# Alain F Zuur

# Elena N Ieno

# Beginner's Guide to

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

Volume II: GAM and zero-inflated models

Published by Highland Statistics Ltd. Highland Statistics Ltd. Newburgh United Kingdom highstat@highstat.com ISBN: 978-0-9571741-4-6 First published in August 2018

© Highland Statistics Ltd.

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Highland Statistics Ltd., 9 St Clair Wynd, Newburgh, United Kingdom), except for brief excerpts in connection with reviews or scholarly analyses. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methods now known or hereafter developed is forbidden. The use in this publication of trade names, trademarks, service marks, and similar terms, whether or 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.

This book is copyrighted material from Highland Statistics Ltd. Scanning this book all or in part and distributing via digital media (including uploading to the internet) without our explicit permission constitutes copyright infringement. Infringing copyright is a criminal offence, and you will be taken to court and risk paying damages and compensation. Highland Statistics Ltd. actively polices against copyright infringement.

Although the authors and publisher (Highland Statistics Ltd., 9 St Clair Wynd, Newburgh, United Kingdom) have taken every care in the preparation and writing of this book, they accept no liability for errors or omissions or for misuse or misunderstandings on the part of any person who uses it. The authors and publisher accept no responsibility for damage, injury, or loss occasioned to any person as a result of relying on material included in, implied, or omitted from this book.

www.highstat.com

Opas and Omas pass away, memories to them stay!

– Alain F Zuur –

To Juan Carlos, Norma and Walter who have constantly encouraged me to achieve more challenging tasks in life. Thank you!

– Elena N Ieno –

#### Preface

Thanks to the likes of Netflix, Amazon Video and Hulu we can now binge-watch a whole television series in one weekend. In front of you, you have Volume II of *Beginner's Guide to Spatial, Temporal, and Spatial-Temporal Ecological Data Analysis with R-INLA*. Volume II, entitled *GAM and zero-inflated models*, is a continuation of Volume I. Volumes I and II consist of a total of 24 chapters. You can binge-read the whole thing in one weekend!

There are analyses in Volume II that we could not perform 10 years ago, simply because the required software did not exist. Thanks to R (R Core Team (2018), R-INLA (Rue et al. 2009) and a large number of packages in R we can now easily apply generalised linear models (GLM), generalised additive models (GAM), generalised linear mixed-effects models (GLMM), and generalised additive mixed-effects models (GAMM) on count data, continuous data, proportional data, and their zero-inflated cousins, with spatial, temporal and spatial-temporal correlation. We can do this for geo-statistical data and for areal data. We can even deal with natural barriers like islands or coastlines.

What was lacking was an explanation and illustration of these techniques for scientists not familiar with, or not interested in, detailed mathematics. There is where our two volumes fill a gap.

The data sets that are analysed in this volume are all real data sets, and each data set comes with its own problems. Some of these data sets were a major challenge even for statisticians to analyse. Yet, they are typical of what biologists tend to sample.

#### Acknowledgements

We are greatly indebted to all scientists who supplied data for this book. Federico Cortés provided the skate data. Alexandre Roulin supplied the owl data. We thank Rijkswaterstaat for allowing us to use the sandeel data. We thank Jagger et al. (2015) for making their data and R code publicly available. We thank Johan Craeymeersch, Ingrid Tulp and Nicola Tien for providing the sandeel data and cooperating on the sandeel chapter. We also thank David Fifield, April Hedd and Carina Gjerdrum for providing the seabird data and assisting with the seabird chapter. And finally, we thank Adel Heenan and Ivor Williams for making their data available online and working together with us on the coral reef chapter.

We greatly appreciate the efforts of those who wrote R (R Development Core Team 2018) and its many packages. This book would not have been possible without the efforts of the R-INLA programmers (Rue et al. 2009; <u>www.r-inla.org</u>; Lindgren et al. 2011). We hope that they will keep up the excellent work.

We would also like to thank Haakon Bakka for commenting on drafts of Chapters 18 and 20.

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 <u>www.highstat.com/books.htm</u>. All R code also may be downloaded from the website for this book. To open the ZIP files with R code, use the password: **< Omitted in online version>** 

### **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

> > August 2018

# Contents

Preface	v
ACKNOWLEDGEMENTS	v
DATA SETS AND R CODE USED IN THIS BOOK	VI
COVER ART	VI
CONTRIBUTORS	XIII
ABOUT THE AUTHORS	XV
17 INTRODUCTION TO VOLUME II	
17.1 Outline of Volume II	
17.2 AVAILABILITY OF THE R CODE AND DATA	
18 ZERO-INFLATED MODELS FOR COUNT DATA IN R-I	NT A
18 ZERO-INFLATED MODELS FOR COUNT DATA IN R-I	
18.1 THE POISSON DENSITY FUNCTION REVISITED	
18.2 THE ZERO-INFLATED POISSON MODEL	
18.2.1 Simulating ZIP data	
18.2.2 The ZIP density function and its mean and variance	
18.2.3 True and false zeros	
18.3 THE ZERO-ALTERED POISSON MODEL	
18.3.1 Two separate analyses	
18.3.2 The ZAP density function and its mean and variance	
18.3.3 ZIP versus ZAP	
18.4 ZIP models in R-INLA using a real skate data set	379
18.4.1 Poisson GLM	379
18.4.2 ZIP model formulation	380
18.4.3 Executing the ZIP model in R-INLA	380
18.4.4 Results of the ZIP model	381
18.4.5 Validation of the ZIP model	
18.4.6 Visualisation of the ZIP model	
18.5 ZAP MODELS IN R-INLA	
18.5.1 ZAP model formulation	
18.5.2 Simple ZAP model implementation in R-INLA	387
18.5.3 Covariates in both parts of the ZAP model:	
Approach 1	389
18.5.4 Covariates in both parts of the ZAP model:	202
Approach 2	
18.5.5 Validation of the ZAP model	
18.5.6 Assessing the number of zeros produced by the model	
18.5.7 Visualisation of the ZAP model	
18.6 ZINB AND ZINB MODELS	

19 SPATIAL CORRELATED AND ZERO-INFLATED SKATE	
DATA	.403
19.1 INTRODUCTION	.403
19.2 DATA EXPLORATION	.405
19.3 STARTING POINT: THE POISSON GLM	.407
19.3.1 Model formulation	.407
19.3.2 Executing the Poisson GLM in R-INLA	.407
19.3.3 Overdispersion	.407
19.3.4 How to continue?	
19.4 FORMULATION OF MODELS WITH SPATIAL CORRELATION	.410
19.4.1 Poisson GLM with spatial correlation	.410
19.4.2 ZAP model with spatial correlation	
19.4.3 ZANB model with spatial correlation	
19.5 CREATING A MESH	
19.5.1 Distances between sampling locations	
19.5.2 Using a non-convex mesh	
19.6 POISSON GLM WITH SPATIAL CORRELATION	.414
19.6.1 Projector matrix	
19.6.2 Defining the spde using PC priors	
19.6.3 Defining the stack	
19.6.4 Defining the formula	.417
19.6.5 Executing the model in R-INLA	
19.6.6 Numerical output of the spatial Poisson GLM	.418
19.6.7 Plotting the spatial random field	.418
19.6.8 Model validation	.421
19.6.9 Simulation study	.424
19.7 ZAP MODEL WITH SPATIAL CORRELATION	.426
19.7.1 Running the model in R-INLA	
19.7.2 Fitted values and Pearson residuals	.428
19.8 ZANB MODEL WITH SPATIAL CORRELATION	.430
19.8.1 Running the model in R-INLA	.430
19.8.2 Getting fitted values and Pearson residuals	.430
19.8.3 Model validation	
19.8.4 Results of the ZANB model with spatial correlation	
19.9 COMPARING ALL MODELLING RESULTS	
19.9.1 DICs and WAICs	
19.9.2 Sample variograms of Pearson residuals	
19.10 MODEL SELECTION APPLIED ON THE ZANB MODEL	
19.10.1 Optimal random structure	
19.10.2 Optimal fixed structure	
19.11 WHAT TO WRITE IN A PAPER	.436
20 GAM WITH CORRELATION AND ZERO INFLATION IN	
R-INLA USING OWL DATA	
20.1 VOCAL BEGGING BEHAVIOUR OF NESTLING BARN OWLS	.439
20.2 DATA EXPLORATION	
20.3 POISSON GLMM	.444

20.4 EASY-TO-UNDERSTAND SMOOTHERS	449
20.4.1 Moving average smoother	
20.4.2 LOESS smoother	
20.4.3 Linear, quadratic and cubic spline regression	
20.4.4 Basis of a smoother	
20.5 More complex smoothers	
20.5.1 Tent smoother	
20.5.2 Penalised tent smoother	
20.5.3 Penalised tent smoother as a mixed-effects model	
20.5.4 B-spline	
20.5.5 Cubic regression spline	
20.5.6 Low-rank thin-plate regression splines	
20.5.7 O'Sullivan spline	
20.6 DEFAULT SMOOTHERS IN R-INLA	.473
20.7 Poisson GAMM in R-INLA	
20.7.1 Model formulation	
20.7.2 Implementation of the tent spline	
20.7.3 Implementation of the B-spline	
20.7.4 Implementation of the cubic regression spline	
20.7.5 Implementation of a low-rank thin-plate	
regression spline	480
20.7.6 Implementation of the O'Sullivan spline	
20.7.7 Default smoothers from R-INLA	
20.7.8 Model validation	
20.7.9 Controlling the amount of smoothing	
20.7.10 Interactions with smoothers	
20.8 ZERO-INFLATED POISSON GAMM IN R-INLA	
20.9 ZIP GAMM WITH SPATIAL CORRELATION	
20.10 Optimal ZIP GAMM	
20.10 OPTIMAL ZIF OANIM	
21 GAM FOR ZERO-INFLATED AND SPATIAL-TEMPORAL	
CORRELATED SANDEEL DATA	
21.1 INTRODUCTION	
21.1.1 Sandeel	499
21.1.2 Variables	
21.2 DATA EXPLORATION	
21.2.1 Spatial position of sampling locations	501
21.2.2 Outliers	503
21.2.3 Collinearity	504
21.2.4 Zero inflation	505
21.2.5 Relationships	505
21.2.6 Data exploration conclusions	506
21.3 MODEL FORMULATION POISSON GLM	506
21.4 IMPLEMENTATION POISSON GLM IN R-INLA	507
21.4.1 R code for a Poisson GLM	507
21.4.2 Model validation	511
21.4.3 Model validation summary – What next?	517

21.5 IMPLEMENTATION POISSON GAM IN R-INLA	517
21.5.1 Which covariate will be a smoother?	518
21.5.2 What type of smoother to use?	519
21.5.3 R code for the Poisson GAM in R-INLA	
21.5.4 Plotting the two smoothers	522
21.5.5 Model validation Poisson GAM	
21.5.6 Summary model validation – What next?	
21.6 POISSON GAM WITH SPACE/TIME DEPENDENCY	
21.6.1 Model formulation	
21.6.2 Poisson GAM with spatial correlation in R-INLA	
21.6.3 Poisson GAM with spatial-temporal correlation	
21.6.4 Model validation	
21.7 Adding interactions for a smoother	
21.8. ZAP MODEL FORMULATION	
21.9 BERNOULLI PART OF THE ZAP GAM	
21.10 ZERO-TRUNCATED POISSON PART OF THE ZAP GAM	
21.11 OPTIMAL ZAP GAM	
21.12 DISCUSSION	
22 ZERO-INFLATED CONTINUOUS SEABIRD DATA	
22.1 INTRODUCTION	
22.2 DATA EXPLORATION	
22.2.1 Spatial locations	
22.2.2 Zero inflation	
22.2.3 Potential outliers	
22.2.4 Collinearity	555
22.2.5 Relationships	
22.3 FAILED APPROACH 1: LINEAR REGRESSION	557
22.3.1 Model formulation	557
22.3.2 Defining covariates and smoothers	558
22.3.3 Executing the regression model in R-INLA	559
22.3.4 Model validation	559
22.4 FAILED APPROACH II: LOGNORMAL DISTRIBUTION	
22.4.1 Linear regression applied on log-transformed data	561
22.4.2 Simulation study for zero inflation	563
22.5 INTRODUCTION TO ZERO-ALTERED GAMMA MODELS	568
22.5.1 ZAG model formulation	568
22.5.2 What is a gamma GLM?	569
22.6 DEFINING THE MESH AND SPATIAL RANDOM FIELD	571
22.6.1 Polygon for land	571
22.6.2 Mesh	
22.6.3 Penalised complexity priors for the spatial	
random field	576
22.7 ZAG MODEL IN R-INLA	
22.7.1 Define smoothers and covariates	578
22.7.2 Bernoulli GAM with spatial correlation	
22.7.3 Gamma GAM with spatial correlation	

22.7.4 Combining both components	590
22.8 SIMULATING ZERO-INFLATED CONTINUOUS DATA	
22.9 DISCUSSION	
23 CORAL REEF DATA AND SPATIAL BARRIER MODELS	
23.1 INTRODUCTION	
23.2 DATA EXPLORATION	
23.3 The beta model	602
23.3.1 Beta distribution	
23.3.2 Beta model	
23.3.3 Beta GAM with spatial correlation for	
the coral reef data	606
23.4 BETA GAM WITH SPATIAL CORRELATION IN R-INLA	
23.4.1 Distances between sites	
23.4.2 Defining a mesh	
23.4.3 Defining the spatial random field	
23.4.4 Defining the smoother and covariates	
23.4.5 Executing the beta GAM in R-INLA	
23.4.6 Extracting fitted values and residuals	
23.4.7 Covariate effects and spatial random field	
23.4.8 Problems with the spatial correlation	
23.5 Shooting holes in the mesh	
23.6 BARRIER MODEL USING SIMULATED DATA	
23.6.1 Polygon for the barrier	
23.6.2 Sampling locations	
23.6.3 Mesh	
23.6.4 Underlying principle of the barrier model	
23.6.5 Triangles inside the barrier	
23.6.6 Barrier polygon based on identified triangles	
23.6.7 Defining the barrier SPDE	
23.6.8 Simulate spatial correlated random effects	
23.6.9 Applying the barrier model on the simulated data	
23.7 BARRIER MODEL APPLIED ON BENTHIC COVERAGE DATA	
23.7.1 Defining a mesh	
23.7.2 Defining the spatial random field	
23.7.3 Defining the stack	
23.7.4 Defining and executing the beta GAM with barrier	
23.7.5 Comparing all models	
23.7.6 Spatial random field for the barrier model 23.8 ZERO-ALTERED BETA GAM	
23.8.1 Introduction	
23.9 DISCUSSION	038
24 ANALYSIS OF AREAL TORNADO DATA	
24.1 AREAL VERSUS GEOSTATISTICAL DATA	
24.2 Who are the neighbours?	
24.3 TORNADO DATA	646

24.4 MODEL SPECIFICATION	648
24.5 CAR CORRELATION	649
24.5.1 Random effects	
24.5.2 A flavour of the underlying mathematics	650
24.6 GAM WITH ICAR IN R-INLA FOR THE TORNADO DATA	654
24.6.1 Covariates in the link function	654
24.6.2 Running the Poisson GLM with iCAR in R-INLA	654
24.6.3 Covariate effects	
24.6.4 Spatial random effects	
24.6.5 Model validation	
24.7 ZERO-INFLATED GAM WITH ICAR CORRELATION	
24.8 CHANGING THE PRIORS AND THE BYM2 MODEL	
24.8.1 Hyperparamer of iCAR	
24.8.2 Penalised complexity prior	
24.8.3 The BYM model for spatial correlation	
28.8.4 The BYM2 model for spatial correlation	
24.9 SPATIAL-TEMPORAL CORRELATION FOR AREAL DATA	
24.9.1 Space + Time	
24.9.2 Four types of interactions	
24.9.3 Type I interaction	
24.9.4 Type II interaction	
24.9.5 Type III interaction	
24.9.6 Type IV interaction	
24.9.7 Discussion	686
APPENDIX A CREATING SPATIAL POLYGONS	687
A.1 FROM MAPS TO SPATIAL POLYGONS	
A.2 MESH WITH A BOUNDARY	
A.3 SPATIAL POLYGON FOR LAND	692
ADDENDLY D OTHED MODEL CTHAT WEDE CONCIDEDED	
APPENDIX B OTHER MODELS THAT WERE CONSIDERED FOR THE SKATE DATA	605
B.1 FORMULATIONS OF THREE MORE MODELS	
B.1.1 ZIP model with spatial correlation	
B.1.1 ZIF model with spatial correlation B.1.2 NB GLM with spatial correlation	
B.1.2 INB OLM with spatial correlation B.1.3 ZINB model with spatial correlation	
B.1.5 ZING model with spatial correlation B.2 Results of NB GLM with spatial correlation	
B.2.1 Fitting the NB GLM with spatial correlation	090
in R-INLA	606
B.2.2 Model validation	
B.2.3 Results	
B.3 OBSERVATION LEVEL RANDOM EFFECTS	
REFERENCES	
INDEX	
RODEXBOOKS BY HIGHLAND STATISTICS LTD	

## Contributors

#### Johan Craeymeersch

Wageningen Marine Research P.O. Box 77, 4400 AB Yerseke The Netherlands

#### Carina Gjerdrum

Canadian Wildlife Service Environment and Climate Change Canada Dartmouth, NS Canada

#### David A. Fifield

Wildlife Research Division Science and Technology Branch Environment and Climate Change Canada Mount Pearl, NL Canada

#### **April Hedd**

Wildlife Research Division Science and Technology Branch Environment and Climate Change Canada Mount Pearl, NL Canada

#### Adel Heenan

School of Ocean Sciences Bangor University Anglesey, LL59 5AB United Kingdom

#### Nicola SH Tien

Wageningen Marine Research P.O Box 68, 1970 AB Ijmuiden The Netherlands **Ingrid Tulp** Wageningen Marine Research P.O Box 68, 1970 AB Ijmuiden The Netherlands

## Ivor D Williams

NOAA Pacific Islands Fisheries Science Center Honolulu, HI 96818 USA

## About the Authors

#### Alain F Zuur

I was born 2 years before man landed on the moon. When I went to high school, I was not good in much except for mathematics and sport. The nice part about mathematics is that you either 'see it' or you don't. Tears for those who don't; unlimited opportunities for those who do!

I went to university in Groningen, which is in The Netherlands. Life is sometimes directed by chance. A guy named 'Duurt' once gave me a ride to the bus station. Even in Dutch 'Duurt' is a funny name! Just 3 seconds before the bus arrived, Duurt shouted, 'Why don't you try NIOZ for a traineeship; biology is fun!' I hadn't fancied biology much at high school. Still, I decided to give it a chance. I took a 4-hour bus trip from Groningen to the island of Texel and knocked on the door of The Netherlands Institute for Sea Research: 'Anyone interested in a mathematician looking for a traineeship?'

I ended up working with some nice folks at NIOZ and they even funded my PhD. I was based in Aberdeen (Scotland) while doing my PhD. The PhD itself was a huge learning curve. Everything I learned at university didn't work for biological data. ARMA models? Waste of time. Fourier analysis? Come back in 500 years. PCA? Doesn't work.

After finishing my PhD in the millennium year, I got a job at the Marine Laboratory in Aberdeen. Not that I fancy Aberdeen, but its surroundings are beautiful. At the Marine Lab the learning curve was even steeper. People came to my office with complicated designs and expected an answer by yesterday. My line manager at the lab made the catastrophic mistake of allowing me to do a 'homer', a private consultancy job. What started with one 'homer' quickly became multiple 'homers', including statistics courses. After a (short) while this became a second full-time job,

so in the year that Dr Who returned on the BBC it was 'adios' and I started to work full time for my own business, Highland Statistics. Core business activities are teaching courses, writing books (and the occasional paper) and consultancy.

Together with my colleague, Elena Ieno, Highland Statistics runs 20–25 statistics courses (covering a wide range of topics)

per year. This sounds like a lot but some of these courses are in very (!) nice places. We wrote three books with Springer, and when we got stuck in Northern Europe because of a volcanic eruption we decided to write a 'Beginner's Guide to ...' book series as self-publishing authors. That 'self-publishing' means paying statistical referees, editors, printing companies, artists, etc. But the best referees were our course participants who asked 10+ questions about every paragraph that we wrote. Writing these books and running these courses makes me feel that I never stopped my PhD work. Virtually every month there are new statistical

methods and new packages in R that are relevant.

Since 2017, my wife Nandani and I are the proud parents of twins Aidan and Naila. Having twins hasn't



slowed me down. Most of this book was written at airports and up in the sky ... and between nap times and feedings!

Alain F Zuur Highland Statistics Ltd. 9 St Clair Wynd AB41 6DZ Newburgh United Kingdom

#### Elena N leno

I have undergraduate and doctoral degrees in biology with a specialisation in marine ecology. I worked on wader feeding ecology on intertidal mudflats and carried out field research in Argentina (South

America) and Texel (The Netherlands). Shortly thereafter I joined a bio-diveristy and ecosystem functioning study group as part of my postdoctoral research experience in Scotland, UK. I was also an honorary research fellow in the School of Biological Sciences, Oceanlab at the University of Aberdeen.



It was when I first began my PhD studies that I became aware of the importance of statistical methods in designing, collecting and analysing data. My career continued with a strong focus on applied statistics. After finishing my postdoc in 2003, I joined Highland Statistics. It was there that I shifted from active research work to being exclusively dedicated to the teaching of statistics to undergraduate and postgraduate biology students.

My experience as a course instructor has taken me to a vast number of countries where I have learned ways to close the gap between statistics and biology. This shared worldwide teaching experience gave me the strength and motivation for writing books as part of a contribution to my team. I am the co-author of 10 books on the analysis of ecological data.

My other interests include traveling, trekking and sharing my passion for wildlife photography and conservation with my partner, Walter.

Elena N Ieno Highland Statistics Ltd. Box No. 82 Avda. Escandinavia 72 Local 6, BQ 4 Gran Alacant 03130 Santa Pola Alicante Spain