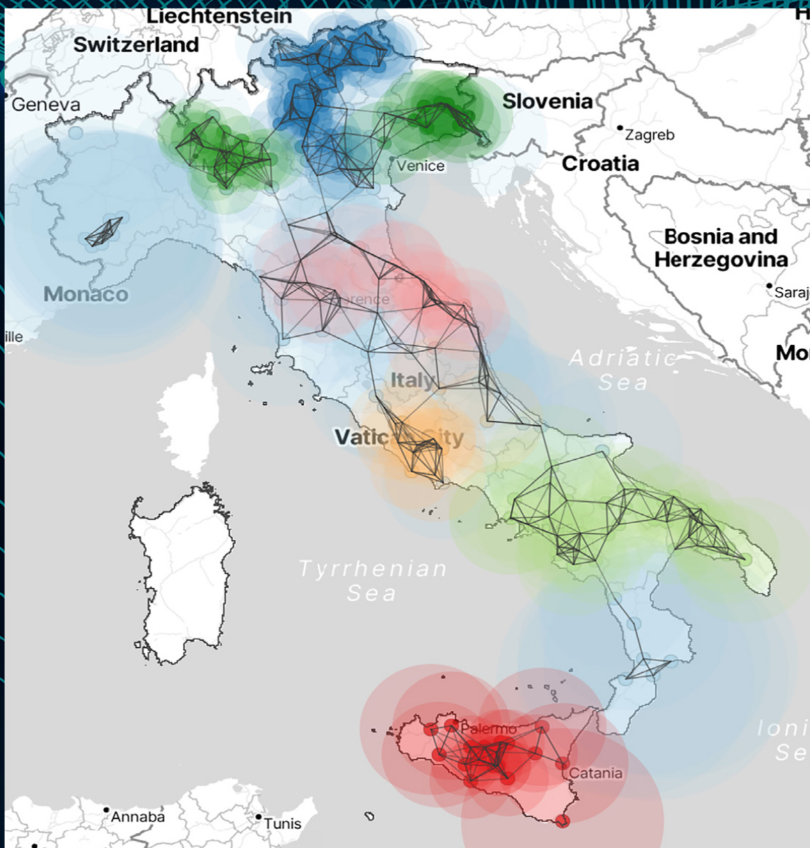
An Introduction to Spatial Data Science with GeoDa

Volume 1

Exploring Spatial Data



Luc Anselin

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

An Introduction to Spatial Data Science with GeoDa

Volume 1 – Exploring Spatial Data

This book is the first in a two-volume series that introduces the field of spatial data science. It offers an accessible overview of the methodology of exploratory spatial data analysis. It also constitutes the definitive user's guide for the widely adopted GeoDa open-source software for spatial analysis. Leveraging a large number of real-world empirical illustrations, readers will gain an understanding of the main concepts and techniques, using dynamic graphics for thematic mapping, statistical graphing, and, most centrally, the analysis of spatial autocorrelation. Key to this analysis is the concept of local indicators of spatial association, pioneered by the author and recently extended to the analysis of multivariate data.

The focus of the book is on intuitive methods to discover interesting patterns in spatial data. It offers a progression from basic data manipulation through description and exploration to the identification of clusters and outliers by means of local spatial autocorrelation analysis. A distinctive approach is to spatialize intrinsically non-spatial methods by means of linking and brushing with a range of map representations, including several that are unique to the GeoDa software. The book also represents the most in-depth treatment of local spatial autocorrelation and its visualization and interpretation by means of GeoDa.

The book is intended for readers interested in going beyond simple mapping of geographical data to gain insight into interesting patterns. Some basic familiarity with statistical concepts is assumed, but no previous knowledge of GIS or mapping is required.

Key Features:

- Includes spatial perspectives on cluster analysis
- Focuses on exploring spatial data
- Supplemented by extensive support with sample data sets and examples on the GeoDaCenter website

This book is both useful as a reference for the software and as a text for students and researchers of spatial data science.

Luc Anselin is the Founding Director of the Center for Spatial Data Science at the University of Chicago, where he is also the Stein-Freiler Distinguished Service Professor of Sociology and the College, as well as a member of the Committee on Data Science. He is the creator of the GeoDa software and an active contributor to the PySAL Python open-source software library for spatial analysis. He has written widely on topics dealing with the methodology of spatial data analysis, including his classic 1988 text on Spatial Econometrics. His work has been recognized by many awards, such as his election to the U.S. National Academy of Science and the American Academy of Arts and Science.

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

An Introduction to Spatial Data Science with GeoDa

Volume 1 – Exploring Spatial Data

Luc Anselin

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Designed cover image: Luc Anselin

First edition published 2024
by CRC Press
2385 NW Executive Center Drive, Suite 320, Boca Raton FL 33431

and by CRC Press
4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

CRC Press is an imprint of Taylor & Francis Group, LLC

Library of Congress Cataloging-in-Publication Data

Names: Anselin, Luc, 1953- author.
Title: An introduction to spatial data science with GeoDa / Luc Anselin.
Description: First edition. | Boca Raton, FL : CRC Press, 2024. | Includes bibliographical references and index. | Contents: Volume 1. Exploring spatial data -- Volume 2. Clustering spatial data. | Summary: “This book is the first in a two-volume series that introduces the field of spatial data science. It offers an accessible overview of the methodology of exploratory spatial data analysis. It also constitutes the definitive user’s guide for the widely adopted GeoDa open source software for spatial analysis. Leveraging a large number of real-world empirical illustrations, readers will gain an understanding of the main concepts and techniques, using dynamic graphics for thematic mapping, statistical graphing, and, most centrally, the analysis of spatial autocorrelation. Key to this analysis is the concept of local indicators of spatial association, pioneered by the author and recently extended to the analysis of multivariate data.”--Provided by publisher.
Identifiers: LCCN 2023048617 | ISBN 9781032229188 (volume 1 ; hardback) | ISBN 9781032229621 (volume 1 ; paperback) | ISBN 9781003274919 (volume 1 ; ebook) | ISBN 9781032713021 (volume 2 ; hardback) | ISBN 9781032713168 (volume 2 ; paperback) | ISBN 9781032713175 (volume 2 ; ebook)
Subjects: LCSH: Spatial analysis (Statistics) | Spatial analysis (Statistics)--Data processing. | GeoDa (Computer file)
Classification: LCC QA278.2 .A565 2024 | DDC 519.5/3--dc23/eng/20240201
LC record available at https://lccn.loc.gov/2023048617

ISBN: 978-1-032-22918-8 (hbk)
ISBN: 978-1-032-22962-1 (pbk)
ISBN: 978-1-003-27491-9 (ebk)

DOI: 10.1201/9781003274919

Typeset in Latin Modern font
by KnowledgeWorks Global Ltd.

Publisher’s note: This book has been prepared from camera-ready copy provided by the authors.

To Emily

Contents

List of Figures

Preface

This two-volume set is the long overdue successor to the *GeoDa Workbook* that I wrote almost twenty years ago (Anselin, 2005a). It was intended to facilitate instruction in spatial analysis and spatial regression by means of the `GeoDa` software (Anselin et al., 2006b). In spite of its age, the workbook is still widely used and much cited, but it is due for a major update.

The update is two-fold. On the one hand, many new methods have been developed or original measures refined. This pertains not only to the spatial autocorrelation indices covered in the original Workbook but also to a collection of newer methods that have become to define *spatial data science*. Secondly, the `GeoDa` software has seen substantial changes to become an open-source and cross-platform ecosystem that encompasses a much wider range of methods than its *legacy* predecessor.

The two volumes outline my vision for an *Introduction to Spatial Data Science*. They include a collection of methods that I view as the core of what is *special* about *spatial* data science, as distinct from applying data science to spatial data. They are not intended to be a comprehensive overview but constitute my personal selection of materials that I see as central to promoting *spatial thinking* through teaching spatial data science.

The level in the current volume is introductory, aimed at my typical audience, which is largely composed of researchers and students (both undergraduate and graduate) who have *not* been exposed to any geographic or spatial concepts or have only limited familiarity with the subject. So, by design, some of the treatment is rudimentary, covering basic concepts in GIS and spatial data manipulation, as well as elementary statistical graphs. I have included this material to keep the books accessible to a larger audience. Readers already familiar with these topics can easily skip to the core techniques.

I believe the two volumes offer a unique perspective, in that they approach the identification of spatial patterns from a number of different standpoints. The first volume includes an in-depth treatment of *local indicators of spatial association*, whereas Volume 2 focuses on *spatial clustering* techniques. The main objective is to indicate where a *spatial* perspective contributes to the broader field of data science and what is unique about it. In addition, the aim is to create an intuition for the type of method that should be applied in different empirical situations. In that sense, the volumes serve both as the complete user guide to the `GeoDa` software and as a *primer* on spatial data science. However, in contrast with the original Workbook, spatial regression methods are not included. Those are covered in Anselin and Rey (2014) and not discussed here.

Most methods contained in the two volumes are treated in more technical detail in the various references provided. With respect to my own work, these include Anselin(1994; 1995; 1996; 1998; 1999; 2005b), Anselin et al. (2002), Anselin et al. (2004), Anselin et al. (2006b), and, more recently, Anselin (2019a; 2019b; 2020), Anselin and Li (2019; 2020) and Anselin et al. (2022). However, a few methods are new and have not been reported elsewhere or are discussed here in greater depth than previously appeared. In this volume, these include the co-location map and the local neighbor match test.

The methods are illustrated with a completely new collection of seven sample data sets that deal with topics ranging from crime, socio-economic determinants of health, and disease spread, to poverty, food insecurity and bank performance. The data pertain not only to the U.S. (Chicago) but also include municipalities in Brazil (the State of Ceará) and in Mexico (the State of Oaxaca), and community banks in Italy. Many of these data sets were used in previous empirical analyses. They are included as built-in *Sample Data* in the latest version of the `GeoDa` software.

The empirical illustrations are based on Version 1.22 of the software, available in Summer 2023. Later versions may include slight changes as well as additional features, but the treatment provided here should remain valid. The software is free, cross-platform and open-source and can be downloaded from https://geodacenter.github.io/download.html.

Acknowledgments

This work would not have existed without the tremendous efforts by the people behind the development of the `GeoDa` software over the past twenty-some years. This started in the early 2000s in the Spatial Analysis Laboratory at the University of Illinois, with major contributions by Ibnu Syabri, supported by the NSF-funded Center for Spatially Integrated Social Science (CSISS). Later, the software development was continued at the GeoDa Center of Arizona State University with Marc McCann as the main software engineer. The last ten years, Xun Li has served as the lead developer of the software. In addition, he has also been a close collaborator on several methodological refinements of the LISA approach. Julia Koschinsky has been on the team for some twenty years, as a constant inspiration and collaborator, starting at UIUC and most recently at the Center for Spatial Data Science of the University of Chicago. Xun and Julia have been instrumental in the migration of `GeoDa` from a closed-source Windows-based desktop software to an open-source and cross-platform ecosystem for exploring spatial data. Julia in particular has been at the forefront of refining the role of ESDA within a scientific reasoning framework, which I have tried to represent in the book.

In addition, I would like to thank Lara Spieker from Taylor & Francis Group for her expert guidance in this project.

Over the years, the research behind the methods covered in this book and the accompanying software development has been funded by grants from the U.S. National Science Foundation, the National Institutes for Health and the Centers for Disease Control, as well as by institutional support by the University of Chicago to the Center for Spatial Data Science.

Finally, Emily has been patiently living with my GeoDa obsession for many years. This book is dedicated to her.

Shelby, MI, Summer 2023

About the Author

Luc Anselin is the Founding Director of the Center for Spatial Data Science at the University of Chicago, where he is also the Stein-Freiler Distinguished Service Professor of Sociology and the College. He previously held faculty appointments at Arizona State University, the University of Illinois at Urbana-Champaign, the University of Texas at Dallas, the Regional Research Institute at West Virginia University, the University of California, Santa Barbara, and The Ohio State University. He also was a visiting professor at Brown University and MIT. He holds a PhD in Regional Science from Cornell University.

Over the past four decades, he has developed new methods for exploratory spatial data analysis and spatial econometrics, including the widely used local indicators of spatial autocorrelation. His 1988 *Spatial Econometrics* text has been cited some 17,000 times. He has implemented these methods into software, including the original SpaceStat software, as well as GeoDa, and as part of the Python PySAL library for spatial analysis.

His work has been recognized by several awards, including election to the U.S. National Academy of Sciences and the American Academy of Arts and Sciences.

1

Introduction

Spatial data are special in that the location of the observations, the *where*, plays a critical role in the methodology required for their analysis. Two aspects in particular distinguish spatial data from the standard independent and identically distributed paradigm (i.i.d.) in statistics and data analysis, i.e., *spatial dependence* and *spatial heterogeneity* (Anselin, 1988; 1990). Spatial dependence refers to the similarity of values observed at neighboring locations, or "everything is related to everything else, but closer places more so," known as Tobler's first law of geography (Tobler, 1970). Spatial heterogeneity is a particular form of structural change associated with *spatial subregions* of the data, i.e., showing a clear break in the spatial distribution of a phenomenon. Both spatial dependence and spatial heterogeneity require a specialized methodology for data analysis, generically referred to as *spatial analysis*.

Spatial data science is an emerging paradigm that extends spatial analysis situated at the interface between spatial statistics and geocomputation. What the term actually encompasses is not settled, and the collection of methods and software tools it represents is also sometimes referred to as geographic data science or geospatial data science (Anselin, 2020; Comber and Brunsdon, 2021; Singleton and Arribas-Bel, 2021; Rey et al., 2023). The concept is closely related to, overlaps somewhat with and has many methods and approaches in common with fields such as geocomputation (Brunsdon and Comber, 2015; Lovelace et al., 2019), cyberGIScience (Wang, 2010; Wang et al., 2013), and, more recently, GeoAI (Janowicz et al., 2020; Gao, 2021).

This two-volume collection is intended as an introduction to the field of spatial data science, emphasizing data exploration and visualization and focusing on the importance of a *spatial* perspective. It represents an attempt to promote spatial thinking in the practice of data science. It is admittedly a selection of methods that reflects my own biases, but it has proven to be an effective collection over many years of teaching and research. The first volume deals with the *exploration* of spatial data, whereas the second volume focuses on *spatial clustering* methods.

The methods covered in both volumes work well for so-called small to medium data settings, but not all of them scale well to *big* data settings. However, some important principles do scale well, like local indicators of spatial association. Even though data sets of very large size have become commonplace and arguably have been the drivers behind a lot of methodological development in modern data science, this is not always relevant for spatial data analysis. The point of departure is often big data (e.g., geo-located social media messages), but eventually, the analysis is carried out at a more spatially aggregate level, where the techniques covered here remain totally relevant.

The methodological approach outlined in this first volume supports an abductive process of exploration, a dynamic interaction between the analyst and the data with the goal of obtaining new insights. The focus is on insights that pertain to *spatial* patterns in the data, such as the *location* of interesting observations (hot spots and cold spots), the presence of

DOI: 10.1201/9781003274919-1

structural breaks in the spatial distribution of the data, and the comparison of such patterns between different variables and over time.

The identification of the patterns is intended to provide cues about the types of processes that may have generated them. It is important to appreciate that exploration is not the same as explanation. In my opinion, exploration nevertheless constitutes an important and necessary step to obtain effective and falsifiable hypotheses to be used in the next stages of the analysis. However, in practice, the line between pure exploration and confirmation (hypothesis testing) is not always that clear, and the process of scientific discovery may move back and forth between the two. I return to this question in more detail in the closing chapter.

The two volumes are both an introduction to the methodology of spatial data science and the definitive guide to the `GeoDa` software. This software represents the implementation of my vision for a gradual progression in the exploration of spatial data, from simple description and mapping to more structured identification of patterns and clusters, culminating with the estimation of spatial regression models. It came at the end of a series of software developments that started in the late 1980s (for a historical overview, see Anselin, 2012).

`GeoDa` is designed to be user-friendly and intuitive, working through a graphical user interface, and therefore it does not require any programming expertise. Similarly, the emphasis in the two volumes is on spatial concepts and how they can be implemented through the software, but it does not deal with geocomputation as such.

A distinctive characteristic of `GeoDa` is the efficient implementation of dynamically linked graphs, in the sense that one or more selected observations in a "view" of the data (a graph or map) are immediately also selected in all the other views, allowing interactive linking and brushing of the data (Anselin et al., 2006b). Since its initial release in 2003 (through the NSF-funded Center for Spatially Integrated Science), the software has been adopted widely for both teaching and research, with close to 600,000 unique downloads at the time of this writing.

In the remainder of this introduction, I first provide a broad overview of the organization of this first volume. This is followed by a quick tour of the `GeoDa` software and a listing of the sample data sets used to illustrate the methods.

1.1 Overview of Volume 1

The first volume is organized into five main parts and an Epilogue as a sixth, offering a progression from basic data manipulation, through description and exploration, to the identification of clusters and outliers by means of spatial autocorrelation analysis. It closes with some reflections on the limits of exploration and its role in scientific discovery. As mentioned, spatial clustering methods are covered in Volume 2.

The six parts are:

- Spatial data wrangling
- EDA and ESDA
- Spatial weights
- Global spatial autocorrelation

- Local spatial autocorrelation
- Epilogue

Part I deals with basic data operations for both tabular and spatial data, covered in two chapters. The material includes a review of the distinctive characteristics of spatial data, how to create spatial layers inside `GeoDa`, as well as essential transformations and data queries. There is also a rudimentary discussion of a range of basic GIS operations, such as projections, converting between points and polygons, and spatial joins. Even though `GeoDa` is not (and not intended to be) a GIS, this functionality has been included over the years in response to user demand.

Part II covers the principles behind exploratory data analysis (EDA) and its spatial counterpart, exploratory spatial data analysis (ESDA). This includes six chapters. Three of these are devoted to map use in various degrees of complexity, starting with basic mapping concepts and moving to statistical maps and maps for rates. The other three chapters deal with conventional (non-spatial) EDA, in the form of univariate and bivariate data exploration, multivariate data exploration and space-time exploration. The core idea here is to leverage linking and brushing between various graphical representations (*views* of the data), which is central to the architecture of `GeoDa`.

The remaining three main parts deal with the topic of spatial autocorrelation. First, in Part III, three chapters are devoted to spatial weights, both contiguity-based and distance-based spatial weights, and various spatial weights operations. These are essential pre-requisites for the computation of the global and local spatial autocorrelation indices covered in Parts IV and V.

Part IV contains three chapters on global spatial autocorrelation, centered around the Moran scatter plot as a visualization device. The basic concepts are covered, as well as more advanced applications and extensions to a bivariate setting. The third chapter provides an overview of some non-parametric techniques, such as a spatial correlogram.

Part V includes an in-depth treatment of local spatial autocorrelation, spread over five chapters. It starts with the introduction of the concept of a LISA and the Local Moran statistic. The second chapter deals with other local spatial autocorrelation statistics, such as the Local Geary and the Getis-Ord statistics. The next two chapters outline extensions to the multivariate domain and to discrete variables. These chapters contain material that was only fairly recently developed. The last chapter of Part V reviews density-based clustering methods applied to point locations, such as DBScan and HDBScan.

The Epilogue offers some thoughts on the limits of the exploratory perspective. This includes an assessment of the role of data exploration in aiding with scientific discovery and scientific reasoning, the limits of spatial analysis, and reproducibility in the exploratory framework as implemented in the `GeoDa` software.

An Appendix includes detailed preference settings for the software and an outline of the complete menu structure. To close, a brief discussion is offered of the new scripting possibilities through the `geodalib` library.

The division of the material in two volumes follows my own teaching practice. The first volume corresponds to what I cover in an *Introduction to Spatial Data Science* course, whereas the second volume matches the content of a *Spatial Cluster Analysis* course. The volumes are also designed to constitute a self-study guide. In fact, a previous version was used as such for remote teaching during the Covid pandemic (in the form of laboratory workbooks, available at https://geodacenter.github.io/documentation.html).

In addition to the material covered in the two volumes, the GeoDaCenter Github site (https://geodacenter.github.io) contains an extensive support infrastructure. This includes detailed documentation and illustrations, as well as a large collection of sample data sets, cookbook examples, and links to a YouTube channel containing lectures and tutorials. Specific software support is provided by means of a list of *frequently asked questions* and *answers to common technical questions*, as well as by the community through the *Google Groups Openspace* list.

1.2 A Quick Tour of GeoDa

Before delving into the specifics of particular methods, I provide a broad overview of the functionality and overall organization of the `GeoDa` software. The complete toolbar with icons corresponding to a collection of related operations is shown in Figure 1.1. Each icon is matched by a menu item, detailed in Appendix B. The menu and user interface can be customized to several languages (details are in Appendix A). The default is English, but options are available for Simplified Chinese, Russian, Spanish, Portuguese and French, with more to come in the future.

With each toolbar icon typically corresponds a drop-down list of specific functions. The structure of the drop-down list matches the menu sub-items (Appendix B).

Figure 1.1: GeoDa toolbar icons

The organization of the toolbar (and menu) follows the same logic as the layout of the parts and chapters in the two books. It represents a progression in the exploration, from left to right, from support functions to queries, description and visualization, and more and more formal methods, ending up with the estimation of actual spatial models in the regression module (not covered here).

A brief overview of each of the major parts is given next. This also includes the spatial clustering functionality, which is discussed more specifically in Volume 2.

1.2.1 Data entry

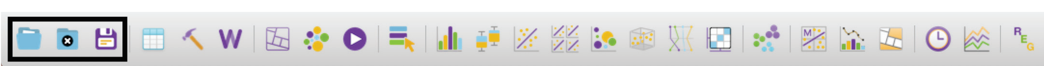


Figure 1.2: Data entry

The three left-most icons, highlighted in Figure 1.2, deal with data entry and general input-output. This includes the loading of spatial and non-spatial (e.g., tabular) data layers from a range of GIS and other file formats (supported through the open-source GDAL library). In addition, it offers connections to spatial databases, such as PostGIS and Oracle Spatial. It also supports a **Save As** function, which allows the software to work as a GIS file format converter. Further details are provided in Chapter 2.

1.2.2 Data manipulation

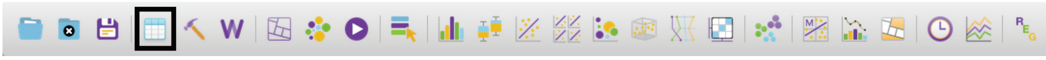


Figure 1.3: Data manipulation/table

Functionality for data manipulation and transformation is provided by the **Table** icon, highlighted in Figure 1.3. This allows new variables to be created, observations selected, queries formulated and includes other data table operations, such as merger and aggregation, detailed in Chapter 2.

1.2.3 GIS operations

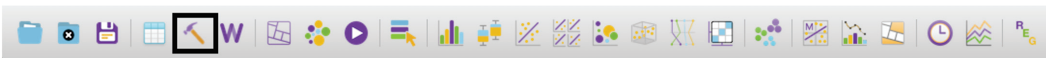


Figure 1.4: GIS operations/tools

Spatial data operations are invoked through the **Tools** icon, highlighted in Figure 1.4. These include many GIS-like operations that were added over the years to provide access to spatial data for users who are not familiar with GIS. For example, point layers can be easily created from tabular data with X,Y coordinates, point in polygon operations support a spatial join, an indicator variable can be used to implement a dissolve application, and reprojection can be readily implemented by means of a **Save As** operation. Specific illustrations are included in Chapter 3.

1.2.4 Weights manager

Figure 1.5: Weights manager

The **Weights Manager** icon, Figure 1.5, contains a final set of functions that are in support of the analytical capabilities. It gives access to a wide range of weight creation and manipulation operations, discussed at length in the chapters of Part III. This includes constructing spatial weights from spatial layers, as well as loading them from external files, summarizing and visualizing their properties, and operations like union and intersection.

1.2.5 Mapping and geovisualization

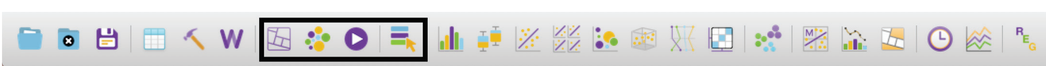


Figure 1.6: Mapping and geovisualization

The mapping and geovisualization functionality is represented by four icons, highlighted in Figure 1.6: the **Map** icon, **Cartogram**, **Map Movie** and **Category Editor**. The mapping function supports all the customary types of choropleth maps, as well as some specialized features, such as extreme value maps, co-location maps and smoothed maps for rates. The cartogram is a specialized type of map that replaces the actual outline of spatial units by a circle, whose area is proportional to a given variable of interest. Animation, in the sense of moving through the locations of observations in increasing or decreasing order of the value for a given variable is implemented by means of the map movie icon. Finally, the category editor provides a way to design custom classifications for use in maps as well as in statistical graphs, such as a histogram. Details are provided in Chapters 4 through 6.

1.2.6 Exploratory data analysis

Figure 1.7: Exploratory data analysis

The next eight icons, grouped in Figure 1.7, contain the functionality for exploratory data analysis and statistical graphs. This includes a **Histogram**, **Box Plot**, **Scatter Plot**, **Scatter Plot Matrix**, **Bubble Chart**, **3D Scatter Plot**, **Parallel Coordinate Plot** and **Conditional Plots**. These provide an array of methods for univariate, bivariate and multivariate exploration. All the graphs are connected to any other open window (graph or map) for instantaneous linking and brushing. This is covered in more detail in Chapters 7 and 8.

1.2.7 Space-time analysis

Figure 1.8: Space-time analysis

The exploration of space-time data, treated in Chapter 9, is invoked by means of the icons on the right, highlighted in Figure 1.8. This includes a **Time Editor**, which is required to transform the cross-sectional observations into a proper (time) sequence. In addition, the **Averages Chart** implements a simple form of treatment analysis, with treatment and controls defined over time and/or across space.

1.2.8 Spatial autocorrelation analysis

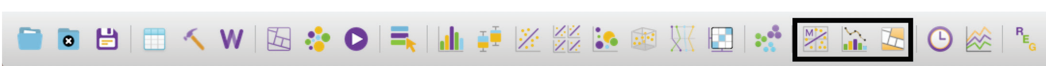


Figure 1.9: Spatial autocorrelation analysis

Spatial autocorrelation analysis is invoked through the three icons highlighted in Figure 1.9. The first two pertain to global spatial autocorrelation. The left-most icon corresponds to various implementations of the Moran scatter plot (Chapters 13 and 14). The middle icon invokes nonparametric approaches to visualize global spatial autocorrelation as a spatial correlogram and distance scatter plot (Chapter 15).

The third icon contains a long list of various implementations of local spatial autocorrelation statistics, including various forms of the Local Moran's I, the Local Geary c, the Getis-Ord statistics and extensions to multivariate settings and discrete variables. The local neighbor match test is a new method based on an explicit assessment of the overlap between locational and attribute similarity. Details are provided in the chapters of Part V.

1.2.9 Cluster analysis

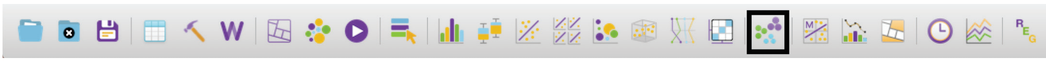


Figure 1.10: Cluster analysis

Finally, cluster analysis is invoked through the icon highlighted in Figure 1.10. An extensive drop-down list also includes the density-based cluster methods DBScan and HDBScan, which are treated in this volume under local spatial autocorrelation (Chapter 20).

The other methods are covered in Volume 2. They include dimension reduction, classic clustering methods and spatially constrained clustering methods. The last items in the drop-down list associated with the cluster icon pertain to the quantitative and visual assessment of cluster validity, including a new cluster match map (see Volume 2).

1.3 Sample Data Sets

As mentioned in the Preface, the methods and software are illustrated by means of empirical examples that use seven new sample data sets. They are available directly from inside the `GeoDa` software through the **Sample Data** tab of the input/output interface (see Figure 2.2).

The specific data sets are:

- *Chicago Carjackings* (n = 1,412)
  - point locations of carjackings in 2020 (Chicago Open Data Portal)
  - see Chapters 2 and 3
- *Ceará Zika*, municipalities in the State of Ceará, Brazil (n = 184)
  - Zika and Microcephaly infections and socio-economic profiles for 2013–2016 (adapted from Amaral et al., 2019)
  - see Chapters 4–6, 10 and Part III of Volume 2 (Spatial Clustering)
- *Oaxaca Development*, municipalities in the State of Oaxaca, Mexico (n = 570)
  - poverty and food insecurity indicators and census variables for 2010 and 2020 (CONEVAL and INEGI) (based on the same original sources as Farah Rivadeneyra, 2017)
  - see Chapters 7–9, 12, 14 and 16–17
- *Italy Community Banks* (n = 261)
  - bank performance indicators for 2011–17 (used by Algeri et al., 2022)
  - see Chapters 11–12, 15 and 20, as well as in Part I of Volume 2 (Dimension Reduction)
- *Chicago Community Areas*, CCA Profiles (n = 77)
  - socio-economic snapshot for Chicago Community Areas in 2020 (American Community Survey from the Chicago Metropolitan Agency for Planning – CMAP – data portal)
  - see Chapters 13 and Chapter 5 of Volume 2 (Hierarchical Clustering Methods)

- *Chicago SDOH*, census Tracts (n = 791)
  - socio-economic determinants of health in 2014 (a subset of the data used in Kolak et al., 2020)
  - see Chapters 18–19 and Chapters 6 and 7 of Volume 2 (Partitioning Clustering Methods and Advanced Clustering Methods)
- *Spirals* (n = 300)
  - canonical data set to test spectral clustering
  - only used in Volume 2 (Chapter 8, Spectral Clustering)

In addition, a few auxiliary files are employed to illustrate basic data handling operations in Chapters 2 and 3, such as a boundary layer for Chicago community areas and input data files in comma-separated text format. These files are available from the GeoDaCenter sample data site at https://geodacenter.github.io/data-and-lab/.

Further details are provided in the context of specific methods.

Part I

Spatial Data Wrangling

2

Basic Data Operations

In this and the following chapter, I introduce the topic of *data wrangling*, i.e., the process of getting data from its raw input into a form that is amenable for analysis. This is often considered to be the most time consuming part of a data science project, taking as much as 80% of the effort (Dasu and Johnson, 2003). Even though the focus in this book is on *analysis* and not on data manipulation per se, I provide a quick overview of the functionality contained in `GeoDa` to assist with these operations. Increasingly, data wrangling has evolved into a field of its own, with a growing number of operations turning into automatic procedures embedded into software (Rattenbury et al., 2017). A detailed discussion of this topic is beyond the scope of the book.

The coverage in this chapter is aimed at novices who are not very familiar with spatial data manipulations. Most of the features illustrated can be readily accomplished by means of dedicated GIS software or by exploiting the spatial data functionality available in the `R` and `Python` worlds. Readers knowledgeable in such operations may want to just skim the materials in order to become familiar with the way they are implemented in `GeoDa`. Alternatively, these operations can be performed outside `GeoDa`, with the end result loaded as a spatial data layer.

In the current chapter, I focus on essential input operations and data manipulations contained in the **Table** functionality. In the next chapter, I consider a range of basic GIS operations pertaining to *spatial* data wrangling.

To illustrate these features, I will use a data set with point locations of car jackings in Chicago in 2020. The *Chicago Carjackings* data layer is available from the **Sample Data** tab in the `GeoDa` file dialog (Figure 2.2).

In addition, in order to replicate the detailed steps used in the illustrations, three original input files are needed as well. These are available from the GeoDa-Center sample data site. They include a simple outline of the community areas, *Chicago_community_areas.shp*, as well as comma delimited (csv) text files with the socio-economic characteristics (*Chicago_CCA_Profiles.csv*), and the coordinates of the car jackings (*Chicago_2020_carjackings.csv*). The sample data site also contains the detailed listing of the variable names.

2.1 Topics Covered

- Load a spatial layer from a range of formats
- Convert between spatial formats
- Create a point layer from coordinates in a table
- Create a grid layer

DOI: 10.1201/9781003274919-2

- Become familiar with the table options
- Use the Calculator Tool to create new variables
- Variable standardization
- Merging tables
- Use the Selection Tool to select observations in a table
- Use a selection shape to select observations in a map

GeoDa Functions

- File > Open
- File > Save
- File > Save As
- Tools > Shape > Points from Table
- Tools > Shape > Create Grid
- Table > Edit Variable Properties
- Table > Add Variable
- Table > Delete Variable(s)
- Table > Rename Variable
- Table > Encode
- Table > Setup Number Formatting
- Table > Move Selected to Top
- Table > Calculator
- Table > Selection Tool
- File > Save Selected As
- Map > Unique Values Map
- Map > Selection Shape
- Map > Save Selection

Toolbar Icons

Figure 2.1: Open | Close | Save | Table | Tools

2.2 Spatial Data

Spatial data are characterized by the combination of two important aspects. First, there is information on variables, just as in any other statistical analysis. In the spatial world, this is referred to as *attribute* information. Typically, it is contained in a *flat* (rectangular) table with observations as rows and variables as columns.

The second aspect of spatial data is special and is referred to as *locational* information. It consists of the precise definition of spatial objects, classified as points, lines or areas (polygons). In essence, the formal characterization of any spatial object boils down to the description of X-Y coordinates of points in space, as well as of a mechanism that spells out how these points are combined into spatial entities.

For a single point, the description simply consists of its coordinates. For areal units, such as census tracts, counties, or states, the associated polygon boundary is defined as a series of line segments, each characterized by the coordinates of their starting and ending points. In other words, what may seem like a continuous boundary, is turned into *discrete* segments.

Traditional data tables have no problem including X and Y coordinates as columns, but as such cannot deal with the boundary definition of irregular spatial units. Since the number of line segments defining an areal boundary can easily vary from observation to observation, there is no efficient way to include this in a fixed number of columns of a flat table. Consequently, a specialized *data structure* is required, typically contained in a geographic information system or GIS.

Several specialized formats have been developed to efficiently combine both the attribute information and the locational information. Such spatial data can be contained in files with a special structure, or in spatially enabled relational data base systems.

I first consider common GIS file formats that can serve as input to `GeoDa`. This is followed by an illustration of simple tabular input of non-spatial files. Finally, a brief overview is given of connections to other input formats.

2.2.1 GIS files

Historically, a wide range of different formats have been developed for GIS data, both proprietary as well as open-source. In addition, there has been considerable effort at standardization, led by the Open Geospatial Consortium (OGC).[1] `GeoDa` leverages the open-source `GDAL` library[2] to support input and output of many of the most popular formats in use today.

While it is impossible to cover all of these specifications in detail, I will illustrate three specific formats here. First is the use of the proprietary *shape file* format of the leading GIS vendor ESRI.[3] In addition, the open-source *GeoJSON* format[4] will be covered, as well as the *Geography Markup Language* of the OGC, a standard XML grammar for defining geographical features.[5]

In `GeoDa`, one can load both polygon and point GIS data, but in the current implementation, line files are *not* supported (e.g., to represent road networks).

2.2.1.1 Spatial file formats

Arguably, the most familiar proprietary spatial data format is the *shape file* format, developed by ESRI. The terminology is a bit confusing, since there is no such thing as *one* shape file, but there is instead a collection of three (or four) files. One file has the extension *.shp*, one *.shx*, one *.dbf* and one *.prj* (with the projection information). The first three are required, the fourth one is optional, but highly recommended. The files should all be in the same directory and have the same file name, except for the file extension.

In the open-source world, an increasingly common format is *GeoJSON*, the geographic augmentation of the JSON standard, which stands from *JavaScript Object Notation*. This format is contained in a text file and is easy for machines to read, due to its highly structured nature.

Finally, the *GML* standard, or Geographic Markup Language, is a XML implementation that prescribes the formal description of geographic features.

[1] https://www.ogc.org
[2] https://gdal.org
[3] https://www.esri.com/library/whitepapers/pdfs/shapefile.pdf
[4] https://geojson.org
[5] https://www.ogc.org/standards/gml

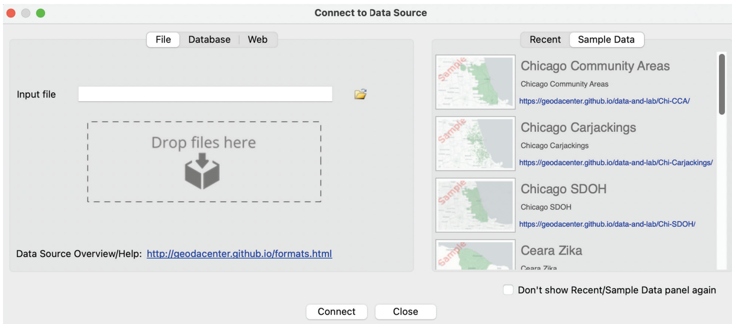


Figure 2.2: Connect to Data Source dialog

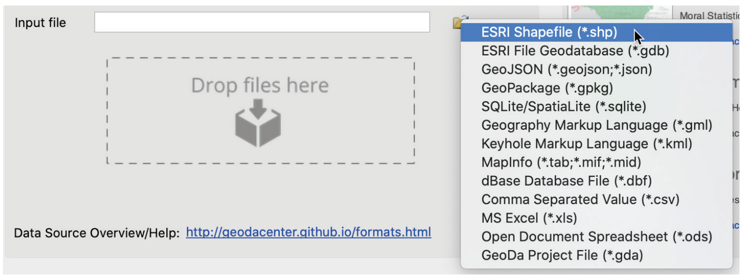


Figure 2.3: Supported spatial file formats

A detailed discussion of the individual formats is beyond the current scope. All are well-documented, with many additional resources available online. Although it is always helpful, there is no need to know the underlying formats in detail in order to use `GeoDa`, since the interaction with the data structures is handled under the hood.

The main file manipulations are invoked from the **File** item in the menu, or by the three left-most icons on the toolbar in Figure 2.1.

2.2.1.2 Polygon layers

Since `GeoDa` is particularly geared to the exploration of areal unit data, the input of a so-called *polygon layer* is illustrated first. Any spatial layer present as a file can be loaded by invoking **File > Open File** from the menu, or by clicking on the left-most **Open** icon on the toolbar in Figure 2.1.

This brings up the **Connect to Data Source** dialog, shown in Figure 2.2. The left panel has **File** as the active input format. Other formats are **Database** and **Web**, which are briefly covered in Section 2.2.3. The right panel shows a series of **Sample Data** data that are included with `GeoDa`. In addition, after some files have been loaded in the current application, the **Recent** panel will contain their file names as well. Files listed in either panel can be loaded by simply clicking on the corresponding icon.

The small folder icon to the right of the **Input file** box brings up a list of supported file formats, as in Figure 2.3. In this first example, the top item in the list is selected, **ESRI Shapefile (*.shp)**.

To illustrate this feature, the four files associated with the *Chicago_community_areas* shape file must be available in a working directory (they must be downloaded from the GeoDaCenter sample data site).

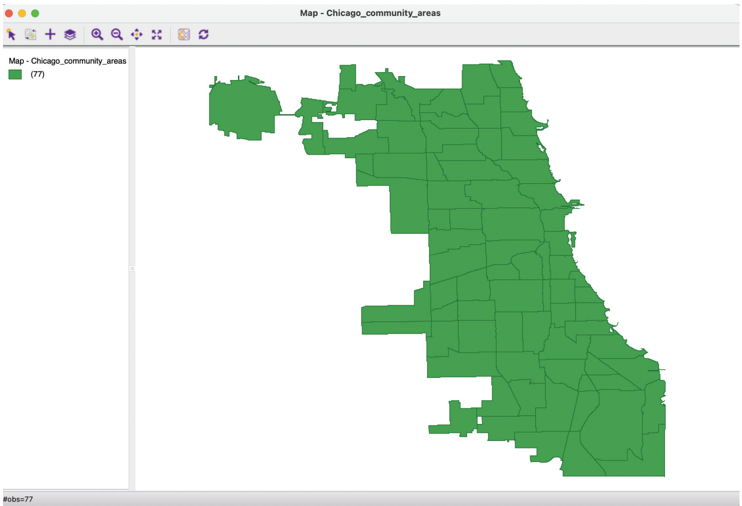


Figure 2.4: Themeless polygon map


```
{
"type": "FeatureCollection",
"name": "Chicago_community_areas",
"crs": { "type": "name", "properties": { "name": "urn:ogc:def:crs:OGC:1.3:CRS84" } },
"features": [
{ "type": "Feature", "properties": { "area_num_1": "35", "area_numbe": "35", "community":
"DOUGLAS", "shape_area": 46004621.158100002, "shape_len": 31027.0545098, "districtno": 7,
"district": "South Side" }, "geometry": { "type": "MultiPolygon", "coordinates":
[ [ [ [ -87.609140876178941, 41.844692502653977 ], [ -87.609148747578075, 41.844661598424032 ],
[ -87.609161120412594, 41.84458961193954 ], [ -87.609167662158384, 41.844517177323162 ],
[ -87.60916860600166, 41.844456260738305 ], [ -87.609150121993977, 41.844238716598113 ],
[ -87.609072412492893, 41.844194738881015 ], [ -87.609006271478208, 41.844106469286963 ],
[ -87.608965021721602, 41.844043457551152 ], [ -87.608915663906146, 41.84395529375054 ],
[ -87.608899801189878, 41.843873616495323 ], [ -87.608867013718623, 41.843804382800478 ],
[ -87.608851434244897, 41.843697606960866 ], [ -87.608810892810936, 41.843571847766412 ],
```


Figure 2.5: Example GeoJSON file contents

Using the navigation dialog and conventions appropriate for each operating system, the shape file can be selected from this directory. This opens a new map window with the spatial layer represented as a themeless choropleth map, as in Figure 2.4. The number of observations is shown in parentheses next to the small green rectangle in the upper-left panel, as well as in the status bar at the bottom (**#obs = 77**).

The current layer is cleared by clicking on the **Close** toolbar icon, the second item on the left in Figure 2.1, or by selecting **File > Close** from the menu. This removes the base map. At this point, the **Close** icon on the toolbar becomes inactive.

A more efficient way to open files is to select the file name in the directory window and to drag it onto the **Drop files here** box in the dialog. Even easier is to load a one of the sample data sets or a recently used one, where a simple click on the associated icon in the **Sample Data** or **Recent** tab suffices.

In contrast to the shape file format, which is binary, a GeoJSON file is simple text and can easily be read by humans. As shown for the *Chicago_community_areas.geojson* file from the sample data site in Figure 2.5 (this file must be downloaded to a working directory), the *locational* information is combined with the attributes. After some header information follows a list of `features`. Each of these contains `properties`, of which the first set consists

```
41.8452665489062 -87.6111225641177 41.8452664389385 -87.6109165492239 41.845266448213 -87.6094061454063
41.8452665039859 -87.6094094918227 41.8452177332683 -87.6093765809228 41.8451533826366 -87.6091408761789
41.844692502654</gml:posList></gml:LinearRing></gml:exterior></gml:Polygon></gml:surfaceMember></
gml:MultiSurface></ogr:geometryProperty>
      <ogr:area_num_1>35</ogr:area_num_1>
      <ogr:area_numbe>35</ogr:area_numbe>
      <ogr:community>DOUGLAS</ogr:community>
      <ogr:shape_area>46004621.1581000015139579772949218750000</ogr:shape_area>
      <ogr:shape_len>31027.0545098000002326443791389465332</ogr:shape_len>
      <ogr:districtno>7</ogr:districtno>
      <ogr:district>South Side</ogr:district>
    </ogr:Chicago_community_areas>
  </ogr:featureMember>
  <ogr:featureMember>
    <ogr:Chicago_community_areas gml:id="Chicago_community_areas.1">
      <gml:boundedBy><gml:Envelope><gml:lowerCorner>-87.6126242724032 41.8168137705722</
gml:lowerCorner><gml:upperCorner>-87.5921528387939 41.8313662468685</gml:upperCorner></gml:Envelope></
gml:boundedBy>
      <ogr:geometryProperty><gml:MultiSurface
gml:id="Chicago_community_areas.geom.1"><gml:surfaceMember><gml:Polygon
gml:id="Chicago_community_areas.geom.1.0"><gml:exterior><gml:LinearRing><gml:posList>-87.5921528387939
41.8169293462668 -87.5923080508337 41.8169321089497 -87.5948918343729 41.8169406679124 -87.5952614717272
41.8169427647923 -87.5959594527106 41.8168331429737 -87.5960713448987 41.8168328321002 -87.5961924032806
41.8168329229749 -87.5962488053834 41.8168329418813 -87.5963998526963 41.8168171023798 -87.5964634709017
41.8169077284035 -87.5968043283174 41.8169268189399 -87.5968828203083 41.8169279822056 -87.5975100001849
```


Figure 2.6: Example GML file contents

of the different variable names with the associated values, just as would be the case in any standard data table. The final item refers to the `geometry`. This includes the `type`, here a `MultiPolygon`, followed by a list of X-Y `coordinates`. In this fashion, the spatial information is integrated with the attribute information.

To view the corresponding map, the *Chicago_community_areas.geojson* file name can be selected in its directory and dragged onto the **Drop files here** box. This brings up the same base map as in Figure 2.4.

2.2.1.3 File format conversion

The file just loaded was originally specified in the GeoJSON format. It can be easily converted to a different format by means of the **File > Save As** functionality. For example, to change it into a GML format file (e.g., for use in a different program), **Geographic Markup Language (*.gml)** can be selected from the drop-down list of available formats, shown in Figure 2.3.

This will yield a text file in the GML XML format, illustrated in Figure 2.6. It shows the characteristic `< >` and `</ >` delimiters of the markup elements, typical of XML files. In the file snippet in the figure, the top lines pertain to the geography of the first polygon, ended by `</ogr:geometryProperty>`. Next follow the actual observations, with variable names and associated values, finally closed off with `</ogr:featureMember>`. After this, a new observation is listed, delineated by the `<ogr:featureMember>` tag, followed by the geographic characteristics. Again, this illustrates how spatial information is combined with attribute information in an efficient file format.

In sum, the **File > Save As** feature in `GeoDa` turns the program into an effective GIS format converter.

2.2.1.4 Point layers

In the same fashion as for polygon layers, spatial data layers containing point locations can be loaded for the file formats listed in Figure 2.3. As before, the file is either selected explicitly from the proper directory, or the file name is dragged directly into the **Drop files here** box in the dialog.

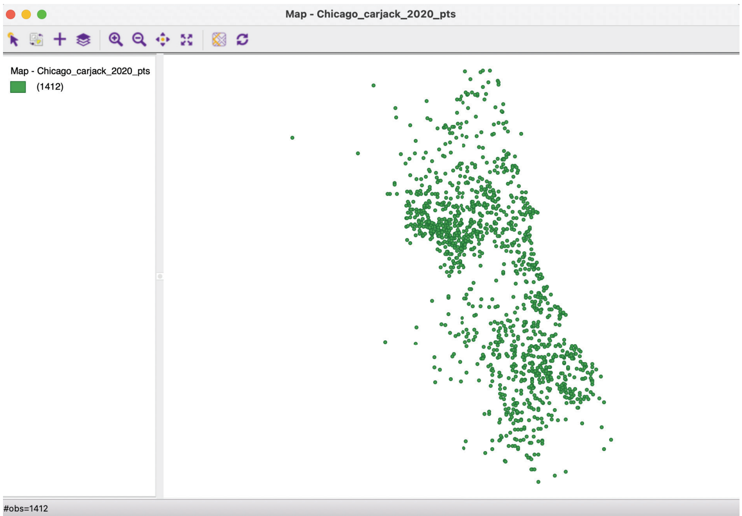


Figure 2.7: Themeless point map

The point map in Figure 2.7 shows the locations of the 1,412 carjackings that occurred in the City of Chicago during the year 2020. It is generated by clicking on the **Chicago Car Jackings** icon in the **Sample Data** tab, or by dragging the file name *Chicago_carjack_2020_pts.shp* from a working directory that contains the shape file.

The shape of the city portrayed by the outline of the points is slightly different from that in the polygon map in Figure 2.4. This is due to a difference in projections: the point map is in the State Plane Illinois East NAD 1983 projection (EPSG 3435), whereas the polygon map uses decimal degrees latitude and longitude (EPSG 4326). This important aspect of spatial data is often a source of confusion for non-GIS specialists. `GeoDa` provides an intuitive interface to deal with projection issues. I return to this topic in Section 2.3.1.1 below and in Section 3.2 in the next chapter.

2.2.2 Tabular files

In addition to GIS files, `GeoDa` can also read regular non-spatial tabular data. While this does not allow for *spatial* analysis (unless coordinates are contained in the table, see Section 2.3.1), all non-spatial operations and graphs are supported. Specifically, all standard techniques of exploratory data analysis (EDA) can be applied, as covered in Chapters 7 and 8 in this Volume. This does not necessitate a map layer.

The data for the Community area socio-economic profiles are contained in a comma-separated file (csv format) on the sample data site (the Chicago CCA Profiles are also available as a spatial layer in the **Sample Data** tab). Selecting this file (after downloading it to a working directory) generates the dialog shown in Figure 2.8.

Since a csv file is pure text, there is no information on the type of the variables included in the file. `GeoDa` tries to guess the type and lists the **Data Type** for each field, as well as a brief preview of the table. At this point, the type can be changed before the data are moved into the actual data table (see also Section 2.4.1.1).

Instead of a base map, which is the default opening window for spatial data, the data table is brought up in a spreadsheet-like format (see also Figure 2.9 for an illustration). In