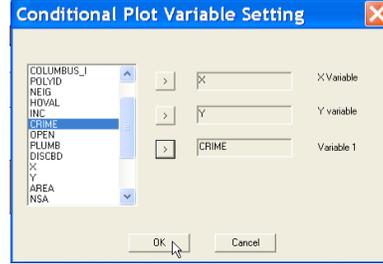
Figure 53: Starting up the conditional plots
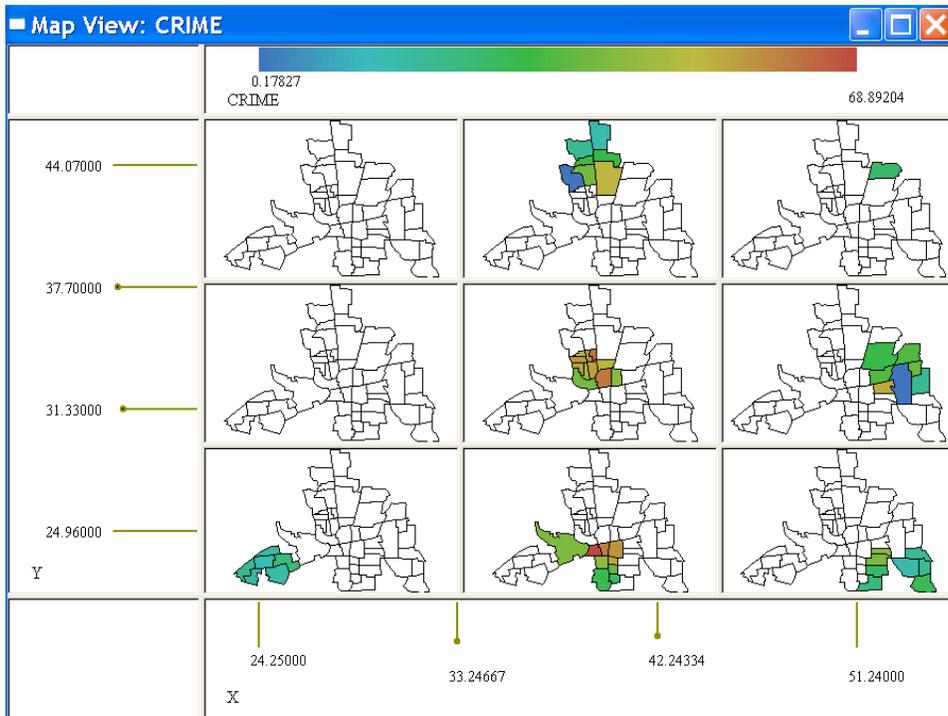
Figure 54: Conditional map plot

The four types of conditional plots are illustrated using the Columbus example and a very simple form of conditioning. The X-axis is for the X coordinates and the Y axis for the Y coordinates. In other words, the nine subplots are for selected locations that fit the specified X-Y range. This is shown in Figure 54 for a choropleth map of the variable CRIME. Note that a continuous color ramp is used for the choropleth map.

The categories of the conditioning variables can be changed by moving the handles on the X-axis to the right or left, and on the Y-axis up or down. This will alter the number of observations falling in each cell and thus highlight how the pattern of the variable under consideration changes in different subsets of the data. For example, in Figure 55, additional neighborhoods are included into the second Y level by moving the Y handle lower. This is easiest to see in the second highest cell on the left hand side, which was empty in Figure 54. Also, moving the handles to the right (X axis) or up (Y axis) collapses the categories together. If this is done for all handles, the plot in the lower left corner will be for the complete data set.
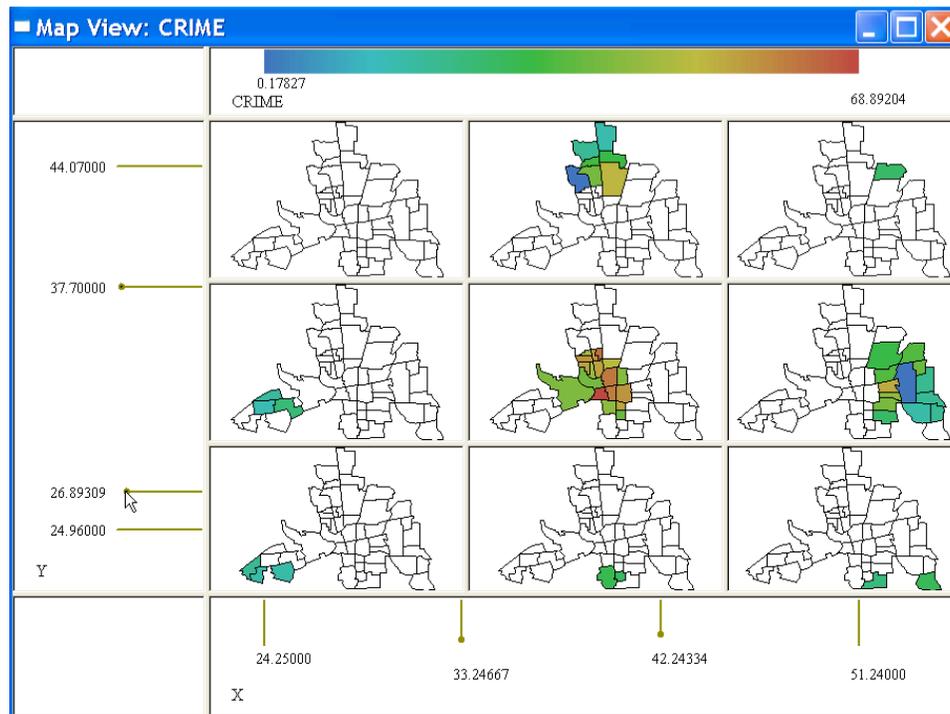
Figure 55: Moving the handles in the conditional plot

In Figure 56, the conditioning is illustrated for the box plots. In each of the cells, a new box plot is drawn, using the range for the complete data set as the reference (the height of the box is the same in each cell, and provides a reference with respect to the complete data set). However, the distribution in each cell is potentially different, with different medians (the red horizontal line), fences and outliers. The box plots follow the new format (see also p. 43) and show the number of observations in each cell in parentheses. When there are fewer than five observations in the cell (as in the upper right corner of Figure 56), no box plot is drawn.
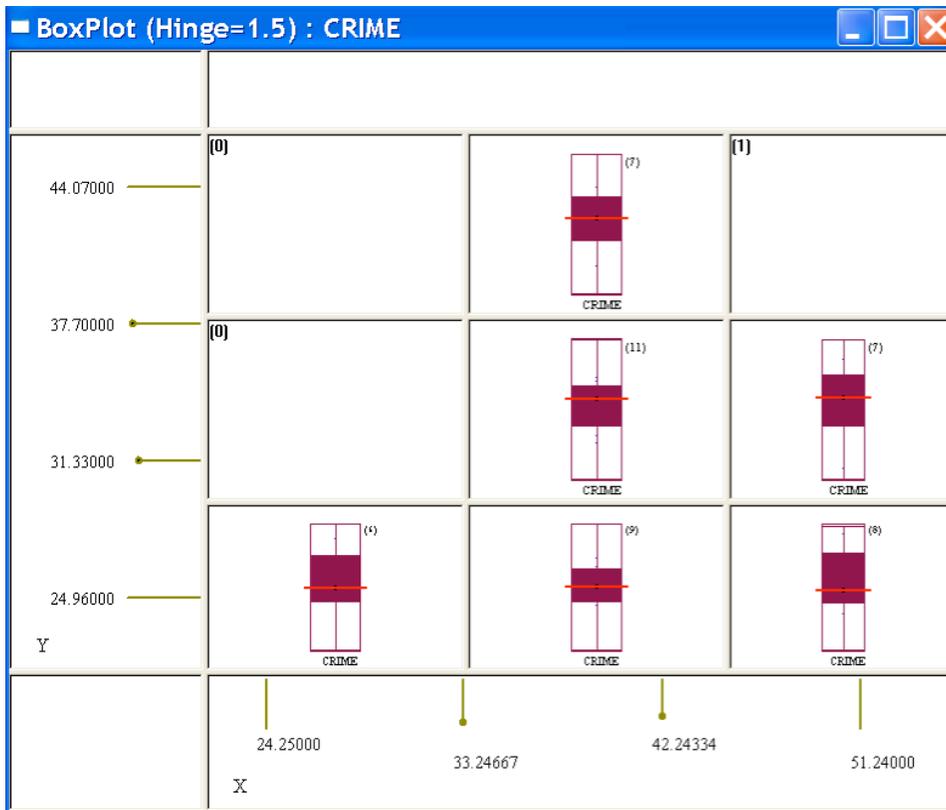
Figure 56: Conditional box plot

A similar approach is taken for the conditional histogram, shown in Figure 57. The categories in the histogram are fixed and pertain to the complete distribution. In order to change this, they need to be adjusted in each cell individually. Each of the cells in the conditional plot shows the

observations that meet the conditioning criteria and where they stack up on the histogram. Each histogram bar shows the number of observations in that class at the top.

Figure 57: Conditional histogram

Finally, the conditional scatter plot is illustrated in Figure 58 for the variables CRIME and INC. In each cell a regression line and its slope are given if at least two observations are present. The location of the points in the plot is always given. For example, in the upper right cell of Figure 58, there is only one observation (one point in the scatter plot). Different slopes in the different cells suggest an interaction effect between the conditioning variables and the linear relation between the two variables considered. If there is no such interaction, then the slopes should be the same in all cells.
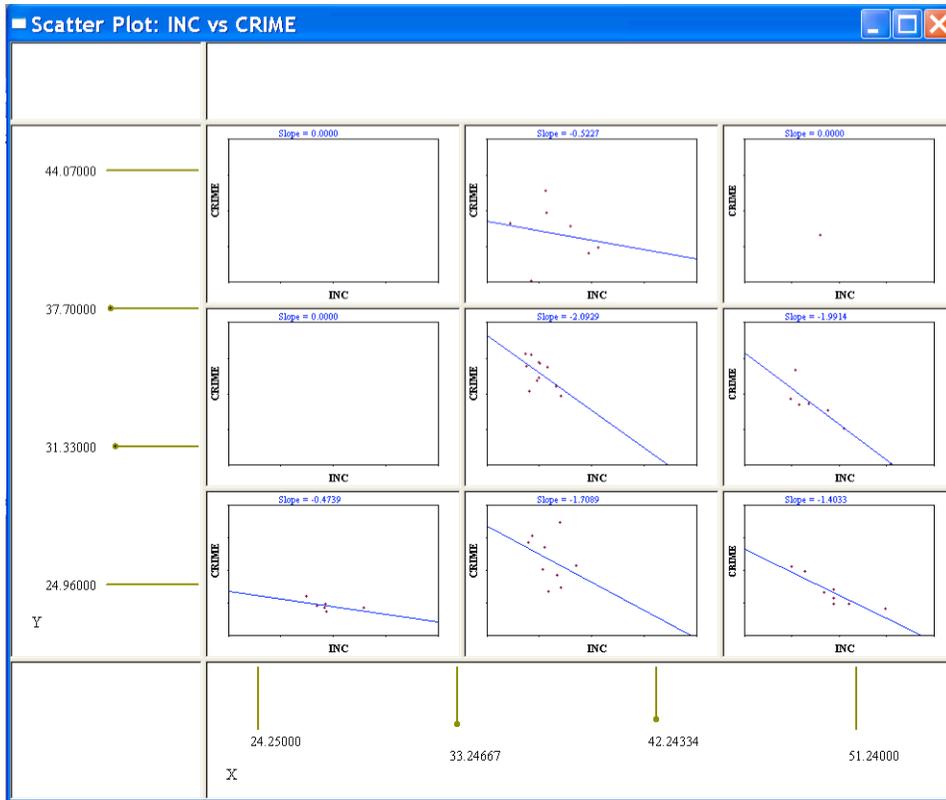
Figure 58: Conditional scatter plot

Histogram

The histogram now uses a different color scheme for the histogram bars. Instead of random color assignment, a continuous color ramp is used, as illustrated in Figure 59.

Box Plot

The box plot has been redesigned as well. Instead of the blue dot to represent the median, this is now shown as a red line that sticks out slightly on both sides of the box. In addition, the number of observations is listed in parentheses at the upper right hand corner, as illustrated in Figure 60.

43
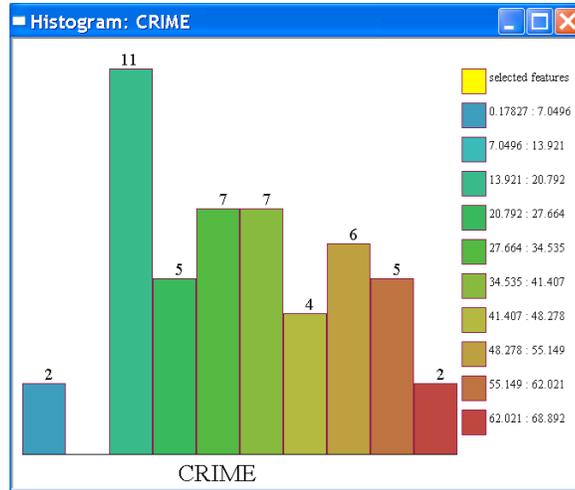
Figure 59: New look histogram
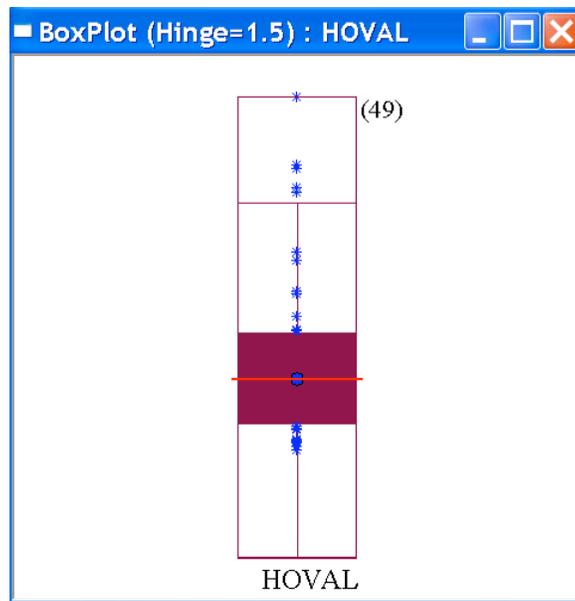
Figure 60: New look box plot

Spatial Regression Analysis

GeoDa now includes some spatial regression functionality. In the current version, this is still fairly limited and experimental, but it works. The user interface in particular is still rudimentary. The basic diagnostics for spatial autocorrelation, heteroskedasticity and non-normality are implemented for the standard ordinary least squares regression. Estimation of spatial lag and spatial error models is supported by means of the Maximum Likelihood method. An extensive overview of the relevant methodology is beyond the scope of this document, but can be found in Anselin and Bera (1998).

The estimation techniques implemented for the Maximum Likelihood approach are based on the algorithms outlined in Smirnov and Anselin (2001). These algorithms were developed to address the estimation of spatial regression models in very large data sets. *GeoDa* has been successfully applied to spatial regression in a data set of 330,000 observations (estimation and inference were complete in a few minutes). A spatial regression using the 3000+ US counties takes a few seconds.

The asymptotic inference consists of a Likelihood Ratio test as well as an estimate of the asymptotic covariance matrix, using a new algorithm developed by Smirnov (2003). All methods use sparse weights of either GAL or GWT format. However, so far, estimation only works for weights that reflect a symmetric spatial arrangement, such as contiguity weights or distance based weights (row-standardized), but *not* for k-nearest neighbor weights.

The regression functionality can be invoked in two different ways. In the opening screen, without loading a shape file, it is activated by selecting `Methods > Regress` (see also p. 9). This is the suggested approach for large data sets (1,000 and up) since it avoids the overhead due to the linking of a large data table. In smaller data sets, the regression can also start within a project, by selecting `Regress` on the main menu. This approach is more appropriate when predicted values and residuals will be used in mapping and further exploratory analysis.

Regression Interface

The `Regress` function starts with a dialog to set some basic parameters for the results and output, as illustrated in Figure 61. The `Report Title` can be ignored, the `Output file name` is the name of the text file to which the results will be written. The default is `Regression.OLS`, which will be the file name used unless a different name is specified, even when the analysis is for a lag or error model.

The next three items determine some additional information that may be included in the output file:

- the `Predicted Value and Residual`: note that this is not the same as the option to save these values to the data table; it only affects the listed output in the output file

- the `Coefficient Variance Matrix`: note that the (asymptotic) standard errors are reported with the coefficient estimates; this option pertains to the complete variance-covariance matrix (including the covariances)

- the `Moran's I z-value`: the default is that this value is *not* reported since the computations involved are substantially slower than those for the Lagrange Multiplier statistics (in Figure 61 this option has been checked)
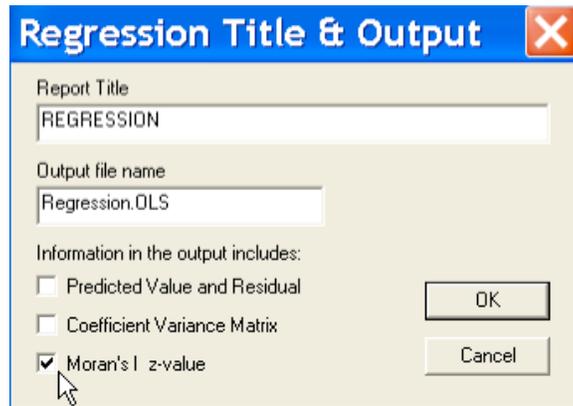
Figure 61: Regression analysis output settings

Clicking the `OK` button will invoke the variable specification dialog for the regression model.

The variable selection dialog is still rudimentary. It uses the >, >>, < and << buttons to move variables from the drop down list to the Dependent Variable box and the Independent Variables list, as shown in Figure 62.
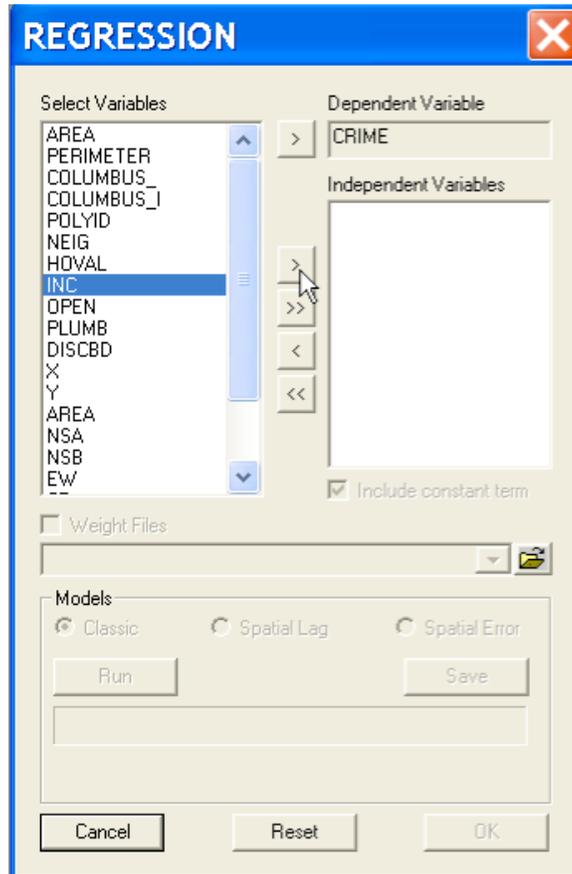
Figure 62: Variable selection for regression analysis

The final step in setting up the regression specification consists of selecting a spatial weights file, as shown in Figure 63. This uses the usual file selection dialog. Also note that a constant term is included by default. Only in very rare circumstances would it make sense to uncheck that box.

The process outlined above is the same for all three regression models: first specify the output options, then select the variables, and close by choosing the spatial weights file.
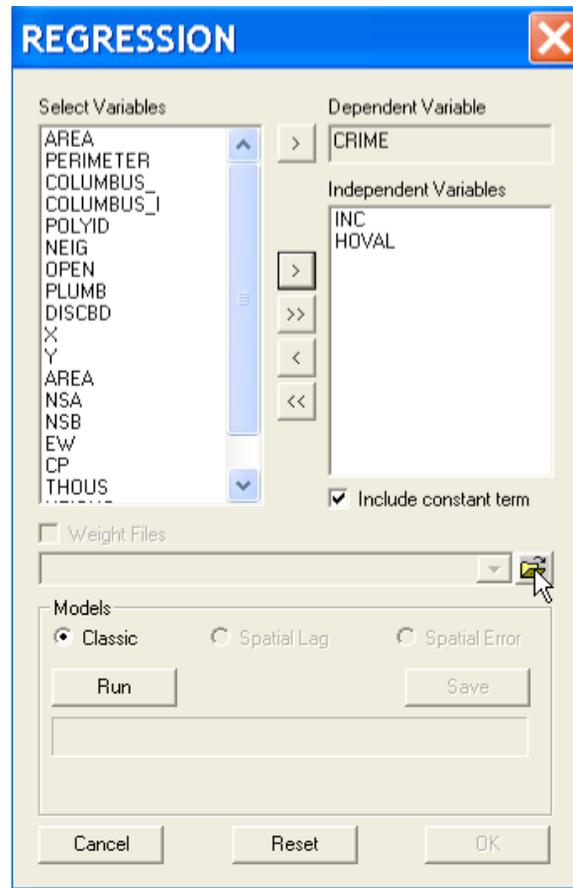
Figure 63: Spatial weights file selection

Ordinary Least Squares with Diagnostics

The model specification is concluded by selecting one of the three radio buttons in the dialog and clicking Run to start the analysis. For example, for the classic regression model this is as shown in Figure 64. After the analysis is over, there are three important choices. The Save button (Figure 65) is to add the predicted value and residuals to the data table (e.g., for mapping or exploratory analysis). This must be done *before* any other options are selected. Clicking this button brings up a dialog to select variable names for these two items, as in Figure 66. Clicking OK finished this process and the new variables are added to the data table (Figure 67).

Checking the OK button (Figure 68) brings up the result file. Note that if several models will be run using the same specification, this button should *not* be checked. In the current setup, once the OK button is selected, the regression analysis is over.

The result file, shown in Figure 69, contains the coefficient estimates, measures of fit and the diagnostics. These include the Bera-Jarque test for non-normality, the Breusch-Pagan, Koenker-Bassett and White tests for heteroskedasticity, and six test statistics for spatial autocorrelation: Moran's I (including a z-value when that option has been checked), Lagrange Multiplier tests for lag, error and both forms, as well as their robust counterparts.
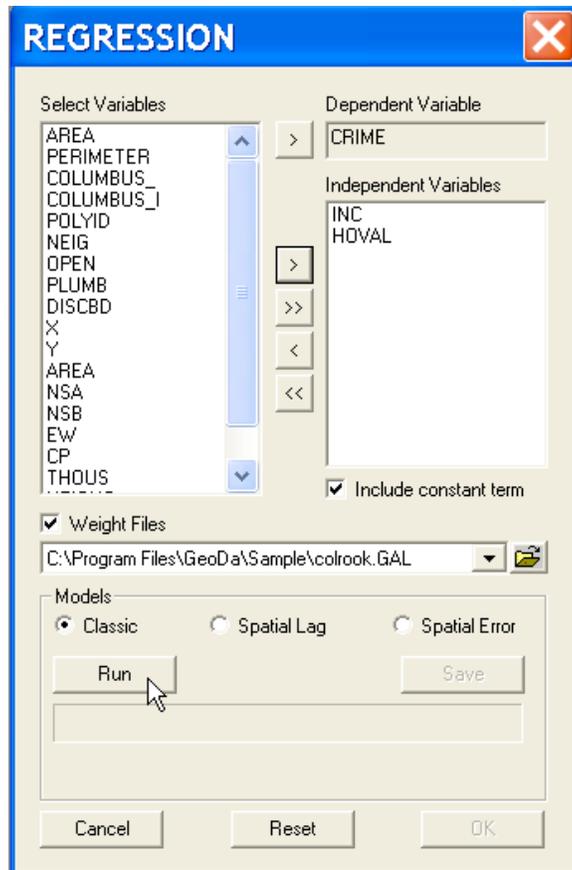
Figure 64: Starting the regression analysis, classic model
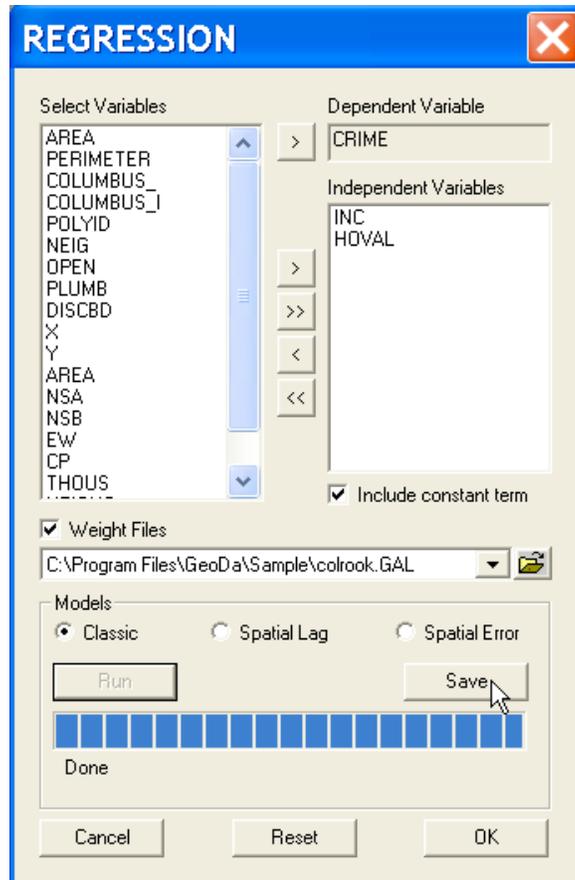
Figure 65: Saving predicted values and residuals, classic model
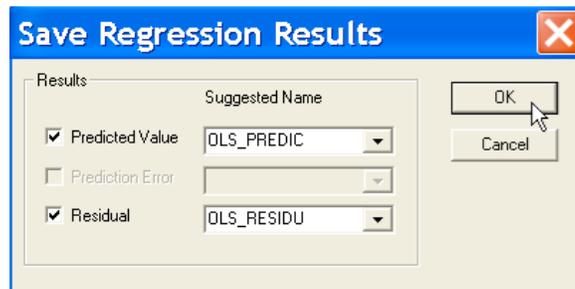
Figure 66: Selecting variable names for saved predicted values and residuals, classic model

NEIGNO	PERIM	OLS_RESIDU	OLS_PREDIC
1005.000000	2.440629	0.346542	15.379438
1001.000000	2.236939	-3.694799	22.496553
1006.000000	2.187547	-5.287394	35.914175
1002.000000	1.427635	-19.985515	52.373275
1007.000000	2.997133	6.447549	44.283961
1008.000000	2.335634	-9.073479	35.140137
1004.000000	2.554577	-34.417722	34.595991

Figure 67: Predicted values and residuals added to table
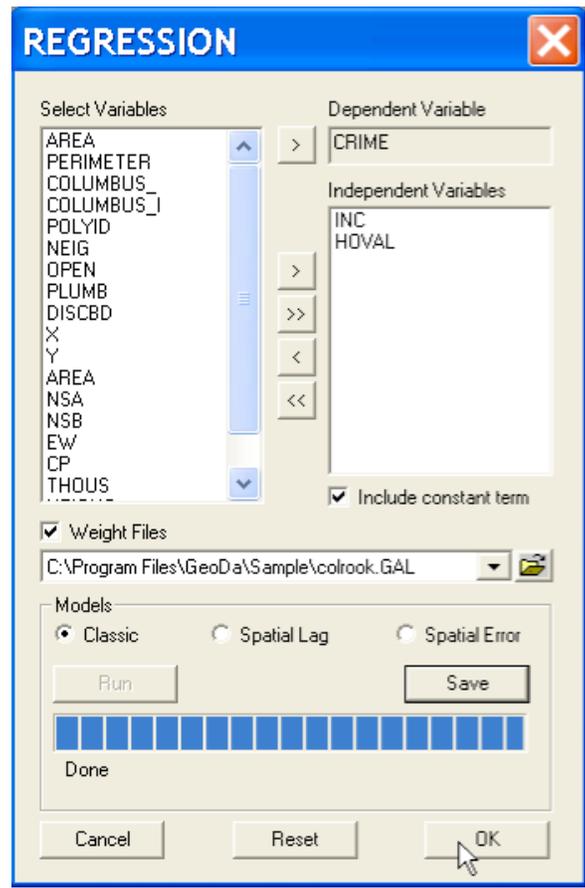
Figure 68: Finishing the regression analysis, classic model

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

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


REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   6.541828
TEST ON NORMALITY OF ERRORS
TEST                   DF        VALUE           PROB
Jarque-Bera            2        1.835753        0.3993663

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                   DF        VALUE           PROB
Breusch-Pagan test     2        7.900442        0.0192505
Koenker-Bassett test   2        5.694088        0.0580156
SPECIFICATION ROBUST TEST
TEST                   DF        VALUE           PROB
White                  5        19.94601        0.0012792

DIAGNOSTICS FOR SPATIAL DEPENDENCE
FOR WEIGHT MATRIX : colrook.GAL  (row-standardized weights)
TEST                        MI/DF     VALUE           PROB
Moran's I (error)           0.249862  2.9376173       0.0033076
Lagrange Multiplier (lag)       1     8.7599071       0.0030792
Robust LM (lag)                 1     3.0721737       0.0796429
Lagrange Multiplier (error)     1     5.8148799       0.0158911
Robust LM (error)               1     0.1271465       0.7214092
Lagrange Multiplier (SARMA)     2     8.8870536       0.0117544
========================= END OF REPORT =============================
```

Figure 69: Results, classic model

Maximum Likelihood in Spatial Lag Model

The process for the spatial lag model is essentially the same as for the classic model. The only difference is that the `Spatial Lag` radio button must be checked, as in Figure 70. The `Save` and `OK` work in the same way. For the spatial lag model, there is a distinction between the residual and the prediction error. The latter is the difference between the observed value and the predicted value that uses only *exogenous* variables, rather than treating the spatial lag Wy as observed.

Figure 70: Spatial regression analysis, lag model

The results reported consist of the estimated coefficients and their asymptotic errors and t-test, measures of fit (log likelihood, AIC and SC), a

Breusch-Pagan test for heteroskedasticity, and a Likelihood Ratio test on the spatial lag parameter.

```
■ Regression.OLS                                                    _ □

REGRESSION
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : COLUMBUS
Spatial Weight      : colrook.GAL
Dependent Variable  :      CRIME  Number of Observations:     49
Mean dependent var  :     35.1288 Number of Variables    :      4
S.D. dependent var  :     16.5605 Degrees of Freedom     :     45
Lag coeff.   (Rho)  :     0.422808

R-squared           :     0.650962 Log likelihood         :     -182.518
Sq. Correlation     : -            Akaike info criterion :      373.035
Sigma-square        :     95.7235  Schwarz criterion      :      380.603
S.E of regression   :     9.78384

-----------------------------------------------------------------------
    Variable    Coefficient     Std.Error    z-value      Probability
-----------------------------------------------------------------------
    W_CRIME      0.4228079      0.1155777     3.658213     0.0002540
    CONSTANT     45.26498       7.175796      6.308008     0.0000000
       INC      -1.036346       0.3052524    -3.395047     0.0006863
     HOVAL      -0.2594178      0.08879673   -2.921479     0.0034839
-----------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                      DF     VALUE        PROB
Breusch-Pagan test                         2     23.57972     0.0000076

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : colrook.GAL
TEST                                      DF     VALUE        PROB
Likelihood Ratio Test                      1     9.719246     0.0018235
========================= END OF REPORT ==============================
```
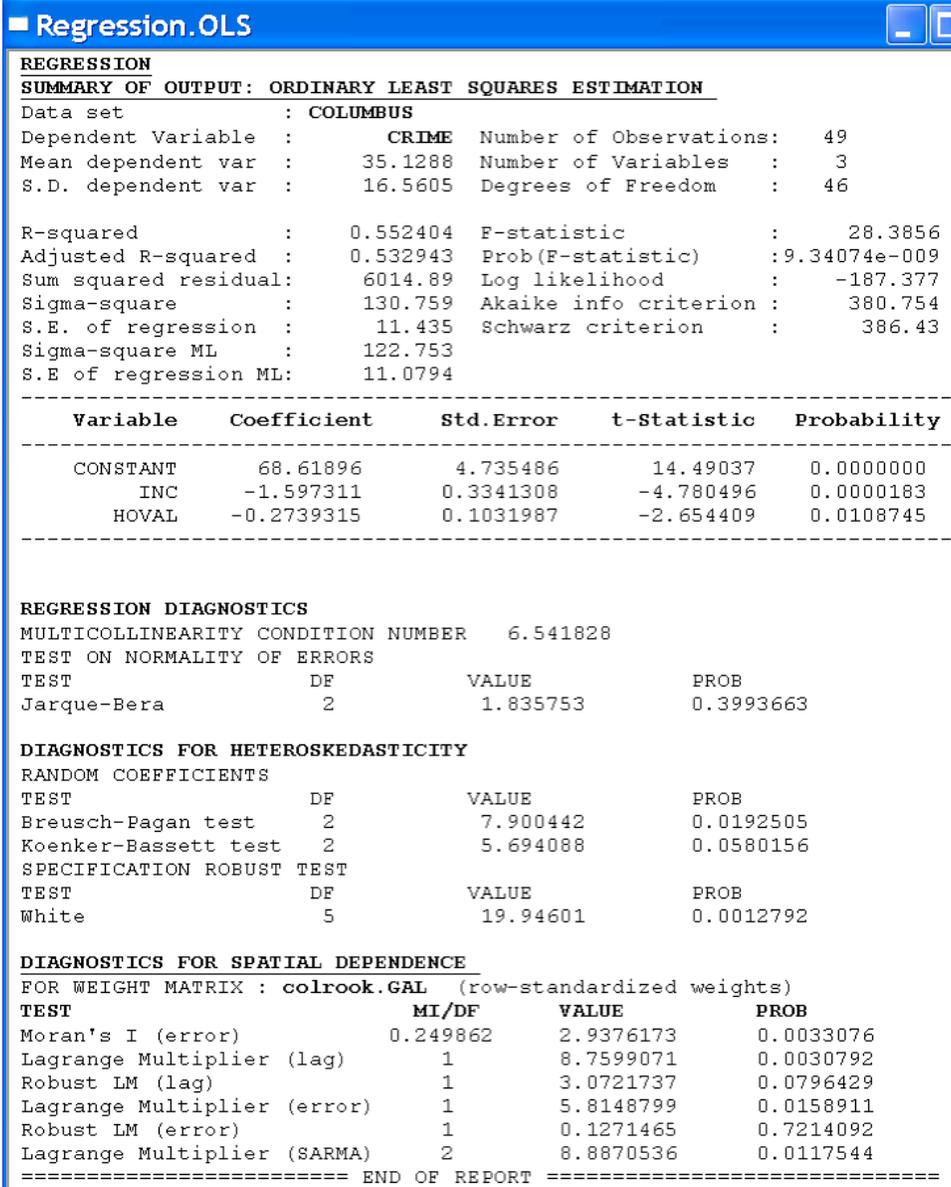
Figure 71: Results, spatial lag model

54

Maximum Likelihood in Spatial Error Model

The process for the spatial lag model is essentially the same as for the classic and lag models. The only difference is that the `Spatial Lag` radio button must be checked, as in Figure 72. The `Save` and `OK` work in the same way. For the spatial error, the prediction error is the difference between observed and predicted y, whereas the "residuals" are the spatially filtered residuals.
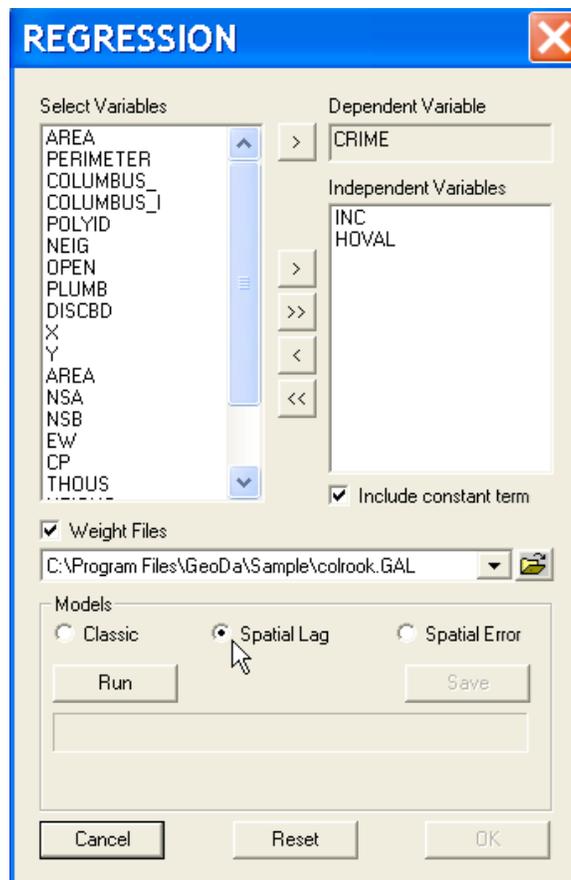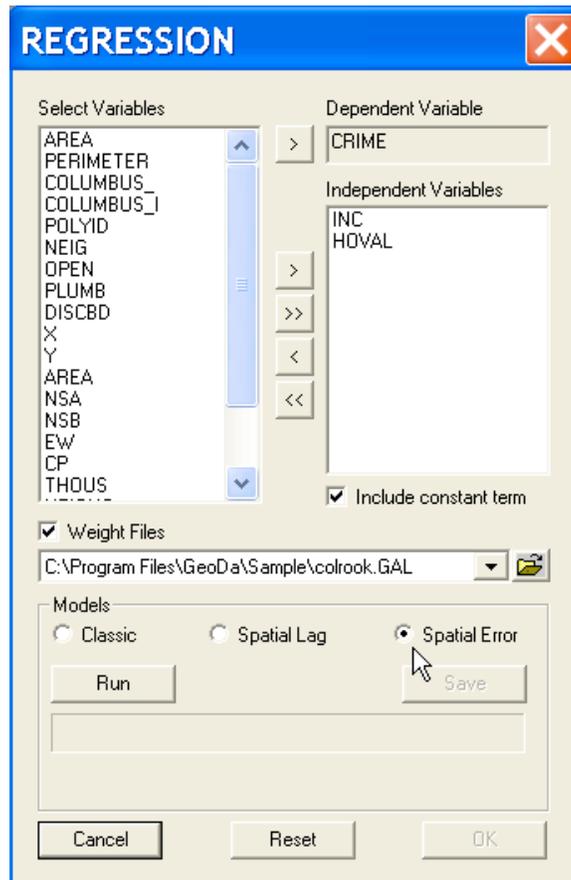

Figure 72: Spatial regression analysis, error model

The results reported consist of the estimated coefficients and their asymptotic errors and t-test, measures of fit (log likelihood, AIC and SC), a Breush-Pagan test for heteroskedasticity, and a Likelihood Ratio test on the spatial error parameter.

```
■ Regression.OLS                                                              ─

REGRESSION
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            :  COLUMBUS
Spatial Weight      :  colrook.GAL
Dependent Variable  :        CRIME   Number of Observations:    49
Mean dependent var  :     35.128824  Number of Variables   :     3
S.D. dependent var  :     16.560476  Degree of Freedom     :    46
Lag coeff. (Lambda) :      0.548474

R-squared           :      0.653718  R-squared (BUSE)      : -
Sq. Correlation     : -             Log likelihood        : -181.894632
Sigma-square        :     94.967743  Akaike info criterion :    369.789
S.E of regression   :      9.74514   Schwarz criterion     :  375.464725

-------------------------------------------------------------------------
     Variable    Coefficient      Std.Error       z-value       Probability
-------------------------------------------------------------------------
     CONSTANT       60.37519         5.32507       11.33791      0.0000000
          INC     -0.9610436       0.3311456       -2.902179     0.0037059
        HOVAL     -0.3031981      0.09264126       -3.27282      0.0010649
       LAMBDA       0.548474       0.1313791       4.174744      0.0000299
-------------------------------------------------------------------------


REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                          DF      VALUE         PROB
Breusch-Pagan test                             2     18.10319     0.0001172

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : colrook.GAL
TEST                                          DF      VALUE         PROB
Likelihood Ratio Test                          1     10.96521     0.0009284
========================= END OF REPORT =============================
```
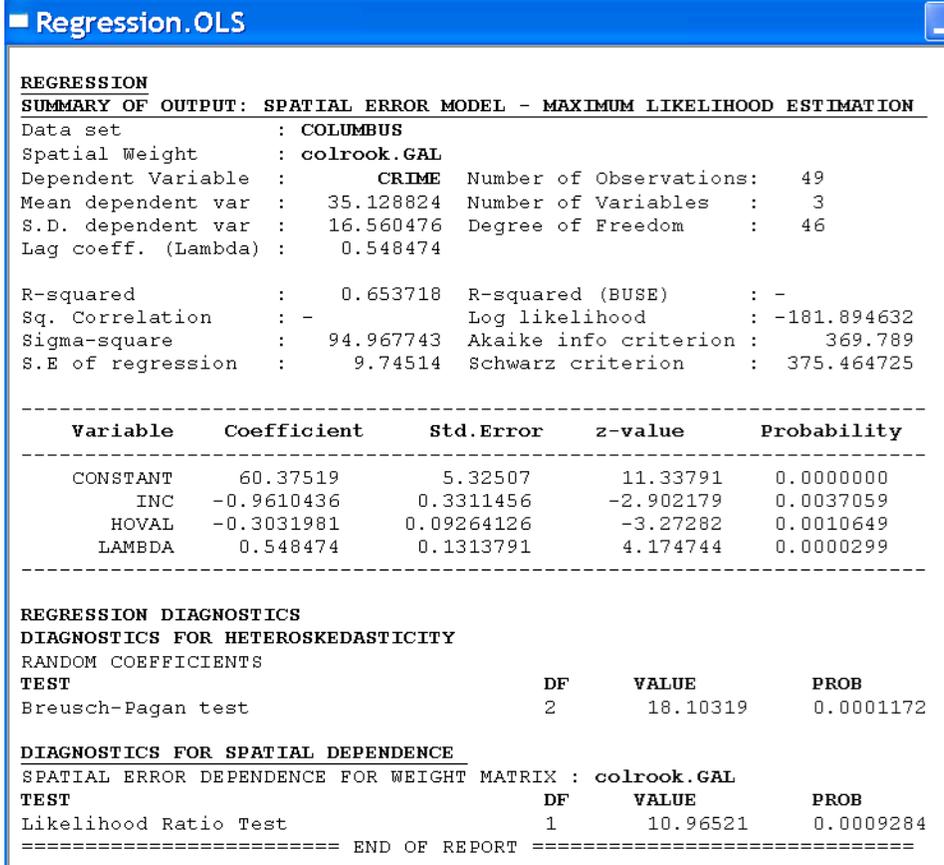
Figure 73: Results, spatial error model

56

Bibliography

Anselin, L. (2003). *GeoDa 0.9 User's Guide*. Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

Anselin, L. and Bera, A. (1998). Spatial dependence in linear regression models with an introduction to spatial econometrics. In Ullah, A. and Giles, D. E., editors, *Handbook of Applied Economic Statistics*, pages 237–289. Marcel Dekker, New York.

Becker, R. A., Cleveland, W., and Shyu, M.-J. (1996). The visual design and control of Trellis displays. *Journal of Computational and Graphical Statistics*, 5:123–155.

Carr, D. B., Chen, J., Bell, S., Pickle, L., and Zhang, Y. (2002). Interactive linked micromap plots and dynamically conditioned choropleth maps. In Anselin, L. and Rey, S., editors, *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*. Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara. CD-ROM.

Dorling, D. (1996). *Area Cartograms: Their Use and Creation*. CATMOG 59, Institute of British Geographers.

Inselberg, A. (1985). The plane with parallel coordinates. *Visual Computer*, 1:69–91.

Smirnov, O. (2003). Computation of the information matrix for models of spatial interaction. Technical report, Regional Economics Applications Laboratory (REAL), University of Illinois, Urbana-Champaign, IL.

Smirnov, O. and Anselin, L. (2001). Fast maximum likelihood estimation of very large spatial autoregressive models: A characteristic polynomial approach. *Computational Statistics and Data Analysis*, 35:301–319.

Wegman, E. J. (1990). Hyperdimensional data analysis using parallel coordinates. *Journal of the American Statistical Association*, 85:664–675.