

Use R!

Series Editors

Robert Gentleman Kurt Hornik Giovanni Parmigiani

More information about this series at <http://www.springer.com/series/6991>

Daniel Borcard • François Gillet • Pierre Legendre

Numerical Ecology with R

Second Edition

 Springer

Daniel Borcard
Université de Montréal
Département de sciences biologiques
Montréal, Québec, Canada H3C 3J7

François Gillet
Université Bourgogne Franche-Comté
UMR Chrono-environnement
Besançon, France

Pierre Legendre
Université de Montréal
Département de sciences biologiques
Montréal, Québec, Canada H3C 3J7

ISSN 2197-5736

ISSN 2197-5744 (electronic)

Use R!

ISBN 978-3-319-71403-5

ISBN 978-3-319-71404-2 (eBook)

<https://doi.org/10.1007/978-3-319-71404-2>

Library of Congress Control Number: 2017961342

© Springer International Publishing AG, part of Springer Nature 2011, 2018

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.

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.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by the registered company Springer International Publishing AG part of Springer Nature.

The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Ecology is sexy. Teaching ecology is therefore the art of presenting a fascinating topic to well-predisposed audiences. It is not easy: the complexities of modern ecological science go well beyond the introductory chapters taught in high schools or the marvellous movies about ecosystems presented on TV. But well-predisposed audiences are ready to make the effort. *Numerical* ecology is another story. For some unclear reasons, a majority of ecology-oriented people are strangely reluctant when it comes to quantifying nature and using mathematical tools to help understand it. As if nature was inherently non-mathematical, which it is certainly not: mathematics is the common language of all sciences. Teachers of biostatistics and numerical ecology thus have to overcome this reluctance: before even beginning to teach the subject itself, they must convince their audience of the interest and necessity of it.

During many decades ecologists, be they students or researchers (in the academic, private or government spheres), used to plan their research and collect data with few, if any, statistical consideration, and then entrusted the “statistical” analyses of their results to a person hired especially for that purpose. That person may well have been a competent statistician, and indeed in many cases the progressive integration of statistics into the whole process of ecological research was triggered by such people. In other cases, however, the end product was a large amount of data summarized using a handful of basic statistics and tests of significance that were far from revealing all the richness of the structures hidden in the data tables. The separation of the ecological and statistical worlds presented many problems. The most important were that the ecologists were unaware of the array of methods available at the time, and the statisticians were unaware of the ecological hypotheses to be tested and the specific requirements of ecological data (the double-zero problem is a good example). Apart from preventing the data to be exploited properly, this double unawareness prevented the development of methods specifically tailored to ecological problems.

The answer to this situation is to form mathematically inclined ecologists. Fortunately, more and more such people have appeared during the recent decades. The result of their work is a huge development of statistical ecology, the availability

of several excellent textbooks, and the increasing awareness of the responsibility of ecologists with regard to the proper design and analysis of their research. This awareness makes the task easier for teachers as well.

Until the first years of this millennium, however, a critical ingredient was still missing for the teaching to be efficient and for the practice of statistics to become generalized among ecologists: a set of standard packages available to everyone, everywhere. A biostatistics or numerical ecology course means nothing without practical exercises. A course linked to commercial software is much better, but it is bound to restrict future applications if the researcher moves and loses access to the software that he or she knows. Furthermore, commercial packages are in most cases written for larger audiences than the community of ecologists and they may not include all the functions required for analysing ecological data. The **R** language resolved that issue, thanks to the dedication of the many researchers who created and freely contributed extensive, well-designed, and well-documented packages. Now the teacher no longer has to say: “this is the way PCA works... on paper;” she or he can say instead: “this is the way PCA works, now I will show you on-screen how to run one, and in a few minutes you will be able to run your own, and do it anywhere in the world on your own data!”

Another fundamental property of the **R** language is that it is meant as a self-learning environment. A book on **R** is therefore bound to follow that philosophy, and must provide the support necessary for anyone wishing to explore the subject by himself or herself. This book has been written to provide a bridge between the theory and practice of numerical ecology, that anyone can cross. Our dearest hope is that it will make many happy teachers and happy ecologists.

Since they are living entities, both the field of numerical ecology and the **R** language evolve. As a result, much has happened in both fields since the publication of the first edition of *Numerical Ecology with R* in 2011. Therefore, it was time not only to update the code provided in the first edition, but also to present new methods, provide more insight into existing ones, offer more examples and a wider array of applications of the major methods. We also took the opportunity to present the code in a more attractive way, generated by R Markdown in RStudio[®], with different colours for functions, objects, arguments and comments.

Our dearest hope is that all this will make many more happy teachers and happy ecologists.

Montréal, QC, Canada
Besançon, France
Montréal, QC, Canada

Daniel Borcard
François Gillet
Pierre Legendre

Contents

1	Introduction	1
1.1	Why Numerical Ecology?	1
1.2	Why R?	2
1.3	Readership and Structure of the Book	2
1.4	How to Use This Book	3
1.5	The Data Sets	4
1.5.1	The Doubs Fish Data	5
1.5.2	The Oribatid Mite Data	7
1.6	A Quick Reminder About Help Sources	7
1.7	Now It Is Time	9
2	Exploratory Data Analysis	11
2.1	Objectives	11
2.2	Data Exploration	11
2.2.1	Data Extraction	11
2.2.2	Species Data: First Contact	12
2.2.3	Species Data: A Closer Look	14
2.2.4	Ecological Data Transformation	21
2.2.5	Environmental Data	28
2.3	Conclusion	34
3	Association Measures and Matrices	35
3.1	Objectives	35
3.2	The Main Categories of Association Measures (Short Overview)	35
3.2.1	Q Mode and R Mode	36
3.2.2	Symmetrical or Asymmetrical Coefficients in Q Mode: The Double-Zero Problem	36
3.2.3	Association Measures for Qualitative or Quantitative Data	37
3.2.4	To Summarize	37

3.3	Q Mode: Computing Dissimilarity Matrices Among Objects	38
3.3.1	Q Mode: Quantitative Species Data	39
3.3.2	Q Mode: Binary (Presence-Absence) Species Data	42
3.3.3	Q Mode: Quantitative Data (Excluding Species Abundances)	46
3.3.4	Q Mode: Binary Data (Excluding Species Presence-Absence Data)	48
3.3.5	Q Mode: Mixed Types Including Categorical (Qualitative Multiclass) Variables	49
3.4	R Mode: Computing Dependence Matrices Among Variables	51
3.4.1	R Mode: Species Abundance Data	52
3.4.2	R Mode: Species Presence-Absence Data	52
3.4.3	R Mode: Quantitative and Ordinal Data (Other than Species Abundances)	53
3.4.4	R Mode: Binary Data (Other than Species Abundance Data)	55
3.5	Pre-transformations for Species Data	55
3.6	Conclusion	57
4	Cluster Analysis	59
4.1	Objectives	59
4.2	Clustering Overview	59
4.3	Hierarchical Clustering Based on Links	62
4.3.1	Single Linkage Agglomerative Clustering	62
4.3.2	Complete Linkage Agglomerative Clustering	64
4.4	Average Agglomerative Clustering	65
4.5	Ward's Minimum Variance Clustering	68
4.6	Flexible Clustering	69
4.7	Interpreting and Comparing Hierarchical Clustering Results	70
4.7.1	Introduction	70
4.7.2	Cophenetic Correlation	71
4.7.3	Looking for Interpretable Clusters	74
4.8	Non-hierarchical Clustering	96
4.8.1	<i>k</i> -means Partitioning	96
4.8.2	Partitioning Around Medoids (PAM)	103
4.9	Comparison with Environmental Data	107
4.9.1	Comparing a Typology with External Data (ANOVA Approach)	107
4.9.2	Comparing Two Typologies (Contingency Table Approach)	111
4.10	Species Assemblages	111
4.10.1	Simple Statistics on Group Contents	111
4.10.2	Kendall's <i>W</i> Coefficient of Concordance	112
4.10.3	Species Assemblages in Presence-Absence Data	115
4.10.4	Species Co-occurrence Network	117

- 4.11 Indicator Species 120
 - 4.11.1 Introduction 120
 - 4.11.2 IndVal: Species Indicator Values 120
 - 4.11.3 Correlation-Type Indices 125
- 4.12 Multivariate Regression Trees (MRT):
Constrained Clustering 126
 - 4.12.1 Introduction 126
 - 4.12.2 Computation (Principle) 127
 - 4.12.3 Application Using Packages `mvpart`
and `MVPARTwrap` 129
 - 4.12.4 Combining MRT and IndVal 134
- 4.13 MRT as a Monothetic Clustering Method 135
- 4.14 Sequential Clustering 138
- 4.15 A Very Different Approach: Fuzzy Clustering 141
 - 4.15.1 Fuzzy *c*-means Using Package
`cluster`'s Function `fanny()` 141
 - 4.15.2 Noise Clustering Using the
`vegclust()` Function 146
- 4.16 Conclusion 150
- 5 Unconstrained Ordination 151**
 - 5.1 Objectives 151
 - 5.2 Ordination Overview 151
 - 5.2.1 Multidimensional Space 151
 - 5.2.2 Ordination in Reduced Space 152
 - 5.3 Principal Component Analysis (PCA) 153
 - 5.3.1 Overview 153
 - 5.3.2 PCA of the Environmental Variables
of the Doubs River Data Using `rda()` 154
 - 5.3.3 PCA on Transformed Species Data 166
 - 5.3.4 Domain of Application of PCA 169
 - 5.3.5 PCA Using Function `PCA.newr()` 170
 - 5.3.6 Imputation of Missing Values in PCA 171
 - 5.4 Correspondence Analysis (CA) 175
 - 5.4.1 Introduction 175
 - 5.4.2 CA Using Function `cca()` of Package `vegan` 176
 - 5.4.3 CA Using Function `CA.newr()` 181
 - 5.4.4 Arch Effect and Detrended Correspondence
Analysis (DCA) 182
 - 5.4.5 Multiple Correspondence Analysis (MCA) 183
 - 5.5 Principal Coordinate Analysis (PCoA) 187
 - 5.5.1 Introduction 187
 - 5.5.2 Application of PCoA to the Doubs Data Set Using
`cmdscale()` and `vegan` 188
 - 5.5.3 Application of PCoA to the Doubs Data Set
Using `pcoa()` 190

5.6	Nonmetric Multidimensional Scaling (NMDS)	193
5.6.1	Introduction	193
5.6.2	Application to the Doubs Fish Data	193
5.6.3	PCoA or NMDS?	196
5.7	Hand-Written PCA Ordination Function	198
6	Canonical Ordination	203
6.1	Objectives	203
6.2	Canonical Ordination Overview	204
6.3	Redundancy Analysis (RDA)	204
6.3.1	Introduction	204
6.3.2	RDA of the Doubs River Data	206
6.3.3	Distance-Based Redundancy Analysis (db-RDA)	249
6.3.4	A Hand-Written RDA Function	253
6.4	Canonical Correspondence Analysis (CCA)	256
6.4.1	Introduction	256
6.4.2	CCA of the Doubs River Data	257
6.5	Linear Discriminant Analysis (LDA)	263
6.5.1	Introduction	263
6.5.2	Discriminant Analysis Using <code>lda()</code>	264
6.6	Other Asymmetric Analyses	268
6.6.1	Principal Response Curves (PRC)	268
6.6.2	Co-correspondence Analysis (CoCA)	271
6.7	Symmetric Analysis of Two (or More) Data Sets	274
6.8	Canonical Correlation Analysis (CCorA)	275
6.8.1	Introduction	275
6.8.2	Canonical Correlation Analysis Using <code>CCorA()</code>	275
6.9	Co-inertia Analysis (CoIA)	277
6.9.1	Introduction	277
6.9.2	Co-inertia Analysis Using Function <code>coinertia()</code> of <code>ade4</code>	278
6.10	Multiple Factor Analysis (MFA)	282
6.10.1	Introduction	282
6.10.2	Multiple Factor Analysis Using <code>FactoMineR</code>	283
6.11	Relating Species Traits and Environment	287
6.11.1	The Fourth-Corner Method	288
6.11.2	RLQ Analysis	290
6.11.3	Application in R	291
6.12	Conclusion	296
7	Spatial Analysis of Ecological Data	299
7.1	Objectives	299
7.2	Spatial Structures and Spatial Analysis: A Short Overview	300
7.2.1	Introduction	300
7.2.2	Induced Spatial Dependence and Spatial Autocorrelation	301
7.2.3	Spatial Scale	302

- 7.2.4 Spatial Heterogeneity 303
- 7.2.5 Spatial Correlation or Autocorrelation Functions
and Spatial Correlograms 303
- 7.2.6 Testing for the Presence of Spatial Correlation:
Conditions 308
- 7.2.7 Modelling Spatial Structures 309
- 7.3 Multivariate Trend-Surface Analysis 309
 - 7.3.1 Introduction 309
 - 7.3.2 Trend-Surface Analysis in Practice 310
- 7.4 Eigenvector-Based Spatial Variables and Spatial Modelling 314
 - 7.4.1 Introduction 314
 - 7.4.2 Distance-Based Moran’s Eigenvector Maps
(dbMEM) and Principal Coordinates of Neighbour
Matrices (PCNM) 315
 - 7.4.3 MEM in a Wider Context: Weights Other than
Geographic Distances 333
 - 7.4.4 MEM with Positive or Negative Spatial Correlation:
Which Ones should Be Used? 348
 - 7.4.5 Asymmetric Eigenvector Maps (AEM):
When Directionality Matters 348
- 7.5 Another Way to Look at Spatial Structures: Multiscale
Ordination (MSO) 355
 - 7.5.1 Principle 355
 - 7.5.2 Application to the Mite Data – Exploratory
Approach 356
 - 7.5.3 Application to the Detrended Mite
and Environmental Data 359
- 7.6 Space-Time Interaction Test in Multivariate ANOVA,
Without Replicates 361
 - 7.6.1 Introduction 361
 - 7.6.2 Testing the Space-Time Interaction with the
sti Functions 364
- 7.7 Conclusion 367
- 8 Community Diversity 369**
 - 8.1 Objectives 369
 - 8.2 The Multiple Facets of Diversity 370
 - 8.2.1 Introduction 370
 - 8.2.2 Species Diversity Measured by a Single Number 370
 - 8.2.3 Taxonomic Diversity Indices in Practice 374
 - 8.3 When Space Matters: Alpha, Beta and Gamma Diversities 379
 - 8.4 Beta Diversity 379
 - 8.4.1 Beta Diversity Measured by a Single Number 379
 - 8.4.2 Beta Diversity as the Variance of the Community
Composition Table: SCBD and LCBD Indices 382

- 8.4.3 Partitioning Beta Diversity into Replacement,
Richness Difference and Nestedness Components 388
- 8.5 Functional Diversity, Functional Composition
and Phylogenetic Diversity of Communities 404
 - 8.5.1 Alpha Functional Diversity 404
 - 8.5.2 Beta Taxonomic, Phylogenetic
and Functional Diversities 408
- 8.6 Conclusion 412
- Bibliography 413**
- Index 427**

About the Authors

Daniel Borcard is lecturer of Biostatistics and Ecology and researcher in Numerical Ecology at Université de Montréal, Québec, Canada. His research interests include Numerical Ecology, Ecology of communities, and Soil Ecology/Zoology.

François Gillet is professor of Community Ecology and Ecological Modelling at Université Bourgogne Franche-Comté, Besançon, France, and visiting professor at École Polytechnique Fédérale de Lausanne, Switzerland. His research deals with the structure, diversity, ecology and dynamics of plant communities.

Pierre Legendre is professor of Quantitative Biology and Ecology at Université de Montréal, fellow of the Royal Society of Canada, and Web of Science Highly Cited Researcher in Environment/Ecology. He is the founder of the field of numerical ecology.

Supplementary Material

All the necessary data files, the scripts used in the chapters, as well as the **R** functions and packages that are not available through the CRAN web site, can be downloaded from our web page <http://adn.biol.umontreal.ca/~numerical ecology/numecolR/>.