1 Introduction

This book is based on material we have successfully used teaching statistics to undergraduates, postgraduates, post-docs and senior scientists working in the field possibly best described as the 'environmental sciences'. The required background for these courses, and therefore this book, is some level of 'familiarity with basic statistics'. This is a very loose phrase, but you should feel reasonably comfortable with concepts such as normal distribution, p-value, correlation, regression and hypothesis testing. Any first-year university undergraduate statistics module should have covered these topics to the depth required.

The book is in two main parts. In the first part, statistical theory is explained in an applied context, and in the second part 17 case study chapters are presented. You may find it useful to start by reading the second part first, identify which chapters are relevant, and then read the corresponding theory chapters from part one. In fact, we anticipate this being the way most people will approach the book: by finding the case study that best matches their own ecological question and then using it and the matching theory chapters to guide their analysis.

1.1 Part 1: Applied statistical theory

In the first part, we discuss several techniques used in analysing ecological data. This part is then divided into six sections:

- Data exploration.
- Regression methods: linear regression, generalised linear modelling (GLM) generalised additive modelling (GAM), linear mixed modelling, generalised least squares (GLS) and multinomial logistic regression.
- Classification and regression tree models.
- Multivariate methods: principal component analysis, redundancy analysis, correspondence analysis, canonical correspondence analysis, principal coordinate analysis, discriminant analysis and (non-metric) multidimensional scaling.
- Time series techniques: auto- and cross-correlations, auto-regressive moving average models, deseasonalising, random walk trends, dynamic factor analysis (DFA), min/max auto-correlation factor analysis (MAFA), chronological clustering.
- Spatial statistics: spatial correlation analysis, regression extensions with spatial neigbourhood structure, variogram analysis and kriging.

The main statistical content of the book starts in Chapter 4 with a detailed discussion of data exploration techniques. This is the essential step in any analysis! In Chapter 5 we move onto a detailed discussion about linear regression and explain how it is the basis for more advanced methods like GLM, GAM, mixed modelling, GLS and multinomial logistic regression. Our experience from courses shows that students find it easier to understand GLM and GAM if you can demonstrate how these methods are an extension of something with which they are already familiar.

In Chapters 6 and 7 we explain GLM and GAM. The GAM chapter is the first chapter where we suggest some readers may wish to skip some of the more mathematical parts, in their first reading of the text. We could have omitted these more mathematical parts entirely, but as most books on GAM can be mathematically difficult for many biological and environmental scientists, we chose to include the key mathematical elements in this text. You would not be the first person to use a GAM in their PhD thesis, and then face a grilling in the viva (oral PhD defence) on 'smoothing splines', 'degrees of freedom', and 'cross-validation'. The optional sections in the GAM chapter will allow you to provide some answers.

Chapter 8 contains mixed modelling and tools to impose an auto-correlation structure on the data using generalised least squares. Our experience suggests that most datasets in biological and environmental studies require mixed modelling and generalised least squares techniques. However, the subject is complicated and few books approach it in a biological or environmental context. In Chapter 9 we discuss regression and classification tree models, which like mixed modelling, are techniques well suited to the difficult datasets often found in ecological and environmental studies.

In Chapters 10 to 15 we discuss multivariate analysis. Although there is only limited controversy over the application of univariate methods, there is considerable debate about the use of multivariate methods. Different authors give equally strong, but very different opinions on when and how a particular multivariate technique should be used. Who is right depends on the underlying ecological questions, and the characteristics of the data. And deciding on the best method and approach is the most important decision you will need to make. Sometimes, more than one method can be applied on your data, and it is here that most of our students start to panic. If two analyses show the same results, then this adds confidence to any real-world decisions based on them, but what if two methods give different results? One option is to cheat and only use the results that give the answer you were hoping for, because you believe this increases your chances of publication or getting a consultancy contract renewed. Alternatively, you can try and understand why the two methods are giving different results. Obviously, we strongly advocate the second approach, as working out why you are getting different results can greatly improve you understanding of the underlying ecological processes under study.

In Chapters 16 and 17 we discuss time series techniques and show how auto-correlation can be added to regression and smoothing models. We also present auto-regressive integrated moving average models with exogenous variables

(ARIMAX) and deseasonaling time series. In Chapter 17 we discuss methods to estimate common trends (MAFA and dynamic factor analysis).

In Chapter 18, spatial statistics is discussed. Variograms, kriging and adding spatial correlation to regression and smoothing techniques are the key techniques.

Using several methods to analyse the same dataset seems to go against the convention of deciding your analysis methods in advance of collecting your data. This is still good advice as it ensures that data are collected in a manner suitable for the chosen analytical technique and prevents 'fishing' through your data until you find a test that happens to give you a nice 'p' value. However, although this approach may well be appropriate for designed experiments, for the types of studies discussed in the case study chapters, it is the ecological question and the structure of the collected data that dictate the analytical approach. The ecological question helps decide on the type of analysis, e.g., univariate, classification, multivariate, time series or spatial methods, and the quality of the data decides the specific method (how many variables, how many zeros, how many observations, do we have linear or non-linear relationships), which can only be addressed by a detailed data exploration. At this point you should decide on the best methodological approach and stick with it. One complication to this approach is that some univariate methods are extensions of other univariate methods and the violation of certain statistical assumptions may force you to switch between methods (e.g., smoothing methods if the residuals show non-linear patterns, or generalised linear modelling if there is violation of homogeneity for count data). This is when it becomes crucial to understand the background provided in the first part of the book

Returning to the question set earlier, in some scientific fields, it is the general belief that you need to formulate your hypothesis in advance and specify every step of the statistical analysis before doing anything else. Although we agree with specifying the hypothesis in advance, deciding on the statistical methods before seeing the data is a luxury we do not have in most ecological and environmental studies. Most of the time you are given a dataset that has been collected by someone else, at some time in the distant past, and with no specific design or question in mind. Even when you are involved in the early stages of an ecological experiment, survey or monitoring programme, it is highly likely that the generated data are so noisy that the pre-specified method ends up unsuitable and you are forced to look at alternatives, As long as you mention in your report, paper or thesis what you did, what worked, and what did not work, we believe this is still a valid approach, and all too often the only approach available.

1.2 Part 2: The case studies

In part two we present 17 case study chapters. The data for each case study are available online (www.highstat.com). Each chapter provides a short data introduction, a data exploration, and full statistical analysis, and they are between 14 and 26 pages long. The aims of the case studies are to:

- Illustrate most of the statistical methods discussed in the theoretical chapters using *real* datasets, where the emphasise is on 'real'. All datasets have been used for PHD theses (9 of the 17 case study chapters), consultancy contracts or scientific papers. These are not toy datasets!
- Provide a range of procedures that can be used to guide your own analysis. Find a chapter with a dataset similar to your own and use the same steps presented in that chapter to lead you through your own analysis.
- Show you, step by step, the decisions that were made during the data analysis. What techniques best suit this particular dataset and the ecological question being asked? How can the data be modified so that a standard statistical technique can be applied? What should you look for when deciding on a certain method; why additive modelling and not linear regression?

Of course the approaches presented in these chapters reflect our own approach to data analysis. Ask another statistician and they may well suggest something different. However, if there is a strong pattern in your data, then regardless of approach it should be detected.

The case studies have been selected to cover a wide diversity of ecological subjects: marine benthos, fisheries, dolphins, turtles, birds, plants, trees, and insects. We have also tried to use data from a range of continents, and the case studies come from Europe, North America, South America, Asia and Africa. We would have liked to include a wider range of data and continents, but finding datasets that were available for publication was difficult. We are therefore especially grateful to the owners of the data for letting us include them in the book and making them available to our readers.

The case study chapters are divided into four groups and follow the same structure as the theory sections:

1. Case study chapters using univariate techniques.
2. Case study chapters using multivariate techniques.
3. Chapters using time series methods.
4. Chapters using spatial statistics.

Each chapter illustrates a particular range of methods, or a specific decision process (for example the choice between regression, additive modelling, GLM or GAM). We also have some chapters where a series of methods are applied.

Univariate case study chapters

The first case study investigates zooplankton measured at two locations in Scotland. The aim of the chapter is to show how to decide among the application of linear regression, additive modelling, generalised linear modelling (with a Poisson distribution) or generalised additive modelling. We also discuss how we used smoothing methods to gain an insight into the required sample size: essential information when designing a cost-effective, but still scientifically robust study.

In the second case study, we go further south, to an estuary in Portugal, where we look at flatfish and habitat relationships. The first time we looked at these data,

it was a "nothing works" experience. Only after a transformation to presence-absence data, were we able to obtain meaningful results. The methods used in this chapter are logistic regression models (both GLM and GAM) and classification tree models.

We then present two case studies using (additive) mixed modelling techniques. The first one is about honeybee pollination in sunflower commercial hybrid seed production (Argentina). We also apply mixed modelling to investigate the abundance of Californian wetland birds (USA) in relation to straw management of rice fields. This chapter shows that if you wrongly ignore auto-correlation in your data, all the parameters can end up as significant. But by including an auto-correlation structure on the error component, the same parameters end up as only borderline significant!

In the fifth and sixth case studies, we apply classification methods. First, we use bird data from a Dutch study. The birds were recorded by radar and surveyed on the ground by biologists. The aim of the analysis was to investigate whether radar can be used to identify bird species, and this was investigated using classification trees. In the other classification chapter, we are back to fish and demonstrate the application of neural networks. The aim of this chapter is to identify discrete populations or stocks of horse mackerel in the northeast Atlantic by using a neural network to analyse parasite presence and abundance data. This chapter is different from the other case studies in that neural networks are not discussed in the preceding theory chapters. However, it is a popular method in some ecological fields and this chapter gives a valuable insight into its value and application.

Multivariate case study chapters

In the seventh case study, the first multivariate one, we look at plant species from the western Montana landscape (USA). Classic multivariate methods like non-metric multidimensional scaling (NMDS) and the Mantel test are used. We also apply GLS on a univariate diversity index and take into account the auto-correlation structure. Several auto-correlation structures are investigated.

In the next case study, we analyse marine benthic data from the Netherlands. This country lies below sea level, and one way of defending the country from the sea is to pump sand onto the beach. This process is also called beach re-nourishment or beach re-charge, and this chapter investigates the impact this process might have on the beach living animals. Several univariate and multivariate methods are considered: GAM, redundancy analysis and variance partitioning. Results of this work were used by the Dutch authorities to improve their coastal management procedures.

In case study nine, Argentinean zoobenthic data are used to show how easy it is to cheat with classic multivariate methods like NMDS and the Mantel test. We also discuss a special transformation to visualise Chord distances in redundancy analysis (RDA).

In the tenth case study, aspects of principal component analysis (PCA) are discussed and illustrated using fatty acid data from stranded dolphins. Covariance vs. correlation, how many PCs and normalisation constraints are all discussed and a

biplot is interpreted. An alternative to PCA that gives simplified interpretations is also included.

In the next chapter, the focus is on morphometric data analysis. PCA, together with landmark data analysis, is applied on skull measurements of turtles.

The twelfth case study, explores Senegalese savanna tree distribution and management using satellite images, and RDA, plus additive modelling to verify the RDA results.

In the last multivariate chapter, Mexican plant data are analysed using canonical correspondence analysis. The role of an invasive species is investigated.

Time series case study chapters

In the first time series case study, Portuguese fisheries landing trends are analysed. The main aim of this chapter is to estimate common trends using DFA and MAFA. In another time series case study chapter, groundfish research survey data from the northwest Atlantic are analysed and the effects of time lags are explored using MAFA and DFA. The techniques are used to highlight the importance of scale and to explore the complex dynamics of a system that has experienced drastic change. In the third time series case study, effects of sea level rise on Dutch salt marsh plant species are analysed using additive mixed modelling and mixed modelling methods that allow for auto-correlation. And we conclude the time series chapters with endangered Hawaiian birds and estimate common trends by applying a form of intervention analysis to detect the effect of management actions.

Spatial case study chapter

The only spatial chapter is on Russian tree data and demonstrates various spatial analysis and interpolation techniques for tree data (including variography, kriging, and regression models with spatial correlation structure).

1.3 Data, software and flowcharts

Chapter 2 discusses data management and software. About 95% of the statistical analyses in this book were carried out in the low-budget software package Brodgar, which can be downloaded from www.brodgar.com. It is a user-friendly 'click-and-go' package that also has a link to the R software (www.r-project.org). However, most statistical analysis can be carried out in other excellent software packages as well, and we discuss our experience with some of them. We also provide flowcharts in this chapter to readers with an overview of the methods.

In Chapter 3, we discuss our experience with teaching the material described in this book and give general recommendations for instructors.