

Use R!

Series Editors:

Robert Gentleman Kurt Hornik Giovanni G. Parmigiani

For further volumes:

<http://www.springer.com/series/6991>

Roger S. Bivand • Edzer Pebesma
Virgilio Gómez-Rubio

Applied Spatial Data Analysis with R

Second Edition

 Springer

Roger S. Bivand
Norwegian School of Economics
Bergen, Norway

Edzer Pebesma
Westfälische Wilhelms-Universität
Münster, Germany

Virgilio Gómez-Rubio
Department of Mathematics
Universidad de Castilla-La Mancha
Albacete, Spain

ISBN 978-1-4614-7617-7 ISBN 978-1-4614-7618-4 (eBook)
DOI 10.1007/978-1-4614-7618-4
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2013938605

© Springer Science+Business Media New York 2013

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed. Exempted from this legal reservation are brief excerpts in connection with reviews or scholarly analysis or material supplied specifically for the purpose of being entered and executed on a computer system, for exclusive use by the purchaser of the work. Duplication of this publication or parts thereof is permitted only under the provisions of the Copyright Law of the Publisher's location, in its current version, and permission for use must always be obtained from Springer. Permissions for use may be obtained through RightsLink at the Copyright Clearance Center. Violations are liable to prosecution under the respective Copyright Law.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

While the advice and information in this book are believed to be true and accurate at the date of publication, neither the authors nor the editors nor the publisher can accept any legal responsibility for any errors or omissions that may be made. The publisher makes no warranty, express or implied, with respect to the material contained herein.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

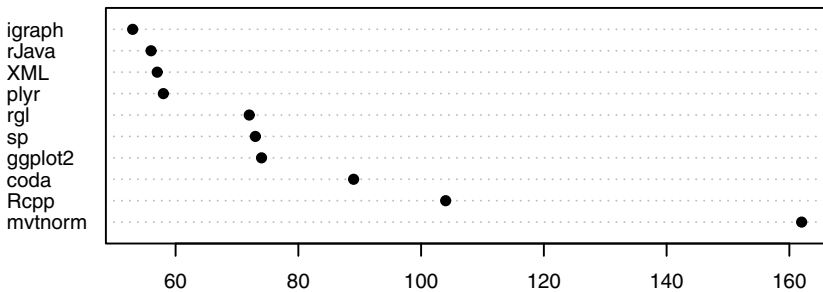
Ewie

Voor Ellen, Ulla en Mandus

A mis padres, Victorina y Virgilio Benigno

Preface (Second Edition)

Ten years ago, the small group of spatial data analysis enthusiasts who met in Vienna at the Distributed Statistical Computing conference mentioned in the preface to the first edition, considered that others might benefit from coordinating software development in our fields. We were in no way prepared for the dramatic and largely unexpected growth in use that software for spatial data analysis with R has seen (R Core Team, 2013). Some of this growth has come from the growth of R as a project, including growth in the use of R within disciplines analysing spatial data.



Analysis as of 10 Mar 2013 covering 4389 total CRAN packages
Limited to top 10 packages, excluding 'Recommended' ones shipped with R

Fig. 1 Direct dependency counts of CRAN packages

We do however feel humbled by the realisation that updating the **sp** package can potentially upset the work of many unsuspecting users and developers. In our first edition, we were proud to include a figure (Fig. 1.1, p.5) showing the dependency tree of **sp**. In revising our book, we have been obliged to drop this figure, as it is illegible at less than poster size. Fig. 1 gives the current top ten ranking of counts of packages depending directly on CRAN

packages, using Dirk Eddelbuettel's code.¹ The number of direct dependencies for **sp** is 73, but if we include indirect dependencies and imports through the dependency tree, we reach 148, and the total number of unique CRAN packages directly or indirectly depending on, importing or suggesting **sp** is 507.²

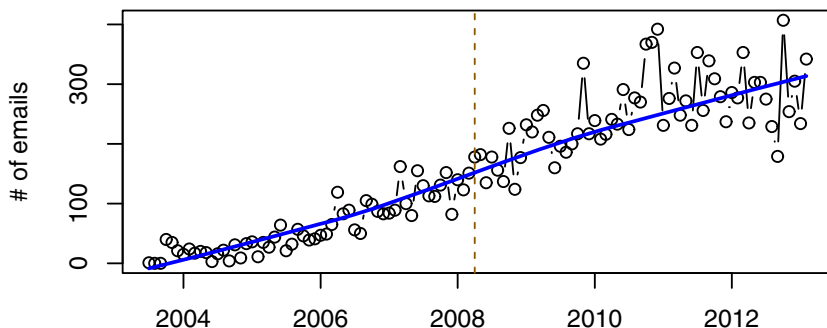


Fig. 2 Monthly numbers of postings to the R-sig-geo mailing list

Another expression for the vitality of the R spatial data analysis community is the development on the numbers of postings on the R-sig-geo mailing list, shown in Fig. 2. Importantly, the proportion of follow-up messages contributed by early participants has fallen, not because they have departed, but because the community has grown. The **raster** package is an important contribution that has taken raster processing and raster algebra to the next level and extended R as a raster GIS; a noticeable amount of list traffic now concerns use of this package. In addition, it seems that many courses have been organised bringing developers and users together, among these the GeoStat courses coordinated by Tom Hengl; we have benefitted from meeting users in the flesh, not only on the list.

Over the few years since the first edition was coming into being, we see clearly that spatial data, and the devices used for its collection, are becoming pervasive. In 2008, we could ask students whether they had access to a Global Positioning System (GPS) receiver. In 2013, we may ask how many GPS receivers students use in smartphones, tablets, vehicle navigation devices, etc. In 2008, Google Earth™ and Google Maps™ with others were seen as resources to be used on computers rather than mobile devices. Now, the failure of a smartphone manufacturer to handle spatial data satisfactorily is a top news story and can prejudice the careers of top managers. Use and handling of spatial data have grown greatly, but perhaps analysis is lagging,

¹ http://dirk.eddelbuettel.com/blog/2012/08/16#counting_cran_depends_followup

² Use is made of functions in the **pkgDepTools** package.

so work to make more and better analytical use of the positional data that is now potentially available should be moved up the agenda.

In revising our book, we have made the incremental changes needed to keep abreast of developments in the packages presented and discussed before. In addition, we have modified Chap. 5 to introduce the new **rgeos** package for handling operations on topologies. We have not attempted to cover the **raster** package in the detail that it deserves, hoping that a **raster** book will appear before long. We have replaced Chap. 6 with a new chapter on representing and handling spatio-temporal data, introducing the **spacetime** package; the chapter from the first edition is now a vignette in **sp**. Since the publication of Cressie and Wikle (2011) has provided new impetus to the analysis of spatio-temporal data, we expect these topics to grow in importance rapidly in coming years. We have also moved the detailed treatment of spatial neighbours to a vignette in the **spdep** package, making room for the presentation of new features now included in **spdep**. In Chap. 10, we have included worked examples of alternatives to WinBUGS for fitting Bayesian hierarchical models, including INLA (Rue et al., 2009) and BayesX and their R interface packages.

Finally, we are pleased that we can now present coloured figures,³ which we hope add to the value of the completed volume; thanks to Hannah Bracken for persevering with us in the revision process. The book website (<http://www.asdar-book.org>) will continue to provide code, data sets, and errata from this edition of our book; we will continue to run the code from the book, with suitable updates where required, nightly on the released version of R.

Bergen, Norway
Münster, Germany
Albacete, Spain

Roger S. Bivand
Edzer Pebesma
Virgilio Gómez-Rubio

³ Although reasonable care has been taken, the rendering of the colours may differ between the published figures and on-screen reproduction in an R session.

Preface (First Edition)

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 journaling working sessions can ensure that we have a record of what we actually did, not what we thought we did. Further, using **Sweave**⁴ – 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).

⁴ <http://www.stat.uni-muenchen.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. Skoien. 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 to 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, Norway
Münster, Germany
London, UK

Roger S. Bivand
Edzer Pebesma
Virgilio Gómez-Rubio

Contents

Preface (Second Edition)	vii
Preface (First Edition)	xi
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
1.2.3 ... and for Reproducible Research?	4
1.3 R and GIS	5
1.3.1 What Is GIS?	5
1.3.2 Service-Oriented Architectures	6
1.3.3 Further Reading on GIS	6
1.4 Types of Spatial Data	8
1.5 Storage and Display	10
1.6 Applied Spatial Data Analysis	11
1.7 R Spatial Resources	14
1.8 Layout of the Book	15
 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	<code>SpatialLines</code>	37
2.6	<code>SpatialPolygons</code>	41
	2.6.1 <code>SpatialPolygonsDataFrame</code> Objects	44
	2.6.2 Holes and Ring Direction	46
2.7	<code>SpatialGrid</code> and <code>SpatialPixel</code> Objects	48
2.8	Raster Objects and the <code>raster</code> Package	54
3	Visualising Spatial Data	59
3.1	The Traditional Plot System	60
	3.1.1 Plotting Points, Lines, Polygons, and Grids	60
	3.1.2 Axes and Layout Elements	61
	3.1.3 Degrees in Axes Labels and Reference Grid	65
	3.1.4 Plot Size, Plotting Area, Map Scale, and Multiple Plots	66
	3.1.5 Plotting Attributes and Map Legends	68
3.2	Trellis/Lattice Plots with <code>spplot</code>	69
	3.2.1 A Straight Trellis Example	70
	3.2.2 Plotting Points, Lines, Polygons, and Grids	70
	3.2.3 Adding Reference and Layout Elements to Plots	73
	3.2.4 Arranging Panel Layout	74
3.3	Alternatives Routes: <code>ggplot</code> , <code>latticeExtra</code>	75
3.4	Interactive Plots	76
	3.4.1 Interacting with Base Graphics	77
	3.4.2 Interacting with <code>spplot</code> and Lattice Plots	78
3.5	Colour Palettes and Class Intervals	79
	3.5.1 Colour Palettes	79
	3.5.2 Class Intervals	79
4	Spatial Data Import and Export	83
4.1	Coordinate Reference Systems	84
	4.1.1 Using the EPSG List	85
	4.1.2 PROJ.4 CRS Specification	86
	4.1.3 Projection and Transformation	88
	4.1.4 Degrees, Minutes, and Seconds	90
4.2	Vector File Formats	91
	4.2.1 Using OGR Drivers in <code>rgdal</code>	92
	4.2.2 Other Import/Export Functions	99
4.3	Raster File Formats	100
	4.3.1 Using GDAL Drivers in <code>rgdal</code>	100
	4.3.2 Other Import/Export Functions	107
4.4	Google Earth™, Google Maps™ and Other Formats	108
4.5	Geographical Resources Analysis Support System (GRASS)..	112
	4.5.1 Broad Street Cholera Data	118
4.6	Other Import/Export Interfaces	122
	4.6.1 Analysis and Visualisation Applications	122

- 4.6.2 TerraLib and aRT 123
- 4.6.3 Other GIS Systems 124
- 4.7 Installing **rgdal** 125
- 5 Further Methods for Handling Spatial Data** 127
 - 5.1 Support 127
 - 5.2 Handling and Combining Features 130
 - 5.2.1 The **rgeos** Package 130
 - 5.2.2 Using **rgeos** 132
 - 5.3 Map Overlay or Spatial Join 140
 - 5.3.1 Spatial Aggregation 142
 - 5.3.2 Using the **raster** Package for Extract Operations 145
 - 5.3.3 Spatial Sampling 146
 - 5.4 Auxiliary Functions 149
- 6 Spatio-Temporal Data** 151
 - 6.1 Introduction 151
 - 6.2 Types of Spatio-Temporal Data 151
 - 6.2.1 Spatial Point or Area, Time Instance or Interval 152
 - 6.2.2 Are Space and Time of *Primary* Interest? 152
 - 6.2.3 Regularity of Space-Time Layouts 152
 - 6.2.4 Do Objects Change Location? 153
 - 6.3 Classes in **spacetime** 154
 - 6.4 Handling Time Series Data with **xts** 155
 - 6.5 Construction of **ST** Objects 156
 - 6.6 Selection, Addition, and Replacement of Attributes 158
 - 6.7 Overlay and Aggregation 159
 - 6.8 Visualisation 161
 - 6.8.1 Multi-panel Plots 161
 - 6.8.2 Space-Time Plots 162
 - 6.8.3 Animated Plots 163
 - 6.8.4 Time Series Plots 164
 - 6.9 Further Packages 164
 - 6.9.1 Handling Spatio-Temporal Data 165
 - 6.9.2 Analysing Spatio-Temporal Data 165
 - 6.10 Outlook 165

Part II Analysing Spatial Data

- 7 Spatial Point Pattern Analysis** 173
 - 7.1 Introduction 173
 - 7.2 Packages for the Analysis of Spatial Point Patterns 174
 - 7.3 Preliminary Analysis of a Point Pattern 178
 - 7.3.1 Complete Spatial Randomness 179
 - 7.3.2 *G* Function: Distance to the Nearest Event 179

7.3.3	<i>F</i> Function: Distance from a Point to the Nearest Event	181
7.4	Statistical Analysis of Spatial Point Processes	182
7.4.1	Homogeneous Poisson Processes	183
7.4.2	Inhomogeneous Poisson Processes	184
7.4.3	Estimation of the Intensity	184
7.4.4	Likelihood of an Inhomogeneous Poisson Process	187
7.4.5	Second-Order Properties	190
7.5	Some Applications in Spatial Epidemiology	192
7.5.1	Case–Control Studies	193
7.5.2	Binary Regression Estimator	198
7.5.3	Binary Regression Using Generalised Additive Models	199
7.5.4	Point Source Pollution	202
7.5.5	Accounting for Confounding and Covariates	206
7.6	Further Methods for the Analysis of Point Patterns	210
8	Interpolation and Geostatistics	213
8.1	Introduction	213
8.2	Exploratory Data Analysis	214
8.3	Non-geostatistical Interpolation Methods	215
8.3.1	Inverse Distance Weighted Interpolation	215
8.3.2	Linear Regression	216
8.4	Estimating Spatial Correlation: The Variogram	217
8.4.1	Exploratory Variogram Analysis	219
8.4.2	Cutoff, Lag Width, Direction Dependence	222
8.4.3	Variogram Modelling	224
8.4.4	Anisotropy	228
8.4.5	Multivariable Variogram Modelling	229
8.4.6	Residual Variogram Modelling	230
8.5	Spatial Prediction	232
8.5.1	Universal, Ordinary, and Simple Kriging	233
8.5.2	Multivariable Prediction: Cokriging	233
8.5.3	Collocated Cokriging	236
8.5.4	Cokriging Contrasts	237
8.5.5	Kriging in a Local Neighbourhood	237
8.5.6	Change of Support: Block Kriging	238
8.5.7	Stratifying the Domain	240
8.5.8	Trend Functions and Their Coefficients	241
8.5.9	Non-linear Transforms of the Response Variable	242
8.5.10	Singular Matrix Errors	243
8.6	Kriging, Filtering, Smoothing	245
8.7	Model Diagnostics	247
8.7.1	Cross Validation Residuals	247
8.7.2	Cross Validation <i>z</i> -Scores	249

- 8.7.3 Multivariable Cross Validation 250
- 8.7.4 Limitations to Cross Validation 250
- 8.8 Geostatistical Simulation 252
 - 8.8.1 Sequential Simulation 252
 - 8.8.2 Non-linear Spatial Aggregation and Block Averages... 254
 - 8.8.3 Multivariable and Indicator Simulation..... 255
- 8.9 Model-Based Geostatistics and Bayesian Approaches 256
- 8.10 Monitoring Network Optimisation 256
- 8.11 Other R Packages for Interpolation and Geostatistics 258
 - 8.11.1 Non-geostatistical Interpolation 258
 - 8.11.2 Spatial 259
 - 8.11.3 RandomFields 259
 - 8.11.4 geoR and geoRglm 259
 - 8.11.5 Fields 260
 - 8.11.6 spBayes 260
- 8.12 Spatio-Temporal Prediction 260

- 9 Modelling Areal Data 263**
 - 9.1 Introduction 263
 - 9.2 Spatial Neighbours and Spatial Weights 266
 - 9.2.1 Neighbour Objects 266
 - 9.2.2 Spatial Weights Objects 269
 - 9.2.3 Handling Spatial Weights Objects 273
 - 9.2.4 Using Weights to Simulate Spatial Autocorrelation ... 274
 - 9.3 Testing for Spatial Autocorrelation 275
 - 9.3.1 Global Tests 278
 - 9.3.2 Local Tests 284
 - 9.4 Fitting Models of Areal Data 288
 - 9.4.1 Spatial Statistics Approaches 290
 - 9.4.2 Spatial Econometrics Approaches..... 303
 - 9.4.3 Other Methods..... 314

- 10 Disease Mapping 319**
 - 10.1 Introduction 320
 - 10.2 Statistical Models 322
 - 10.2.1 Poisson-Gamma Model..... 323
 - 10.2.2 Log-Normal Model 325
 - 10.2.3 Marshall’s Global EB Estimator..... 326
 - 10.3 Spatially Structured Statistical Models 328
 - 10.4 Bayesian Hierarchical Models 330
 - 10.4.1 The Poisson-Gamma Model Revisited..... 332
 - 10.4.2 Spatial Models 336
 - 10.5 Geoadditive Models..... 345
 - 10.6 Detection of Clusters of Disease 347
 - 10.6.1 Testing the Homogeneity of the Relative Risks 348
 - 10.6.2 Moran’s *I* Test of Spatial Autocorrelation 350
 - 10.6.3 Tango’s Test of General Clustering 351

- 10.6.4 Detection of the Location of a Cluster 352
- 10.6.5 Geographical Analysis Machine 353
- 10.6.6 Kulldorff’s Statistic..... 353
- 10.6.7 Stone’s Test for Localised Clusters..... 355
- 10.7 Spatio-Temporal Disease Mapping 356
 - 10.7.1 Introduction 356
 - 10.7.2 Spatio-Temporal Modelling of Disease..... 357
- 10.8 Other Topics in Disease Mapping..... 361

- Afterword** 363
 - R and Package Versions Used 364
 - Data Sets Used 364

- References** 367

- Subject Index** 387

- Functions Index** 401