GeoDa: An Introduction to Spatial Data Analysis[*]

Luc Anselin, Ibnu Syabri and Youngihn Kho
Spatial Analysis Laboratory
Department of Agricultural and Consumer Economics
University of Illinois, Urbana-Champaign
Urbana, IL 61801
USA
anselin@uiuc.edu, syabri@uiuc.edu, kho@uiuc.edu

May 5, 2004

Abstract

This paper presents an overview of *GeoDa*$^{\text{TM}}$, a free software program intended to serve as a user-friendly and graphical introduction to spatial analysis for non-GIS specialists. It includes functionality ranging from simple mapping to exploratory data analysis, the visualization of global and local spatial autocorrelation, and spatial regression. A key feature of *GeoDa* is an interactive environment that combines maps with statistical graphics, using the technology of dynamically linked windows. A brief review of the software design is given, as well as some illustrative examples that highlight distinctive features of the program in applications dealing with public health, economic development, real estate analysis and criminology.

Key Words: geovisualization, exploratory spatial data analysis, spatial outliers, smoothing, spatial autocorrelation, spatial regression.

1 Introduction

The development of specialized software for spatial data analysis has seen rapid growth since the lack of such tools was lamented in the late 1980s by Haining

[*] This research was supported in part by US 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. In addition, this research was made possible in part through a Cooperative Agreement between the Center for Disease Control and Prevention (CDC) and the Association of Teachers of Preventive Medicine (ATPM), award number TS-1125. The contents of the paper are the responsibility of the authors and do not necessarily reflect the official views of NSF, NCI, the CDC or ATPM. Special thanks go to Oleg Smirnov for his assistance with the implementation of the spatial regression routines, and to Julie Le Gallo and Julia Koschinsky for preparing, respectively, the data set for the European convergence study and for the Seattle house prices. *GeoDa*$^{\text{TM}}$ is a trademark of Luc Anselin.

(1989) and cited as a major impediment to the adoption and use of spatial statistics by GIS researchers. Initially, attention tended to focus on conceptual issues, such as how to integrate spatial statistical methods and a GIS environment (loosely vs. tightly coupled, embedded vs. modular, etc.), and which techniques would be most fruitfully included in such a framework. Familiar reviews of these issues are represented in, among others, Anselin and Getis (1992), Goodchild et al. (1992), Fischer and Nijkamp (1993), Fotheringham and Rogerson (1993, 1994), Fischer et al. (1996), and Fischer and Getis (1997). Today, the situation is quite different, and a fairly substantial collection of spatial data analysis software is readily available, ranging from niche programs, customized scripts and extensions for commercial statistical and GIS packages, to a burgeoning open source effort using software environments such as R, Java and Python. This is exemplified by the growing contents of the software tools clearing house maintained by the U.S.-based Center for Spatially Integrated Social Science (CSISS).[1]

CSISS was established in 1999 as a research infrastructure project funded by the U.S. National Science Foundation in order to promote a spatial analytical perspective in the social sciences (Goodchild et al. 2000). It was readily recognized that a major instrument in disseminating and facilitating spatial data analysis would be an easy to use, visual and interactive software package, aimed at the non-GIS user and requiring as little as possible in terms of other software (such as GIS or statistical packages). *GeoDa* is the outcome of this effort. It is envisaged as an “introduction to spatial data analysis” where the latter is taken to consist of visualization, exploration and explanation of *interesting* patterns in geographic data.

The main objective of the software is to provide the user with a natural path through an empirical spatial data analysis exercise, starting with simple mapping and geovisualization, moving on to exploration, spatial autocorrelation analysis, and ending up with spatial regression. In many respects, *GeoDa* is a reinvention of the original *SpaceStat* package (Anselin 1992), which by now has become quite dated, with only a rudimentary user interface, an antiquated architecture and performance constraints for medium and large data sets. The software was redesigned and rewritten from scratch, around the central concept of dynamically linked graphics. This means that different “views” of the data are represented as graphs, maps or tables with selected observations in one highlighted in all. In that respect, *GeoDa* is similar to a number of other modern spatial data analysis software tools, although it is quite distinct in its combination of user friendliness with an extensive range of incorporated methods. A few illustrative comparisons will help clarify its position in the current spatial analysis software landscape.

In terms of the range of spatial statistical techniques included, *GeoDa* is most alike to the collection of functions developed in the open source R environment. For example, descriptive spatial autocorrelation measures, rate smoothing and spatial regression are included in the *spdep* package, as described by Bivand and

[1] See http://www.csiss.org/clearinghouse/.

Gebhardt (2000), Bivand (2002a,b), and Bivand and Portnov (2004). In contrast to R, *GeoDa* is completely driven by a point and click interface and does not require any programming. It also has more extensive mapping capability (still somewhat experimental in R) and full linking and brushing in dynamic graphics, which is currently not possible in R due to limitations in its architecture. On the other hand, *GeoDa* is not (yet) customizable or extensible by the user, which is one of the strengths of the R environment. In that sense, the two are seen as highly complementary, ideally with more sophisticated users "graduating" to R after being introduced to the techniques in *GeoDa*.[2]

The use of dynamic linking and brushing as a central organizing technique for data visualization has a strong tradition in exploratory data analysis (EDA), going back to the notion of linked scatterplot brushing (Stuetzle 1987), and various methods for dynamic graphics outlined in Cleveland and McGill (1988). In geographical analysis, the concept of "geographic brushing" was introduced by Monmonier (1989) and made operational in the *Spider/Regard* toolboxes of Haslett, Unwin and associates (Haslett et al. 1990, Unwin 1994). Several modern toolkits for exploratory spatial data analysis (ESDA) also incorporate dynamic linking, and, to a lesser extent, brushing. Some of these rely on interaction with a GIS for the map component, such as the linked frameworks combining XGobi or XploRe with ArcView (Cook et al. 1996, 1997, Symanzik et al. 2000), the SAGE toolbox, which uses ArcInfo (Wise et al. 2001), and the DynESDA extension for ArcView (Anselin 2000), *GeoDa*'s immediate predecessor. Linking in these implementations is constrained by the architecture of the GIS, which limits the linking process to a single map (in *GeoDa*, there is no limit on the number of linked maps). In this respect, *GeoDa* is similar to other freestanding modern implementations of ESDA, such as the cartographic data visualizer, or *cdv* (Dykes 1997), GeoVISTA Studio (Takatsuka and Gahegan 2002) and STARS (Rey and Janikas 2004). These all include functionality for dynamic linking, and to a lesser extent, brushing. They are built in open source programming environments, such as Tkl/Tk (cdv), Java (GeoVISTA Studio) or Python (STARS) and thus easily extensible and customizable. In contrast, *GeoDa* is (still) a closed box, but of these packages it provides the most extensive and flexible form of dynamic linking and brushing for both graphs and maps.

Common spatial autocorrelation statistics, such as Moran's I and even the Local Moran are increasingly part of spatial analysis software, ranging from CrimeStat (Levine 2004), to the *spdep* and *DCluster* packages available on the open source Comprehensive R Archive Network (CRAN),[3] as well as commercial packages, such as the spatial statistics toolbox of the forthcoming release of ArcGIS 9.0 (ESRI 2004). However, at this point in time, none of these include the range and ease of construction of spatial weights, or the capacity to carry out sensitivity analysis and visualization of these statistics contained in *GeoDa*. Apart from the R *spdep* package, *Geoda* is the only one to contain functionality

[2] Note that the CSISS spatial tools project is an active participant in the development of spatial data analysis methods in R, see, e.g., http://sal.agecon.uiuc.edu/csiss/Rgeo/

[3] http://cran.r-project.org/

for spatial regression modeling among the software mentioned here.

A prototype version of the software (known as *DynESDA*) has been in limited circulation since early 2001 (Anselin et al. 2002a,b), but the first official release of a beta version of *GeoDa* occurred on February 5, 2002. The program is available for free and can be downloaded from the CSISS software tools web site (http://sal.agecon.uiuc.edu/geoda_main.php).The most recent version, 0.9.5-i, was released in January 2003. The software has been well received for both teaching and research use and has a rapidly growing body of users. For example, after slightly more than a year since the initial release (i.e., as of the end of April 2004), the number of registered users exceeds 1,800, while increasing at a rate of about 150 new users per month.

In the remainder of the paper, we first outline the design and briefly review the overall functionality of *GeoDa*. This is followed by a series of illustrative examples, highlighting features of the mapping and geovisualization capabilities, exploration in multivariate EDA, spatial autocorrelation analysis, and spatial regression. The paper closes with some comments regarding future directions in the development of the software.

2 Design and Functionality

The design of *GeoDa* consists of an interactive environment that combines maps with statistical graphs, using the technology of dynamically linked windows. It is geared to the analysis of *discrete* geospatial data, i.e., objects characterized by their location in space either as points (point coordinates) or polygons (polygon boundary coordinates). The current version adheres to ESRI's shape file as the standard for storing spatial information. It contains functionality to read and write such files, as well as to convert ascii text input files for point coordinates or boundary file coordinates to the shape file format. It uses ESRI's MapObjects LT2 technology for spatial data access, mapping and querying. The analytical functionality is implemented in a modular fashion, as a collection of C++ classes with associated methods.

In broad terms, the functionality can be classified into six categories:

- *spatial data manipulation and utilities*: data input, output, and conversion
- *data transformation*: variable transformations and creation of new variables
- *mapping*: choropleth maps, cartogram and map animation
- *EDA*: statistical graphics
- *spatial autocorrelation*: global and local spatial autocorrelation statistics, with inference and visualization
- *spatial regression*: diagnostics and maximum likelihood estimation of linear spatial regression models

The full set of functions is listed in Table 1 and is documented in detail in the *GeoDa* User's Guides (Anselin 2003, 2004).[4]

The software implementation consists of two important components: the user interface and graphics windows on the one hand, and the computational engine on the other hand. In the current version, all graphic windows are based on Microsoft Foundation Classes (MFC) and thus are limited to MS Windows platforms.[5] In contrast, the computational engine (including statistical operations, randomization, and spatial regression) is pure C++ code and largely cross platform.

The bulk of the graphical interface implements five basic classes of windows: histogram, box plot, scatter plot (including the Moran scatter plot), map and grid (for the table selection and calculations). The choropleth maps, including the significance and cluster maps for the local indicators of spatial autocorrelation (LISA) are derived from MapObjects classes. Three additional types of maps were developed from scratch and do not use MapObjects: the map movie (map animation), the cartogram, and the conditional maps. The three dimensional scatter plot is implemented with the OpenGL library.

The functionality of *GeoDa* is invoked either through menu items or directly by clicking toolbar buttons, as illustrated in Figure 1. A number of specific applications are highlighted in the following sections, focusing on some distinctive features of the software.

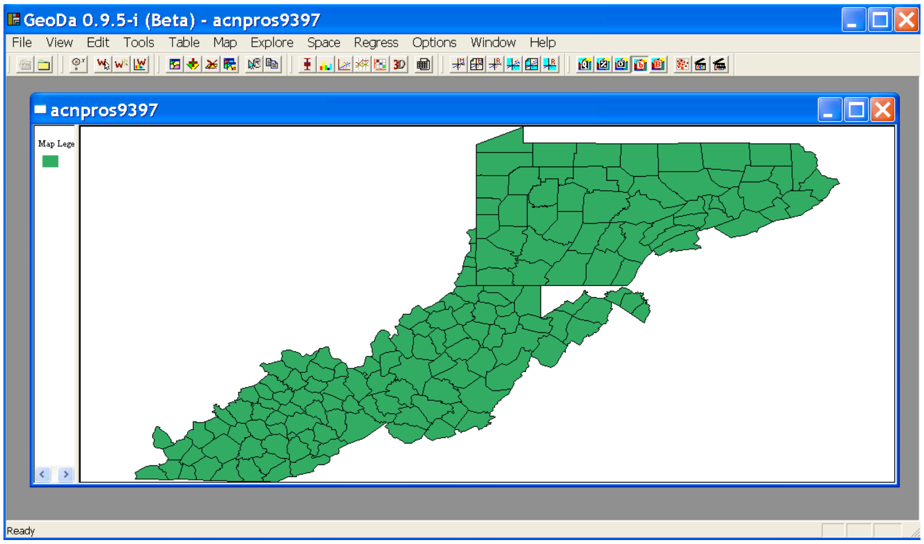


Figure 1: The opening screen with menu items and toolbar buttons

[4] A Quicktime movie with a demonstration of the main features can be found at http://sal.agecon.uiuc.edu/movies/GeoDaDemo.mov.

[5] Ongoing development concerns the porting of all MFC based classes to a cross-platform architecture, using wxWindows. See also Section 7.

Table 1: *GeoDa* Functionality Overview

Category	Functions
Spatial Data	data input from shape file (point, polygon)
	data input from text (to point or polygon shape)
	data output to text (data or shape file)
	create grid polygon shape file from text input
	centroid computation
	Thiessen polygons
Data Transformation	variable transformation (log, exp, etc.)
	queries, dummy variables (regime variables)
	variable algebra (addition, multiplication, etc.)
	spatial lag variable construction
	rate calculation and rate smoothing
	data table join
Mapping	generic quantile choropleth map
	standard deviational map
	percentile map
	outlier map (box map)
	circular cartogram
	map movie
	conditional maps
	smoothed rate map (EB, spatial smoother)
	excess rate map (standardized mortality rate, SMR)
EDA	histogram
	box plot
	scatter plot
	parallel coordinate plot
	three-dimensional scatter plot
	conditional plot (histogram, box plot, scatter plot)
Spatial Autocorrelation	spatial weights creation (rook, queen, distance, k-nearest)
	higher order spatial weights
	spatial weights characteristics (connectedness histogram)
	Moran scatterplot with inference
	bivariate Moran scatterplot with inference
	Moran scatterplot for rates (EB standardization)
	Local Moran significance map
	Local Moran cluster map
	bivariate Local Moran
	Local Moran for rates (EB standardization)
Spatial Regression	OLS with diagnostics (e.g., LM test, Moran's I)
	Maximum Likelihood spatial lag model
	Maximum Likelihood spatial error model
	predicted value map
	residual map

3 Mapping and Geovisualization

The bulk of the mapping and geovisualization functionality consists of a collection of specialized choropleth maps, focused on highlighting outliers in the data, so-called *box maps* (Anselin 1999). In addition, considerable capability is included to deal with the intrinsic variance instability of rates, in the form of empirical Bayes (EB) or spatial smoothers.[6] As mentioned in Section 2, the mapping operations use the classes contained in ESRI's MapObjects, extended with the capability for linking and brushing. *GeoDa* also includes a circular cartogram,[7] map animation in the form of a map movie, and conditional maps. The latter are nine micro choropleth maps constructed by conditioning on three intervals for two conditioning variables, using the principles outlined in Becker et al. (1996) and Carr et al. (2002).[8] In contrast to the traditional choropleth maps, the cartogram, map movie and conditional maps do not use MapObjects classes, and were developed from scratch.

We illustrate the rate smoothing procedure, outlier maps and linking operations. The objective in this analysis is to identify locations that have elevated mortality rates and to assess the sensitivity of the designation as outlier to the effect of rate smoothing. Using data on prostate cancer mortality in 156 counties contained in the Appalachian Cancer Network (ACN), for the period 1993-97, we construct a box map by specifying the number of deaths as the numerator and the population as the denominator.[9] The resulting map for the crude rates (i.e., without any adjustments for differing age distributions or other relevant factors) is shown as the upper-left panel in Figure 2. Three counties are identified as *outliers* and shown in dark red.[10] These match the outliers *selected* in the box plot in the lower-left panel of the figure. The *linking* of all maps and graphs results in those counties also being cross-hatched on the maps.

The upper-right panel in the Figure represents a smoothed rate map, where the rates were transformed by means of an Empirical Bayes procedure to remove the effect of the varying population at risk. As a result, the original outliers are no longer, but a different county is identified as having elevated risk. Also, a lower outlier is found as well, shown as dark blue in the box map.[11] Note that the upper outlier is barely distinguishable, due to the small area of the county in question. This is a common problem when working with admininistrative units. In order to remove the potentially misleading effect of area on the perception of interesting patterns, a circular cartogram is shown in the lower-right panel

[6] The EB procedure is due to Clayton and Kaldor (1987), see also Marshall (1991) and Bailey and Gatrell (1995), pp. 303-308. For an alternative recent software implementation, see Anselin et al. (2004). Spatial smoothing is discussed at length in Kafadar (1996).

[7] The cartogram is constructed using the non-linear cellular automata algorithm due to Dorling (1996).

[8] The conditional maps are part of a larger set of conditional plots, which includes histograms, box plots and scatter plots.

[9] Data obtained from the the National Cancer Institute SEER site (Surveillance, Epidemiology and End Results), http://seer.cancer.gov/seerstat/.

[10] The respective counties are Cumberland, KY, Pocahontas, WV, and Forest, PA.

[11] The new upper outlier is Ohio county, WV, the lower outlier is Centre county, PA.

of Figure 2, where the area of the circles is proportional to the value of the EB smoothed rate. The upper outlier is shown as a red circle, the lower outlier as a blue circle. The yellow circles are the counties that were outliers in the crude rate map, highlighted here as a result of linking with the other maps and graphs.[12]

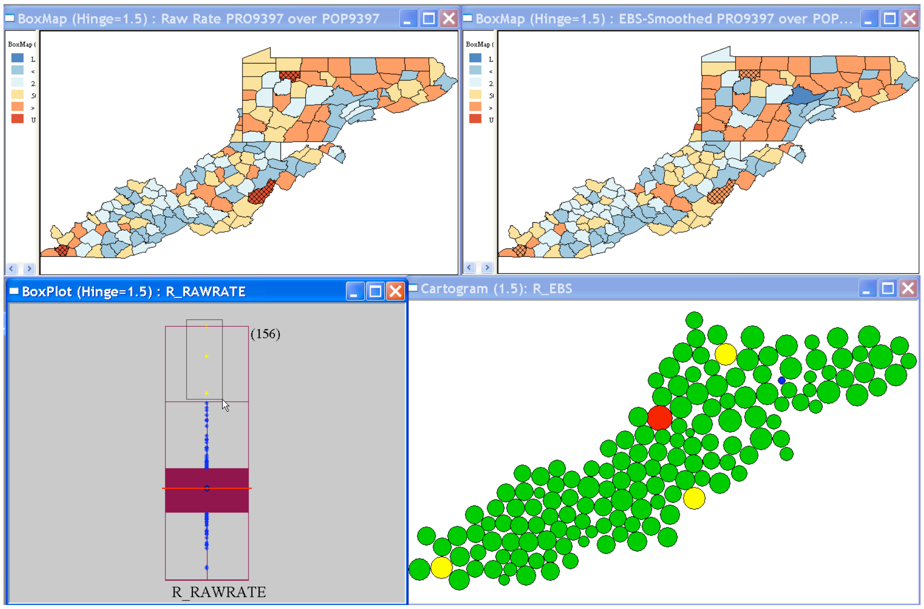

BoxMap (Hinge=1.5) : Raw Rate PRO9397 over POP9397
BoxMap (Hinge=1.5) : EBS-Smoothed PRO9397 over POP...
BoxPlot (Hinge=1.5) : R_RAWRATE
(156)
R_RAWRATE
Cartogram (1.5): R_EBS

Figure 2: Linked box maps, box plot and cartogram, raw and smoothed prostate cancer mortality rates.

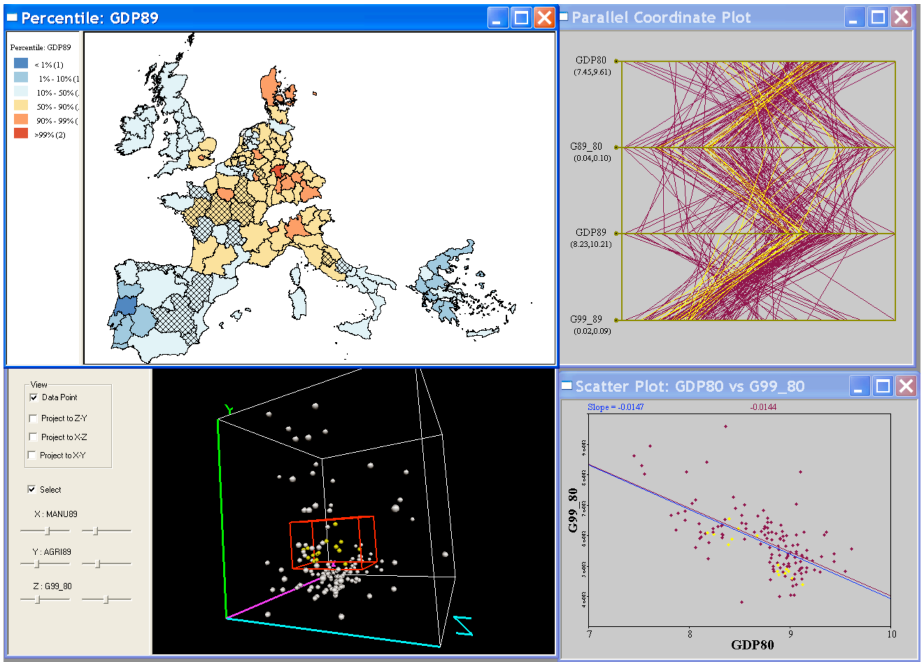
4 Multivariate EDA

Multivariate exploratory data analysis is implemented in *GeoDa* through linking and brushing between a collection of statistical graphs. These include the usual histogram, box plot and scatter plot, but also a parallel coordinate plot (PCP) and three-dimensional scatter plot, as well as conditional plots (conditional histogram, box plot and scatter plot).

We illustrate some of this functionality with an exploration of the relationships between economic growth and initial development, typical of the recent “spatial” regional convergence literature (for an overview, see Rey 2004). We use economic data over the period 1980-1999 for 145 European regions, most of

[12] Note that the outliers identified may be misleading since the rate analyzed is not adjusted for differences in age distribution. In other words, the outliers shown may simply be counties with a larger proportion of older males. A much more detailed analysis is necessary before any policy conclusions may be drawn.

them at the NUTS II level of spatial aggregation, except for a few at the NUTS I level (for Luxembourg and the United Kingdom).[13]



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

☑ Select

X : MANU89

Y : AGRI89

Z : G99_80

Figure 3: Multivariate exploratory data analysis with linking and brushing.

Figure 3 illustrates the various linked plots and map. The left-hand panel contains a simple percentile map (GDP per capital in 1989), and a three-dimensional scatter plot (for the percent agricultural and manufacturing employment in 1989 as well as the GDP growth rate over the period 1980-99). In the top right-hand panel is a PCP for the growth rates in the two periods of interest (1980-89 and 1989-99) and the GDP per capita in the base year, the typical components of a convergence regression. In the bottom of the right-hand panel is a simple scatter plot of the growth rate in the full period (1980-99) on the base year GDP.

Both plots on the right hand side illustrate the typical empirical phenomenon that higher GDP at the start of the period is associated with a lower growth rate. However, as demonstrated in the PCP (some of the lines suggest a positive relation between GDP and growth rate), the pattern is not uniform and there

[13] The data are from the most recent version of the NewCronos Regio database by Eurostat. NUTS stands for "Nomenclature of Territorial Units for Statistics" and contains the definition of administrative regions in the EU member states. NUTS II level regions are roughly comparable to counties in the U.S. context and are available for all but two countries. Luxembourg constitutes only a single region. For the United Kingdom, data is not available at the NUTS II level, since these regions do not correspond to local governmental units.

is a suggestion of heterogeneity. A further exploration of this heterogeneity can be carried out by brushing any one of these graphs. For example, in Figure 3, a selection box in the three-dimensional scatter plot is moved around (*brushing*) which highlights the selected observations in the map (cross-hatched) and in the PCP, clearly showing opposite patterns in subsets of the selection. Furthermore, in the scatter plot, the slope of the regression line can be recalculated for a subset of the data without the selected locations, to assess the sensitivity of the slope to those observations. In the example shown here, the effect on convergence over the whole period is minimal (-0.147 vs. -0.144), but other selections show a more pronounced effect. Further exploration of these patterns does suggest a degree of spatial heterogeneity in the convergence results (for a detailed investigation, see Le Gallo and Dall'erba 2003).

5 Spatial Autocorrelation Analysis

Spatial autocorrelation analysis includes tests and visualization of both global (test for *clustering*) and local (test for *clusters*) Moran's I statistic. The global test is visualized by means of a Moran scatterplot (Anselin 1996), in which the slope of the regression line corresponds to Moran's I. Significance is based on a permutation test. The traditional univariate Moran scatterplot has been extended to depict bivariate spatial autocorrelation as well, i.e., the correlation between one variable at a location, and a different variable at the neighboring locations (Anselin et al. 2002a). In addition, there also is an option to standardize rates for the potentially biasing effect of variance instability (see Assunção and Reis 1999).

Local analysis is based on the Local Moran statistic (Anselin 1995), visualized in the form of significance and cluster maps. It also includes several options for sensitivity analysis, such as changing the number of permutations (to as many as 9999), re-running the permutations several times, and changing the significance cut off value. This provides an ad hoc approach to assess the sensitivity of the results to problems due to multiple comparisons (i.e., how stable is the indication of clusters or outliers when the significance barrier is lowered).

The maps depict the locations with significant Local Moran statistics (LISA significance maps) and classify those locations by type of association (LISA cluster maps). Both types of maps are available for brushing and linking. In addition to these two maps, the standard output of a LISA analysis includes a Moran scatter plot and a box plot depicting the distribution of the local statistic. Similar to the Moran scatter plot, the LISA concept has also been extended to a bivariate setup and includes an option to standardize for variance instability of rates.

The functionality for spatial autocorrelation analysis is rounded out by a range of operations to construct spatial weights, using either boundary files (contiguity based) or point locations (distance based). A connectivity histogram helps in identifying potential problems with the neighbor structure, such as "islands" (locations without neighbors).

We illustrate spatial autocorrelation analysis with a study of the spatial distribution of 692 house sales prices for 1997 in Seattle, WA. This is part of a broader investigation into the effect of subsidized housing on the real estate market.[14] For the purposes of this example, we only focus on the univariate spatial distribution, and the location of any significant clusters or spatial outliers in the data.

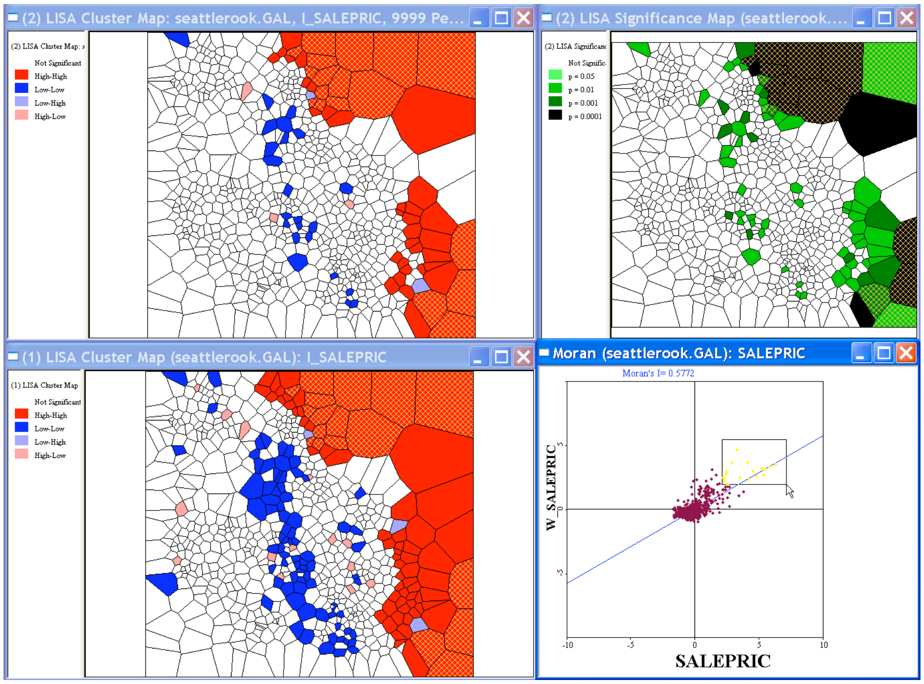


Figure 4: LISA cluster maps and significance maps.

The original house sales data are for point locations, which, for the purposes of this analysis are converted to Thiessen polygons. This allows a definition of "neighbor" based on common boundaries between the Thiessen polygons. On the left hand panel of Figure 4, two LISA cluster maps are shown, depicting the locations of significant Local Moran's I statistics, classified by type of spatial association. The dark red and dark blue locations are indications of spatial *clusters* (respectively, high surrounded by high, and low surrounded by low).[15] In contrast, the light red and light blue are indications of *spatial outliers* (respectively, high surrounded by low, and low surrounded by high). The bottom

[14] The data are from the King County (Washington State) Department of Assessments.

[15] More precisely, the locations highlighted show the "core" of a cluster. The cluster itself can be thought of as consisting of the core as well as the neighbors. Clearly some of these clusters are overlapping.

map uses the default significance of $p = 0.05$, whereas the top map is based on $p = 0.01$ (after carrying out 9999 permutations). The matching *significance map* is in the top right hand panel of Figure 4. Significance is indicated by darker shades of green, with the darkest corresponding to $p = 0.0001$. Note how the tighter significance criterion eleminates some (but not that many) locations from the map. In the bottom right hand panel of the Figure, the corresponding Moran scatterplot is shown, with the most extreme "high-high" locations selected. These are shown as cross-hatched polygons in the maps, and almost all obtain highly significant (at $p = 0.0001$) local Moran's I statistics.

The overall pattern depicts a cluster of high priced houses on the East side, with a cluster of low priced houses following an axis through the center. Put in context, this is not surprising, since the East side represents houses with a lake view, while the center cluster follows a highway axis and generally corresponds with a lower income neighborhood. Interestingly, the pattern is not uniform, and several spatial outliers can be distinguished. Further investigation of these patterns would require a full hedonic regression analysis.

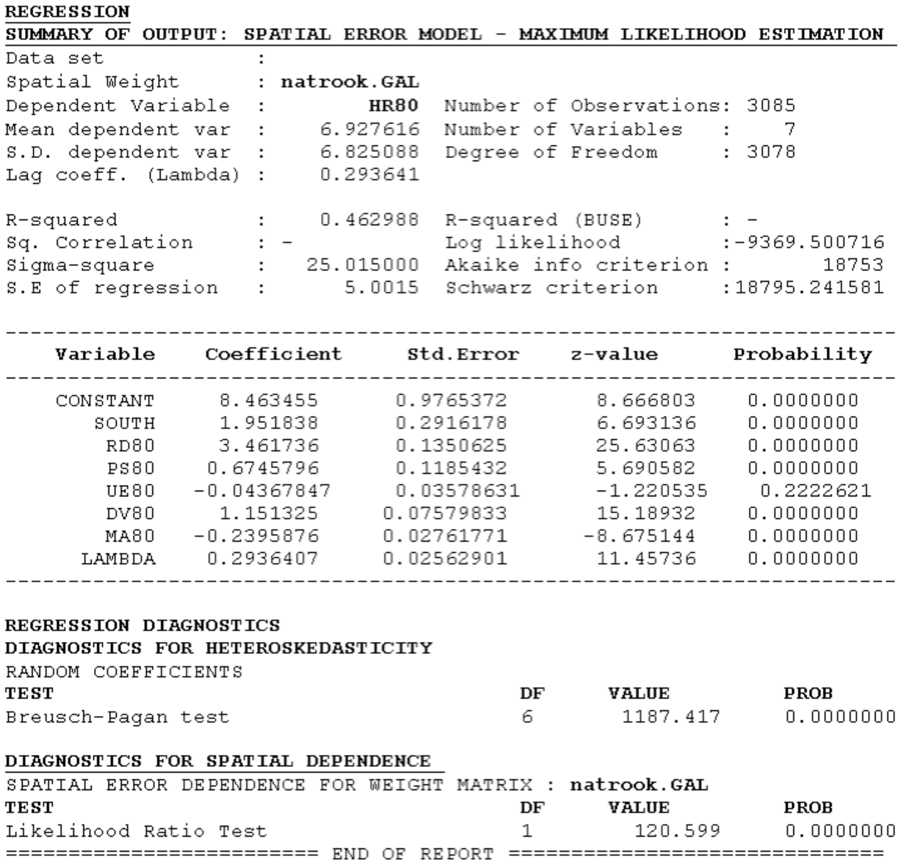
6 Spatial Regression

As of version 0.9.5-i, *GeoDa* also includes a limited degree of spatial regression functionality. The basic diagnostics for spatial autocorrelation, heteroskedasticity and non-normality are implemented for the standard ordinary least squares regression. Estimation of the spatial lag and spatial error models is supported by means of the Maximum Likelihood (ML) method (see Anselin and Bera 1998, for a review of the technical issues). In addition to the estimation itself, predicted values and residuals are calculated and made available for mapping.

The ML estimation in *GeoDa* distinguishes itself by the use of extremely efficient algorithms, that allow the estimation of models for very large data sets. The standard eigenvalue simplification is used (Ord 1975) for data sets up to 1,000 observations. Beyond that, the sparse algorithm of Smirnov and Anselin (2001) is used, which exploits the characteristic polynomial associated with the spatial weights matrix. This algorithm allows estimation of very large data sets in reasonable time. In addition, *GeoDa* implements the recent algorithm of Smirnov (2003) to compute the asymptotic variance matrix for all the model coefficients (i.e., including both the spatial and non-spatial coefficients). This involves the inversion of a matrix of the dimensions of the data sets. To date, *GeoDa* is the only software that provides such estimates for large data sets.

All estimation methods employ sparse spatial weights, but they are currently constrained to weights that are intrinsically symmetric (e.g., excluding k-nearest neighbor weights). The regression routines have been successfully applied to real data sets of more than 300,000 observations (with estimation and inference completed in a few minutes). By comparison, a spatial regression for the 3000+ US counties takes a few seconds.

We illustrate the spatial regression capabilities with a partial replication and extension of the homicide model used in Baller et al. (2001) and Messner and



```
REGRESSION
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            :
Spatial Weight      : natrook.GAL
Dependent Variable  :        HR80  Number of Observations: 3085
Mean dependent var  :    6.927616  Number of Variables   :    7
S.D. dependent var  :    6.825088  Degree of Freedom     : 3078
Lag coeff. (Lambda) :    0.293641

R-squared           :    0.462988  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        :-9369.500716
Sigma-square        :   25.015000  Akaike info criterion :       18753
S.E of regression   :      5.0015  Schwarz criterion     :18795.241581

-----------------------------------------------------------------------
    Variable    Coefficient     Std.Error    z-value      Probability
-----------------------------------------------------------------------
    CONSTANT     8.463455      0.9765372       8.666803    0.0000000
       SOUTH     1.951838      0.2916178       6.693136    0.0000000
        RD80     3.461736      0.1350625       25.63063    0.0000000
        PS80    0.6745796      0.1185432       5.690582    0.0000000
        UE80  -0.04367847      0.03578631     -1.220535    0.2222621
        DV80     1.151325     0.07579833       15.18932    0.0000000
        MA80   -0.2395876     0.02761771      -8.675144    0.0000000
      LAMBDA    0.2936407     0.02562901       11.45736    0.0000000
-----------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                        DF      VALUE          PROB
Breusch-Pagan test                          6        1187.417      0.0000000

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : natrook.GAL
TEST                                        DF      VALUE          PROB
Likelihood Ratio Test                       1         120.599      0.0000000
======================== END OF REPORT ==============================
```

Figure 5: Maximum Likelihood estimation of the spatial error model.

Anselin (2004). These studies assessed the extent to which a classic regression specification, well-known in the ciminology literature, is robust to the explicit consideration of spatial effects. The model relates county homicide rates to a number of socio-economic explanatory variables. In the original study, a full ML analysis of all US continental counties was precluded by the constraints on the eigenvalue-based SpaceStat routines. Instead, attention focused on two subsets of the data containing 1412 counties in the US South and 1673 counties in the non-South.

In Figure 5, we show the result of the ML estimation of a spatial error model of county homicide rates for the complete set of 3085 continental US counties in 1980. The explanatory variables are the same as before: a Southern dummy variable, a resource deprivation index, a population structure indicator, unemployment rate, divorce rate and median age.[16]

The results confirm a strong positive and significant spatial autoregressive coefficient ($\hat{\lambda} = 0.29$). Relative to the OLS results (e.g., Messner and Anselin 2004, Table 7.1., p. 137), the coefficient for unemployment has become insignificant, illustrating the misleading effect spatial error autocorrelation may have on inference using OLS estimates. The model diagnostics also suggest a continued presence of problems with heteroskedasticity. However, *GeoDa* currently does not include functionality to deal with this.

7 Future Directions

GeoDa is a work in progress and still under active development. This development proceeds along three fronts. First and foremost is an effort to make the code cross-platform and open source. This requires considerable change in the graphical interface, moving from the Microsoft Foundation Classes (MFC) that are standard in the various MS Windows flavors, to a cross-platform alternative. The current efforts use wxWindows, which operates on the same code base with a native GUI flavor in Windows, MacOS X and Linux/Unix. Making the code open source is currently precluded by the reliance on proprietary code in ESRI's MapObjects. Moreover, this involves more than simply making the source code available, but entails considerable reorganization and streamlining of code (refactoring), to make it possible for the community to effectively participate in the development process.

A second strand of development concerns the spatial regression functionality. While currently still fairly rudimentary, the inclusion of estimators other than ML and the extension to models for spatial panel data are in progress. Finally, the functionality for ESDA itself is being extended to data models other than the discrete locations in the "lattice" case. Specifically, exploratory variography is being added, as well as the exploration of patterns in flow data.

Given its initial rate of adoption, there is a strong indication that *GeoDa* is indeed providing the "introduction to spatial data analysis" that makes it possible for growing numbers of social scientists to be exposed to an explicit spatial perspective. Future development of the software should enhance this capability and it is hoped that the move to an open source environment will involve an international community of like minded developers in this venture.

References

Anselin, L. (1992). *SpaceStat, a Software Program for Analysis of Spatial Data.* National Center for Geographic Information and Analysis (NCGIA), Univer-

[16] See the original papers for technical details and data sources. In Baller et al. (2001), a different set of spatial weights was used than in this example, but the conclusions of the specification tests are the same. Specifically, using the county contiguity, the robust Lagrange Multiplier tests are 1.24 for the Lag alternative, and 24.88 for the Error alternative, strongly suggesting the latter as the proper alternative.

sity of California, Santa Barbara, CA.

Anselin, L. (1995). Local indicators of spatial association — LISA. *Geographical Analysis*, 27:93–115.

Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Fischer, M., Scholten, H., and Unwin, D., editors, *Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences*, pages 111–125. Taylor and Francis, London.

Anselin, L. (1999). Interactive techniques and exploratory spatial data analysis. In Longley, P. A., Goodchild, M. F., Maguire, D. J., and Rhind, D. W., editors, *Geographical Information Systems: Principles, Techniques, Management and Applications*, pages 251–264. John Wiley, New York, NY.

Anselin, L. (2000). Computing environments for spatial data analysis. *Journal of Geographical Systems*, 2(3):201–220.

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. (2004). *GeoDa 0.95i Release Notes*. Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

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

Anselin, L. and Getis, A. (1992). Spatial statistical analysis and geographic information systems. *The Annals of Regional Science*, 26:19–33.

Anselin, L., Kim, Y.-W., and Syabri, I. (2004). Web-based spatial analysis tools for the exploration of spatial data. *Journal of Geographical Systems*, 6. forthcoming.

Anselin, L., Syabri, I., and Smirnov, O. (2002a). Visualizing multivariate spatial correlation with dynamically linked windows. In Anselin, L. and Rey, S., editors, *New Tools for Spatial Data Analysis: Proceedings of the Specialist Meeting*. Center for Spatially Integrated Social Science (CSISS), University of California, Santa Barbara. CD-ROM.

Anselin, L., Syabri, I., Smirnov, O., and Ren, Y. (2002b). Visualizing spatial autocorrelation with dynamically linked windows. *Computing Science and Statistics*, 33. CD-ROM.

Assunção, R. and Reis, E. A. (1999). A new proposal to adjust Moran's I for population density. *Statistics in Medicine*, 18:2147–2161.

Bailey, T. C. and Gatrell, A. C. (1995). *Interactive Spatial Data Analysis*. John Wiley and Sons, New York, NY.

Baller, R., Anselin, L., Messner, S., Deane, G., and Hawkins, D. (2001). Structural covariates of U.S. county homicide rates: Incorporating spatial effects. *Criminology*, 39(3):561–590.

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.

Bivand, R. (2002a). Implementing spatial data analysis software tools in R. 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.

Bivand, R. (2002b). Spatial econometrics functions in R: Classes and methods. *Journal of Geographical Systems*, 4(4):405–421.

Bivand, R. and Gebhardt, A. (2000). Implementing functions for spatial statistical analysis using the R language. *Journal of Geographical Systems*, 2(3):307–317.

Bivand, R. S. and Portnov, B. A. (2004). Exploring spatial data analysis techniques using R: The case of observations with no neighbors. In Anselin, L., Florax, R. J., and Rey, S. J., editors, *Advances in Spatial Econometrics: Methodology, Tools and Applications*, pages 121–142. Springer-Verlag, Berlin.

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.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–681.

Cleveland, W. S. and McGill, M. (1988). *Dynamic Graphics for Statistics*. Wadsworth, Pacific Grove, CA.

Cook, D., Majure, J., Symanzik, J., and Cressie, N. (1996). Dynamic graphics in a GIS: A platform for analyzing and exploring multivariate spatial data. *Computational Statistics*, 11:467–480.

Cook, D., Symanzik, J., Majure, J. J., and Cressie, N. (1997). Dynamic graphics in a GIS: More examples using linked software. *Computers and Geosciences*, 23:371–385.

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

Dykes, J. A. (1997). Exploring spatial data representation with dynamic graphics. *Computers and Geosciences*, 23:345–370.

ESRI (2004). *An Overview of the Spatial Statistics Toolbox. ArcGIS 9.0 Online Help System (ArcGIS 9.0 Desktop, Release 9.0, June 2004)*. Environmental Systems Research Institute, Redlands, CA.

Fischer, M. and Nijkamp, P. (1993). *Geographic Information Systems, Spatial Modelling and Policy Evaluation*. Springer-Verlag, Berlin.

Fischer, M. M. and Getis, A. (1997). *Recent Development in Spatial Analysis*. Springer-Verlag, Berlin.

Fischer, M. M., Scholten, H. J., and Unwin, D. (1996). *Spatial Analytical Perspectives on GIS*. Taylor and Francis, London.

Fotheringham, A. S. and Rogerson, P. (1993). GIS and spatial analytical problems. *International Journal of Geographical Information Systems*, 7:3–19.

Fotheringham, A. S. and Rogerson, P. (1994). *Spatial Analysis and GIS*. Taylor and Francis, London.

Goodchild, M. F., Anselin, L., Appelbaum, R., and Harthorn, B. (2000). Toward spatially integrated social science. *International Regional Science Review*, 23(2):139–159.

Goodchild, M. F., Haining, R. P., Wise, S., and others (1992). Integrating GIS and spatial analysis — problems and possibilities. *International Journal of Geographical Information Systems*, 6:407–423.

Haining, R. (1989). Geography and spatial statistics: Current positions, future developments. In Macmillan, B., editor, *Remodelling Geography*, pages 191–203. Basil Blackwell, Oxford.

Haslett, J., Wills, G., and Unwin, A. (1990). SPIDER — an interactive statistical tool for the analysis of spatially distributed data. *International Journal of Geographic Information Systems*, 4:285–296.

Kafadar, K. (1996). Smoothing geographical data, particularly rates of disease. *Statistics in Medicine*, 15:2539–2560.

Le Gallo, J. and Dall'erba, S. (2003). Evaluating the temporal and spatial heterogeneity of the European convergence process, 1980–1999. Technical report, Université Montesquieu-Bordeaux IV, Pessac Cedex, France.

Levine, N. (2004). The CrimeStat program: Characteristics, use and audience. *Geographical Analysis*. Forthcoming.

Marshall, R. J. (1991). Mapping disease and mortality rates using Empirical Bayes estimators. *Applied Statistics*, 40:283–294.

Messner, S. F. and Anselin, L. (2004). Spatial analyses of homicide with areal data. In Goodchild, M. and Janelle, D., editors, *Spatially Integrated Social Science*, pages 127–144. Oxford University Press, New York, NY.

Monmonier, M. (1989). Geographic brushing: Enhancing exploratory analysis of the scatterplot matrix. *Geographical Analysis*, 21:81–4.

Ord, J. K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70:120–126.

Rey, S. J. (2004). Spatial analysis of regional income inequality. In Goodchild, M. F. and Janelle, D., editors, *Spatially Integrated Social Science*, pages 280–299. Oxford University Press, Oxford.

Rey, S. J. and Janikas, M. V. (2004). STARS: Space-time analysis of regional systems. *Geographical Analysis*. forthcoming.

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.

Stuetzle, W. (1987). Plot windows. *Journal of the American Statistical Association*, 82:466–475.

Symanzik, J., Cook, D., Lewin-Koh, N., Majure, J. J., and Megretskaia, I. (2000). Linking ArcView and XGobi: Insight behind the front end. *Journal of Computational and Graphical Statistics*, 9(3):470–490.

Takatsuka, M. and Gahegan, M. (2002). GeoVISTA Studio: A codeless visual programming environment for geoscientific data analysis and visualization. *Computers and Geosciences*, 28:1131–1141.

Unwin, A. (1994). REGARDing geographic data. In Dirschedl, P. and Osterman, R., editors, *Computational Statistics*, pages 345–354. Physica Verlag, Heidelberg.

Wise, S., Haining, R., and Ma, J. (2001). Providing spatial statistical data analysis functionality for the GIS user: the SAGE project. *International Journal of Geographic Information Science*, 15(3):239–254.