Use R!

Series Editors:
Robert Gentleman Kurt Hornik Giovanni Parmigiani

Use R!

Albert: Bayesian Computation with R
Bivand/Pebesma/Gómez-Rubio: Applied Spatial Data Analysis with R
Cook/Swayne: Interactive and Dynamic Graphics for Data Analysis:
With R and GGobi
Hahne/Huber/Gentleman/Falcon: Bioconductor Case Studies
Paradis: Analysis of Phylogenetics and Evolution with R
Pfaff: Analysis of Integrated and Cointegrated Time Series with R
Sarkar: Lattice: Multivariate Data Visualization with R
Spector: Data Manipulation with R

Roger S. Bivand • Edzer J. Pebesma
Virgilio Gómez-Rubio

Applied Spatial Data Analysis with R

Springer

Roger S. Bivand
Norwegian School of Economics
and Business Administration
Breiviksveien 40
5045 Bergen
Norway

Virgilio Gómez-Rubio
Department of Epidemiology
and Public Health
Imperial College London
St. Mary's Campus
Norfolk Place
London W2 1PG
United Kingdom

Edzer J. Pebesma
University of Utrecht
Department of Physical Geography
3508 TC Utrecht
Netherlands

Series Editors:
Robert Gentleman
Program in Computational Biology
Division of Public Health Sciences
Fred Hutchinson Cancer Research Center
1100 Fairview Ave. N, M2-B876
Seattle, Washington 98109-1024
USA

Kurt Hornik
Department für Statistik und Mathematik
Wirtschaftsuniversität Wien Augasse 2-6
A-1090 Wien
Austria

Giovanni Parmigiani
The Sidney Kimmel Comprehensive Cancer
Center at Johns Hopkins University
550 North Broadway
Baltimore, MD 21205-2011
USA

ISBN 978-0-387-78170-9 e-ISBN 978-0-387-78171-6
DOI 10.1007/978-0-387-78171-6

Library of Congress Control Number: 2008931196

© 2008 Springer Science+Business Media, LLC
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

springer.com

Ewie

Voor Ellen, Ulla en Mandus

A mis padres, Victorina y Virgilio Benigno

Preface

We began writing this book in parallel with developing software for handling and analysing spatial data with R (R Development Core Team, 2008). Although the book is now complete, software development will continue, in the R community fashion, of rich and satisfying interaction with users around the world, of rapid releases to resolve problems, and of the usual joys and frustrations of getting things done. There is little doubt that without pressure from users, the development of R would not have reached its present scale, and the same applies to analysing spatial data analysis with R.

It would, however, not be sufficient to describe the development of the R project mainly in terms of narrowly defined utility. In addition to being a community project concerned with the development of world-class data analysis software implementations, it promotes specific choices with regard to how data analysis is carried out. R is open source not only because open source software development, including the dynamics of broad and inclusive user and developer communities, is arguably an attractive and successful development model.

R is also, or perhaps chiefly, open source because the analysis of empirical and simulated data in science should be reproducible. As working researchers, we are all too aware of the possibility of reaching inappropriate conclusions in good faith because of user error or misjudgement. When the results of research really matter, as in public health, in climate change, and in many other fields involving spatial data, good research practice dictates that someone else should be, at least in principle, able to check the results. Open source software means that the methods used can, if required, be audited, and journalling working sessions can ensure that we have a record of what we actually did, not what we thought we did. Further, using `Sweave`[1] – a tool that permits the embedding of R code for complete data analyses in documents – throughout this book has provided crucial support (Leisch, 2002; Leisch and Rossini, 2003).

[1] `http://www.statistik.lmu.de/~leisch/Sweave/`.

We acknowledge our debt to the members of R-core for their continuing commitment to the R project. In particular, the leadership and example of Professor Brian Ripley has been important to us, although our admitted 'muddling through' contrasts with his peerless attention to detail. His interested support at the Distributed Statistical Computing conference in Vienna in 2003 helped us to see that encouraging spatial data analysis in R was a project worth pursuing. Kurt Hornik's dedication to keep the Comprehensive R Archive Network running smoothly, providing package maintainers with superb, almost 24/7, service, and his dry humour when we blunder, have meant that the useR community is provided with contributed software in an unequalled fashion. We are also grateful to Martin Mächler for his help in setting up and hosting the R-Sig-Geo mailing list, without which we would have not had a channel for fostering the R spatial community.

We also owe a great debt to users participating in discussions on the mailing list, sometimes for specific suggestions, often for fruitful questions, and occasionally for perceptive bug reports or contributions. Other users contact us directly, again with valuable input that leads both to a better understanding on our part of their research realities and to the improvement of the software involved. Finally, participants at R spatial courses, workshops, and tutorials have been patient and constructive.

We are also indebted to colleagues who have contributed to improving the final manuscript by commenting on earlier drafts and pointing out better procedures to follow in some examples. In particular, we would like to mention Juanjo Abellán, Nicky Best, Peter J. Diggle, Paul Hiemstra, Rebeca Ramis, Paulo J. Ribeiro Jr., Barry Rowlingson, and Jon O. Skøien. We are also grateful to colleagues for agreeing to our use of their data sets. Support from Luc Anselin has been important over a long period, including a very fruitful CSISS workshop in Santa Barbara in 2002. Work by colleagues, such as the first book known to us on using R for spatial data analysis (Kopczewska, 2006), provided further incentives both to simplify the software and complete its description. Without John Kimmel's patient encouragement, it is unlikely that we would have finished this book.

Even though we have benefitted from the help and advice of so many people, there are bound to be things we have not yet grasped – so remaining mistakes and omissions remain our sole responsibility. We would be grateful for messages pointing out errors in this book; errata will be posted on the book website (`http://www.asdar-book.org`).

Bergen — *Roger S. Bivand*
Münster — *Edzer J. Pebesma*
London — *Virgilio Gómez-Rubio*
April 2008

Contents

Preface VII

1 *Hello World*: Introducing Spatial Data 1
1.1 Applied Spatial Data Analysis 1
1.2 Why Do We Use R 2
1.2.1 ... In General? 2
1.2.2 ... for Spatial Data Analysis? 3
1.3 R and GIS 4
1.3.1 What is GIS? 4
1.3.2 Service-Oriented Architectures 6
1.3.3 Further Reading on GIS 6
1.4 Types of Spatial Data 7
1.5 Storage and Display 10
1.6 Applied Spatial Data Analysis 11
1.7 R Spatial Resources 13
1.7.1 Online Resources 14
1.7.2 Layout of the Book 14

Part I Handling Spatial Data in R

2 Classes for Spatial Data in R 21
2.1 Introduction 21
2.2 Classes and Methods in R 23
2.3 `Spatial` Objects 28
2.4 `SpatialPoints` 30
2.4.1 Methods 31
2.4.2 Data Frames for Spatial Point Data 33
2.5 `SpatialLines` 38

2.6 SpatialPolygons ... 41
2.6.1 SpatialPolygonsDataFrame Objects ... 44
2.6.2 Holes and Ring Direction ... 46
2.7 SpatialGrid and SpatialPixel Objects ... 47

3 Visualising Spatial Data ... 57
3.1 The Traditional Plot System ... 58
3.1.1 Plotting Points, Lines, Polygons, and Grids ... 58
3.1.2 Axes and Layout Elements ... 60
3.1.3 Degrees in Axes Labels and Reference Grid ... 64
3.1.4 Plot Size, Plotting Area, Map Scale, and Multiple Plots ... 65
3.1.5 Plotting Attributes and Map Legends ... 66
3.2 Trellis/Lattice Plots with spplot ... 68
3.2.1 A Straight Trellis Example ... 68
3.2.2 Plotting Points, Lines, Polygons, and Grids ... 70
3.2.3 Adding Reference and Layout Elements to Plots ... 72
3.2.4 Arranging Panel Layout ... 73
3.3 Interacting with Plots ... 74
3.3.1 Interacting with Base Graphics ... 74
3.3.2 Interacting with spplot and Lattice Plots ... 76
3.4 Colour Palettes and Class Intervals ... 76
3.4.1 Colour Palettes ... 76
3.4.2 Class Intervals ... 77

4 Spatial Data Import and Export ... 81
4.1 Coordinate Reference Systems ... 82
4.1.1 Using the EPSG List ... 83
4.1.2 PROJ.4 CRS Specification ... 84
4.1.3 Projection and Transformation ... 85
4.1.4 Degrees, Minutes, and Seconds ... 87
4.2 Vector File Formats ... 88
4.2.1 Using OGR Drivers in rgdal ... 89
4.2.2 Other Import/Export Functions ... 93
4.3 Raster File Formats ... 93
4.3.1 Using GDAL Drivers in rgdal ... 94
4.3.2 Writing a Google Earth™ Image Overlay ... 97
4.3.3 Other Import/Export Functions ... 98
4.4 Grass ... 99
4.4.1 Broad Street Cholera Data ... 104
4.5 Other Import/Export Interfaces ... 106
4.5.1 Analysis and Visualisation Applications ... 108
4.5.2 TerraLib and aRT ... 108
4.5.3 Other GIS and Web Mapping Systems ... 110
4.6 Installing rgdal ... 111

5 Further Methods for Handling Spatial Data 113
5.1 Support 113
5.2 Overlay 116
5.3 Spatial Sampling 118
5.4 Checking Topologies 120
5.4.1 Dissolving Polygons 121
5.4.2 Checking Hole Status 122
5.5 Combining Spatial Data 123
5.5.1 Combining Positional Data 123
5.5.2 Combining Attribute Data 124
5.6 Auxiliary Functions 126

6 Customising Spatial Data Classes and Methods 127
6.1 Programming with Classes and Methods 127
6.1.1 `S3`-Style Classes and Methods 129
6.1.2 `S4`-Style Classes and Methods 130
6.2 Animal Track Data in Package Trip 130
6.2.1 Generic and Constructor Functions 131
6.2.2 Methods for Trip Objects 133
6.3 Multi-Point Data: `SpatialMultiPoints` 134
6.4 Hexagonal Grids 137
6.5 Spatio-Temporal Grids 140
6.6 Analysing Spatial Monte Carlo Simulations 144
6.7 Processing Massive Grids 146

Part II Analysing Spatial Data

7 Spatial Point Pattern Analysis 155
7.1 Introduction 155
7.2 Packages for the Analysis of Spatial Point Patterns 156
7.3 Preliminary Analysis of a Point Pattern 160
7.3.1 Complete Spatial Randomness 160
7.3.2 G Function: Distance to the Nearest Event 161
7.3.3 F Function: Distance from a Point to the Nearest Event 162
7.4 Statistical Analysis of Spatial Point Processes 163
7.4.1 Homogeneous Poisson Processes 164
7.4.2 Inhomogeneous Poisson Processes 165
7.4.3 Estimation of the Intensity 165
7.4.4 Likelihood of an Inhomogeneous Poisson Process 168
7.4.5 Second-Order Properties 171
7.5 Some Applications in Spatial Epidemiology 172
7.5.1 Case–Control Studies 173
7.5.2 Binary Regression Estimator 178

7.5.3 Binary Regression Using Generalised Additive Models . . . 180
7.5.4 Point Source Pollution . . . 182
7.5.5 Accounting for Confounding and Covariates . . . 186
7.6 Further Methods for the Analysis of Point Patterns . . . 190

8 Interpolation and Geostatistics . . . 191
8.1 Introduction . . . 191
8.2 Exploratory Data Analysis . . . 192
8.3 Non-Geostatistical Interpolation Methods . . . 193
8.3.1 Inverse Distance Weighted Interpolation . . . 193
8.3.2 Linear Regression . . . 194
8.4 Estimating Spatial Correlation: The Variogram . . . 195
8.4.1 Exploratory Variogram Analysis . . . 196
8.4.2 Cutoff, Lag Width, Direction Dependence . . . 200
8.4.3 Variogram Modelling . . . 201
8.4.4 Anisotropy . . . 205
8.4.5 Multivariable Variogram Modelling . . . 206
8.4.6 Residual Variogram Modelling . . . 208
8.5 Spatial Prediction . . . 209
8.5.1 Universal, Ordinary, and Simple Kriging . . . 209
8.5.2 Multivariable Prediction: Cokriging . . . 210
8.5.3 Collocated Cokriging . . . 212
8.5.4 Cokriging Contrasts . . . 213
8.5.5 Kriging in a Local Neighbourhood . . . 213
8.5.6 Change of Support: Block Kriging . . . 215
8.5.7 Stratifying the Domain . . . 216
8.5.8 Trend Functions and their Coefficients . . . 217
8.5.9 Non-Linear Transforms of the Response Variable . . . 218
8.5.10 Singular Matrix Errors . . . 220
8.6 Model Diagnostics . . . 221
8.6.1 Cross Validation Residuals . . . 222
8.6.2 Cross Validation z-Scores . . . 223
8.6.3 Multivariable Cross Validation . . . 225
8.6.4 Limitations to Cross Validation . . . 225
8.7 Geostatistical Simulation . . . 226
8.7.1 Sequential Simulation . . . 227
8.7.2 Non-Linear Spatial Aggregation and Block Averages . . . 229
8.7.3 Multivariable and Indicator Simulation . . . 230
8.8 Model-Based Geostatistics and Bayesian Approaches . . . 230
8.9 Monitoring Network Optimization . . . 231
8.10 Other R Packages for Interpolation and Geostatistics . . . 233
8.10.1 Non-Geostatistical Interpolation . . . 233
8.10.2 spatial . . . 233
8.10.3 RandomFields . . . 234
8.10.4 geoR and geoRglm . . . 235
8.10.5 fields . . . 235

9 Areal Data and Spatial Autocorrelation 237
9.1 Introduction 237
9.2 Spatial Neighbours 239
9.2.1 Neighbour Objects 240
9.2.2 Creating Contiguity Neighbours 242
9.2.3 Creating Graph-Based Neighbours 244
9.2.4 Distance-Based Neighbours 246
9.2.5 Higher-Order Neighbours 249
9.2.6 Grid Neighbours 250
9.3 Spatial Weights 251
9.3.1 Spatial Weights Styles 251
9.3.2 General Spatial Weights 253
9.3.3 Importing, Converting, and Exporting Spatial Neighbours and Weights 255
9.3.4 Using Weights to Simulate Spatial Autocorrelation 257
9.3.5 Manipulating Spatial Weights 258
9.4 Spatial Autocorrelation: Tests 258
9.4.1 Global Tests 261
9.4.2 Local Tests 268

10 Modelling Areal Data 273
10.1 Introduction 273
10.2 Spatial Statistics Approaches 274
10.2.1 Simultaneous Autoregressive Models 277
10.2.2 Conditional Autoregressive Models 282
10.2.3 Fitting Spatial Regression Models 284
10.3 Mixed-Effects Models 287
10.4 Spatial Econometrics Approaches 289
10.5 Other Methods 296
10.5.1 GAM, GEE, GLMM 297
10.5.2 Moran Eigenvectors 302
10.5.3 Geographically Weighted Regression 305

11 Disease Mapping 311
11.1 Introduction 312
11.2 Statistical Models 314
11.2.1 Poisson-Gamma Model 315
11.2.2 Log-Normal Model 316
11.2.3 Marshall's Global EB Estimator 318
11.3 Spatially Structured Statistical Models 319
11.4 Bayesian Hierarchical Models 321
11.4.1 The Poisson-Gamma Model Revisited 322
11.4.2 Spatial Models 325
11.5 Detection of Clusters of Disease 332
11.5.1 Testing the Homogeneity of the Relative Risks 333
11.5.2 Moran's I Test of Spatial Autocorrelation 335

11.5.3 Tango's Test of General Clustering 335
11.5.4 Detection of the Location of a Cluster 337
11.5.5 Geographical Analysis Machine 337
11.5.6 Kulldorff's Statistic 338
11.5.7 Stone's Test for Localised Clusters 340
11.6 Other Topics in Disease Mapping 341

Afterword 343
R and Package Versions Used 344
Data Sets Used 344

References 347

Subject Index 361

Functions Index 371