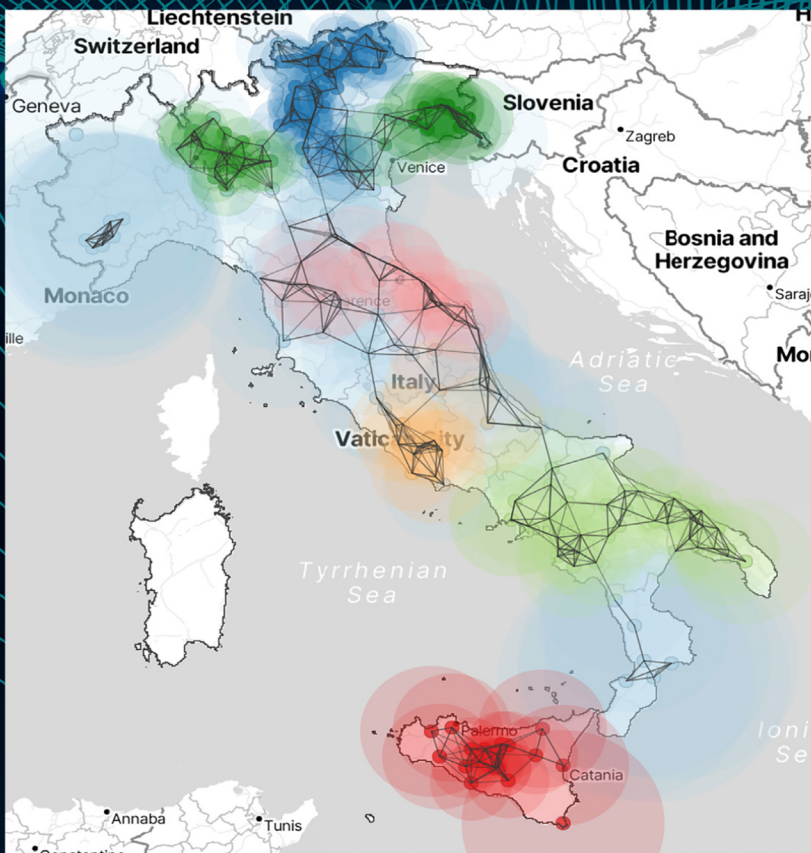


An Introduction to Spatial Data Science with GeoDa

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

© 2024 Luc Anselin

Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged, please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, access www.copyright.com or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. For works that are not available on CCC, please contact mpkbookspermissions@tandf.co.uk

Trademark notice: Product or corporate names may be trademarks or registered trademarks and are used only for identification and explanation without intent to infringe.

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](https://doi.org/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



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

Contents

List of Figures	xv
Preface	xxiii
Acknowledgments	xxv
About the Author	xxvii
1 Introduction	1
1.1 Overview of Volume 1	2
1.2 A Quick Tour of GeoDa	4
1.2.1 Data entry	4
1.2.2 Data manipulation	5
1.2.3 GIS operations	5
1.2.4 Weights manager	5
1.2.5 Mapping and geovisualization	5
1.2.6 Exploratory data analysis	6
1.2.7 Space-time analysis	6
1.2.8 Spatial autocorrelation analysis	6
1.2.9 Cluster analysis	7
1.3 Sample Data Sets	7
I Spatial Data Wrangling	9
2 Basic Data Operations	11
2.1 Topics Covered	11
2.2 Spatial Data	12
2.2.1 GIS files	13
2.2.2 Tabular files	17
2.2.3 Other spatial data input	18
2.3 Creating Spatial Layers	19
2.3.1 Point layers from coordinates	19
2.3.2 Grid	22
2.4 Table Manipulations	22
2.4.1 Variable properties	23
2.4.2 Calculator	26
2.4.3 Merging tables	29
2.5 Queries	30
2.5.1 Selection Tool	31
2.5.2 Indicator variable	33
2.5.3 Save selected observations	33
2.5.4 Spatial selection	34

3	GIS Operations	37
3.1	Topics Covered	37
3.2	Projections	38
3.2.1	Coordinate reference system	39
3.2.2	Selecting a projection	39
3.2.3	Reprojection	41
3.3	Converting Between Points and Polygons	41
3.3.1	Mean centers and centroids	42
3.3.2	Tessellations	44
3.4	Minimum Spanning Tree	45
3.4.1	Concept	45
3.4.2	Minimum Spanning Tree options	49
3.5	Aggregation	49
3.5.1	Dissolve	50
3.5.2	Aggregation in table	51
3.6	Multi-Layer Support	52
3.6.1	Loading multiple layers	52
3.6.2	Automatic reprojection	54
3.6.3	Selection in multiple layers	54
3.7	Spatial Join	55
3.7.1	Spatial assign	56
3.7.2	Spatial count	57
3.8	Linked Multi-Layers	58
3.8.1	Specifying an inter-layer linkage	59
3.8.2	Visualizing linked selections	59
II	EDA and ESDA	61
4	Geovisualization	63
4.1	Topics Covered	64
4.2	From EDA to ESDA	64
4.2.1	Exploratory data analysis	65
4.2.2	Mapping as exploration	66
4.2.3	Exploratory spatial data analysis	66
4.2.4	Linking and brushing	66
4.3	Thematic Maps – Overview	67
4.3.1	Choropleth map	67
4.3.2	Map classification	67
4.3.3	Legend and color	68
4.3.4	Implementation	69
4.4	Common Map Classifications	69
4.4.1	Quantile map	69
4.4.2	Equal intervals map	71
4.4.3	Natural breaks map	72
4.5	Map options	73
4.5.1	The map legend	74
4.5.2	The map toolbar	75
4.5.3	Map base layer (base map)	75
4.5.4	Saving the classification as a categorical variable	78
4.5.5	Saving the map as an image	79

4.5.6	Other map options	80
4.6	Custom Classifications	81
4.6.1	Category editor	81
4.6.2	Applications of custom categories	83
4.6.3	Saving the custom categories – the Project File	83
5	Statistical Maps	87
5.1	Topics Covered	87
5.2	Extreme Value Maps	88
5.2.1	Percentile map	88
5.2.2	Box map	89
5.2.3	Standard deviation map	91
5.3	Mapping Categorical Variables	92
5.3.1	Unique values map	93
5.3.2	Co-location map	94
5.4	Cartogram	96
5.4.1	Principle	96
5.4.2	Creating a cartogram	97
5.5	Map Animation	98
6	Maps for Rates	101
6.1	Topics Covered	101
6.2	Choropleth Maps for Rates	102
6.2.1	Spatially extensive and spatially intensive variables	102
6.2.2	Raw rate map	104
6.3	Excess Risk – SMR – LQ	106
6.3.1	Relative risk	106
6.3.2	Excess risk map	106
6.4	Rate Smoothing	108
6.4.1	Variance instability	109
6.4.2	Borrowing strength	109
6.4.3	Empirical Bayes smoothed rate map	112
7	Univariate and Bivariate Data Exploration	115
7.1	Topics Covered	115
7.2	Analyzing the Distribution of a Single Variable	116
7.2.1	Histogram	116
7.2.2	Box plot	121
7.3	Bivariate Analysis – The Scatter Plot	123
7.3.1	Scatter plot basics	123
7.3.2	Smoother option – local regression	128
7.4	Scatter Plot Matrix	130
7.4.1	Implementation	130
7.5	Spatial Heterogeneity	134
7.5.1	Averages chart	134
7.5.2	Brushing the scatter plot	138
8	Multivariate Data Exploration	141
8.1	Topics Covered	141
8.2	The Curse of Dimensionality	142
8.2.1	The empty space phenomenon	143

8.3	Three Variables: Bubble Chart and 3-D Scatter Plot	144
8.3.1	Bubble chart	144
8.3.2	3-D scatter plot	148
8.4	Conditional Plots	151
8.4.1	Implementation	152
8.4.2	Conditional statistical graphs	153
8.4.3	Conditional maps	155
8.5	Parallel Coordinate Plot	156
8.5.1	Implementation	156
8.5.2	Clusters and outliers in PCP	159
9	Space-Time Exploration	161
9.1	Topics Covered	161
9.2	Time Editor	162
9.2.1	Creating grouped variables	162
9.3	Time Player – Comparative Statics	166
9.3.1	Box plot over time	166
9.3.2	Scatter plot with a time-lagged variable	167
9.3.3	Thematic map over time	169
9.4	Treatment Effect Analysis – Averages Chart	171
9.4.1	Difference in means	172
9.4.2	Difference in difference	172
III	Spatial Weights	177
10	Contiguity-Based Spatial Weights	179
10.1	Topics Covered	179
10.2	The Concept of Spatial Weights	180
10.2.1	Spatial weights matrix	181
10.2.2	Types of contiguity	183
10.2.3	Higher-order contiguity	185
10.2.4	Practical considerations	187
10.3	Creating Contiguity Weights	188
10.3.1	Weights manager	188
10.3.2	Rook contiguity	189
10.3.3	Queen contiguity	191
10.3.4	Higher-order contiguity	192
10.3.5	Project file	192
10.3.6	Using existing weights	192
10.3.7	Space-time weights	193
10.4	Weights Characteristics	196
10.4.1	Summary characteristics	196
10.4.2	Connectivity histogram	196
10.4.3	Connectivity map	198
10.4.4	Connectivity graph	199
10.4.5	Connectivity option	200
11	Distance-Based Spatial Weights	203
11.1	Topics Covered	203
11.2	Distance Metrics	204

11.2.1	Distance in a Cartesian coordinate system	204
11.2.2	Great circle distance	205
11.2.3	General distance	205
11.3	Distance-Based Weights	206
11.3.1	Distance-band weights	206
11.3.2	K-nearest neighbor weights	207
11.3.3	General distance weights	208
11.4	Implementation	209
11.4.1	Distance-band weights	211
11.4.2	Isolates	213
11.4.3	K-nearest neighbor weights	215
11.4.4	General distance matrix	216
11.5	Broadening the Concept of Contiguity	217
11.5.1	Contiguity-based weights for points	218
11.5.2	Distance-based weights for polygons	220
11.5.3	Block weights	220
11.5.4	Space-time weights	222
11.6	Manipulating Weights	223
11.6.1	Set operations	223
11.6.2	Make symmetric	224
12	Special Weights Operations	225
12.1	Topics Covered	225
12.2	Spatial Weights as Distance Functions	226
12.2.1	Inverse distance weights	226
12.2.2	Kernel weights	229
12.3	Spatial Transformations	233
12.3.1	Spatially lagged variables	233
12.3.2	Using inverse distance weights – <i>potential</i> measure	236
12.3.3	Using kernel weights	238
12.4	Spatial Weights in Rate Smoothing	239
12.4.1	Spatial rate smoothing	240
12.4.2	Spatial Empirical Bayes smoothing	242
IV	Global Spatial Autocorrelation	245
13	Spatial Autocorrelation	247
13.1	Topics Covered	247
13.2	Spatial Randomness	248
13.2.1	Parametric approach	248
13.2.2	Randomization	249
13.2.3	Simulating spatial randomness	250
13.3	Positive and negative spatial autocorrelation	250
13.3.1	Positive spatial autocorrelation	250
13.3.2	Negative Spatial Autocorrelation	251
13.4	Spatial Autocorrelation Statistic	252
13.4.1	Attribute similarity	252
13.4.2	Locational similarity	253
13.4.3	Examples of spatial autocorrelation statistics	253
13.5	Moran’s I	254

13.5.1	Statistic	254
13.5.2	Inference	255
13.5.3	Interpretation	257
13.5.4	Moran scatter plot	259
13.6	Visualizing spatial autocorrelation with the Moran scatter plot	261
13.6.1	Creating a Moran scatter plot	261
13.6.2	Moran scatter plot options	262
13.6.3	Brushing the Moran scatter plot	264
14	Advanced Global Spatial Autocorrelation	267
14.1	Topics Covered	267
14.2	Specialized Moran Scatter Plots	267
14.2.1	Differential Moran scatter plot	268
14.2.2	Moran scatter plot for EB rates	271
14.3	Bivariate Spatial Correlation	273
14.3.1	Bivariate Moran scatter plot	273
14.4	The Many Faces of Space-Time Correlation	276
14.4.1	Serial (temporal) correlation	276
14.4.2	Serial (temporal) correlation between spatial lags	276
14.4.3	Space-time regression	277
14.4.4	Serial and space-time regression	279
15	Nonparametric Spatial Autocorrelation	281
15.1	Topics Covered	281
15.2	Non-Parametric Approaches	282
15.3	Spatial Correlogram	282
15.3.1	Creating a spatial correlogram	283
15.3.2	Spatial correlogram options	285
15.4	Smoothed Distance Scatter Plot	288
15.4.1	Multivariate extension	290
15.4.2	Creating a smoothed distance scatter plot	290
15.4.3	Smoothed distance scatter plot options	291
V	Local Spatial Autocorrelation	295
16	LISA and Local Moran	297
16.1	Topics Covered	297
16.2	LISA Principle	298
16.3	Local Moran	298
16.3.1	Formulation	299
16.3.2	Implementation	300
16.4	Clusters and Spatial Outliers	302
16.4.1	Clusters	302
16.4.2	Spatial outliers	304
16.5	Significance and Interpretation	305
16.5.1	Multiple comparisons	306
16.5.2	Interpretation of significance	309
16.5.3	Interpretation of clusters	310
16.6	Conditional Local Cluster Maps	311

17 Other Local Spatial Autocorrelation Statistics	313
17.1 Topics Covered	313
17.2 Extensions of the Local Moran	314
17.2.1 Median Local Moran	314
17.2.2 Differential Local Moran	318
17.2.3 Local Moran with EB Rate	320
17.3 Local Geary	322
17.3.1 Implementation	323
17.3.2 Clusters and spatial outliers	325
17.3.3 Comparing Local Geary and Local Moran	326
17.4 Getis-Ord Statistics	329
17.4.1 Implementation	329
17.4.2 Clusters and outliers	330
17.4.3 Comparing G statistics and Local Moran	331
17.5 Which Local Statistic to Use?	333
18 Multivariate Local Spatial Autocorrelation	335
18.1 Topics Covered	335
18.2 The Multivariate Spatial Autocorrelation Problem	335
18.3 Bivariate Local Moran	336
18.3.1 Implementation	337
18.3.2 Interpretation	340
18.4 Multivariate Local Geary	340
18.4.1 Implementation	342
18.4.2 Interpretation	344
18.5 Local Neighbor Match Test	345
18.5.1 Implementation	346
18.5.2 Local neighbor match and Multivariate Local Geary	349
19 LISA for Discrete Variables	351
19.1 Topics Covered	351
19.2 Univariate Local Join Count Statistic	351
19.2.1 Implementation	353
19.3 Bivariate Local Join Count Statistic	355
19.3.1 Implementation	356
19.4 Co-Location Local Join Count Statistic	357
19.4.1 Implementation	357
19.5 Quantile LISA	358
19.5.1 Implementation	359
19.5.2 Univariate Quantile LISA	360
19.5.3 Bivariate and Multivariate Quantile LISA	360
20 Density-Based Clustering Methods	363
20.1 Topics Covered	364
20.2 Heat Map	364
20.2.1 Implementation	364
20.3 DBSCAN	366
20.3.1 Important concepts	366
20.3.2 DBSCAN algorithm	368
20.3.3 Implementation	370
20.4 DBSCAN*	373

20.4.1	Mutual reachability distance	373
20.4.2	Implementation	375
20.5	HDBSCAN	376
20.5.1	Important concepts	377
20.5.2	HDBSCAN algorithm	379
20.5.3	Cluster membership	382
20.5.4	Outliers	382
20.5.5	Implementation	383
VI	Epilogue	389
21	Postscript – The Limits of Exploration	391
A	Appendix A – GeoDa Preference Settings	395
B	Appendix B – Menu Structure	399
C	Appendix C – Scripting with GeoDa via the geodalib Library	403
	Bibliography	405
	Index	417

List of Figures

1.1	GeoDa toolbar icons	4
1.2	Data entry	4
1.3	Data manipulation/table	5
1.4	GIS operations/tools	5
1.5	Weights manager	5
1.6	Mapping and geovisualization	5
1.7	Exploratory data analysis	6
1.8	Space-time analysis	6
1.9	Spatial autocorrelation analysis	6
1.10	Cluster analysis	7
2.1	Open Close Save Table Tools	12
2.2	Connect to Data Source dialog	14
2.3	Supported spatial file formats	14
2.4	Themeless polygon map	15
2.5	Example GeoJSON file contents	15
2.6	Example GML file contents	16
2.7	Themeless point map	17
2.8	CSV format file input dialog	18
2.9	Coordinate variables in data table	19
2.10	Specifying the point coordinates	19
2.11	Point layer from coordinates	20
2.12	Blank CRS	20
2.13	CRS with projection information	21
2.14	Grid creation dialog	21
2.15	Grid layer over Chicago community areas	22
2.16	Table options	23
2.17	Edit variable properties	24
2.18	Create a date/time variable	25
2.19	Date/time format	25
2.20	Customizing the date/time format	26
2.21	Calculator interface	27
2.22	Merge dialog	29
2.23	Merged tables	30
2.24	Selection Tool	31
2.25	Selected observations in themeless map	32
2.26	Selected observations in new map	32
2.27	Car thefts by month	33
2.28	Selection on the map	34
2.29	Selection on a map category	35
3.1	UTM zones for North America (source: GISGeography)	40

3.2	Load CRS from another layer	41
3.3	New CRS from another layer	41
3.4	Reprojected Chicago community areas layer	42
3.5	Display mean centers on map	43
3.6	Add mean centers to table	43
3.7	Chicago community area mean center layer	44
3.8	Display Thiessen polygons	45
3.9	Connectivity graph of mean centers	46
3.10	Minimum spanning tree for connectivity graph of mean centers	47
3.11	Toy example point coordinates	47
3.12	Inter-point distance matrix	48
3.13	Connectivity graph	48
3.14	Initial steps in Prim's algorithm	48
3.15	Minimum spanning tree	49
3.16	Dissolve dialog	50
3.17	Dissolved districts map	51
3.18	Dissolved districts table	51
3.19	Map layer settings	52
3.20	Polygon layer on top with opacity=0	53
3.21	Points reprojected to lat-lon	54
3.22	Select on current map layer (community areas)	55
3.23	Spatial join dialog – areal unit identifier for point	56
3.24	Community area identifier (SA) in point table	56
3.25	Point in polygon mismatch	57
3.26	Spatial join dialog – aggregation options	57
3.27	Spatial count in polygon area table	58
3.28	Set association dialog	59
3.29	Linked selection	59
4.1	Maps and Rates Cartogram Map Movie Category Editor	64
4.2	Mapping options	68
4.3	Quintile map for housing index, Ceará	69
4.4	Tied values in a quantile map	70
4.5	Equal intervals map for housing index, Ceará	71
4.6	Equal intervals map and histogram	72
4.7	Natural breaks map for housing index, Ceará	73
4.8	Map options	74
4.9	Legend options	74
4.10	Map toolbar	75
4.11	Map base layer options	76
4.12	Stamen TonerLight base map, default map transparency	76
4.13	Stamen TonerLight base map, 0.40 map transparency	77
4.14	ESRI WorldStreetMap base map, 0.40 map transparency	78
4.15	Saved Image As dialog	79
4.16	Map without outlines visible	80
4.17	Category editor interface	82
4.18	Quintile maps for sanitation and infrastructure	84
4.19	Thematic maps for sanitation and infrastructure using custom classification	84
4.20	Custom category definition in project file	85
5.1	Percentile map	88

5.2	Box map, hinge = 1.5, with box plot	89
5.3	Outliers in box plot and box map	90
5.4	Standard deviation map	91
5.5	Unique values map, Zika indicator	92
5.6	Co-location map variable selection, Zika and Microcephaly indicators	93
5.7	Co-location map, Zika and Microcephaly incidence	94
5.8	Selection of co-located observations	95
5.9	Cartogram	96
5.10	Linked Cartogram and map	97
5.11	Animation control panel	98
5.12	Linked animation sanitation-infrastructure	99
6.1	Population and population density	103
6.2	GDP and GDP per capita	103
6.3	Rate map interface	104
6.4	GDP per capita as a rate map	104
6.5	Rate calculation in table	105
6.6	GDP per capita in table	105
6.7	GDP per capita excess risk	107
6.8	GDP per capita excess risk in table	108
6.9	Crude and EB rate map	112
6.10	Effect of smoothing – small populations	113
6.11	Effect of smoothing – large populations	113
7.1	Histogram Box Plot Scatter Plot Scatter Plot Matrix	116
7.2	Histogram for percent poverty 2020	117
7.3	Histogram for percent poverty 2020 – 12 bins	117
7.4	Bar chart for settlement categories	118
7.5	Descriptive statistics	119
7.6	Linking between histogram and map	120
7.7	Linking between map and histogram	120
7.8	Box plot for population change 2020–2010	121
7.9	Box map for population change 2020–2010 (upper outliers selected)	122
7.10	Scatter plot of food insecurity on poverty	124
7.11	Regime regression	125
7.12	Correlation	126
7.13	Default LOWESS local regression fit	127
7.14	Default LOWESS local regression fit with bandwidth 0.6	128
7.15	Default LOWESS local regression fit with bandwidth 0.05	129
7.16	Scatter plot matrix	131
7.17	Scatter plot correlation matrix	132
7.18	Scatter plot matrix with LOWESS fit	133
7.19	Food insecurity 2020 in Valles Centrales	134
7.20	Averages Chart – Valles Centrales	135
7.21	Map brushing and the averages chart – 1	136
7.22	Map brushing and the averages chart – 2	136
7.23	Map brushing and the averages chart – 3	137
7.24	Map brushing and the scatter plot – 1	138
7.25	Map brushing and the scatter plot – 2	138
7.26	Map brushing and the scatter plot – 3	139

8.1	Bubble Chart 3D Scatter Plot Parallel Coordinate Plot Conditional Plot	142
8.2	One dimension	143
8.3	Two dimensions	143
8.4	Discrete evaluation points in three variable dimensions	143
8.5	Two dimensions	144
8.6	Three dimensions	144
8.7	Bubble chart – default settings: education, basic services, extreme poverty	145
8.8	Bubble chart – bubble size adjusted: education, basic services, extreme poverty	146
8.9	Bubble chart – no color: education, basic services, extreme poverty	147
8.10	Bubble chart – categories: education, basic services, region	147
8.11	3D scatter plot: education, basic services, extreme poverty	148
8.12	Interacting with 3D scatter plot	149
8.13	Selection in the 3D scatter plot	150
8.14	Brushing and linking with the 3D scatter plot	151
8.15	Conditional scatter plot – 3 by 3	152
8.16	Conditional scatter plot – unique values	153
8.17	Conditional box plot – lack of health access by region	154
8.18	Conditional box map – 2 by 2	155
8.19	Parallel coordinate plot	157
8.20	Parallel coordinate plot – standardized variables	157
8.21	Brushing the PCP	158
8.22	Brushing map and PCP	158
8.23	Clusters of observations in PCP	159
8.24	Outlier observation in PCP	159
9.1	Time Averages Chart	162
9.2	Time Editor interface	163
9.3	Grouped variables in data table	163
9.4	Grouped variables in project file	165
9.5	Time Player dialog	166
9.6	Box plot over time	167
9.7	Box plot for two time periods	168
9.8	Time lagged scatter plot variables	169
9.9	Scatter plots with time lagged variables	169
9.10	Thematic maps for car ownership in 2000	170
9.11	Thematic maps for car ownership in 2010	170
9.12	Thematic maps for car ownership in 2020	170
9.13	Selected municipalities with altitude in two upper quantiles	171
9.14	Difference in means, selected and unselected, static, in 2000 and 2020	172
9.15	Difference in difference setup, 2000 to 2020	173
9.16	Difference in difference results	174
9.17	Difference in means, unselected, dynamic, 2000 to 2020	174
10.1	Weights Manager	180
10.2	Example spatial layout	181
10.3	Contiguity as a network	182
10.4	Regular grid layout	184
10.5	Rook contiguity	184
10.6	Queen contiguity	184
10.7	Second-order contiguity	186

10.8	Higher order contiguity as multiple steps in the graph	186
10.9	Weights Manager interface	188
10.10	Rook contiguity in the Weights File Creation interface	189
10.11	Rook weights summary interface	190
10.12	GAL file contents	190
10.13	Summary properties of first-order queen contiguity	191
10.14	Summary properties of second-order queen contiguity	193
10.15	Summary properties of inclusive second-order queen contiguity	193
10.16	Weights description in Project File	194
10.17	Second-order contiguity weights loaded from file	195
10.18	Saved space-time table as csv file	195
10.19	Space-time GAL contiguity file	196
10.20	Rook contiguity histogram	197
10.21	Rook and queen contiguity histograms	197
10.22	Linked contiguity histogram and map	198
10.23	Connectivity map for queen contiguity	199
10.24	Connectivity graph for queen contiguity	200
10.25	Connectivity graph for queen contiguity without map	200
11.1	Nearest neighbor properties	207
11.2	Italian community bank point layer	209
11.3	Distance-based weights in the Weights File Creation interface	210
11.4	Default distance-based weights summary properties	210
11.5	GWT weights file format	211
11.6	Connectivity histogram for default distance weights	212
11.7	Connectivity graph for default distance weights	213
11.8	Weights summary properties for distance 73 km	214
11.9	Connectivity graph for distance band 73 km	215
11.10	Weights summary properties for 6 k-nearest neighbors	216
11.11	Connectivity graph for 6 k-nearest neighbors	217
11.12	Contiguity for points from Thiessen polygons	218
11.13	Weights summary properties for point contiguity (Thiessen polygons)	218
11.14	Connectivity histogram for point contiguity (Thiessen polygons)	219
11.15	Connectivity graph for point contiguity (Thiessen polygons)	220
11.16	Block weights in the Weights File Creation interface	221
11.17	Connectivity graph for block weights by region	221
11.18	Space-time weights GAL file	222
11.19	Connectivity graph for intersection distance 73 km and knn 6	224
12.1	Summary properties of inverse distance weights	228
12.2	GWT files with distance for original and rescaled coordinates	229
12.3	Kernel weights in Weights File Creation interface	231
12.4	Properties of kernel weights	232
12.5	KWT file	233
12.6	Spatial lag in calculator	234
12.7	Spatially lagged variables added to table	235
12.8	Variables computed with inverse distance weights added to table	237
12.9	Disability rate: raw rate and EB smoothed rate	239
12.10	Spatial rate map: first- and second-order queen contiguity	241
12.11	Spatial rate map: k-nearest neighbor and inverse distance weights	242
12.12	Spatial rate map: inverse distance weights from Calculator	242

12.13	Spatial Empirical Bayes smoothed rate map using region block weights . . .	243
13.1	Moran Scatter Plot Spatial Correlogram Cluster Maps	248
13.2	Spatially random observations – i.i.d	249
13.3	Spatially random observations – true and randomized per capita income .	249
13.4	True and randomized per capita income – histograms	250
13.5	Positive spatial autocorrelation	251
13.6	Negative spatial autocorrelation	251
13.7	Reference distribution, Moran’s I for per capita income	257
13.8	Reference distribution, Moran’s I for randomized per capita income	258
13.9	Moran scatter plot, community area per capita income	259
13.10	Positive spatial autocorrelation, high-high	260
13.11	Negative spatial autocorrelation, low-high	261
13.12	Moran scatter plot LOWESS smoother	263
13.13	Brushing the Moran scatter plot	264
13.14	Brushing and linking the Moran scatter plot	265
13.15	Static spatial weights in Moran scatter plot	265
14.1	Access to health care in 2010 and 2020	269
14.2	Change in access to health care between 2010 and 2020	269
14.3	Moran scatter plot for access to health care in 2010 and 2020	270
14.4	Differential Moran scatter plot for access to health care 2020–2010	271
14.5	Moran scatter plot for raw rate and EB Moran scatter plot	273
14.6	Bivariate Moran scatter plot for access to health care 2020 and its spatial lag in 2010	275
14.7	Correlation between access to health care in 2020 and 2010	277
14.8	Correlation between the spatial lag of access to health care in 2020 and the spatial lag in 2010	278
14.9	Correlation between access to health care in 2020 and its spatial lag in 2010	278
14.10	OLS regression of pPHA 2020 on pPHA 2010 and W PHA 2010	279
15.1	Moran Scatter Plot Spatial Correlogram Cluster Maps	282
15.2	Spatial distribution of loan loss provisions – Italian banks 2016	283
15.3	Correlogram parameter selection	284
15.4	Default spatial correlogram – LLP (2016)	285
15.5	Spatial correlogram – LLP, 150 km cut off, 10 bins	286
15.6	Spatial correlogram – LLP, 150 km cut off, 15 bins	287
15.7	Smoothed distance scatter plot	289
15.8	Smoothed distance scatter plot with 150 km distance cut-off	291
15.9	Smoothed distance scatter plot with 150 km distance cut-off and customized Y-Axis	292
15.10	Smoothed distance scatter plot with 150 km distance cut-off and span 0.5	293
15.11	Smoothed distance scatter plot with 150 km distance cut-off, span 0.5, linear fit	293
15.12	Bivariate smoothed distance scatter plot	294
16.1	Moran Scatter Plot Spatial Correlogram Cluster Maps	298
16.2	Significance map and cluster map	300
16.3	Significance map and cluster map – 99999 permutations	301
16.4	High-High observations in Moran scatter plot	303
16.5	Significant High-High observations in Moran scatter plot	303

16.6 High-High cluster location 304

16.7 Significant High-Low observations in Moran scatter plot 305

16.8 High-Low spatial outlier location 305

16.9 Significance map – $p < 0.01$ 306

16.10 Cluster map – $p < 0.01$ 306

16.11 Cluster map – Bonferroni bound 308

16.12 False Discovery Rate calculation 309

16.13 Cluster map – FDR 309

16.14 Cores and neighbors 310

16.15 Conditional cluster map 311

17.1 Significance maps conventional and Median Local Moran 315

17.2 High-High clusters in Median Local Moran 316

17.3 Low-Low clusters in Median Local Moran 316

17.4 Low-High spatial outliers in Median Local Moran 317

17.5 High-Low spatial outliers in Median Local Moran 317

17.6 Cluster maps conventional and Median Local Moran 317

17.7 Cluster maps for Local Moran in 2010 and 2020 319

17.8 Differential Local Moran significance and cluster maps 319

17.9 Significance maps – raw rate and EB rate 321

17.10 Cluster maps – raw rate and EB rate 321

17.11 Local Geary significance and cluster maps 323

17.12 Local Geary – High-High clusters 324

17.13 Local Geary – Low-Low clusters 324

17.14 Local Geary – Other clusters 324

17.15 Local Geary – Spatial outliers 325

17.16 Local Geary and Local Moran cluster maps 326

17.17 Local Geary and Local Moran – High-High clusters 327

17.18 Local Geary and Local Moran – Low-Low clusters 327

17.19 Local Geary and Local Moran – Other positive clusters 327

17.20 Local Geary and Local Moran – Spatial outliers 328

17.21 Local Moran and Local Geary – Spatial outliers 328

17.22 G_i^* significance map and cluster map 330

17.23 G_i^* High-High clusters 331

17.24 G_i^* Low-Low clusters 331

17.25 G_i^* and Local Moran High-High clusters 332

17.26 G_i^* and Local Moran Low-Low clusters 332

17.27 G_i^* and Local Moran Low-High spatial outliers 333

17.28 G_i^* and Local Moran High-Low spatial outliers 333

18.1 Box maps – Child Poverty and Crowded Housing 338

18.2 Co-location of Child Poverty and Crowded Housing 338

18.3 Local Moran cluster map – Child Poverty and Crowded Housing 339

18.4 Co-location of Local Moran for Child Poverty and Crowded Housing 339

18.5 Bivariate Local Moran cluster map – Child Poverty and Crowded Housing 340

18.6 Box map and Local Geary cluster map – Uninsured 342

18.7 Local Geary cluster maps – Child Poverty, Crowded Housing 343

18.8 Co-location of Local Geary 343

18.9 Multivariate Local Geary cluster map – Child Poverty, Crowded Housing,
Uninsured 344

18.10 Multivariate Local Geary cluster map and Univariate Local Geary co-location 344

18.11	Multivariate Local Geary cluster in PCP	345
18.12	K nearest neighbor connectivity graphs	346
18.13	Intersection of K nearest neighbor connectivity graphs	347
18.14	Local neighbor match cardinality map	348
18.15	Local neighbor match significance map	348
18.16	Local neighbor significant matches	349
18.17	Local neighbor match and Multivariate Local Geary	349
19.1	Unique values Black-Hispanic tracts	353
19.2	Local join count significance map	354
19.3	Bivariate local join count significance map	356
19.4	Local join count spatial outlier	356
19.5	Co-location local join count	358
19.6	Univariate Quantile LISA	360
19.7	Spatial outliers with Quantile LISA	361
19.8	Multivariate Quantile LISA and Multivariate Local Geary	362
20.1	Clusters > DBSCAN HDBSCAN	364
20.2	Default heat map	365
20.3	Heat map with bandwidth 73 km	366
20.4	DBSCAN Core, Border and Noise points	367
20.5	Connectivity graph for Eps = 20	368
20.6	DBSCAN clusters for Eps = 20 and MinPts = 4	368
20.7	DBSCAN clustering settings	370
20.8	Default DBSCAN cluster map	371
20.9	DBSCAN cluster map, Eps=50, Min Pts=10	372
20.10	DBSCAN-Star dendrogram, Eps=50, Min Pts=10	374
20.11	DBSCAN-Star cluster map, Eps=50, Min Pts=10	375
20.12	Exploring the dendrogram	376
20.13	Level set	377
20.14	Core distance and lambda for each point	379
20.15	Minimum spanning tree connectivity graph	380
20.16	Pruning the MST	381
20.17	Cluster stability	381
20.18	HDBSCAN cluster map	384
20.19	HDBSCAN dendrogram	384
20.20	HDBSCAN condensed tree	385
20.21	HDBSCAN cluster membership	385
20.22	HDBSCAN core distance heat map	386
A.1	Preferences settings – System	396
A.2	Preferences settings – Data	397
C.1	libgeoda architecture	404

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



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

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.



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

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 Brunson, 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 (Brunson 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

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 **geodolib** 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



Taylor & Francis

Taylor & Francis Group

<http://taylorandfrancis.com>

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

- 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

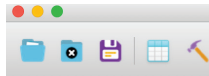


Figure 2.1: Open | Close | Save | Table | Tools

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 library² 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* format⁴ 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>

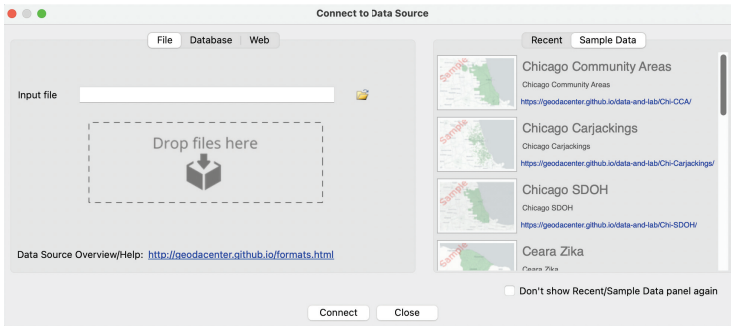


Figure 2.2: Connect to Data Source dialog

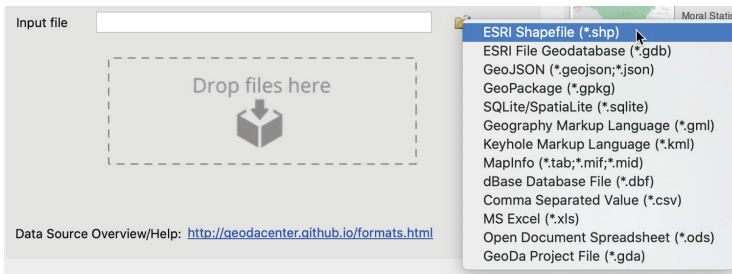


Figure 2.3: Supported spatial file formats

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).

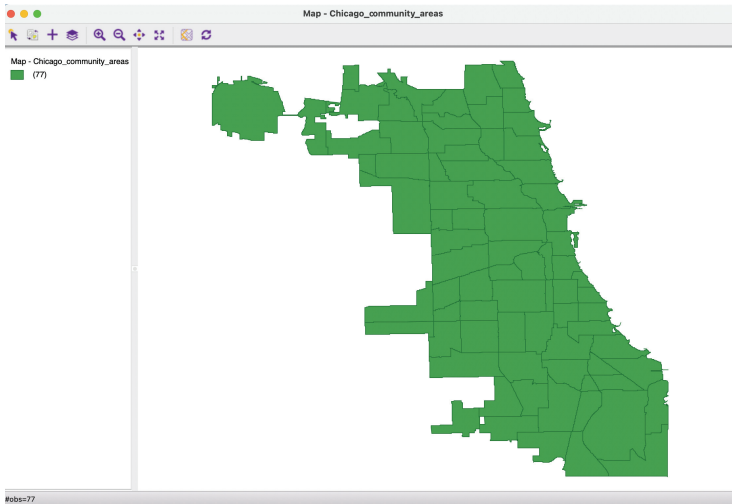


Figure 2.4: Themeless polygon map

```
{
  "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>D0UGLAS</ogr:community>
  <ogr:shape_area>46004621.1581000015139579772949218750000</ogr:shape_area>
  <ogr:shape_len>31027.054509800002326443791389465332</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.

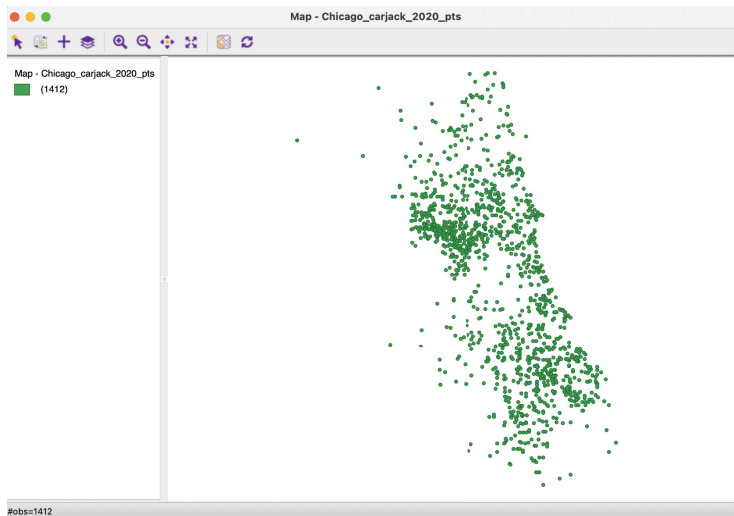


Figure 2.7: Themeless point map

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