Statistics for Biology and Health

Alain F. Zuur
Elena N. Ieno · Graham M. Smith

Analysing Ecological Data

Springer

Statistics for Biology and Health

Alain F. Zuur
Elena N. Ieno · Graham M. Smith

Analysing Ecological Data

Springer

Statistics for Biology and Health

Series Editors
M. Gail, K. Krickeberg, J. Samet, A. Tsiatis, W. Wong

Statistics for Biology and Health

Bacchieri/Cioppa: Fundamentals of Clinical Research
Borchers/Buckland/Zucchini: Estimating Animal Abundance: Closed Populations
Burzykowski/Molenberghs/Buyse: The Evaluation of Surrogate Endpoints
Everitt/Rabe-Hesketh: Analyzing Medical Data Using S-PLUS
Ewens/Grant: Statistical Methods in Bioinformatics: An Introduction, 2nd ed.
Gentleman/Carey/Huber/Irizarry/Dudoit: Bioinformatics and Computational Biology Solutions Using R and Bioconductor
Hougaard: Analysis of Multivariate Survival Data
Keyfitz/Caswell: Applied Mathematical Demography, 3rd ed.
Klein/Moeschberger: Survival Analysis: Techniques for Censored and Truncated Data, 2nd ed.
Kleinbaum/Klein: Logistic Regression: A Self-Learning Text, 2nd ed.
Kleinbaum/Klein: Survival Analysis: A Self-Learning Text, 2nd ed.
Lange: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.
Manton/Singer/Suzman: Forecasting the Health of Elderly Populations
Martinussen/Scheike: Dynamic Regression Models for Survival Data
Moyé: Multiple Analyses in Clinical Trials: Fundamentals for Investigators
Nielšen: Statistical Methods in Molecular Evolution
Parmigiani/Garrett/Irizarry/Zeger: The Analysis of Gene Expression Data: Methods and Software
Proschan/Lan/Wittes: Statistical Monitoring of Clinical Trials: A Unified Approach
Siegmund/Yakir: The Statistics of Gene Mapping
Simon/Korn/McShane/Radmacher/Wright/Zhao: Design and Analysis of DNA Microarray Investigations
Sorensen/Gianola: Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics
Stallard/Manton/Cohen: Forecasting Product Liability Claims: Epidemiology and Modeling in the Manville Asbestos Case
Sun: The Statistical Analysis of Interval-censored Failure Time Data
Therneau/Grambsch: Modeling Survival Data: Extending the Cox Model
Ting: Dose Finding in Drug Development
Vittinghoff/Glidden/Shiboski/McCulloch: Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models
Wu/Ma/Casella: Statistical Genetics of Quantitative Traits: Linkage, Map and QTL
Zhang/Singer: Recursive Partitioning in the Health Sciences
Zuur/Ieno/Smith: Analysing Ecological Data

Alain F. Zuur
Elena N. Ieno
Graham M. Smith

Analysing Ecological Data

Springer

Alain F. Zuur
Highland Statistics Ltd.
Newburgh AB41 6FN
UNITED KINGDOM
highstat@highstat.com

Elena N. Ieno
Highland Statistics Ltd.
Newburgh AB41 6FN
UNITED KINGDOM
bio@highstat.com

Graham M. Smith
School of Science and the Environment
Bath Spa University
Bath BA2 9BN
UNITED KINGDOM
g.m.smith@bathspa.ac.uk

Library of Congress Control Number: 2006933720

ISBN-10: 0-387-45967-7
ISBN-13: 978-0-387-45967-7
e-ISBN-10: 0-387-45972-3
e-ISBN-13: 978-0-387-45972-1

Printed on acid-free paper.

Printed in the United States of America

9 8 7 6 5 4 3 2 1

springer.com

To Asterix, Juultje and Poek, for paying more attention to my laptop

To Norma and Juan Carlos, and to Antonio (d' Aieta) who showed me that it was worthwhile crossing the great waters...

To Moira, for accepting all the hours shared with my computer that I should have been sharing with her

Preface

'*Which test should I apply?*' During the many years of working with ecologists, biologists and other environmental scientists, this is probably the question that the authors of this book hear the most often. The answer is always the same and along the lines of 'What are your underlying questions?', 'What do you want to show?'. The answers to these questions provide the starting point for a detailed discussion on the ecological background and purpose of the study. This then gives the basis for deciding on the most appropriate analytical approach. Therefore, a better starting point for an ecologist is to avoid the phrase 'test' and think in terms of 'analysis'. A test refers to something simple and unified that gives a clear answer in the form of a p-value: something rarely appropriate for ecological data. In practice, one has to apply a data exploration, check assumptions, validate the models, perhaps apply a series of methods, and most importantly, interpret the results in terms of the underlying ecology and the ecological questions being investigated.

Ecology is a quantitative science trying to answer difficult questions about the complex world we live in. Most ecologists are aware of these complexities, but few are fully equipped with the statistical sophistication and understanding to deal with them.

Even data gathered from apparently simple ecological research can require a level of statistical awareness rarely taught at the undergraduate or even the postgraduate level. There is little enough time to teach the essentials of ecology, let alone finding the time to teach 'advanced' statistics. Hopefully, for post graduates moving into academia there will be some advanced statistical support available, but many ecologist end up working in government, a voluntary organisation or consultancy where statistical support is minimal.

Although, the authors of this book believe that a quantitative approach is at the core of being a good ecologist, they also appreciate how challenging many ecologists find statistics. This book is therefore aimed at three levels of reader.

At one level it is aimed at making ecologists aware of how important it is to design scientifically robust ecological experiments or monitoring programmes, and the importance of selecting the best analytical technique. For these readers we hope the book, in particular the case studies, will encourage them to develop their personal statistical skills, or convince them they need statistical support.

On the next level it is aimed at the statistically literate ecologist, who may not be fully aware of the techniques we discuss, or when to use them. Hopefully, we have explained things well enough for these readers to feel confident enough to use some of the techniques we describe. Often these techniques are presented in a

fairly impenetrable manner, even for the statistically aware ecologist, and we have tried to make our presentation as 'ecologist friendly' as possible.

Finally, we hope the book will be of value to statisticians, whether they have a background in ecology or statistics. Ecological data can be particularly challenging to analyse, and we hope that providing an insight into our approach, together with the detailed case studies, will be of value to statistician readers, regardless of their background and expertise.

Overall, however, we hope this book will contribute in some small way to improving the collection and analysis of ecological data and improve the quality of environmental decision making.

After reading this book, you should be able to apply the following process: 'These are my questions', 'This is my statistical approach', 'Here is proof that I did it all correct (model validation)', 'This is what the data show' and 'Here is the ecological interpretation'.

Acknowledgement

A large part of the material in this book has been used by the first two authors as course material for MSc and PhD students, post-docs, scientists, both as academic and non-academic courses. We are greatly indepted to all 1200–1500 course participants who helped improve the material between 2000 and 2005 by asking questions and commenting on the material.

We would also like to thank a series of persons who commented on parts of this book: Ian Jolliffe, Anatoly Saveliev, Barry O'Neill, Neil Campbell, Graham Pierce, Ian Tuck, Alex Douglas, Pam Sikkink, Toby Marthews, Adrian Bowman, and six anonymous reviewers and the copy-editor. Their criticisms, comments, help and suggestions have greatly improved this book.

The first author would like to thank Rob Fryer and FRS Marine Laboratory for providing the flexibility to start the foundation of this book.

We would also like to thank the people and organizations who donated data for the theory chapters. The acknowledgement for the unpublished squid data (donated by Graham Pierce, University of Aberdeen) used in Chapters 4 and 7 is as follows. Data collection was financed by the European Commission under the following projects: FAR MA 1.146, AIR1-CT92-0573, FAIR CT 1520, Study Project 96/081, Study project 97/107, Study Project 99/063, and Q5CA-2002-00962. We would like to thank Roy Mendelssohn (NOAA/NMFS) for giving us a copy of the data used in Mendelssohn and Schwing (2002). The raw data are summaries calculated from the COADS dataset. The COADS references are Slutz et al. (1985) and Woodruff et al. (1987). We thank Jaap van der Meer (NIOZ) for allowing us to use the Balgzand data, The Bahamas National Trust and Greenforce Andros Island Marine Study for providing the Bahamas fisheries dataset, Chris Elphick (University of Connecticut) for the sparrow data, and Hrafnkell Eiríksson (Marine Research Institute, Reykjavik) for the Icelandic Nephrops time series. The public domain SRTM data used in Chapter 19 were taken from the U.S. Geological Survey, EROS Data Center, Sioux Falls, SD. We thank Steve Hare (University of Washington) for allowing us to use the 100 biological and physical time series

from the North Pacific Ocean in Chapter 17. A small part of Chapter 13 is based on Zuur (1999, unpublished PhD thesis), which was partly financed by the EU project DYNAMO (FAIR-CT95-0710).

A big 'thank you' is also due to the large number of folks who wrote R (www.r-project.org) and its many libraries. We made a lot of use of the lattice, regression, GLM, GAM (mgcv) and mixed modelling libraries (nlme). This thank you is probably also on behalf of the readers of this book as everything we did can be done in R.

Finally, we would like to thank John Kimmel for giving us the opportunity to write this book, and his support during the entire process. On to the next book.

Alain F. Zuur
Elena N. Ieno
Graham M. Smith

February 2007

Contents

Contributors ..... **xix**

1 Introduction **1**
1.1 Part 1: Applied statistical theory 1
1.2 Part 2: The case studies 3
1.3 Data, software and flowcharts 6

2 Data management and software **7**
2.1 Introduction 7
2.2 Data management 8
2.3 Data preparation 9
2.4 Statistical software 13

3 Advice for teachers **17**
3.1 Introduction 17

4 Exploration **23**
4.1 The first steps 24
4.2 Outliers, transformations and standardisations 38
4.3 A final thought on data exploration 47

5 Linear regression **49**
5.1 Bivariate linear regression 49
5.2 Multiple linear regression 67
5.3 Partial linear regression 73

6 Generalised linear modelling **79**
6.1 Poisson regression 79
6.2 Logistic regression 88

7 Additive and generalised additive modelling **97**
7.1 Introduction 97
7.2 The additive model 101
7.3 Example of an additive model 102
7.4 Estimate the smoother and amount of smoothing 104
7.5 Additive models with multiple explanatory variables 108

7.6 Choosing the amount of smoothing 112
7.7 Model selection and validation 115
7.8 Generalised additive modelling 120
7.9 Where to go from here 124

8 Introduction to mixed modelling 125
8.1 Introduction 125
8.2 The random intercept and slope model 128
8.3 Model selection and validation 130
8.4 A bit of theory 135
8.5 Another mixed modelling example 137
8.6 Additive mixed modelling 140

9 Univariate tree models 143
9.1 Introduction 143
9.2 Pruning the tree 149
9.3 Classification trees 152
9.4 A detailed example: Ditch data 152

10 Measures of association 163
10.1 Introduction 163
10.2 Association between sites: Q analysis 164
10.3 Association among species: R analysis 171
10.4 Q and R analysis: Concluding remarks 176
10.5 Hypothesis testing with measures of association 179

11 Ordination — First encounter 189
11.1 Bray–Curtis ordination 189

12 Principal component analysis and redundancy analysis 193
12.1 The underlying principle of PCA 193
12.2 PCA: Two easy explanations 194
12.3 PCA: Two technical explanations 196
12.4 Example of PCA 197
12.5 The biplot 200
12.6 General remarks 205
12.7 Chord and Hellinger transformations 206
12.8 Explanatory variables 208
12.9 Redundancy analysis 210
12.10 Partial RDA and variance partitioning 219
12.11 PCA regression to deal with collinearity 221

13 Correspondence analysis and canonical correspondence analysis 225
13.1 Gaussian regression and extensions 225
13.2 Three rationales for correspondence analysis 231
13.3 From RGR to CCA 238

13.4 Understanding the CCA triplot 240
13.5 When to use PCA, CA, RDA or CCA 242
13.6 Problems with CA and CCA 243

14 Introduction to discriminant analysis 245
14.1 Introduction 245
14.2 Assumptions 248
14.3 Example 250
14.4 The mathematics 254
14.5 The numerical output for the sparrow data 255

15 Principal coordinate analysis and non-metric multidimensional scaling 259
15.1 Principal coordinate analysis 259
15.2 Non-metric multidimensional scaling 261

16 Time series analysis — Introduction 265
16.1 Using what we have already seen before 265
16.2 Auto-regressive integrated moving average models with exogenous variables 281

17 Common trends and sudden changes 289
17.1 Repeated LOESS smoothing 289
17.2 Identifying the seasonal component 293
17.3 Common trends: MAFA 299
17.4 Common trends: Dynamic factor analysis 303
17.5 Sudden changes: Chronological clustering 315

18 Analysis and modelling of lattice data 321
18.1 Lattice data 321
18.2 Numerical representation of the lattice structure 323
18.3 Spatial correlation 327
18.4 Modelling lattice data 331
18.5 More exotic models 334
18.6 Summary 338

19 Spatially continuous data analysis and modelling 341
19.1 Spatially continuous data 341
19.2 Geostatistical functions and assumptions 342
19.3 Exploratory variography analysis 346
19.4 Geostatistical modelling: Kriging 358
19.5 A full spatial analysis of the bird radar data 363

20 Univariate methods to analyse abundance of decapod larvae 373
20.1 Introduction 373
20.2 The data 374
20.3 Data exploration 377

20.4 Linear regression results 379
20.5 Additive modelling results 381
20.6 How many samples to take? 383
20.7 Discussion 385

21 Analysing presence and absence data for flatfish distribution in the Tagus estuary, Portugal 389
21.1 Introduction 389
21.2 Data and materials 390
21.3 Data exploration 392
21.4 Classification trees 395
21.5 Generalised additive modelling 397
21.6 Generalised linear modelling 398
21.7 Discussion 401

22 Crop pollination by honeybees in Argentina using additive mixed modelling 403
22.1 Introduction 403
22.2 Experimental setup 404
22.3 Abstracting the information 404
22.4 First steps of the analyses: Data exploration 407
22.5 Additive mixed modelling 408
22.6 Discussion and conclusions 414

23 Investigating the effects of rice farming on aquatic birds with mixed modelling 417
23.1 Introduction 417
23.2 The data 419
23.3 Getting familiar with the data: Exploration 420
23.4 Building a mixed model 424
23.5 The optimal model in terms of random components 427
23.6 Validating the optimal linear mixed model 430
23.7 More numerical output for the optimal model 431
23.8 Discussion 433

24 Classification trees and radar detection of birds for North Sea wind farms 435
24.1 Introduction 435
24.2 From radars to data 436
24.3 Classification trees 438
24.4 A tree for the birds 440
24.5 A tree for birds, clutter and more clutter 445
24.6 Discussion and conclusions 447

25 Fish stock identification through neural network analysis of parasite fauna 449

25.1 Introduction 449
25.2 Horse mackerel in the northeast Atlantic 450
25.3 Neural networks 452
25.4 Collection of data 455
25.5 Data exploration 456
25.6 Neural network results 457
25.7 Discussion 460

26 Monitoring for change: Using generalised least squares, non-metric multidimensional scaling, and the Mantel test on western Montana grasslands 463
26.1 Introduction 463
26.2 The data 464
26.3 Data exploration 467
26.4 Linear regression results 472
26.5 Generalised least squares results 476
26.6 Multivariate analysis results 479
26.7 Discussion 483

27 Univariate and multivariate analysis applied on a Dutch sandy beach community 485
27.1 Introduction 485
27.2 The variables 486
27.3 Analysing the data using univariate methods 487
27.4 Analysing the data using multivariate methods 494
27.5 Discussion and conclusions 499

28 Multivariate analyses of South-American zoobenthic species — spoilt for choice 503
28.1 Introduction and the underlying questions 503
28.2 Study site and sample collection. 504
28.3 Data exploration 506
28.4 The Mantel test approach 509
28.5 The transformation plus RDA approach 512
28.6 Discussion and conclusions 512

29 Principal component analysis applied to harbour porpoise fatty acid data 515
29.1 Introduction 515
29.2 The data 515
29.3 Principal component analysis 517
29.4 Data exploration 518
29.5 Principal component analysis results 518
29.6 Simpler alternatives to PCA 524
29.7 Discussion 526