The GeoDa Book

Exploring Spatial Data

Luc Anselin

ISBN: 0-9863421-2-2
ISBN-13: 978-0-9863421-2-7
GeoDa Press LLC, Chicago, IL

Chapter 1

Introduction

It has been more than ten years since I published a Workbook that contained a series of exercises to illustrate the functionality of *Legacy* `GeoDa`, i.e., version 0.9.5-i of the software (Anselin 2005a). Since then, the software has evolved considerably, from a Windows-only program that relied heavily on ESRI's MapObjects library, to a collection of cross-platform and open source code, built on top of modern C++ libraries such as `wxWidgets`, `Boost` and `GDAL`.

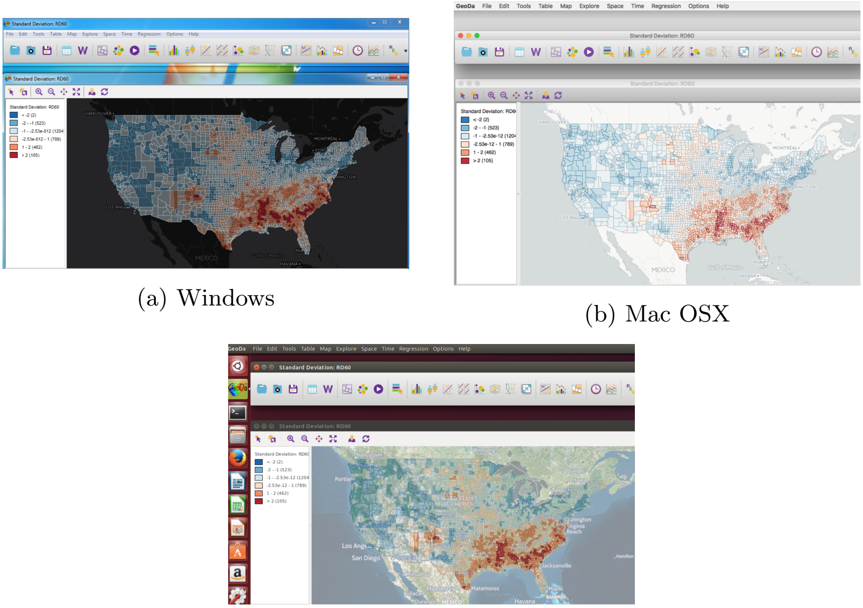
This book constitutes the definitive guide to the `GeoDa` software as well as an introduction to spatial data science. In that sense it differs fundamentally from the 2005 Workbook, which was almost exclusively focused on the software and assumed that the explanation for the methods was obtained elsewhere. Here, both aspects are combined.

`GeoDa` was designed to be a software tool that facilitates the exploration and analysis of geospatial data as a progression from simple description and visualization to structured exploration and formal modeling. This book follows the same logic. It moves through the various phases of an exploration of spatial data and explains the relevant methods in conjunction with their implementation in the software.

In this introductory chapter, I provide some background on the history and motivation for the development of `GeoDa`. After instructions on how to install the program, I give a quick tour of the functionality. The chapter closes with a presentation of the organization of the book.

1.1 Background and Motivation

`GeoDa` was first released in 2002 by the NSF-funded Center for Spatially Integrated Social Science. It built upon a long history of software development for spatial analysis, going back to the `SpaceStat` package (Anselin 1991), which was the first freestanding package to implement spatial autocorrelation and spatial econometric functionality. An extensive review of the history of spatial



(a) Windows

(b) Mac OSX

(c) Ubuntu

Figure 1.1: Cross-platform GeoDa

analytical software is beyond the current scope, but can be found in Anselin (2005b) and Anselin (2012).

A slightly revised version of GeoDa appeared in 2003 as version 0.9.5-i (see Anselin 2003, 2004). This version remained unchanged until it was replaced by the first beta releases of the open source GeoDa in 2009 (then referred to as OpenGeoda). Version 0.9.5-i, which we now refer to as *Legacy* GeoDa was written in C++, but only ran on the Windows XP operating system and relied on the ESRI MapObjects library for much of its mapping functionality. It was one of the first desktop programs to fully implement dynamic linking and brushing for all its open windows (maps, tables and statistical graphs), and it quickly gained a large user base.

As outlined in detail in Anselin et al. (2006), Legacy GeoDa emphasized geovisualization (e.g., outlier maps, rate smoothing, cartogram, map animation), exploratory data analysis (e.g., statistical graphs, parallel coordinate plots, conditional plots), and especially exploratory spatial data analysis (ESDA). A central role was reserved for the computation and visualization of global and local spatial autocorrelation statistics. It also contained a large number of utility functions to construct spatial data sets (e.g., shape files from point coordinates, Thiessen polygons, centroids), and to create and manipulate spatial weights (e.g., contiguity weights, distance band weights, k-nearest neighbor weights). While not its primary focus, a limited set of spatial regression functionality was also included in GeoDa from the beginning.

In 2005, a decision was made to fundamentally change the architecture of the software and to make the project open source as well as cross-platform.

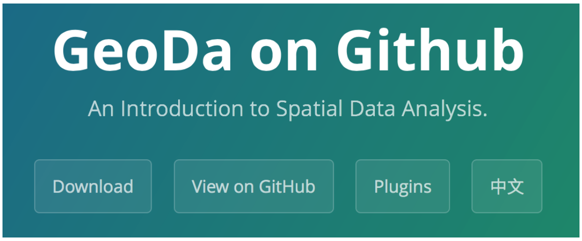


Figure 1.2: GeoDa on Github, http://geodacenter.github.io.

This required a total redesign of the code, still using C++, but now employing standard open source libraries, such as the C++ Standard Template Library (STL), wxWidgets for the cross-platform graphical user interface, the Boost C++ library for computational geometry, and the OGR Simple Features Library to access a wide range of data sources (included as of Version 1.6). The resulting product was first released in beta version in 2009 (Anselin and McCann 2009). The official Version 1.0 followed in October 2011 under the open source GPL 3.0 license. The program works identically in Windows, Mac OS X and Linux, but takes on the native look and feel of each operating system, as shown in Figure 1.1.

The most recent open source GeoDa includes all the functionality of Legacy GeoDa, but it also contains several new features, such as support for a wide range of spatial data formats (leveraging the OGR library), data base connections, realistic background layers for maps, as well as advanced functionality, such as lowess smoothers, a scatter plot matrix, non-parametric spatial correlograms, space-time exploration and limited treatment effects analysis. In addition, the user interface was redesigned and the underlying data structures have been made much more robust. We return to a more detailed overview of the functionality in Section 1.3. The version covered in this book is 1.8 or later, released in Spring 2016.

1.2 Installing GeoDa

Installing the GeoDa software is relatively straightforward. The main access point is the GeoDa Github site, http://geodacenter.github.io, shown in Figure 1.2. Binaries are provided for both Windows and Mac OSX operating systems, as well as for the Ubuntu Linux distribution. The complete source code is available as well, with instructions for compilation under different platforms.

1.2.1 Binaries

Access to the latest binaries is obtained by clicking the Download button in the Github web interface (the left-most button in Figure 1.2). This provides binaries for the Windows, Mac OSX and Ubuntu Linux operating systems.

Figure 1.3: GeoDa logo.

Windows binaries are available for both 32 bit and 64 bit versions of the operating system, and cover XP, Vista, as well as Windows 7, 8, 8.1 and 10. After downloading the GeoDa Setup Wizard, some straightforward steps (illustrated on the web site) lead to the executable.

Installation under Mac OSX is even simpler. Clicking on the link starts downloading the program. Once downloaded, the executable can be moved to the Applications folder and the icon added to the Dock.

The Ubuntu binaries are downloaded as a file with a filename such as GeoDa-version-Ubuntu-64bit.deb, where version corresponds to the GeoDa version, e.g., 1.8.6.[1] Note that each version of Ubuntu has its own separate binary. After launching the *.deb file, select Install in the Ubuntu Software Center dialog. This may result in a warning that the "package is of bad quality," which can easily be overridden by selecting Ignore and Install.

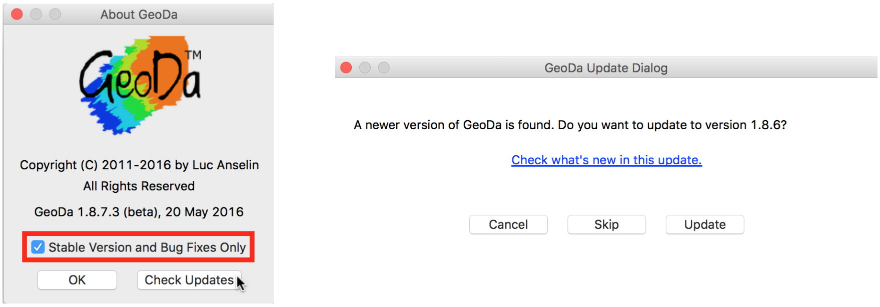
With the executable in the proper directory, the program is launched by clicking on the executable or on the GeoDa icon (shown in Figure 1.3). Mac OSX users may need to right click on the GeoDa icon first and confirm that they want to use the downloaded software.[2] This will only need to be done once, afterwards the usual double click suffices to launch the program.

1.2.2 Compiling from Source

The GeoDa source code is at https://github.com/GeoDaCenter/geoda. Specific instructions for compilation are contained in the BuildTools directory at https://github.com/GeoDaCenter/geoda/tree/master/BuildTools, for Windows, Mac OSX, and the Linux Ubuntu and Centos distributions. The associated readme files provide the steps needed to compile the source and associated libraries. In addition to instructions for the base installation, this also outlines how to compile the OGR plugins as well as the commercial plugins for ESRI's Arc SDE and Oracle spatial. Warning: this is not for the faint of heart and assumes a more than basic familiarity with the operating system, the compiler options, make files, etc.

[1] The file name contains Ubuntu-32bit for the 32 bit version of the operating system.

[2] In Mac OSX 10.8 (Mountain Lion) and later, the "Security and Privacy" General options must be set to "Allow applications downloaded from: Mac App Store and identified developers." The default setting is "Mac App Store" only, which will not allow GeoDa to run. It is straightforward to change the options in the Mac OSX Preference panel. The first time the program is launched, Mac OSX 10.8 will likely give a warning about the developer not being authorized. After making sure the System Preferences are set correctly, the program can be opened by using CTRL-Click and selecting "Open."



Copyright (C) 2011-2016 by Luc Anselin
All Rights Reserved
GeoDa 1.8.7.3 (beta), 20 May 2016

☑ Stable Version and Bug Fixes Only

A newer version of GeoDa is found. Do you want to update to version 1.8.6?

Check what's new in this update.

(a) Check on updates

(b) New GeoDa version available

Figure 1.4: GeoDa automatic update check feature

1.2.2.1 Dependencies

GeoDa is released under a GPL 3.0 open source software license.[3] The software is written in C++ and relies on several open source libraries and some specific open source code.[4] This includes the following widely used libraries:

- the wxWidgets cross-platform GUI library (http://www.wxwidgets.org)
- the general Boost C++ libraries (http://www.boost.org) as well as the Boost.Polygon Voronoi library
- the GDAL libraries (http://www.gdal.org)
- the CLAPACK linear algebra libraries (http://www.netlib.org/clapack)
- the ANN approximate nearest neighbor library (http://www.cs.umd.edu/~mount/ANN/)

In addition, source code is borrowed from the following sources:

- FastArea.c++ for area computation (https://github.com/erich666/jgt-code/blob/master/Volume_07/Number_2/Sunday2002/FastArea.c++)
- logger.h for logging during debugging (http://accu.org/index.php/journals/1304)

1.2.3 Automatic Updates

Since version 1.8, GeoDa automatically checks for available updates upon launching the program, as long as an active internet connection is present. This feature works for Windows and Mac OSX installations, but not for Linux. The frequency of the updates is set in the About dialog. In Figure 1.4a, the default setting is illustrated by the highlighted area. The checked box implies that only bug fixes and new stable versions will be uploaded. With the box unchecked, all changes will be uploaded, even for intermediate development versions. For most users, the default setting should be appropriate.

[3] See http://www.gnu.org/licenses/gpl-3.0.en.html for specifics.

[4] For an up to date set of the dependencies, see the list given at http://geodacenter.github.io.

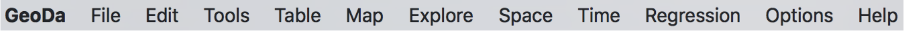


Figure 1.5: `GeoDa` menu.

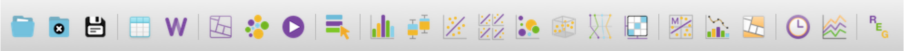


Figure 1.6: `GeoDa` floating toolbar.

Updates can also be checked at any point (not just upon the launch of the program), by opening the `About` dialog and clicking on the `Check Updates` button. If the version is current, a dialog will appear with that message.

If a more recent version of `GeoDa` is available, the `GeoDa Update Dialog` appears, as shown in Figure 1.4b. One can read the list of changes included in the update, skip the new version, or select the `Update` button to download the latest version. After downloading, a restart is required.[5]

1.3 A Quick Tour of `GeoDa`

After launching the program (by double clicking on the `GeoDa` icon), the `GeoDa Menu` and floating `Toolbar` appear on the screen, as shown in Figures 1.5 and 1.6. The toolbar was redesigned in Version 1.8 and may look a little unfamiliar to users of older versions of `GeoDa`. However, the logic behind it remains the same. Each toolbar icon (and its associated options) corresponds to an item in the menu (and submenus). Specific functionality is invoked by clicking on the appropriate toolbar icon and selecting the desired option. Alternatively, the menu structure can be followed until the selected option is reached.

Table 1.1, expanded from Table 1 in Anselin et al. (2006), contains an exhaustive overview of the functionality in `GeoDa`. It is organized along major categories of operations: data input, data export, spatial data creation, data query, variable transformations, choropleth mapping, statistical maps, smoothed rate maps, exploratory data analysis (EDA), spatial autocorrelation, space-time exploration, exploring heterogeneity, and spatial regression.

In the remainder of this Section, I provide an alternative classification of the functionality, focused on specific subsets of icons on the toolbar. The review is intended for those users that are new to `GeoDa` and is intended to give a general sense of what the program can do. Others can skip directly to the relevant chapters. I cover, in turn, data entry, data manipulation, the weights manager, mapping and geovisualization, exploratory data analysis, spatial autocorrelation analysis, space-time analysis, and spatial regression. In the toolbar (Figure 1.6), these functional groupings are separated by a subtle vertical line.

[5] Note that after the new version is installed, the setting for update checks is always reset to the default (as in Figure 1.4a). So, if the preferred option is to download every change, then the box on the dialog needs to be explicitly unchecked (since the update is a new installation, the previous settings cannot be remembered). The setting will be in place for as long as that version of `GeoDa` is being used.

Table 1.1: GeoDa functionality overview

Category	Functions
Data Input	reading spatial data file formats (supported by OGR) connecting to spatial data bases connecting to web feature services (WFS) loading data from CartoDB loading data using Project files joining tables
Data Export	export data to a different format export selected observations
Spatial Data Creation	point layers from x-y coordinates point layers from polygon centroids point layers from polygon mean centers Thiessen polygons from point layers rectangular grid layers
Data Query	map query linking and brushing logical query design creating indicator variables
Variable Transformations	adding/deleting variables transformations (log, exp, etc.) variable algebra random numbers permutation of observations creating spatial lag variables rate calculation and rate smoothing editing variable properties editing table cells
Choropleth Mapping	quantile map equal interval map natural breaks map unique value map saving map categories saving the map as an image zooming and panning adding a base layer customizing the legend
Statistical Maps	standard deviational map box map percentile map cartogram animation

Table 1.1. Continued

Category	Functions
Smoothed Rate Maps	 crude rate map excess risk map Empirical Bayes smoothing spatial rate smoothing spatial Empirical Bayes smoothing saving calculated rates
Exploratory Data Analysis	 histogram box plot scatter plot scatter plot matrix lowess smoothing bubble chart 3D scatter plot parallel coordinate plot conditional plots and maps
Spatial Autocorrelation	 spatial weights manager contiguity weights higher order contiguity distance band weights k-nearest neighbor weights weights properties adding neighbors to a selection Moran scatter plot Moran scatter plot for rates spatial correlogram local Moran maps local Moran rate maps local G statistic saving local statistics
Space-Time Exploration	 time editor creating space-time variables saving a space-time table saving space-time weights time player, animating plots and maps differential global spatial autocorrelation differential local spatial autocorrelation saving time lags
Exploring Heterogeneity	 Chow test averages tool, assessing structural change difference in differences (DID) analysis saving dummy variables

Table 1.1. Continued

Category	Functions
Spatial Regression	OLS with spatial diagnostics ML spatial lag model ML spatial error model spatial analysis of variance pooled space-time regression DID with spatial effects predicted value maps residual maps

1.3.1 Data Entry

The first functional module deals with loading data, saving files, and opening and closing projects, which corresponds roughly to the contents of the `File` menu. This is invoked through the three left-most icons on the toolbar, as highlighted in Figure 1.7. Clicking on the left-most (`Open Project`) icon brings up a data loading dialog through which a range of spatial file formats can be accessed, including ESRI shape files, MapInfo files, JSON files, GML files, etc., as well as tabular data (containing `x-y` coordinates) in formats including csv, dBase and Excel. Also, connections can be established to spatial databases, and data can be obtained from web feature services (WFS) and CartoDB data bases. This is explored further in Chapter 2. The other two icons `Close` a project, i.e., remove the current data from memory, and `Save` the current project. The `Save As` functionality (from the `File` menu) also serves as a handy file format converter, since one can load a file in one spatial format and save it (export it) in a different format. The `File` menu also lets one save a new data set with a subset of (selected) observations and create and inspect a `Project File` (for details, see Chapter 2).

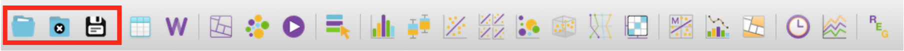


Figure 1.7: Data entry

1.3.2 Data Manipulation

The second set of functions, invoked by the `Table` icon in the toolbar, contains an extensive set of methods to create, edit and transform variables. This constitutes an important aspect of what is referred to as *data munging* in the recent data science literature. The icon is highlighted in Figure 1.8. In addition to tools to manipulate variables, the `Table` functionality also includes methods

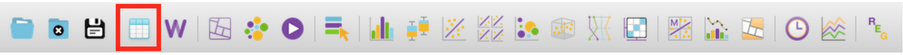


Figure 1.8: Table functionality

to join a different data set, query the data, select observations in a given value range for a variable, and create associated indicator variables. Details are given in Chapters 4 and 5.

1.3.3 Weights Manager

Spatial weights are an essential tool to carry out spatial autocorrelation and spatial regression analysis. The weights manager in GeoDa combines under a single icon (the W in the toolbar) the functionality that previously (in versions of GeoDa before 1.8) was represented by three toolbar icons that stood for Create Weights, Open Weights File, and Connectivity Histogram (characteristics of the weights). The icon, highlighted in Figure 1.9, brings up a dialog that controls the creation, saving, loading and inspection of the properties of spatial weights. Both contiguity-based (including higher order contiguity) and distance-based weights (distance band and k-nearest neighbor weights) can be created for point and polygon layers. The characteristics include a contiguity histogram (a histogram representing the distribution of the neighbor cardinality) and an interactive connectivity map, which highlights the neighbors for each location as the cursor moves over the map. Once saved, the information pertaining to the weights is stored in the Project File. As long as this Project File is saved, then the spatial weights and their characteristics are automatically loaded the next time the project is opened. This functionality can also be accessed through the Tools item in the Menu. The topic of spatial weights is explored in more depth in Chapters 18 through 20.

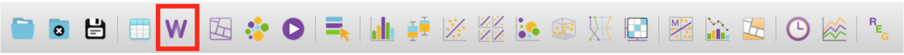


Figure 1.9: Weights functionality

1.3.4 Mapping and Geovisualization

The next set of four icons represent the mapping and geovisualization functionality in GeoDa, comprised of choropleth mapping, cartograms, map animation and the Category Editor, a dialog to interactively construct custom break points for maps and graphs. The icons are highlighted in Figure 1.10. They correspond to the items contained in the Map menu (except for the Category Editor, which is accessed as a right click Option in the map window). GeoDa is not intended to be an advanced cartographic package, and the maps are therefore designed primarily for exploration, with interactivity in mind, rather than refined graphical representation. The first icon invokes a full range of choropleth map types. Basic maps include classifications based on quantiles,

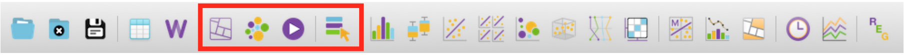


Figure 1.10: Mapping and geovisualization functionality

equal intervals, and natural breaks, as well as unique value maps. This is illustrated more extensively in Chapter 7. In addition, a number of so-called statistical maps are provided, such as standard deviational maps, box maps and percentile maps, reviewed in Chapter 9. A more specialized functionality pertains to various methods to deal with maps for proportions or rates, including a number of procedures that smooth the rates to address their inherent variance instability (see Chapter 12 for technical details). The various map types correspond to items in the drop-down menu that appears after clicking on the mapping icon.

In addition to portraying data associated with a spatial point or polygon layer that is loaded from a GIS file, `GeoDa` can now also add a realistic basemap behind the layer, based on map tiles provided by CartoDB or Nokia (this requires an internet connection, for details, see Chapter 8). Unlike the other options, the base map layer is invoked by selecting an icon in the map window (similar to zooming and panning), and cannot be accessed from the `Menu`.

A circular cartogram is created by means of the second icon in the group, or, alternatively, by selecting the matching item in the `Map` menu. Details on the principles behind the cartogram and its particular implementation in `GeoDa` are given in Chapter 10. The third icon starts the animation dialog. This allows one to move through the data (not just for a map, but also for any other open window containing graphs) from low to high values (or high to low) highlighting the corresponding data points either individually or cumulatively. It is an intuitive and visual way to assess clustering of the data. Further details are provided in Chapter 11.

Finally, the `Category Editor` icon (the right-most in the group) brings up a dialog to create break points for a legend other than the traditional choropleth map break points (such as quantiles, equal intervals, etc.). This is especially helpful when maps or graphs are compared over time and the same intervals need to be used to facilitate comparison. Similarly, in many context, institutional factors dictate specific break points, such as those associated with income categories in social experiments. The (named) custom breakpoints can be used in any graph or map and their definition is saved in the `Project File`. A detailed account of the operation of the `Category Editor` is given in Chapter 13.

1.3.5 Exploratory Data Analysis (EDA)

The largest group of icons corresponds to the (non-spatial) exploratory data analysis functionality, as highlighted in Figure 1.11. The group is comprised of eight icons, representing the histogram, box plot, scatter plot, scatter plot matrix, bubble plot, three-dimensional scatter plot, parallel coordinate plot and conditional plots. This functionality can also be accessed through the

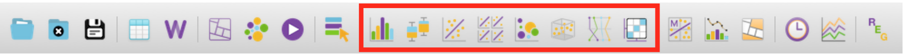


Figure 1.11: Exploratory Data Analysis – EDA

Explore menu. Details are provided in Chapters 14 through 17.

1.3.6 Spatial Autocorrelation Analysis

The group of three icons to the right of EDA cover spatial autocorrelation analysis, respectively, global spatial autocorrelation, non-parametric spatial autocorrelation, and local spatial autocorrelation. The icons are highlighted in Figure 1.12. The same functionality can also be accessed through the Space menu. Global spatial autocorrelation is visualized by means of the Moran scatter plot, with special standardization available for analyzing rates. Technical details are given in Chapter 21. The second icon invokes the calculation of a spatial correlogram, or non-parametric measure of spatial autocorrelation, new with Version 1.8 of GeoDa, and further discussed in Chapter 22. The rightmost icon represents local spatial autocorrelation, and includes both the local Moran implementation (with special allowance for rate instability) and the local G statistics. These are covered in Chapters 23 and 24. In addition to these familiar cross-sectional methods, GeoDa also implements so-called *differential* spatial autocorrelation analysis, both global and local. In differential analysis, the measure of spatial autocorrelation pertains to the temporal first differences of a variable, which are computed on the fly (see Chapter 27). This is part of the space-time analytical functionality, the core of which is considered in the next Section.

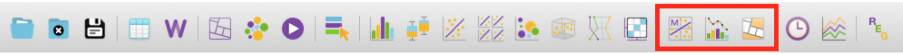


Figure 1.12: Spatial autocorrelation analysis

1.3.7 Space-Time Analysis

The next to last group of icons on the GeoDa toolbar consists of two icons that provide access to space-time analysis. The first invokes the Time Manager, the second the Averages Chart (new with Version 1.8). The two icons are highlighted in Figure 1.13. The Time Manager brings up the Time Editor and the Time Player, both of which can also be accessed via the Time menu. These functions allow for the creation of space-time (*grouped*) variables from a cross-sectional data set with observation on a variable at different points in time (each time period corresponding to a different column in the data table). The Time Player then uses the space-time variables to move through all the open maps and charts over time, which allows for comparative statics analysis. This is elaborated upon in Chapters 25 and 26.

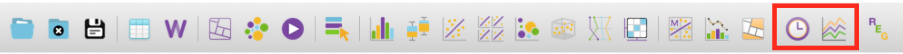


Figure 1.13: Space-time analysis

The Averages Chart is a visual tool to assess structural breaks, particularly in the context of treatment effects analysis. The effect of breaks on the mean of a variable (hence the name `Averages Chart`) can be assessed both cross-sectionally (spatial heterogeneity) as well as over time. It includes both visual inspection as well as rudimentary tests for structural stability and difference in differences regression. Technical details are provided in Chapters 29 and 30. The `Averages Chart` can also be invoked by means of the `Explore` menu.

1.3.8 Spatial Regression

The right-most icon on the toolbar invokes the interface to carry out spatial `Regression` analysis. It is shown highlighted in Figure 1.14. The counterpart of the toolbar icon is the `Regression` item in the menu. Clicking on the icon or the menu item brings up a dialog through which the dependent and explanatory variables of the regression model are specified, the spatial weights selected (for diagnostics and/or spatial models) and the desired estimation chosen (classic model, spatial lag or spatial error). Since Version 1.8, `GeoDa` also includes functionality to estimate spatial regression models for pooled space-time data. A special case is the difference in differences regression with spatial effects. Technical details are provided in Chapters 31 through 36.

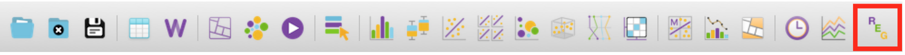


Figure 1.14: Spatial regression analysis

1.4 Organization of the Book

The materials in the book are grouped into seven parts, following a progression that starts with basic description and moves on to data exploration, assessing spatial autocorrelation and detecting spatial heterogeneity. The end point of this evolution is a spatial regression analysis. While most of the discussion pertains to cross-sectional data, one part is specifically devoted to exploration in both space and time.

The first part, *Getting Started*, consists of five chapters covering essential data handling. This outlines basic data input and output, creating spatial data (e.g., from point coordinates), and converting between points and polygons. In addition, much of the `Table` functionality is covered, such as manipulating variables (creating new variables, transformations), and querying in a table and on the map. The last chapter in this part outlines the basics of dynamic interacting with the data through linking and brushing.

An appendix provides details on the data sets used in the examples.