Alain F Zuur
Elena N Ieno
Anatoly A Saveliev

Beginner's Guide to

Spatial, Temporal, and Spatial-Temporal Ecological Data Analysis with R-INLA

Volume I: Using GLM and GLMM

Published by Highland Statistics Ltd.
Highland Statistics Ltd.
Newburgh
United Kingdom
highstat@highstat.com

ISBN: 978-0-9571741-9-1
First published in February 2017

www.highstat.com

I would like to thank my wife Nandani for giving me the most beautiful present a man can wish for. They say that storks bring babies. It is true!

– Alain F Zuur –

To Norma, Juan Carlos, and Walter for their constant support

– Elena N Ieno –

Special thanks to my wife, friends, and colleagues who make my life more interesting

– Anatoly A Saveliev –

Preface

The authors of this book have been giving statistics courses to ecologists for 15 years. We have taught more than 8,000 scientists. During our courses we cover topics such as R, data exploration, data visualizsation, multiple linear regression, generalised linear models, linear mixed-effects models, generalised linear mixed-effects modelling (GLMM), generalised additive models (GAM), generalised additive mixed-effects models (GAMM), Bayesian analysis and MCMC, and multivariate analysis, among many other topics. Over the years a large number of participants have asked us to teach a module that covers the analysis of spatial, temporal, and spatial-temporal data. Although random effects in GLMM and GAMM can be used to deal with dependency, such an approach is not optimal for spatial, temporal or spatial-temporal data. Although there were various tools available in R, they either required expertise knowledge or required extensive computing time (e.g. MCMC in WinBUGS or OpenBUGS). We therefore elected to stay away from teaching and writing about spatial, temporal, and spatial-temporal data analysis.

It was only after we became aware of material described in Lindgren et al. (2011) that we realised that GLMs and GLMMs, and all their zero-inflated cousins and smoothing cousins, can be extended to spatial, temporal, and spatial-temporal data.

Unfortunately, the literature describing the approach (Integrated Nested Laplace Approximation, abbreviated as INLA) is rather technical. A book published in 2015 by Blangiardo and Cameletti helped us understand the INLA world better. Although we find it an excellent book, it still requires a fair amount of statistical knowledge in order to fully comprehend the material.

Availability of the software package R-INLA has put the application of GLMs and GLMMs on spatial, temporal, and spatial-temporal data within the reach of every scientist. We therefore decided to extend our *Beginner's Guide* book series with a book on the use of R-INLA to analyse spatial, temporal, and spatial-temporal data.

Acknowledgements

We are greatly indebted to all scientists who supplied data for this book. Alexandre Roulin supplied the owl data. Bob Steidl provided the osprey data. Robert Cruikshanks allowed us to use the Irish pH dataset. Christophe Barbraud gave us the Adelie penguins data. Boudjéma Samraoui provided the White Storks data. Juan Timi gave us the Brazilian sand perch parasite data. Yusuke Fukuda provided the crocodile data. Matias Maggi supplied the honey bee mites data. Allesandro Ligas provided the crayfish data. Mette Mauritzen gave us the polar bears data. Chris Smeenk supplied the sperm whale data. Michael Reed gave us the Hawaiian bird data, and Helen Sofaer supplied the orange-crowned

warbler data. We also thank the following authors for making their data publically available: Petty et al. (2015) for the subnivium temperature data, Hopkins et al. (2013) for the chimpanzees data, Sturrock et al. (2015) for the otolith plaice data, Irl et al. (2015) for the plant richness data on La Palma (and who kindly emailed a modified data set), Crozier et al. (2011) for the sockeye salmon, Etheridge et al. (1998) for the historical CO_2 records, and Muller and van Woesik (2014) for the white-pox disease data.

We greatly appreciate the efforts of those who wrote R (R Development Core Team 2016) and its many packages. This book would not have been possible without the efforts of the R-INLA programmers (Rue et al. 2009; www.r-inla.org; Lindgren et al. 2011). We hope that they will keep up the excellent work.

We thank Joseph Hilbe and Thierry Onkelinx for helpful comments on an earlier draft. Special thanks to Christine Andreasen for editing this book.

Data sets and R code used in this book

All data sets used in this book may be downloaded from www.highstat.com/books.htm. All R code also may be downloaded from the website for this book. To open the ZIP file with R code, use the password **[omitted in this file]**

Cover art

The cover drawing is by Jon Thompson (www.yellowbirdgallery.org). Mr Thompson was born in 1939 to Irish parents and has lived most of his life in Scotland. In the 1980s, he was drawn to the Orkney Islands. He is continually inspired by the landscape and bird life of Orkney. He has been creating bird art for 30 years in a variety of media, including drawing, painting, sculpture, and jewellery, never attempting to reproduce nature, but to draw parallels with it. A close-up view of a bird feather is all the inspiration he needs.

Alain F Zuur,
Newburgh, Scotland

Elena N Ieno,
Alicante, Spain

Anatoly A Saveliev,
Kazan, Russia

February 2017

Contents

1 OVERVIEW OF THIS BOOK 1
1.1 VOLUMES I AND II 1
1.1.1 Volume I 1
1.1.2 Volume II 1
1.2 WHAT TYPE OF SPATIAL DATA DO WE ANALYSE IN THIS BOOK? 1
1.2.1 Areal and lattice data 1
1.2.2 Geostatistical data 2
1.2.3 Spatial point pattern data 3
1.3 OUTLINE OF THIS BOOK 3
1.4 PREREQUISITES 4
1.5 AVAILABILITY OF R CODE AND DATA 4

2 RECOGNISING STATISTICAL DEPENDENCY 5
2.1 PSEUDOREPLICATION 5
2.2 LINEAR REGRESSION APPLIED TO SPATIAL DATA 7
2.2.1 Irish pH data 7
2.2.2 Protocol from Zuur et al. (2016) 8
2.2.3 Visualisation of the experimental design 9
2.2.4 Data exploration 9
2.2.5 Dependency 12
2.2.6 Statistical model 15
2.2.7 Fit the model 16
2.2.8 Model validation 17
2.3 GAM APPLIED TO TEMPORAL DATA 21
2.3.1 Subnivium temperature data 21
2.3.2 Sources of dependency 22
2.3.3 The model 23
2.3.4 Model validation 24
2.4 GLMM APPLIED ON HIERARCHICAL AND SPATIAL DATA 26
2.5 TECHNICALITIES 28
2.5.1 Matrix notation 28
2.5.2 How is dependency causing problems? 31
2.6 DISCUSSION 32

3 TIME SERIES AND GLS 33
3.1 OSPREYS 33
3.2 COVARIANCE AND CORRELATION COEFFICIENTS 33
3.3 LINEAR REGRESSION MODEL 35
3.4 FOCUSSING ON THE RESIDUAL COVARIANCE MATRIX 35
3.5 DEPENDENCY AND THE COVARIANCE MATRIX 36
3.6 GLS: DEALING WITH TEMPORAL DEPENDENCY 39
3.6.1 Adelie penguins 39
3.6.2 Do we have dependency? 40
3.6.3 Formulation of the linear regression model 40

3.6.4 Application of the linear regression model *41*
3.6.5 R code for acf and variogram *45*
3.6.6 Formulation of the GLS model *46*
3.6.7 Implementation using the gls function *50*
3.7 MULTIPLE TIME SERIES 51
3.8 DISCUSSION 53

4 SPATIAL DATA AND GLS 55
4.1 VARIOGRAM MODELS FOR SPATIAL DEPENDENCY 55
4.2 APPLICATION ON THE IRISH pH DATA 57
4.3 MATÉRN CORRELATION FUNCTION 59

5 LINEAR MIXED EFFECTS MODELS AND DEPENDENCY 61
5.1 WHITE STORKS 61
5.2 CONSIDERING THE DATA (WRONGLY) AS ONE-WAY NESTED 62
5.3 FITTING THE ONE-WAY NESTED MODEL USING LMER 65
5.4 MODEL VALIDATION 67
5.5 SKETCHING THE FITTED VALUES 68
5.6 CONSIDERING THE DATA (CORRECTLY) AS TWO-WAY NESTED 69
5.7 APPLICATIONS TO SPATIAL AND TEMPORAL DATA 72
5.8 DIFFERENCE WITH THE AR1 PROCESS APPROACH 72

6 MODELLING SPACE EXPLICITLY 73
6.1 MODEL FORMULATION 73
6.2 COVARIANCE MATRIX OF THE SPATIAL RANDOM EFFECT 75
6.3 SPATIAL-TEMPORAL CORRELATION* 79

7 INTRODUCTION TO BAYESIAN STATISTICS 83
7.1 WHY GO BAYESIAN? 83
7.2 GENERAL PROBABILITY RULES 84
7.3 THE MEAN OF A DISTRIBUTION* 85
7.4 BAYES' THEOREM AGAIN 87
7.5 CONJUGATE PRIORS 88
7.6 MARKOV CHAIN MONTE CARLO SIMULATION 93
7.6.1 Underlying idea *93*
7.6.2 Installing JAGS and R2jags *94*
7.6.3 Flowchart for running a model in JAGS *94*
7.6.4 Preparing the data for JAGS *95*
7.6.5 JAGS code *96*
7.6.6 Initial values and parameters to save *98*
7.6.7 Running JAGS *99*
7.6.8 Accessing numerical output from JAGS *100*
7.6.9 Assess mixing *100*
7.6.10 Posterior information *101*
7.7 INTEGRATED NESTED LAPLACE APPROXIMATION* 103
7.7.1 Joint posterior distribution *103*
7.7.2 Marginal distributions *105*
7.7.3 Back to high school *107*
7.7.4 INLA *109*

7.8 EXAMPLE USING R-INLA 110
7.9 DISCUSSION 114

8 MULTIPLE LINEAR REGRESSION IN R-INLA 115
8.1 INTRODUCTION 115
8.2 DATA EXPLORATION 116
8.3 MODEL FORMULATION 117
8.4 LINEAR REGRESSION RESULTS 117
8.4.1 Executing the model in R-INLA 117
8.4.2 Output for the betas 117
8.4.3 Output for the hyper-parameters 119
8.4.4 Fitted model 123
8.5 MODEL VALIDATION 123
8.6 MODEL SELECTION 126
8.6.1 Should we do it? 126
8.6.2 Using the DIC 126
8.6.3 Out of sample prediction 131
8.6.4 Posterior predictive check 133
8.7 VISUALISING THE MODEL 135

9 MIXED EFFECTS MODELLING IN R-INLA TO ANALYSE OTOLITH DATA 139
9.1 OTOLITHS IN PLAICE 139
9.2 MODEL FORMULATION 140
9.3 DEPENDENCY 140
9.4 DATA EXPLORATION 141
9.5 RUNNING THE MODEL IN R-INLA 143
9.6 MODEL VALIDATION 146
9.7 MODEL SELECTION 149
9.8 MODEL INTERPRETATION 149
9.8.1 Option 1 for prediction: Adding extra data 150
9.8.2 Option 2 for prediction: Using the inla.make.lincombs 153
9.8.3 Adding extra data or inla.make.lincombs? 155
9.9 MULTIPLE RANDOM EFFECTS 155
9.10 CHANGING PRIORS OF FIXED PARAMETERS 156
9.11 CHANGING PRIORS OF HYPERPARAMETERS 158
9.12 SHOULD WE CHANGE PRIORS? 164

10 POISSON, NEGATIVE BINOMIAL, BINOMIAL AND GAMMA GLMS IN R-INLA 165
10.1 POISSON AND NEGATIVE BINOMIAL GLMS IN R-INLA 165
10.1.1 Introduction 165
10.1.2 Poisson GLM in R-INLA 166
10.1.3 Negative binomial GLM in R-INLA 172
10.1.4 Model selection for the NB GLM 175
10.1.5 Visualisation of the NB GLM 177
10.2 BERNOULLI AND BINOMIAL GLM 180
10.2.1 Bernoulli GLM 181

10.2.2 Model selection with the marginal likelihood *184*
10.2.3 Binomial GLM *185*
10.3 GAMMA GLM 187

11 MATÉRN CORRELATION AND SPDE 191
11.1 CONTINUOUS GAUSSIAN FIELD 191
11.2 MODELS THAT WE HAVE IN MIND 191
11.3 MATÉRN CORRELATION 192
11.4 SPDE APPROACH 197

12 LINEAR REGRESSION MODEL WITH SPATIAL DEPENDENCY FOR THE IRISH PH DATA 205
12.1 INTRODUCTION 205
12.2 MODEL FORMULATION 205
12.3 LINEAR REGRESSION RESULTS 206
12.4 MODEL VALIDATION 207
12.5 ADDING SPATIAL CORRELATION TO THE MODEL 208
12.6 DEFINING THE MESH FOR THE IRISH PH DATA 212
12.7 DEFINE THE WEIGHT FACTORS A_{IK} 216
12.8 DEFINE THE SPDE 218
12.9 DEFINE THE SPATIAL FIELD 218
12.10 DEFINE THE STACK 218
12.11 DEFINE THE FORMULA FOR THE SPATIAL MODEL 221
12.12 EXECUTE THE SPATIAL MODEL IN R 221
12.13 RESULTS 222
12.14 MODEL SELECTION 227
12.15 MODEL VALIDATION 228
12.16 MODEL INTERPRETATION 228
12.17 DETAILED INFORMATION ABOUT THE STACK* 232
12.17.1 Stack for the fitted model again *232*
12.17.2 Stack for the new covariate values *234*
12.17.3 Combine the two stacks *236*
12.17.4 Run the model *236*

13 SPATIAL POISSON MODELS APPLIED TO PLANT DIVERSITY 239
13.1 INTRODUCTION 239
13.2 DATA EXPLORATION 239
13.2.1 Sampling locations *239*
13.2.2 Outliers *241*
13.2.3 Collinearity *242*
13.2.4 Relationships *243*
13.2.5 Numbers of zeros *244*
13.2.6 Conclusions data exploration *244*
13.3 MODEL FORMULATION 244
13.4 GLM RESULTS 245
13.5 ADDING SPATIAL CORRELATION TO THE MODEL 248
13.5.1 Model formulation *248*

13.5.2 Mesh *248*
13.5.3 Projector matrix *253*
13.5.4 SPDE *254*
13.5.5 Spatial field *254*
13.5.6 Stack *254*
13.5.7 Formula *255*
13.5.8 Run R-INLA *255*
13.5.9 Inspect results *256*
13.6 SIMULATING FROM THE MODEL 262
13.7 WHAT TO WRITE IN A PAPER 265

14 TIME-SERIES ANALYSIS IN R-INLA 267
14.1 SIMULATION STUDY 267
14.2 TRENDS IN MIGRATION DATES OF SOCKEYE SALMON 269
14.2.1 Applying a random walk trend model *269*
14.2.2 Posterior distribution of the sigmas *272*
14.2.3 Covariates and trends *273*
14.2.4 Making the trend smoother *274*
14.3 TRENDS IN POLAR BEAR MOVEMENTS 280
14.4 TRENDS IN WHALE STRANDINGS 283
14.5 MULTIVARIATE TIME SERIES FOR HAWAIIAN BIRDS 285
14.5.1 Importing and preparing the data *285*
14.5.2 Data exploration *286*
14.5.3 Model formulation *287*
14.5.4 Executing the models *288*
14.5.5 Mixing Poisson and negative binomial distributions *295*
14.6 AR1 TRENDS 297
14.6.1 AR1 trend for regularly spaced time-series data *297*
14.6.2 AR1 trend for irregularly spaced time-series data *299*

15 SPATIAL-TEMPORAL MODELS FOR ORANGE CROWNED WARBLERS COUNT DATA 307
15.1 INTRODUCTION 307
15.2 POISSON GLM 308
15.3 MODEL WITH SPATIAL CORRELATION 312
15.4 SPATIAL-TEMPORAL CORRELATION: AR1 318
15.4.1 Why do it? *318*
15.4.2 Explanation of the model *318*
15.4.4 Simulating a spatial-temporal AR random field *320*
15.4.5 Implementation of AR1 model in R-INLA *323*
15.4.6 More detailed information on the code *326*
15.5 SPATIAL-TEMPORAL CORRELATION: EXCHANGEABLE 328
15.6 SPATIAL-TEMPORAL CORRELATION: REPLICATED 329
15.7 SIMULATION STUDY 330
15.8 DISCUSSION 333

16 SPATIAL-TEMPORAL BERNOULLI MODELS FOR CORAL DISEASE DATA 335
16.1 INTRODUCTION 335
16.2 BERNOULLI MODEL IN R-INLA 336
16.3 SPATIAL CORRELATED BERNOULLI MODEL 338
16.4 SPATIAL-TEMPORAL CORRELATED BERNOULLI MODEL 342

REFERENCES 347

INDEX 353

OTHER BOOKS 357