

Spatial Data Analysis

Theory and Practice

Spatial Data Analysis: Theory and Practice provides a broad-ranging treatment of the field of spatial data analysis. It begins with an overview of spatial data analysis and the importance of location (place, context and space) in scientific and policy-related research. Covering fundamental problems concerning how attributes in geographical space are represented to the latest methods of exploratory spatial data analysis and spatial modelling, it is designed to take the reader through the key areas that underpin the analysis of spatial data, providing a platform from which to view and critically appreciate many of the key areas of the field. Parts of the text are accessible to undergraduate and master's level students, but it also contains sufficient challenging material that it will be of interest to geographers, social scientists and economists, environmental scientists and statisticians, whose research takes them into the area of spatial analysis.

ROBERT HAINING is Professor of Human Geography at the University of Cambridge. He has published extensively in the field of spatial data analysis, with particular reference to applications in the areas of economic geography, medical geography and the geography of crime. His previous book, *Spatial Data Analysis in the Social and Environmental Sciences* (Cambridge University Press, 1993) was well received and cited internationally.

Cambridge University Press

0521773199 - Spatial Data Analysis: Theory and Practice - Robert Haining

Frontmatter

[More information](#)

Spatial Data Analysis

Theory and Practice

ROBERT HAINING
University of Cambridge



Cambridge University Press
0521773199 - Spatial Data Analysis: Theory and Practice - Robert Haining
Frontmatter
[More information](#)

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa
<http://www.cambridge.org>

© Robert Haining 2003

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2003

Printed in the United Kingdom at the University Press, Cambridge

Typeface Lexicon No. 2 10/14 pt and Lexicon No. 1 *System L^AT_EX 2_ε* [TB]

A catalogue record for this book is available from the British Library

Library of Congress Cataloguing in Publication data

Haining, Robert P.

Spatial data analysis: theory and practice / Robert Haining.

p. cm.

Includes bibliographical references and index.

ISBN 0-521-77319-9 – ISBN 0-521-77437-3 (pbk.)

1. Spatial analysis (Statistics) 2. Geology – Statistical methods – Data processing. I. Title.

QA278.2 .H345 2003

001.4'22–dc 2002031242

ISBN 0 521 77437 3 paperback

The publisher has used its best endeavours to ensure that the URLs for external websites referred to in this book are correct and active at the time of going to press. However, the publisher has no responsibility for the websites and can make no guarantee that a site will remain live or that the content is or will remain appropriate.

To my wife, Rachel, and our children,
Celia, Sarah and Mark

Contents

Preface xv
Acknowledgements xvii

Introduction 1

- 0.1 About the book 1
- 0.2 What is spatial data analysis? 4
- 0.3 Motivation for the book 5
- 0.4 Organization 8
- 0.5 The spatial data matrix 10

Part A The context for spatial data analysis

1 Spatial data analysis: scientific and policy context 15

- 1.1 Spatial data analysis in science 15
 - 1.1.1 Generic issues of place, context and space in scientific explanation 16
 - (a) Location as place and context 16
 - (b) Location and spatial relationships 18
 - 1.1.2 Spatial processes 21
- 1.2 Place and space in specific areas of scientific explanation 22
 - 1.2.1 Defining spatial subdisciplines 22
 - 1.2.2 Examples: selected research areas 24
 - (a) Environmental criminology 24
 - (b) Geographical and environmental (spatial) epidemiology 26
 - (c) Regional economics and the new economic geography 29

viii Contents

- (d) Urban studies 31
 - (e) Environmental sciences 32
 - 1.2.3 Spatial data analysis in problem solving 33
- 1.3 Spatial data analysis in the policy area 36
- 1.4 Some examples of problems that arise in analysing spatial data 40
 - 1.4.1 Description and map interpretation 40
 - 1.4.2 Information redundancy 41
 - 1.4.3 Modelling 41
- 1.5 Concluding remarks 41
- 2 The nature of spatial data 43**
 - 2.1 The spatial data matrix: conceptualization and representation issues 44
 - 2.1.1 Geographic space: objects, fields and geometric representations 44
 - 2.1.2 Geographic space: spatial dependence in attribute values 46
 - 2.1.3 Variables 47
 - (a) Classifying variables 48
 - (b) Levels of measurement 50
 - 2.1.4 Sample or population? 51
 - 2.2 The spatial data matrix: its form 54
 - 2.3 The spatial data matrix: its quality 57
 - 2.3.1 Model quality 58
 - (a) Attribute representation 59
 - (b) Spatial representation: general considerations 59
 - (c) Spatial representation: resolution and aggregation 61
 - 2.3.2 Data quality 61
 - (a) Accuracy 63
 - (b) Resolution 67
 - (c) Consistency 70
 - (d) Completeness 71
 - 2.4 Quantifying spatial dependence 74
 - (a) Fields: data from two-dimensional continuous space 74
 - (b) Objects: data from two-dimensional discrete space 79
 - 2.5 Concluding remarks 87

Part B **Spatial data: obtaining data and quality issues**

3 Obtaining spatial data through sampling 91

3.1 Sources of spatial data 91

3.2 Spatial sampling 93

3.2.1 The purpose and conduct of spatial sampling 93

3.2.2 Design- and model-based approaches to spatial sampling 96

(a) Design-based approach to sampling 96

(b) Model-based approach to sampling 98

(c) Comparative comments 99

3.2.3 Sampling plans 100

3.2.4 Selected sampling problems 103

(a) Design-based estimation of the population mean 103

(b) Model-based estimation of means 106

(c) Spatial prediction 107

(d) Sampling to identify extreme values or detect rare events 108

3.3 Maps through simulation 113

4 Data quality: implications for spatial data analysis 116

4.1 Errors in data and spatial data analysis 116

4.1.1 Models for measurement error 116

(a) Independent error models 117

(b) Spatially correlated error models 118

4.1.2 Gross errors 119

(a) Distributional outliers 119

(b) Spatial outliers 122

(c) Testing for outliers in large data sets 123

4.1.3 Error propagation 124

4.2 Data resolution and spatial data analysis 127

4.2.1 Variable precision and tests of significance 128

4.2.2 The change of support problem 129

(a) Change of support in geostatistics 129

(b) Areal interpolation 131

4.2.3 Analysing relationships using aggregate data 138

(a) Ecological inference: parameter estimation 141

(b) Ecological inference in environmental epidemiology: identifying valid hypotheses 147

(c) The modifiable areal units problem (MAUP) 150

x Contents

- 4.3 Data consistency and spatial data analysis 151
- 4.4 Data completeness and spatial data analysis 152
 - 4.4.1 The missing-data problem 154
 - (a) Approaches to analysis when data are missing 156
 - (b) Approaches to analysis when spatial data are missing 159
 - 4.4.2 Spatial interpolation, spatial prediction 164
 - 4.4.3 Boundaries, weights matrices and data completeness 174
- 4.5 Concluding remarks 177

Part C **The exploratory analysis of spatial data**

- 5 Exploratory spatial data analysis: conceptual models 181**
 - 5.1 EDA and ESDA 181
 - 5.2 Conceptual models of spatial variation 183
 - (a) The regional model 183
 - (b) Spatial 'rough' and 'smooth' 184
 - (c) Scales of spatial variation 185
- 6 Exploratory spatial data analysis: visualization methods 188**
 - 6.1 Data visualization and exploratory data analysis 188
 - 6.1.1 Data visualization: approaches and tasks 189
 - 6.1.2 Data visualization: developments through computers 192
 - 6.1.3 Data visualization: selected techniques 193
 - 6.2 Visualizing spatial data 194
 - 6.2.1 Data preparation issues for aggregated data: variable values 194
 - 6.2.2 Data preparation issues for aggregated data: the spatial framework 199
 - (a) Non-spatial approaches to region building 200
 - (b) Spatial approaches to region building 201
 - (c) Design criteria for region building 203
 - 6.2.3 Special issues in the visualization of spatial data 206
 - 6.3 Data visualization and exploratory spatial data analysis 210
 - 6.3.1 Spatial data visualization: selected techniques for univariate data 211
 - (a) Methods for data associated with point or area objects 211
 - (b) Methods for data from a continuous surface 215
 - 6.3.2 Spatial data visualization: selected techniques for bi- and multi-variate data 218

6.3.3	Uptake of breast cancer screening in Sheffield	219
6.4	Concluding remarks	225
7	Exploratory spatial data analysis: numerical methods	226
7.1	Smoothing methods	227
7.1.1	Resistant smoothing of graph plots	227
7.1.2	Resistant description of spatial dependencies	228
7.1.3	Map smoothing	228
	(a) Simple mean and median smoothers	230
	(b) Introducing distance weighting	230
	(c) Smoothing rates	232
	(d) Non-linear smoothing: headbanging	234
	(e) Non-linear smoothing: median polishing	236
	(f) Some comparative examples	237
7.2	The exploratory identification of global map properties: overall clustering	237
7.2.1	Clustering in area data	242
7.2.2	Clustering in a marked point pattern	247
7.3	The exploratory identification of local map properties	250
7.3.1	Cluster detection	251
	(a) Area data	251
	(b) Inhomogeneous point data	259
7.3.2	Focused tests	263
7.4	Map comparison	265
	(a) Bivariate association	265
	(b) Spatial association	268
Part D	Hypothesis testing and spatial autocorrelation	
8	Hypothesis testing in the presence of spatial dependence	273
8.1	Spatial autocorrelation and testing the mean of a spatial data set	275
8.2	Spatial autocorrelation and tests of bivariate association	278
8.2.1	Pearson's product moment correlation coefficient	278
8.2.2	Chi-square tests for contingency tables	283
Part E	Modelling spatial data	
9	Models for the statistical analysis of spatial data	289
9.1	Descriptive models	292
9.1.1	Models for large-scale spatial variation	293

xii Contents

9.1.2	Models for small-scale spatial variation	293
	(a) Models for data from a surface	293
	(b) Models for continuous-valued area data	297
	(c) Models for discrete-valued area data	304
9.1.3	Models with several scales of spatial variation	306
9.1.4	Hierarchical Bayesian models	307
9.2	Explanatory models	312
9.2.1	Models for continuous-valued response variables: normal regression models	312
9.2.2	Models for discrete-valued area data: generalized linear models	316
9.2.3	Hierarchical models	
	(a) Adding covariates to hierarchical Bayesian models	320
	(b) Modelling spatial context: multi-level models	321
10	Statistical modelling of spatial variation: descriptive modelling	325
10.1	Models for representing spatial variation	325
10.1.1	Models for continuous-valued variables	326
	(a) Trend surface models with independent errors	326
	(b) Semi-variogram and covariance models	327
	(c) Trend surface models with spatially correlated errors	331
10.1.2	Models for discrete-valued variables	334
10.2	Some general problems in modelling spatial variation	338
10.3	Hierarchical Bayesian models	339
11	Statistical modelling of spatial variation: explanatory modelling	350
11.1	Methodologies for spatial data modelling	350
11.1.1	The 'classical' approach	350
11.1.2	The econometric approach	353
	(a) A general spatial specification	355
	(b) Two models of spatial pricing	356
11.1.3	A 'data-driven' methodology	358
11.2	Some applications of linear modelling of spatial data	358
11.2.1	Testing for regional income convergence	359
11.2.2	Models for binary responses	361
	(a) A logistic model with spatial lags on the covariates	361
	(b) Autologistic models with covariates	364
11.2.3	Multi-level modelling	365

11.2.4 Bayesian modelling of burglaries in Sheffield	367
11.2.5 Bayesian modelling of children excluded from school	376
11.3 Concluding comments	378
Appendix I Software	379
Appendix II Cambridgeshire lung cancer data	381
Appendix III Sheffield burglary data	385
Appendix IV Children excluded from school: Sheffield	391
<i>References</i>	<i>394</i>
<i>Index</i>	<i>424</i>

Preface

Interest in analysing spatial data has grown considerably in the scientific research community. This reflects the existence of well-formulated questions or hypothesis in which location plays a role, of spatial data of sufficient quality, of appropriate statistical methodology.

In writing this book I have drawn on a number of scientific and also policy-related fields to illustrate the scale of interest – actual and potential – in analysing spatial data. In seeking to provide this overview of the field I have given a prominent place to two fields of research: Geographic Information Science (GISc) and applied spatial statistics.

It is important as part of the process of understanding the results of spatial data analysis to define the relationship between geographic reality and how that reality is captured in a digital database in the form of a data matrix containing both attribute data and data on locations. The usefulness of operations on that data matrix – revising or improving an initial representation (e.g. spatial smoothing), testing hypotheses (e.g. does this map pattern contain spatial clusters of events?) or fitting models (e.g. to explain offence patterns or health outcomes in terms of socio-economic covariates) – will depend on how well the reality that is being represented has been captured in the data matrix. Awareness of this link is important and insights can be drawn from the GISc literature.

I have drawn on developments in spatial statistics which can be applied to data collected from continuous surfaces and from regions partitioned into sub-areas (e.g. a city divided into wards or enumeration districts). In covering this material I have attempted to draw out the important ideas whilst directing the reader to specialist sources and original papers. This book is not an exhaustive treatment of all areas of spatial statistics (it does not cover point processes), nor of all areas of spatial analysis (it does not include cartographic modelling).

Implementing a programme of spatial data analysis is greatly assisted if supporting software is available. Geographic information systems (GIS) software are now widely used to handle spatial data and there is a growing quantity of software some of it linked to GIS for implementing spatial statistical methods. The appendix directs the reader to some relevant software.

Readership

This book brings together techniques and models for analysing spatial data in a way that I hope is accessible to a wide readership, whilst still being of interest to the research community.

Parts of this book have been tried out on year 2 geography undergraduates at the University of Cambridge in an eight-hour lecture course that introduced them to certain areas of geographic information science and methods of spatial analysis. The parts used are chapters 1, 2, sections 3.1, 3.2.1, 3.2.3, 3.2.4(a) from chapter 3, selected sections from chapter 4 (e.g. detecting errors and outliers, areal interpolation problems), selected sections from chapter 7 (section 7.1.3, map smoothing) and some selected examples on modelling and mapping output using the normal linear regression model. In associated practicals simple methods for hot spot detection are applied (the first part of section 7.3.1(a)) together with logistic regression for modelling (along the lines of section 11.2.2(a)).

Parts of the book have been tried out on postgraduate students on a one year M.Phil. in Geographic Information Systems and Remote Sensing at Cambridge. One 16-hour course was on general methods of spatial analysis but particularly for data from continuous surfaces. In addition to some of the foundation material covered in chapters 1 to 4 there was an extended treatment of the material in section 4.4.2 with particular reference to kriging with Gaussian data (including estimation and modelling of the semi-variogram taken from chapter 10 and the references therein). A second 16-hour course dealt with exploratory spatial data analysis and spatial modelling with reference to the analysis of crime and health data. This focused on area data. The material in chapter 7 was included with an introduction provided by the conceptual frameworks described in chapter 5. The part of the course on modelling took selected material from chapter 9 and drew on examples referred to in that chapter and chapter 11.

Acknowledgements

This book has taken shape over the last two years at the University of Cambridge but has its roots in teaching and research that go back over many years most significantly to my time at the University of Sheffield. In one sense at least the book dates back to the early 1970s and a one-off lecture given by Michael Dacey at Northwestern University on spatial autocorrelation. That lecture was my introduction to the problems of analysing spatial data. Michael Goodchild invited me to spend some time at the NCGIA in Santa Barbara in the later 1980s and this too proved very formative.

I am grateful to friends and colleagues over the years with whom I have worked. The University of Sheffield had the foresight in the mid 1990s to invest in a research centre – the Sheffield Centre for Geographic Information and Spatial Analysis. This opened up opportunities for me to work on a range of different problems both theoretical and applied and fostered numerous collaborations both within the University and with local agencies. I would like to thank in particular Max Craglia, Ian Masser and Steve Wise in working with me to establish SCGISA and with whom I have undertaken many projects and had many interesting discussions.

I have had the benefit of working with many excellent researchers and in particular I would like to acknowledge Judith Bush, Paul Brindley, Vania Ceccato, Sue Collins, Andrew Costello, Young-Hoon Kim, Jingsheng Ma, Xiaoming Ning, Paola Signoretta and Dawn Thompson. At Cambridge I am working with Jane Law and together we are learning to apply Bayesian methodology to crime and health data, using the WinBUGS program. The examples in chapters 10 and 11 owe a great deal to her hard work. Jane and I are also working to encourage interest in these methods in agencies in the Cambridge region.

Sections on error propagation, missing-data estimation and spatial sampling have benefited from research collaborations with Giuseppe Arbia, Bob

xviii Acknowledgements

Bennett, Dan Griffith, Luis Flores and Jinfeng Wang. Some of the visualization material and exploratory data analysis has benefited from two ESRC projects undertaken with Steve Wise. I have worked with Max Craglia on several projects with strong policy dimensions, notably two recent Home Office projects. Some of the material on spatial modelling has benefited from collaboration with Eric Sheppard and Paul Plummer on modelling price variation in interdependent markets. My interest in applications of spatial analysis methods to problems in the areas of health studies have been stimulated by projects with Marcus Blake, Judith Bush and Dawn Thompson; also with David Hall and Ravi Maheswaran at Sheffield and recently with Andy Cliff with whom I have done some work on American measles data. Martin Kulldorff has kindly given me advice on the use of his scan test. My more recent interest in the application of spatial analysis methods to data in criminology, as well as drawing my attention to relevant literature, owe much to advice from Tony Bottoms, Andrew Costello and Paul Wiles. Thanks also to the many people who have drawn my attention to a wide range of relevant literature – apologies for not including them all by name.

Parts of this book have been tested on undergraduate and postgraduate students at the University of Cambridge. My thanks to them for sitting through the ‘first draft’. My thanks also to three anonymous readers who saw the first part of this book and made many excellent suggestions.

My thanks to Phil Stickler in the Cartography Laboratory at the University of Cambridge for drawing up the figures. Thanks also to the editorial and production guidance of Tracey Sanderson, Carol Miller and Anne Rix at Cambridge University Press.

Thanks to the following for allowing me to use their data in the examples: Dawn Thompson and Sheffield Health for the breast cancer screening data; South Yorkshire Police for several crime data sets, including the burglary and victimized offender data sets; Sheffield children’s services unit for the data on children excluded from school; James Reid for the updated ward boundary data for Cambridgeshire; Sara Godward of the Cancer Intelligence Unit for the Cambridgeshire lung cancer data, Andy Cliff for the US measles data.

Finally my thanks to my mother and father who have given me such encouragement over the years. This book is dedicated in particular to Rachel, my wife and ‘best friend’, for all her support and not least her willingness and enthusiasm to upsticks and try something new and different on occasions too numerous to count.

Copyright acknowledgements

Some figures in this book display boundary material which is copyright of the Crown, the ED-LINE consortium. Ordnance Survey data supplied by EDINA Digimap (a JISC supplied service) were also used. Some of the figures in this book are reproduced with the kind permission of the original publishers. These are as follows and with full references in the bibliography.

- Figure 3.5: Kluwer Academic Publishers. From *Geo-ENVII Geostatistics for Environmental Applications*, edited by J. Gomez-Hernandez, A. Soares and R. Frodevaux (1998) Savelieva et al. Conditional stochastic co-simulations of the Chernobyl fallout, fig. 14, p. 463.
- Figure 4.3: Pion Limited, London. From M. Tranmer and D.G. Steel (1998) Investigating the ecological fallacy. *Environment and Planning, A*, 30, fig. 1, p. 827 and fig. 2, p. 830.
- Figure 6.2: Taylor and Francis Limited, London. From M. Craglia, R. Haining and P. Wiles (2000) A comparative evaluation of approaches to urban crime pattern analysis. *Urban Studies* 37, fig. 5, p. 725.
- Figure 6.3: Oxford University Press. From *Journal of Public Health Medicine* (1994) R. Haining, S. Wise and M. Blake. Constructing regions for small area health analysis, fig. 3, p. 433.
- Figure 6.4: Kluwer Academic Publishers. From *Mathematical Geology*, 20 (1989) M. Oliver and R. Webster. A geostatistical basis for spatial weighting in multivariate classification, fig. 3, pp. 15–35.
- Figure 6.7: Kluwer Academic Publishers. From G. Verly et al. (1984) *Geostatistics for Natural Resources Characterization*. N. Cressie towards resistant geostatistics, fig. 8, p. 33.
- Figures 6.8 to 6.12: Springer-Verlag. From *Journal of Geographical Systems*, 2 (2000) R. Haining et al. Providing scientific visualization for spatial data analysis, pp. 121–40.
- Figure 7.1: John Wiley and Sons Limited. From *Statistics in Medicine* (1999) K. Kafadar. Simultaneous smoothing and adjusting mortality rates in US counties. Figs. 2(a)–2(d). Thanks also to Dr Kafadar for providing the original digital version of these figures.
- Figure 7.3 and 7.7: Routledge. From *GIS and Health*, edited by A. Gatrell and M. Loytonen. M. Kulldorff. Statistical methods for spatial epidemiology, figs. 4.1 and 4.2.
- Figure 7.8: John Wiley and Sons Limited. From *Statistics in Medicine* (1988) R. Stone. Investigations of excess environmental risks around putative sources. Fig. 3.

xx Acknowledgements

- Figures 8.1 and 8.2: Ohio State University Press, Columbus, Ohio. From *Geographical Analysis* (1991) R. Haining. Bivariate correlation with spatial data Figs. 2, 3 and 5
- Figure 8.3: International Biometric Society. From *Biometrics* (1997). A. Cerioli Modified test of independence in 2×2 tables with spatial data, Fig. 2, pp. 619–28.
- Figure 10.1: Ohio State University Press, Columbus, Ohio. From *Geographical Analysis* (1994) D. Griffith et al. Heterogeneity of attribute sampling error in spatial data sets, Fig. 1, p. 31a.
- Figure 11.2: Taylor and Francis Limited, London. From *Urban Studies* (2001) M. Craglia et al. Modelling high intensity crime areas in English cities p. 1931.