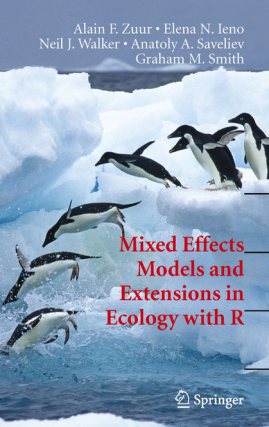


Alain F. Zuur • Elena N. Ieno
Neil J. Walker • Anatoly A. Saveliev
Graham M. Smith

A photograph of several penguins in a polar environment. Some are on large white ice floes, while others are swimming in the blue water. The background shows more ice and a clear sky.

**Mixed Effects
Models and
Extensions in
Ecology with R**

 Springer

Statistics for Biology and Health

Series Editors:

M. Gail

K. Krickeberg

J. M. 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
- Duchateau/Janssen*: The Frailty Model
- 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*: Survival Analysis: A Self-Learning Text, 2nd ed.
- Kleinbaum/Klein*: Logistic Regression: A Self-Learning Text, 2nd ed.
- Lange*: Mathematical and Statistical Methods for Genetic Analysis, 2nd ed.
- Lazar*: The Statistical Analysis of Functional MRI Data
- Manton/Singer/Suzman*: Forecasting the Health of Elderly Populations
- Martiniusen/Scheike*: Dynamic Regression Models for Survival Data
- Moyé*: Multiple Analyses in Clinical Trials: Fundamentals for Investigators
- Nielsen*: Statistical Methods in Molecular Evolution
- O'Quigley*: Proportional Hazards Regression
- Parmigiani/Garrett/Irizarry/Zeger*: The Analysis of Gene Expression Data: Methods and Software
- Prochan/LanWittes*: 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, Maps, and QTL
- Zhang/Singer*: Recursive Partitioning in the Health Sciences
- Zuur/Ieno/Smith*: Analysing Ecological Data
- Zuur/Ieno/Walker/Saveliev/Smith*: Mixed Effects Models and Extensions in Ecology with R

Alain F. Zuur · Elena N. Ieno · Neil J. Walker ·
Anatoly A. Saveliev · Graham M. Smith

Mixed Effects Models and Extensions in Ecology with R

 Springer

Alain F. Zuur
Highland Statistics Ltd.
Newburgh
United Kingdom
highstat@highstat.com

Elena N. Ieno
Highland Statistics Ltd.
Newburgh
United Kingdom
bio@highstat.com

Neil J. Walker
Central Science Laboratory
Gloucester
United Kingdom
n.walker@csl.gov.uk

Anatoly A. Saveliev
Kazan State University
Kazan
Russia
saa@ksu.ru

Graham M. Smith
Bath Spa University
Bath
United Kingdom
graham.smith@myotis.co.uk

Series Editors

M. Gail
National Cancer Institute
Rockville, MD 20892
USA

K. Krickeberg
Le Chatelet
F-63270 Manglieu
France

J. Samet
Department of Preventive
Medicine
Keck School of Medicine
University of Southern
California
1441 Eastlake Ave. Room
4436, MC 9175
Los Angeles, CA 90089

A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA

W. Wong
Department of Statistics
Stanford University
Stanford, CA 94305-4065
USA

ISSN 1431-8776
ISBN 978-0-387-87457-9
DOI 10.1007/978-0-387-87458-6

e-ISBN 978-0-387-87458-6

Library of Congress Control Number: 2008942429

© Springer Science+Business Media, LLC 2009

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

*Thanks to my parents for sharing the burden
of my university fees – Alain F. Zuur*

*To my friends, colleagues, and former students
who are actively committed to the protection and
care of the environment – Elena N. Ieno*

*Thanks to my wife Tatiana for her patience
and moral support – Anatoly A. Saveliev*

*I would like to thank all family and friends for
help and support through times good and bad
during the writing of this book – Neil J. Walker*

*To my parents who, even now, continue to support
me in everything I do – Graham M. Smith*

Preface

No sooner, it seems, had our first book *Analysing Ecological Data* gone to print, than we embarked on the writing of the nearly 600 page text you are now holding. This proved to be a labour of love of sorts – we felt that there were certain issues sufficiently common in the analysis of ecological data that merited more detailed description and analysis. Thus the present book can be seen as a ‘sequel’ to *Analysing Ecological Data* but with much greater emphasis on these very issues so commonly encountered in the collection of, and analysis of, ecological data. In particular, we look at different ways of analysing nested data, heterogeneity of variance, spatial and temporal correlation, and zero-inflated data.

The original plan was to write a text of about 350 pages, but to do justice to the sheer range of problems and ideas we have well exceeded that original target (as you can see!). Such is the scope of applied statistics in ecology. In particular, partly on the back of reviewer’s comments, we have included a chapter on Bayesian Monte-Carlo Markov-Chain applications in generalized linear modelling. We hope this serves as an informative introduction (but no more than an introduction!) to this interesting and increasingly relevant area of statistics.

We received lots of positive feedback on the approach and style we used in *Analysing Ecological Data*, especially the combination of case studies and a theory section. We have therefore followed the same approach with this book. This time, however, we have provided the R code used for the analysis. Most of this R code is included in the text, but where the code was particularly long, it is only available from the book’s website at www.highstat.com. In the case studies, we also included advice on what to write in a paper.

Newburgh, United Kingdom
Newburgh, United Kingdom
Gloucester, United Kingdom
Kazan, Russia
Bath, United Kingdom
December 2008

Alain F. Zuur
Elena N. Ieno
Neil J. Walker
Anatoly A. Saveliev
Graham M. Smith

Acknowledgements

The material in this book has been taught in various courses in 2007 and 2008, and we are greatly in debt to all participants who helped improving the material. We would also like to thank a number of people who read and commented on parts of earlier drafts, namely Chris Elphick, Alex Douglas, and Graham Pierce. The manuscript was reviewed by Loveday Conquest (University of Washington), Sarah Goslee (USDA), Thomas Kneib (LMU Munich), Bret Larget (University of Wisconsin), Ruth Salway (University of Bath), Jing Hua Zhao (University of Cambridge), and several anonymous referees. We thank them all for their positive, encouraging, and useful reviews. Their comments and criticisms greatly improved the book.

The most difficult part of writing a book is finding public domain data which can be used in theory chapters. We are particularly thankful to the following persons for donating data sets. Sonia Mendes and Graham Pierce for the whale data, Gerard Janssen for the benthic data, Pam Sikkink for the grassland data, Graham Pierce and Jennifer Smith for the squid data, Alexandre Roulin for the barn owl data, Michael Reed and Chris Elphick for the Hawaiian bird data, Tatiana Rogova for the Volzhsko-Kamsky forestry data, Robert Cruikshanks, Mary Kelly-Quinn and John O'Halloran for the Irish (sodium dominance index) river data, Chris Elphick for the sparrow and California bird data, Michael Penston for the sea lice data, Joaquín Vicente and Christian Gortázar for the wild boar and deer data, Ken Mackenzie for the cod data, and António Mira for the snake data. The proper references are given in the text. We also would like to thank all people involved in the case study chapters; they are credited where relevant.

Michelle Cronin provided the seal photo on the back cover, Joaquin Vicente the deer photo, and Malena Sabatino gave us the bee photo. The photograph of the koalas was provided by Australian Koala Foundation (www.savethekoala.com). The photo on the front cover is from © Wayne Lynch/Arcticphoto.com.

Finally, we would like to thank John Kimmel for giving us the opportunity to write this book and for patiently accepting the 6-month delay. Up to the next book.

Contents

1	Introduction	1
1.1	What Is in the Book?	1
1.1.1	To Include or Not to Include GLM and GAM	3
1.1.2	Case Studies	4
1.1.3	Flowchart of the Content	4
1.2	Software	5
1.3	How to Use This Book If You Are an Instructor	6
1.4	What We Did Not Do and Why	6
1.5	How to Cite R and Associated Packages	7
1.6	Our R Programming Style	8
1.7	Getting Data into R	9
1.7.1	Data in a Package	10
2	Limitations of Linear Regression Applied on Ecological Data	11
2.1	Data Exploration	12
2.1.1	Cleveland Dotplots	12
2.1.2	Pairplots	14
2.1.3	Boxplots	15
2.1.4	xyplot from the Lattice Package	15
2.2	The Linear Regression Model	17
2.3	Violating the Assumptions; Exception or Rule?	19
2.3.1	Introduction	19
2.3.2	Normality	19
2.3.3	Heterogeneity	20
2.3.4	Fixed X	21
2.3.5	Independence	21
2.3.6	Example 1; Wedge Clam Data	22
2.3.7	Example 2; Moby's Teeth	26
2.3.8	Example 3; Nereis	28
2.3.9	Example 4; Pelagic Bioluminescence	30
2.4	Where to Go from Here	31

- 3 Things Are Not Always Linear; Additive Modelling 35**
 - 3.1 Introduction 35
 - 3.2 Additive Modelling 36
 - 3.2.1 GAM in gam and GAM in mgcv 37
 - 3.2.2 GAM in gam with LOESS 38
 - 3.2.3 GAM in mgcv with Cubic Regression Splines 42
 - 3.3 Technical Details of GAM in mgcv 44
 - 3.3.1 A (Little) Bit More Technical Information
on Regression Splines 47
 - 3.3.2 Smoothing Splines Alias Penalised Splines 49
 - 3.3.3 Cross-Validation 51
 - 3.3.4 Additive Models with Multiple Explanatory Variables . . . 53
 - 3.3.5 Two More Things 53
 - 3.4 GAM Example 1; Bioluminescent Data for Two Stations 55
 - 3.4.1 Interaction Between a Continuous and Nominal Variable . . 59
 - 3.5 GAM Example 2: Dealing with Collinearity 63
 - 3.6 Inference 66
 - 3.7 Summary and Where to Go from Here? 67

- 4 Dealing with Heterogeneity 71**
 - 4.1 Dealing with Heterogeneity 72
 - 4.1.1 Linear Regression Applied on Squid 72
 - 4.1.2 The Fixed Variance Structure 74
 - 4.1.3 The VarIdent Variance Structure 75
 - 4.1.4 The varPower Variance Structure 78
 - 4.1.5 The varExp Variance Structure 80
 - 4.1.6 The varConstPower Variance Structure 80
 - 4.1.7 The varComb Variance Structure 81
 - 4.1.8 Overview of All Variance Structures 82
 - 4.1.9 Graphical Validation of the Optimal Model 84
 - 4.2 Benthic Biodiversity Experiment 86
 - 4.2.1 Linear Regression Applied on the Benthic
Biodiversity Data 86
 - 4.2.2 GLS Applied on the Benthic Biodiversity Data 89
 - 4.2.3 A Protocol 90
 - 4.2.4 Application of the Protocol on the Benthic Biodiversity
Data 92

- 5 Mixed Effects Modelling for Nested Data 101**
 - 5.1 Introduction 101
 - 5.2 2-Stage Analysis Method 103
 - 5.3 The Linear Mixed Effects Model 105
 - 5.3.1 Introduction 105
 - 5.3.2 The Random Intercept Model 106
 - 5.3.3 The Random Intercept and Slope Model 109
 - 5.3.4 Random Effects Model 111

- 5.4 Induced Correlations 112
 - 5.4.1 Intraclass Correlation Coefficient 114
- 5.5 The Marginal Model 114
- 5.6 Maximum Likelihood and REML Estimation 116
 - 5.6.1 Illustration of Difference Between ML and REML 119
- 5.7 Model Selection in (Additive) Mixed Effects Modelling 120
- 5.8 RIKZ Data: Good Versus Bad Model Selection 122
 - 5.8.1 The Wrong Approach 122
 - 5.8.2 The Good Approach 127
- 5.9 Model Validation 128
- 5.10 Begging Behaviour of Nestling Barn Owls 129
 - 5.10.1 Step 1 of the Protocol: Linear Regression 130
 - 5.10.2 Step 2 of the Protocol: Fit the Model with GLS 132
 - 5.10.3 Step 3 of the Protocol: Choose a Variance Structure 132
 - 5.10.4 Step 4: Fit the Model 133
 - 5.10.5 Step 5 of the Protocol: Compare New Model with Old Model 133
 - 5.10.6 Step 6 of the Protocol: Everything Ok? 134
 - 5.10.7 Steps 7 and 8 of the Protocol: The Optimal Fixed Structure 135
 - 5.10.8 Step 9 of the Protocol: Refit with REML and Validate the Model 137
 - 5.10.9 Step 10 of the Protocol 139
 - 5.10.10 Sorry, We are Not Done Yet 139

- 6 Violation of Independence – Part I 143**
 - 6.1 Temporal Correlation and Linear Regression 143
 - 6.1.1 ARMA Error Structures 150
 - 6.2 Linear Regression Model and Multivariate Time Series 152
 - 6.3 Owl Sibling Negotiation Data 158

- 7 Violation of Independence – Part II 161**
 - 7.1 Tools to Detect Violation of Independence 161
 - 7.2 Adding Spatial Correlation Structures to the Model 166
 - 7.3 Revisiting the Hawaiian Birds 171
 - 7.4 Nitrogen Isotope Ratios in Whales 172
 - 7.4.1 Moby 172
 - 7.4.2 All Whales 174
 - 7.5 Spatial Correlation due to a Missing Covariate 177
 - 7.6 Short Godwits Time Series 182
 - 7.6.1 Description of the Data 182
 - 7.6.2 Data Exploration 183
 - 7.6.3 Linear Regression 184
 - 7.6.4 Protocol Time 186
 - 7.6.5 Why All the Fuss? 190

- 8 Meet the Exponential Family** 193
 - 8.1 Introduction 193
 - 8.2 The Normal Distribution 194
 - 8.3 The Poisson Distribution 196
 - 8.3.1 Preparation for the Offset in GLM 198
 - 8.4 The Negative Binomial Distribution 199
 - 8.5 The Gamma Distribution 201
 - 8.6 The Bernoulli and Binomial Distributions 202
 - 8.7 The Natural Exponential Family 204
 - 8.7.1 Which Distribution to Select? 205
 - 8.8 Zero Truncated Distributions for Count Data 206

- 9 GLM and GAM for Count Data** 209
 - 9.1 Introduction 209
 - 9.2 Gaussian Linear Regression as a GLM 210
 - 9.3 Introducing Poisson GLM with an Artificial Example 211
 - 9.4 Likelihood Criterion 213
 - 9.5 Introducing the Poisson GLM with a Real Example 215
 - 9.5.1 Introduction 215
 - 9.5.2 R Code and Results 216
 - 9.5.3 Deviance 217
 - 9.5.4 Sketching the Fitted Values 218
 - 9.6 Model Selection in a GLM 220
 - 9.6.1 Introduction 220
 - 9.6.2 R Code and Output 220
 - 9.6.3 Options for Finding the Optimal Model 221
 - 9.6.4 The Drop1 Command 222
 - 9.6.5 Two Ways of Using the Anova Command 223
 - 9.6.6 Results 223
 - 9.7 Overdispersion 224
 - 9.7.1 Introduction 224
 - 9.7.2 Causes and Solutions for Overdispersion 224
 - 9.7.3 Quick Fix: Dealing with Overdispersion in
a Poisson GLM 225
 - 9.7.4 R Code and Numerical Output 226
 - 9.7.5 Model Selection in Quasi-Poisson 227
 - 9.8 Model Validation in a Poisson GLM 228
 - 9.8.1 Pearson Residuals 229
 - 9.8.2 Deviance Residuals 229
 - 9.8.3 Which One to Use? 230
 - 9.8.4 What to Plot? 230
 - 9.9 Illustration of Model Validation in Quasi-Poisson GLM 231
 - 9.10 Negative Binomial GLM 233
 - 9.10.1 Introduction 233
 - 9.10.2 Results 236

- 9.11 GAM 238
 - 9.11.1 Distribution of larval Sea Lice Around Scottish Fish Farms 239
- 10 GLM and GAM for Absence–Presence and Proportional Data 245**
 - 10.1 Introduction 245
 - 10.2 GLM for Absence–Presence Data 246
 - 10.2.1 Tuberculosis in Wild Boar 246
 - 10.2.2 Parasites in Cod 252
 - 10.3 GLM for Proportional Data 254
 - 10.4 GAM for Absence–Presence Data 258
 - 10.5 Where to Go from Here? 259
- 11 Zero-Truncated and Zero-Inflated Models for Count Data 261**
 - 11.1 Introduction 261
 - 11.2 Zero-Truncated Data 263
 - 11.2.1 The Underlying Mathematics for Truncated Models 263
 - 11.2.2 Illustration of Poisson and NB Truncated Models 265
 - 11.3 Too Many Zeros 269
 - 11.3.1 Sources of Zeros 270
 - 11.3.2 Sources of Zeros for the Cod Parasite Data 271
 - 11.3.3 Two-Part Models Versus Mixture Models, and Hippos 271
 - 11.4 ZIP and ZINB Models 274
 - 11.4.1 Mathematics of the ZIP and ZINB 274
 - 11.4.2 Example of ZIP and ZINB Models 278
 - 11.5 ZAP and ZANB Models, Alias Hurdle Models 286
 - 11.5.1 Mathematics of the ZAP and ZANB 287
 - 11.5.2 Example of ZAP and ZANB 288
 - 11.6 Comparing Poisson, Quasi-Poisson, NB, ZIP, ZINB, ZAP and ZANB GLMs 291
 - 11.7 Flowchart and Where to Go from Here 293
- 12 Generalised Estimation Equations 295**
 - 12.1 GLM: Ignoring the Dependence Structure 295
 - 12.1.1 The California Bird Data 295
 - 12.1.2 The Owl Data 299
 - 12.1.3 The Deer Data 300
 - 12.2 Specifying the GEE 302
 - 12.2.1 Introduction 302
 - 12.2.2 Step 1 of the GEE: Systematic Component and Link Function 303
 - 12.2.3 Step 2 of the GEE: The Variance 304
 - 12.2.4 Step 3 of the GEE: The Association Structure 304
 - 12.3 Why All the Fuss? 309
 - 12.3.1 A Bit of Maths 310

- 12.4 Association for Binary Data 313
- 12.5 Examples of GEE 314
 - 12.5.1 A GEE for the California Birds. 314
 - 12.5.2 A GEE for the Owls 316
 - 12.5.3 A GEE for the Deer Data. 319
- 12.6 Concluding Remarks 320

- 13 GLMM and GAMM 323**
 - 13.1 Setting the Scene for Binomial GLMM 324
 - 13.2 GLMM and GAMM for Binomial and Poisson Data 327
 - 13.2.1 Deer Data 327
 - 13.2.2 The Owl Data Revisited. 333
 - 13.2.3 A Word of Warning 339
 - 13.3 The Underlying Mathematics in GLMM 339

- 14 Estimating Trends for Antarctic Birds in Relation to Climate Change 343**
 A.F. Zuur, C. Barbraud, E.N. Ieno, H. Weimerskirch, G.M. Smith, and N.J. Walker
 - 14.1 Introduction 343
 - 14.1.1 Explanatory Variables 344
 - 14.2 Data Exploration 345
 - 14.3 Trends and Auto-correlation. 350
 - 14.4 Using Ice Extent as an Explanatory Variable 352
 - 14.5 SOI and Differences Between Arrival and Laying Dates 354
 - 14.6 Discussion 360
 - 14.7 What to Report in a Paper 361

- 15 Large-Scale Impacts of Land-Use Change in a Scottish Farming Catchment 363**
 A.F. Zuur, D. Raffaelli, A.A. Saveliev, N.J. Walker, E.N. Ieno, and G.M. Smith
 - 15.1 Introduction 363
 - 15.2 Data Exploration 365
 - 15.3 Estimation of Trends for the Bird Data 367
 - 15.3.1 Model Validation 368
 - 15.3.2 Failed Approach 1 372
 - 15.3.3 Failed Approach 2 373
 - 15.3.4 Assume Homogeneity? 374
 - 15.4 Dealing with Independence 374
 - 15.5 To Transform or Not to Transform. 378
 - 15.6 Birds and Explanatory Variables 378
 - 15.7 Conclusions 380
 - 15.8 What to Write in a Paper 381

16 Negative Binomial GAM and GAMM to Analyse Amphibian Roadkills 383
 A.F. Zuur, A. Mira, F. Carvalho, E.N. Ieno, A.A. Saveliev, G.M. Smith, and N.J. Walker

16.1 Introduction 383
 16.1.1 Roadkills 383

16.2 Data Exploration 385

16.3 GAM 389

16.4 Understanding What the Negative Binomial is Doing 394

16.5 GAMM: Adding Spatial Correlation 396

16.6 Discussion 397

16.7 What to Write in a Paper 397

17 Additive Mixed Modelling Applied on Deep-Sea Pelagic Bioluminescent Organisms 399
 A.F. Zuur, I.G. Priede, E.N. Ieno, G.M. Smith, A.A. Saveliev, and N.J. Walker

17.1 Biological Introduction 399

17.2 The Data and Underlying Questions 401

17.3 Construction of Multi-panel Plots for Grouped Data 402
 17.3.1 Approach 1 402
 17.3.2 Approach 2 407
 17.3.3 Approach 3 408

17.4 Estimating Common Patterns Using Additive Mixed Modelling 410
 17.4.1 One Smoothing Curve for All Stations 410
 17.4.2 Four Smoothers; One for Each Month 414
 17.4.3 Smoothing Curves for Groups Based on Geographical Distances 417
 17.4.4 Smoothing Curves for Groups Based on Source Correlations 418

17.5 Choosing the Best Model 419

17.6 Discussion 420

17.7 What to Write in a Paper 421

18 Additive Mixed Modelling Applied on Phytoplankton Time Series Data 423
 A.F. Zuur, M.J. Latuhihin, E.N. Ieno, J.G. Baretta-Bekker, G.M. Smith, and N.J. Walker

18.1 Introduction 423
 18.1.1 Biological Background of the Project 424

18.2 Data Exploration 427

18.3 A Statistical Data Analysis Strategy for DIN 429

18.4 Results for Temperature 439

18.5 Results for DIAT1 441

18.6 Comparing Phytoplankton and Environmental Trends 443