# 3 Advice for teachers

## 3.1 Introduction

In this chapter, we discuss our experience in teaching some of the material described in this book. Our first piece of advice is to avoid explaining too many statistical techniques in one course. When we started teaching statistics we tried to teach univariate, multivariate and time series methods in five days to between 8 and 100 biologists and environmental scientists. We did this in the form of in-house courses, open courses and university courses. The audiences in institutional in-house and open courses typically consisted of senior scientist, post-docs, PhD-students and a few brave MSc students. The university courses had between 50 and 100 PhD or MSc students. The courses covered modules from data exploration, regression, generalised linear modelling, generalised additive modelling, multivariate analysis (non-metric multidimensional scaling, principal component analysis, correspondence analysis, canonical correspondence analysis, redundancy analysis) and time series. Although these 'show-me-all' courses were popular, the actual amount of information that participants were able to fully understand was far less than we had hoped for. It was just too much information for five days (40 hours).

We now teach several modules across all levels of expertise, including large groups (up to 100) of undergraduate and first-year PhD-students. The teaching for the univariate modules is now broken down into the following components (1 day is 8 hours):

- Data exploration (1 day). The emphasis in this module is on outliers and data transformations. Most students will be obsessed by the idea of normality of data (Chapter 4). This idea is so strong that some will even make QQ-plots of explanatory variables. We suggest instructors emphasise that normality means normality at each X value (Chapter 4). As not all datasets have enough replicates to make histograms for the data at each X value, it may be better to convince the students to first apply linear regression and then test the model for normally distributed residuals. It may also be an option to expose students to the idea that normality is one underlying assumption of linear regression, but it is not the most essential one (Chapter 5). We suggest using no more than three datasets as students complain that it is difficult to remember all the details of the data. Any dataset used in the case study chapters would be suit-

able. The emphasis of this module should be outlier detection, transformations and collinearity between explanatory variables. Do not forget to explain the concepts (and differences) of interaction, collinearity and confounding!

- Regression (1 day). In every course we ask the participants the question: 'Do you understand linear regression?' Three quarters will respond positively, but most will fail to identify the four underlying assumptions of linear regression, and many will have never seen a graph of residuals plotted against explanatory variables. Because generalised linear modelling (GLM) and generalised additive modelling (GAM) are considerably easier to explain once the students understand the principle of Figures 5.5 and 5.6, we suggest you spend a lot of time establishing a sound understanding of regression. The challenge is to get the students to understand the numerical output and how to apply a proper model validation process. Homogeneity and residuals versus each explanatory variables are the key points here. The flowchart in Figure 3.1 is useful in this stage as it explains to the student how to decide among linear regression, GLM and GAM.
- Generalised linear modelling (1 day Poisson and half a day logistic regression). Many students find this subject difficult, especially with logistic regression. We suggest starting the students on an exercise with only one explanatory variable so that they can see the model fit in the form of a logistic curve.
- Generalised additive modelling (1 day). The second part of the GAM chapter is slightly more technical than the first part. If the audience consists of biologists, we suggest explaining only the underlying concepts of splines, degrees of freedom and cross-validation. We chose to present the more technical aspects of GAM in this book simply because it is difficult to find it in textbooks for biologists. Some PhD students that we have supervised were interrogated during their viva on GAM.
- Mixed modelling and generalised least squares (1 day). Having completed this book, we realised that the data in nearly every case study could have been analysed with mixed modelling or generalised least squares (GLS). In fact, most ecological datasets may require mixed modelling methods! Yet, it is highly complicated and difficult to explain to biologists.

The multivariate material also takes five days (40 hours) to explain. Relevant topics are measures of association (1 day), ANOSIM and the Mantel test (1 day), principal component analysis and redundancy analysis (1 day), correspondence analysis and canonical correspondence analysis (0.5 days) and half a day for methods like non-metric multidimensional scaling and discriminant analysis. We strongly advise combining the multivariate module with a day on data exploration. The important aspect of the multivariate module is to teach students when to choose a particular method, and this is mainly driven by the underlying question, together with the quality of the data (e.g., a species data matrix with many zero observations and patchy species means non-metric multidimensional scaling).

Figure 3.3 gives an overview on how most of the multivariate techniques discussed in this book fit together. Various case study chapters emphasise the options and choices that have to be made. Note that various other methods exist, e.g., dis-

tance based redundancy analysis (Legendre and Anderson 1999), but these are not discussed in this book.

Time series analysis is a difficult, but popular, subject. We tend to start with data exploration (1 day), followed by a brief summary of linear regression (0.5 day), and then we introduce the idea of having an auto-correlation structure on the data using generalised least squares (0.5 days). Once students are used to the concept of auto-correlation, topics such as auto-correlation, cross-correlation and auto-regressive moving average models can be introduced (0.75 day). Seasonality is another important subject and takes at least two hours to explain. More specialised methods like min/max auto-correlation factor analysis and dynamic factor analysis are popular, and it is important to emphasise the differences between them. It might also be an option to explain generalised additive modelling and regression methods and the possibility of adding an auto-correlation structure on the errors, leading to methods like generalised additive mixed modelling, generalised least squares and generalised linear mixed modelling (which is not discussed in this book).

As to the spatial modules, we tend to teach these methods in a follow-up course to students who are familiar with GAMs, regression, etc. Knowledge of time series, or at least the concept of auto-correlation, helps. Figure 3.3 shows a decision tree that can be used to explain how to decide upon the most appropriate method. The final decision on choice of method comes down to the small technical details and is part of the model validation process. This is expanded in the case study chapters.

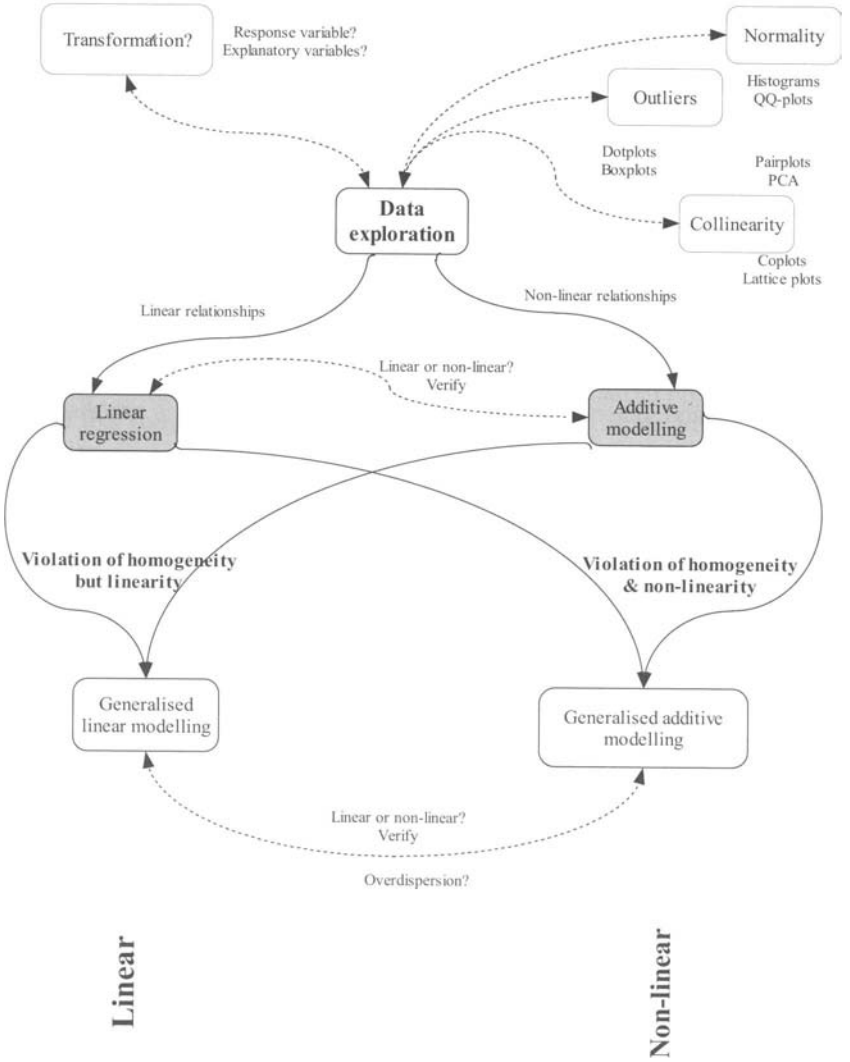


Figure 3.1. Flowchart showing how linear regression, additive modelling, generalised linear modelling (using the Poisson distribution and log-link function) and generalised additive modelling (using the Poisson distribution) are related to each other. In linear regression, violation of homogeneity means that the GLM with a Poisson distribution may be used. Normality but non-linear relationships (as detected for example by a graph of the residuals versus each explanatory variable) means that additive modelling can be applied. Non-linear relationships and violation of the normality assumption means a GAM with a Poisson distribution. The graph will change if another link function or distribution is used.

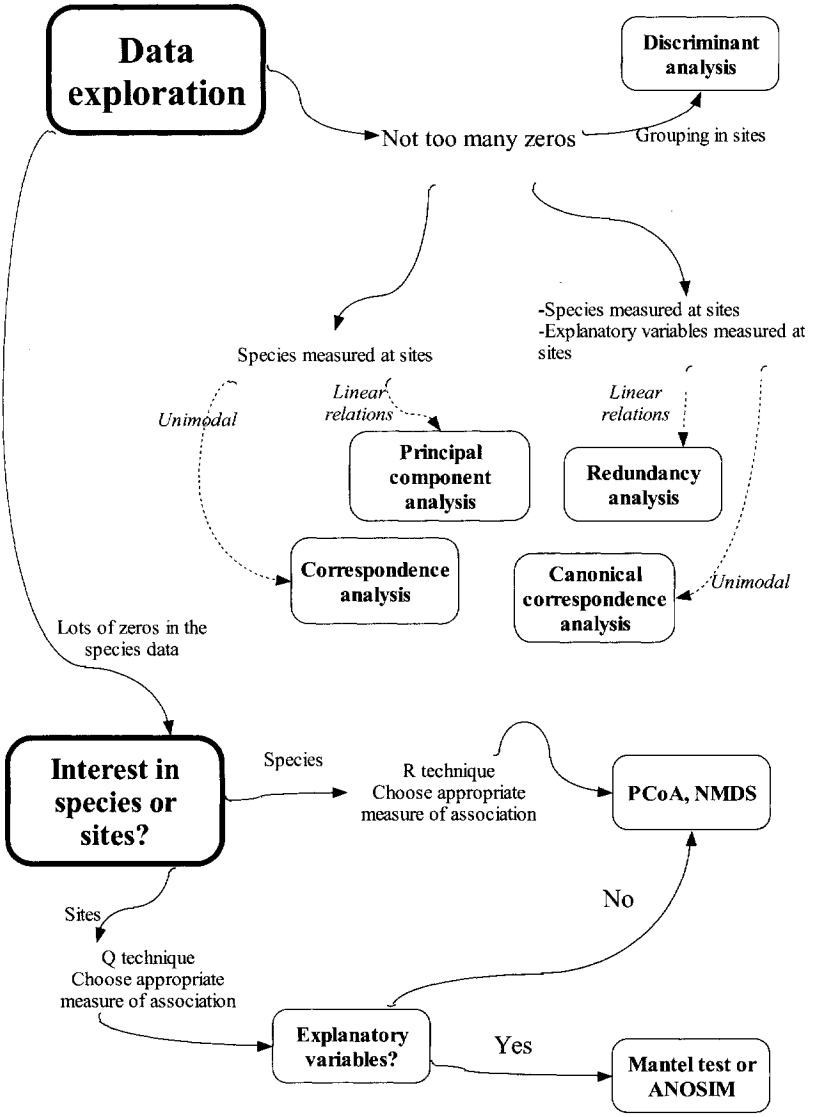


Figure 3.2. Overview of multivariate analysis methods discussed in this book. We assume that the data matrix consists of species measured at sites. If this is not the case (e.g., the data contain chemical variables), it becomes more tempting to apply principal component analysis, correspondence analysis, redundancy analysis or canonical correspondence analysis. If the species data matrix contains many zeros and double zeros, this needs to be taken into account, by choosing an appropriate association matrix and using principal co-ordinate analysis (PCoA), non-metric multidimensional scaling (NMDS), the Mantel test or ANOSIM.

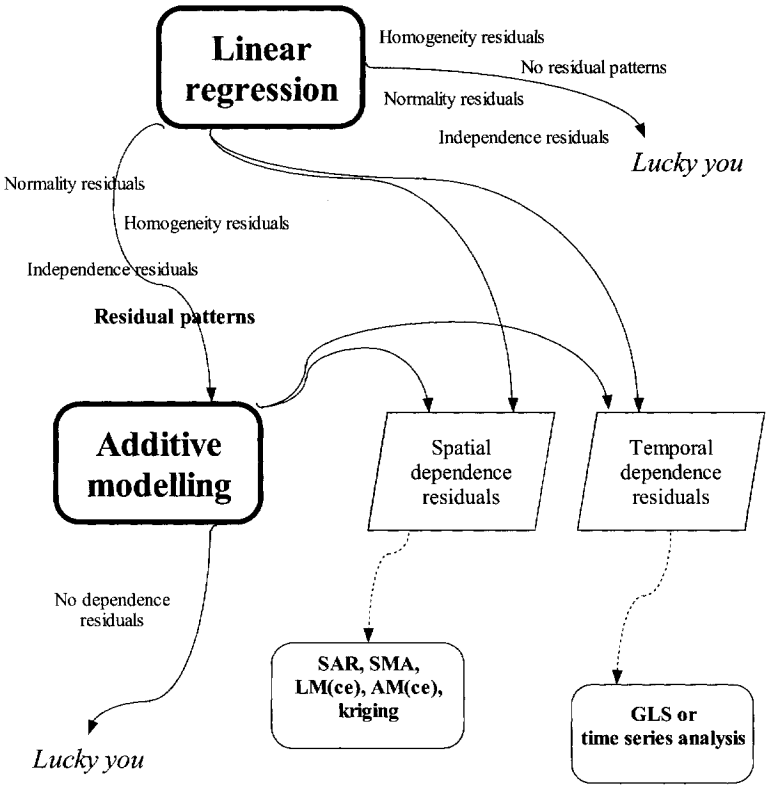


Figure 3.3. Decision tree to choose between linear regression models and additive models, and in case of temporal dependence GLS or time series methods can be applied. If there is spatial dependence in the additive or linear regression models, then SAR, SMA, LM(ce), AM(ce) or kriging techniques should be applied.