

Model-based thinking for community ecology

David I. Warton · Scott D. Foster ·
Glenn De'ath · Jakub Stoklosa · Piers K. Dunstan

Received: 29 January 2014 / Accepted: 30 May 2014 / Published online: 19 November 2014
© Springer Science+Business Media Dordrecht 2014

Abstract In this paper, a case is made for the use of model-based approaches for the analysis of community data. This involves the direct specification of a statistical model for the observed multivariate data. Recent advances in statistical modelling mean that it is now possible to build models that are appropriate for the data which address key ecological questions in a statistically coherent manner. Key advantages of this approach include interpretability, flexibility, and efficiency, which we explain in detail and illustrate by example. The steps in a model-based approach to analysis are outlined, with an emphasis on key features arising in a multivariate context. A key distinction in the model-based approach is the emphasis on

diagnostic checking to ensure that the model provides reasonable agreement with the observed data. Two examples are presented that illustrate how the model-based approach can provide insights into ecological problems not previously available. In the first example, we test for a treatment effect in a study where different sites had different sampling intensities, which was handled by adding an offset term to the model. In the second example, we incorporate trait information into a model for ordinal response in order to identify the main reasons why species differ in their environmental response.

Keywords Community-level modelling · Four-corner problem · Model checking · Multivariate analysis · Ordination · Species distribution models

Communicated by P. R. Minchin and J. Oksanen.

D. I. Warton (✉) · J. Stoklosa
School of Mathematics and Statistics and Evolution &
Ecology Research Centre, The University of New South
Wales, Sydney, NSW 2052, Australia
e-mail: David.Warton@unsw.edu.au

S. D. Foster · P. K. Dunstan
CSIRO's Wealth from Oceans Flagship, Hobart, TAS,
Australia

S. D. Foster
CSIRO's Division of Computational Informatics, Hobart,
TAS, Australia

G. De'ath
Australian Institute of Marine Science, Cape Ferguson,
QLD, Australia

Introduction

In ecology, a core concern historically has been the study of how community structure changes in response to changes in the environment. A key tool in such studies has been abundance data simultaneously collected on a suite of taxa to make inferences about communities at particular locations. We refer to these data as multivariate abundance data. In this paper, we will review analysis methods that involve specifying a statistical model for the observed multivariate abundance data (hereafter “model-based approach”). This

is a difficult problem because the number of taxa may be large and often exceeds the number of sites sampled, and because the data often have a large proportion of zeros, typically rendering classical multivariate techniques (Anderson 2003) inappropriate. Some important first steps towards model-based approaches to multivariate analysis were made by David Goodall and contemporaries in the development of Gaussian ordination (Gauch et al. 1974; Goodall and Johnson 1982), although the method only specified a model for mean response, and stopped short of specifying a plausible distribution for the observed abundance data.

Over the last half century, technological advances and improvements in computational power have facilitated extraordinary changes in both the theory and practice of statistics. A conspicuous example known to many plant ecologists is Markov Chain Monte Carlo (MCMC) methods, which readily enable Bayesian inference for complex ecological models (e.g. Cressie et al. 2009). But exciting advances have been made in non-Bayesian analysis also. In the context of community analyses, changes in technology are driving a movement towards model-based approaches to multivariate analysis, Bayesian and non-Bayesian, with recent years seeing the development of many such tools for use with multivariate abundance data (ter Braak et al. 2003; Gelfand et al. 2005; Elith and Leathwick 2007; Yee 2010; Dunstan et al. 2011; Ives and Helmus 2011; Wang et al. 2012; Foster et al. 2013; Pledger and Arnold 2014). These methods are typically computationally intensive, but have a number of advantages, including the flexibility to tailor models more closely to the properties of the data and to the research questions of interest. We believe model-based approaches are a significant new movement which can provide important new insights into community structure not previously available.

A particular area where model-based approaches are developing rapidly is in the species distribution modelling literature, where model-based approaches are widely used in single-species gradient modelling (Elith and Leathwick 2009). Extensions of such methods to the multivariate case enable simultaneous inference across an assemblage of species (Ovaskainen and Soininen 2011, for example), and study of the reasons why species differ in their response to gradients (such as in Pollock et al. 2012). Such extensions are often referred to as community-level

models (Ferrier and Guisan 2006), and are somewhat synonymous with model-based approaches to multivariate analysis as we describe here.

While the model-based paradigm is a central thrust of modern statistical science, approaches to multivariate analysis in ecology previously used are a significant departure from this, typically being based on: (1) matrices of pairwise dissimilarities (Anderson 2001, for example), (2) generalisations of correspondence analysis (ter Braak 1986), or (3) redundancy analysis (Legendre and Legendre 2012). All these methods are purely algorithmic, in the sense that they are defined via algorithms rather than via models, they do not explicitly take into account the statistical properties (i.e. inherently random nature) of the data, and the link to theory or testable hypotheses is weak. Not accounting for the statistical properties of data can have important consequences, when the properties of the data interact with the method; for example, when strong mean–variance trends are present but unaccounted for (Warton et al. 2012). Not having the capacity to explicitly link to theory also can have important consequences, with opportunities missed for insight into ecological structure and process. If methods based on pairwise dissimilarities are the only ones considered then the research questions that can be reliably answered are constrained; questions regarding a species', or group of species', the presence or abundance at a site are difficult to answer in such a framework. The main reason for the departure from a model-based approach, as stated in Anderson (2001) and elsewhere, is that suitable models had not been designed to handle multivariate abundance data. This may have been true in the past, but it is not true any longer.

In this paper, we make the case for the use of model-based approaches to multivariate analysis of community data, and advantages of this approach are outlined, from both the ecological and statistical perspectives. In our opinion, model-based approaches improve our understanding of community ecology by explicitly accounting for the uncertainty inherent in ecological data, by a clear specification of the assumptions of the analysis, and by offering the ability to make formal inference. The steps in a model-based approach to analysis are explained, with an emphasis on key features arising in a multivariate context. Examples are presented that illustrate how a model-based approach can provide insights into ecological problems not previously available.

Table 1 Examples of (a) Model-based software packages on R that can handle multivariate data (b) Common distributions for different response types

Purpose	Software	Example reference
(a) Model-based software on R		
Constrained ordination, predictive modelling	VGAM	Yee (2010)
Hypothesis testing for multivariate data	mvabund	Wang et al. (2012)
Classifying species by environmental response	speciesMix	Hui et al. (2013)
Classifying sites by species and environmental data	RCPmod	Foster et al. (2013)
Classifying sites and species without environmental data	from author	Pledger and Arnold (2014)
Predictive modelling	lme4 and other GLMM software	Ives and Helmus (2011)
Data type	Distribution (link)	Example reference
(b) Example distributions that can be used for responses of different types		
Presence/absence	Binomial (logit)	Dunstan et al. (2011)
Count	Poisson or negative binomial (log)	Wang et al. (2012)
Ordinal	Multinomial (logit; “Proportional odds”)	Sect. 3.2
Biomass	Tweedie (log)	Dunstan et al. (2013)

In principle, any distribution in (b) can be used in combination with models in (a), although different packages have different functionalities

Model-based approaches for ecological questions

A model-based approach to analysis is defined in this paper as one which involves the specification of a statistical model for the data that were *observed*, in order to explicitly answer ecological questions. Modelling the observed data (or more precisely, the underlying data-generating mechanism) is a key distinction compared to analysis methods which start with summary statistics calculated from the observed data, such as pairwise measures of dissimilarity (Anderson 2001, for example). A statistical model has an explicit mathematical representation that accounts for the systematic variation in the data (the signal) as well as the randomness in the data (noise). The systematic variation is what the analyst usually wants to extract from the data, while the randomness is an inherent property of all ecological data that needs to be accounted for in order to see the signal. This allows a clear understanding of when the data supports our hypothesis and when it does not. A model-based approach is already widely used in ecology for univariate analysis (e.g. Clark 2007; Bolker et al. 2009; Cressie et al. 2009; Zuur et al. 2010), and until recently, multivariate analysis has seemingly been an exception to the rule. However, a number of model-based approaches to multivariate analysis have

recently been proposed, as summarised in Table 1a. Use of such a statistical model has a number of direct benefits, including interpretability, flexibility and efficiency, as explained below.

Interpretability

When the observed data have been modelled directly, model outcomes can be interpreted in the context of the observed data and inferences can be made about the processes underpinning what is observed. Important ecological relationships can be directly quantified and important quantities can be predicted (under current and future scenarios). Further, any analysis tool will work better in some situations than others, and an advantage of models is that they involve clearly stating assumptions and hence identifying the conditions under which they can be expected to work well. Such assumptions can be checked during data analysis.

Flexibility

A variety of different types of data can be modelled, and a variety of different types of research questions can be answered, using readily available modelling tools. There is the capacity to explicitly incorporate

ecological theory into models (e.g. Etienne 2007; Clark 2007), and one has the capacity to choose between competing theories (model selection) as well as to generate new ecological theories from data.

Efficiency

Formal inference is enabled through well-established statistical theory, and when the model provides a good fit to data, this theory typically offers optimality properties (Anderson 2003) suggestive of good performance. Desirable statistical properties, including high efficiency compared to alternatives, have been demonstrated empirically for model-based approaches to multivariate analysis in ecology (Warton 2011; Warton et al. 2012).

The process of model-based analyses is structured, sequential (Fig. 1) and well-established (e.g. Neter et al. 1996). A key point of distinction for a model-based approach to algorithmic approaches is the pivotal role of data properties in informing the method of analysis ultimately used. Specifically, the properties of the data determine the type of model proposed, and diagnostic tools are used to gauge how well the model reflects the data, as described later. The diagnostic step is very important as it indicates potential improvements to the model (Fig. 1) such that model assumptions better reflect data properties (e.g. accounting for overdispersion), which ultimately produces a more robust and defensible inference.

A second important point of difference from algorithmic approaches concerns the strength of the connection between the research question and the analysis method. The research question is typically embedded explicitly in the model used for analysis—for example, as a parameter to be estimated. Conclusions that are drawn from analysis then can involve specific, quantitative answers to the research question (“formal inference”). An explicit connection between an analysis approach and the research question is typically not possible without specifying a model, and thus conclusions drawn from algorithmic methods of analysis can be somewhat opaque.

The steps in a model-based analysis, as described in Fig. 1, will be discussed in the following sections—these same steps also apply to Bayesian analyses, but with some additional considerations concerning prior specification.

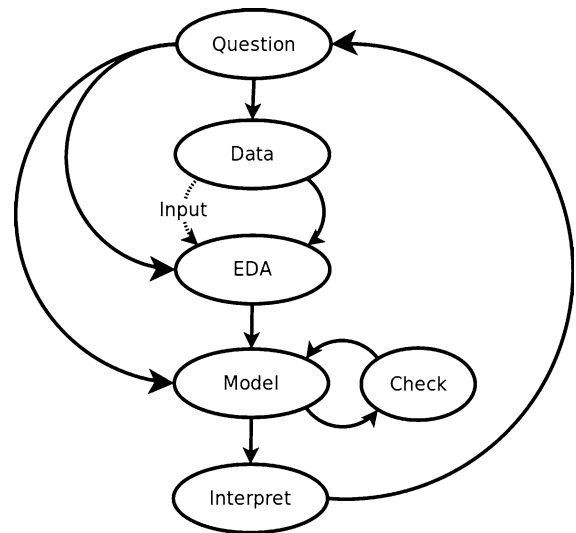


Fig. 1 Flow of model-based analysis. Each ellipse refers to understanding an aspect of the modelling process. EDA stands for “exploratory data analysis”. See Sect. 2 for details on any of these analysis steps

Identifying the ecological question

The motivation for any ecological study can usually be framed in terms of answering one or two key questions. Correctly identifying the question(s) and formalising them within a statistical framework focuses on questions that can be resolved using data. The questions, once specified, not only determine how the data should be collected or selected for analysis, but they also influence how the data should be analysed.

The number of possible questions is limited only by our imagination, and our understanding of community ecology. However, four common types of analysis objective, which have different implications for analysis methodology are: (1) Testing of a hypothesis concerning ecological theory developed prior to data collection (*a priori*); (2) Determining which of a set of predictors best characterise species responses; (3) Prediction of the values of a variable of ecological importance, typically in areas that do not have observed data (e.g. mapping species richness in a geographic region, Sousa et al. 2006); and (4) Exploring the nature of the relationship between environmental variables and abundance.

Sometimes there are multiple analysis objectives, implying multiple analyses of the same data. This is usually permissible but should be done cautiously—in

particular, in the case of testing hypotheses (1) note the hypotheses need to be specified *a priori* and not derived from the data used for testing, otherwise inferences will be invalid (typically, with grossly inflated Type I Errors). A similar caution also applies to estimation of model coefficients and coverage probabilities for subsequent confidence intervals—they are biased when the coefficients for which inference is desired have not been specified *a priori*, but have instead been chosen from a broader set of potential model terms.

Data

The second key component to a model-based approach is the data that can be used to answer the ecological question. In statistical terms, data are often categorised as either *response* or *predictor* variables. The response variable(s) are those of primary interest, and in this paper, are measures of abundance (e.g. counts, percentage cover, biomass and semi-quantitative ratings) or the presence/absence for each taxon. Properties of the response variable(s) are typically key to determining how the data are ultimately analysed. Predictor variables can take many forms. They can include variables implied by the study design, environmental variables directly measured in the field and sometimes the presence/abundance of other species. For example, if an experiment was performed using a randomised complete block experimental design, then between block variation should be accounted for in the model.

Exploratory data analysis (EDA)

Prior to fitting the model, it is advisable to do some preliminary analyses to identify the key properties of data that need to be incorporated into a model. This is often referred to as exploratory data analysis (EDA) after Tukey (1977). Failure to explore data prior to analysis was recently highlighted by Steel et al. (2013) as the first of a set of common statistical errors in ecology.

Often unusual or unexpected data occur in a data set, and these need to be queried to see if they represent reality or are generated from an error of one kind or another. Sometimes, predictor variables are strongly skewed and require transformation prior to modelling, to ensure that values in the tail of the distribution do not exert undue influence. Note that, this does not have

the same unwanted consequences as transforming the response variables can (e.g. O’Hara and Kotze 2010; Warton and Hui 2011).

It is important during EDA to try to stay “close to the data”, and given that the observed data are to be modelled, it is properties of the observed data that need to be queried. As such, bivariate scatterplots and boxplots of the observed data can detect patterns that might be missed by more abstract visualisation tools such as ordination diagrams (Warton 2008).

Choice of model

The precise model used for analysis depends on both the question being addressed and the data at hand.

How the ecological question informs model choice

A given dataset can be analysed in different ways, depending on the ecological question. A simple example familiar to many readers will be a set of paired measurements of a quantitative variable. Such a dataset could be analysed using a *t* test (to compare means), linear regression (to predict one from the other) or a major axis (to test for agreement), amongst other approaches. The objective of the study has a key role in guiding the type of analysis to be undertaken.

Four common types of objective were listed previously, and each of these corresponds to a different approach to analysis: (1) *Hypothesis testing* When there is an *a priori* hypothesis to test; (2) *Model selection* Determining which predictors best characterise species response; (3) *Predictive modelling* To predict the values of a response variable; (4) *Description* To explore and describe the main patterns in response, typically without formal inference.

How the data informs model choice

A number of different properties of data affect the type of model used, and this section focuses on two key properties of multivariate abundance data: (1) the distribution of each of the response variables (the “abundance” property) and (2) correlation between responses (the “multivariate” property).

When thinking about what type of distribution to specify in the model, the key consideration is to try to accurately model the response variable(s), in our case, the multivariate abundances. The reason we are

primarily interested in the response variable in regression is that we typically condition on predictors, e.g. given that mean temperature at a site is 20 degrees Celsius, what species are we likely to find. This conditioning has the consequence that predictors are not treated as random, and hence we can focus on constructing a plausible statistical model for the multivariate abundances alone, irrespective of the distribution of predictors.

When analysing multivariate abundance data, the distribution of each response variable typically contains many zeros, because not all taxa are observed at all sites. Such data also tends to exhibit a strong mean–variance relationship (Warton et al. 2012), which arises partly as a result of the large number of zeros in the data. This strong mean–variance relationship is a key property of the data (which we can think of this as the “abundance” property) which needs to be accounted for in analysis. Precisely how it is accounted for depends on what form the data arise in—as counts, presence/absence, biomass, *etc.*, and some example statistical distributions are listed in Table 1b. Note that one distribution that is conspicuously missing from Table 1b is the normal distribution. The normal distribution assumes a constant mean–variance relationship, which is rarely plausible for multivariate abundance data under any transformation, given the high frequency of zeros.

A second key property of multivariate abundances is correlation across taxa (the “multivariate” property). Classical regression modelling approaches assume observations are conditionally independent—that is, beyond similarities inferred by site characteristics, knowing the value of a response variable at a site gives no useful information for predicting the values of any other response. Community ecology offers an important exception to this when modelling multi-species data—species often interact, so when data on multiple taxa are collected from a single site, they should in the first instance be assumed to be correlated.

Correlation may arise not only across taxa, but also across measurements within taxa, due to spatial or temporal autocorrelation (Cressie et al. 2009). Even when there is thought to be little autocorrelation in the response, it can arise when predictor variables omitted from the model themselves exhibit autocorrelation.

Key tools for handling correlation in non-normal data include generalised linear mixed models (Bolker

et al. 2009) and generalised estimating equations (Warton 2011). But a particular difficulty in community ecology is that typically there are many variables (because many different taxa are observed)—this is a very difficult problem to handle statistically, because unless some constraints are imposed on the form of correlation in the data, the number of pairwise interactions between species explodes to unmanageable numbers very quickly, e.g. if 100 taxa are observed, then there are almost 5000 possible pairwise correlations between taxa. A key opportunity in community ecology is the development of parsimonious models for multi-taxa correlation that can be used to specify realistic, fully parametric models for multivariate abundance data. Early attempts (for example, the random site effect in Jamil et al. 2013, which induces equal correlations between all pairs of species) are perhaps overly simplistic and cannot claim to offer plausible models for species interaction.

This difficult issue has historically been circumvented in multivariate ecology using design-based inference (Manly 2007), i.e. taking correlation into account at the inference stage and not in the original model specification. This is achieved by resampling rows of correlated observations (Anderson 2001) across (independent) sites, to ensure community-level inferences are valid for correlated data, even when the correlation has not been correctly accounted for. Recently, the `mvabund` package for R (Wang et al. 2012) used this approach in a model-based context, where a model describes the environmental response of each taxon, then rows (sites) are resampled to ensure valid inference, despite correlation across taxa. Such an approach is suitable for hypothesis testing, but in the context of variable selection and predictive modelling, cross-validation is perhaps better suited to the same purpose—predictive performance on test data could be used as the criterion for choosing between competing models, as in Hui et al. (2013). By putting all (correlated) responses from a site in the same test/training group, subsequent inferences are robust to correlation, as was the case previously for row resampling.

Design-based approaches can be extended to handle spatial or temporal correlation, by grouping observations at different sites into blocks of related sites which are then treated as sampling units. This is known as block resampling (Lahiri 2003) or block cross-validation, and it can ensure valid inference if observations have negligible correlation across

blocks. The importance of block cross-validation has recently been recognised in the species distribution modelling literature (Wenger and Olden 2012).

Checking the model

A key step to ensure valid and robust inference is to perform model diagnostics, i.e. viewing the data through a lens defined by the model. If the model fails to fit the data, then these diagnostic checks should highlight this fact. Some well-known tools for this purpose include residual plots, prediction to hold-out samples, comparison of competing models using information criteria (Burnham and Anderson 1998) and more controversially (Shuster 2005), goodness-of-fit tests. Diagnostic checks can be constructed for all the model types and can be tailored to inspect particular aspects of the model (e.g. Foster and Bravington 2011).

Residual plots are widely used in univariate least squares regression to assess model assumptions (*sensu* Draper and Smith 1998). What many readers will not realise is that these same principles can be used for any parametric model, including multivariate ones. A very general definition of residuals was proposed by Dunn and Smyth (1996), which they referred to as randomised quantile residuals. These residuals come *exactly* from a standard normal distribution, if the model fitted is exactly correct. Surprisingly, this result remains true even when the original response data are discrete or present/absent, as seen later. Code to construct Dunn–Smyth residuals plots in common settings is available in the `mvabund` package (Wang et al. 2012), and we recommend such plots as an excellent starting point when checking model assumptions.

If fitting a model using Bayesian estimation, then an additional consideration that applies at this point is checking the priors that were put on parameters, e.g. via posterior probability checking (Gelman et al. 2013). If using simulation approaches for model fitting, such as Markov Chain Monte Carlo, an additional consideration is checking that the algorithm used to fit the model has converged, for which diagnostic tools are also generally available (Gelman et al. 2013).

Interpretation

The final step in the flow diagram of Fig. 1 involves studying analysis outcomes and how they relate to the

original research question. It is worth noting that all analysis outputs are estimated with uncertainty. Repeating the survey or experiment that generated the data would lead to different data hence non-identical estimates. Thus, output should always be interpreted jointly with estimates of uncertainty (e.g. using confidence intervals for parameter estimates). Fortunately, such estimates of uncertainty are routinely provided in statistical modelling output, an important point of difference from some algorithmic approaches to analysis. Steel et al. (2013) review other common issues regarding the interpretation step.

Examples

Vegetation restoration and invertebrate communities

This dataset comes from a survey of the effects of vegetation restoration efforts on invertebrate communities, analysed in Warton and Hudson (2004, data from Anthony Pik). Invertebrates were sampled at eight sites subjected to vegetation restoration efforts, and at two additional control sites not subject to restoration. Invertebrates were used as bioindicators, i.e. to measure success of restoration efforts. It was of interest to determine if and how invertebrate communities differ between control and treatment sites. Invertebrates were sampled in pitfall traps—five pitfalls being set out at each site. Unfortunately, only four pitfalls were recovered from one of the sites. Invertebrates were classified in order and aggregated across samples for analysis, and the ensuing dataset consisted of abundance counts of 35 orders of terrestrial invertebrates across the 10 sites.

Objective

Is there an effect of restoration efforts on invertebrate communities? If such an effect was detected, the next obvious step would be to study the nature of the effect.

Data

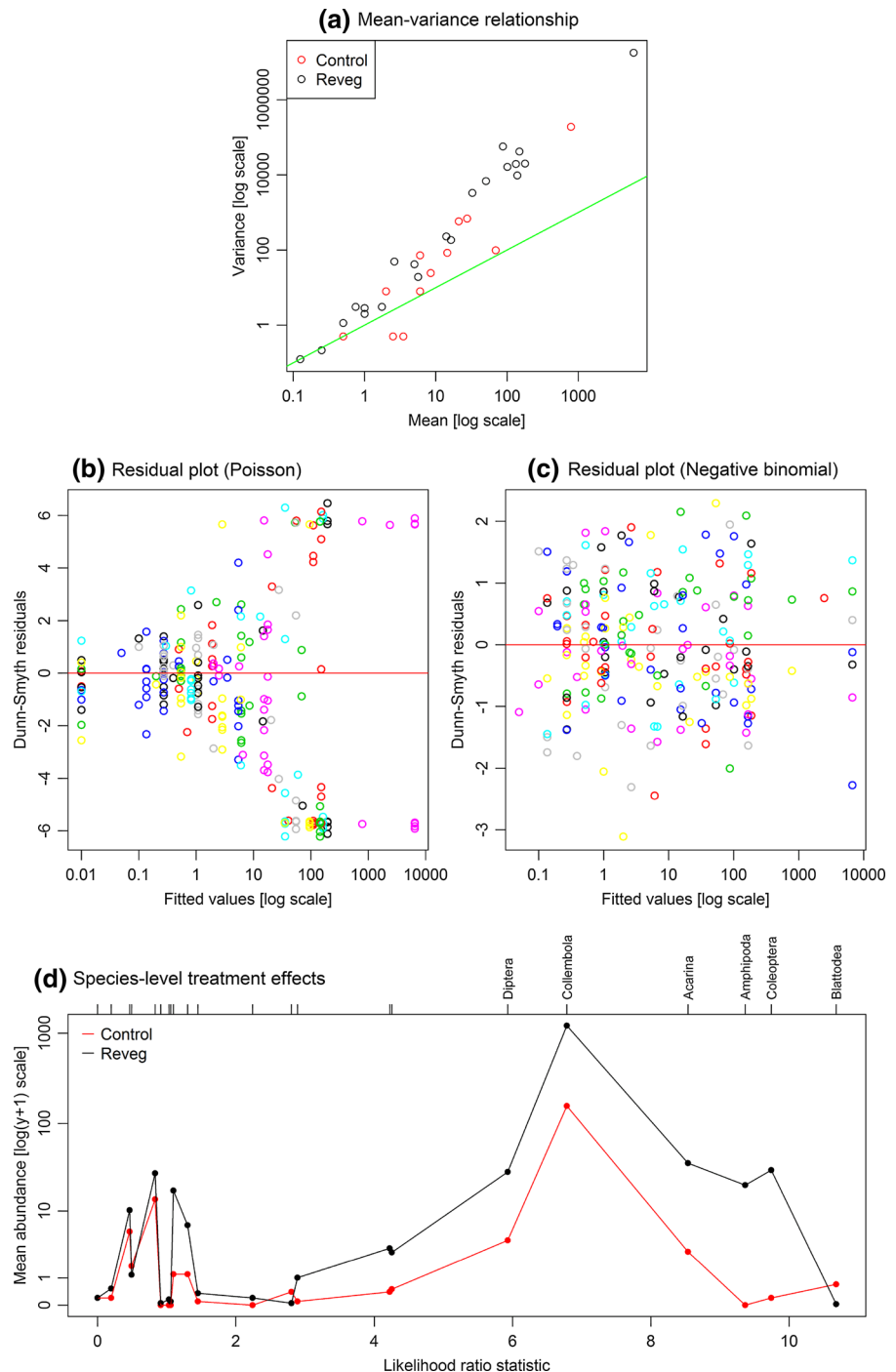
Counts of invertebrates classified to the ordinal level, where in each site, invertebrates were collected in 4–5 pitfall traps. Because data were counts, a Poisson or

negative binomial distribution was used. Because a different number of pitfall traps was used at each site, the sampling intensity varied and needed to be taken into account for analysis. This was done by including an offset in the model, equal to the logarithm of the number of pitfall traps collected at each site.

Fig. 2 Analysis of the vegetation restoration abundance data. **a** Mean–variance relationship, each point representing the sample mean and variance of a species in a treatment group; Residual plots for log-linear models fitted assuming the data are **b** Poisson and **c** negative binomial in distribution; **d** species-level treatment effects on invertebrate abundance, with mean abundance in each treatment (y-axis) plotted against univariate test statistic (likelihood ratio statistic, $-2 \log A$, x-axis) for each invertebrate order. In **a**, there is a curvilinear trend and points tend to be above the one-to-one line. In **b**, there is a fan-shape pattern, no longer evident in **c**. **a–c** all suggest data are overdispersed compared to the Poisson distribution, and better modelled by the negative binomial. In **d**, the orders with strongest evidence of a treatment effect appear in the *right half* of the plot, and for most these, there was a substantial increase in abundance after vegetation restoration efforts

Exploratory data analysis

The mean–variance relationship in the data was visualised by plotting sample variance against sample mean for each treatment group and each taxonomic group (Fig. 2a). A strong increasing pattern was



observed, with the variance often much larger than the mean, and suggesting data were overdispersed. Thus, we consider a negative binomial model in the analyses below.

Many species were rare, with 11 taxonomic groups found only once in the 10 sites (singleton groups). We removed such groups on the grounds that they provide negligible information about the effect of treatment (ensuing sensitivity analyses confirmed these species had negligible effect on results). This left 24 species at ten sites, so the data were highly multivariate, and conventional approaches to modelling correlation between response variables were not possible.

Model The primary question of interest involved testing the *a priori* hypothesis of no effect of restoration efforts on invertebrates. As such, this question is most naturally addressed using a hypothesis testing framework. The R package *mvabund* (Wang et al. 2012) was developed for precisely this purpose—to provide a suite of model-based tools for testing multivariate hypotheses. We used the `manyglm` function (Wang et al. 2012) to fit negative binomial regressions to each invertebrate order, with an offset for the number of pitfall traps at a site. Likelihood ratio statistics were calculated for each order to test the effect of the restoration treatment. As a community-level measure of treatment effect, we summed the likelihood ratio statistics (“sum-of-LR” statistic, Warton et al. 2012). To account for correlation, in abundance across taxa we used the PIT-trap, a bootstrap method operating on probability integral transform residuals (effectively, Dunn–Smyth residuals). The PIT-trap was recently developed and demonstrated to have good performance when resampling abundance data with many zeros (Warton and Wang in review). We estimated *P* values from 999 bootstrap resamples.

Check Plots of Dunn–Smyth residuals against fitted values (Fig. 2c) exhibited little pattern, and hence there was little evidence of departure from model assumptions. In contrast, we also fitted a Poisson distribution to the data and observed a fan-shaped pattern in the ensuing residual plot (Fig. 2b). This suggested that the mean–variance relationship would not have been adequately modelled by a Poisson distribution. In particular, there was greater sample variation in larger counts (i.e. overdispersion) relative to the Poisson.

Interpretation The global test statistic was $-2 \log \Lambda = 78.3$, with $P = 0.016$, suggesting good evidence of an effect of restoration on invertebrate communities. A plot of mean abundance for each treatment group against the univariate likelihood ratio statistic suggested there were six invertebrate orders that exhibited evidence of a treatment effect—*Blattodea*, *Coleoptera*, *Amphipoda*, *Acarina*, *Collembola* and *Diptera* (Fig. 2d). These six invertebrate orders all had unadjusted univariate *P* values that were significant at the 0.05 level, and collectively they accounted for 65 % of the deviance explained by restoration. Therefore, they were our main target in terms of understanding the nature of the restoration effect. *Blattodea* had the strongest evidence of an effect, being absent from all eight restored sites but present, albeit in low abundance, at both the control sites. In the remaining five orders with evidence of a treatment effect, there was a substantial increase in mean abundance (7-fold or more) following restoration efforts. Amphipods were completely absent from both the control sites.

Species traits in dune meadows

The second example is the well-known Dutch Dune Meadow data set from Jongman et al. (1987), consisting of semi-quantitative abundances of 28 plant species sampled at 20 sites. While initial analyses of this dataset (Jongman et al. 1987) related plant species abundance to environmental variables only, an additional matrix of species trait data is available as online supplementary materials for Jamil et al. (2013). We will make use of this trait dataset to try to understand why response to environmental variables differs across species.

Objective

Which species traits explain interspecific differences in environmental response?

Data

Abundances were recorded using the Domin scale, a variation on Braun–Blanquet scale which categorises species abundance on an ordinal scale between 0 (completely absent) and 10 (present everywhere). This suggests some form of ordinal regression is

appropriate, and so we used a proportional odds model. Five environmental variables and five species trait variables were available for use in the models.

Exploratory data analysis

Quantitative predictor variables were typically right-skewed, and thus all were log-transformed prior to analysis, which has the advantage of putting them on the proportional scale.

Model

A proportional odds model was used since the response variable was ordinal.

The proportional odds model included main effects terms for environmental and species trait variables, and importantly, their interaction. The interaction between environmental and trait variables was of primary interest, this term explains how interspecific variation in environmental response can be explained by differences in functional traits across species. This general approach has been independently proposed by a few authors (Pollock et al. 2012; Jamil et al. 2013; Brown et al. 2014), and has interesting connections to the fourth-corner problem (Jamil et al. 2013; Brown et al. 2014).

In addition to the trait effects, we considered whether to include in the model a random species intercept term, and random species \times environment interaction terms, to soak up variation in abundance not explained by traits.

These models were fitted using the `c1m` and `c1mm` functions from the `ordinal` package on R (Christensen 2013).

Rather than having a specific hypothesis to test concerning interactions, the objective is better described as variable selection, i.e. *which* interactions of environmental and trait variables are useful for predicting abundance.

Forward selection was used to screen the predictor variables to include in analyses. This suggested that only five of the predictor variables were useful in predicting abundance—one environmental variable (management type) and four species trait variables (specific leaf area, height, leaf dry matter concentration and lifespan). Linear terms for each of these variables appeared to be sufficient, as no quadratic terms were selected in stepwise procedures.

Given that we only considered one environmental variable and four species traits, there were four possible types of interaction that could be included in the model, and again we used forward selection to choose an interpretable model (Draper and Smith 1998). To account for correlation between species, we chose the final model along the forward selection path by cross-validation, where sites were randomly allocated to test/training groups. The validity of the method requires sites to be independent, but not species within sites. Our cross-validation procedure computed the log-likelihood for test observations using a 90:10 training:test split (i.e. assigning two sites at random to the test sample), and averaging results over 50 different choices of test sample.

Check

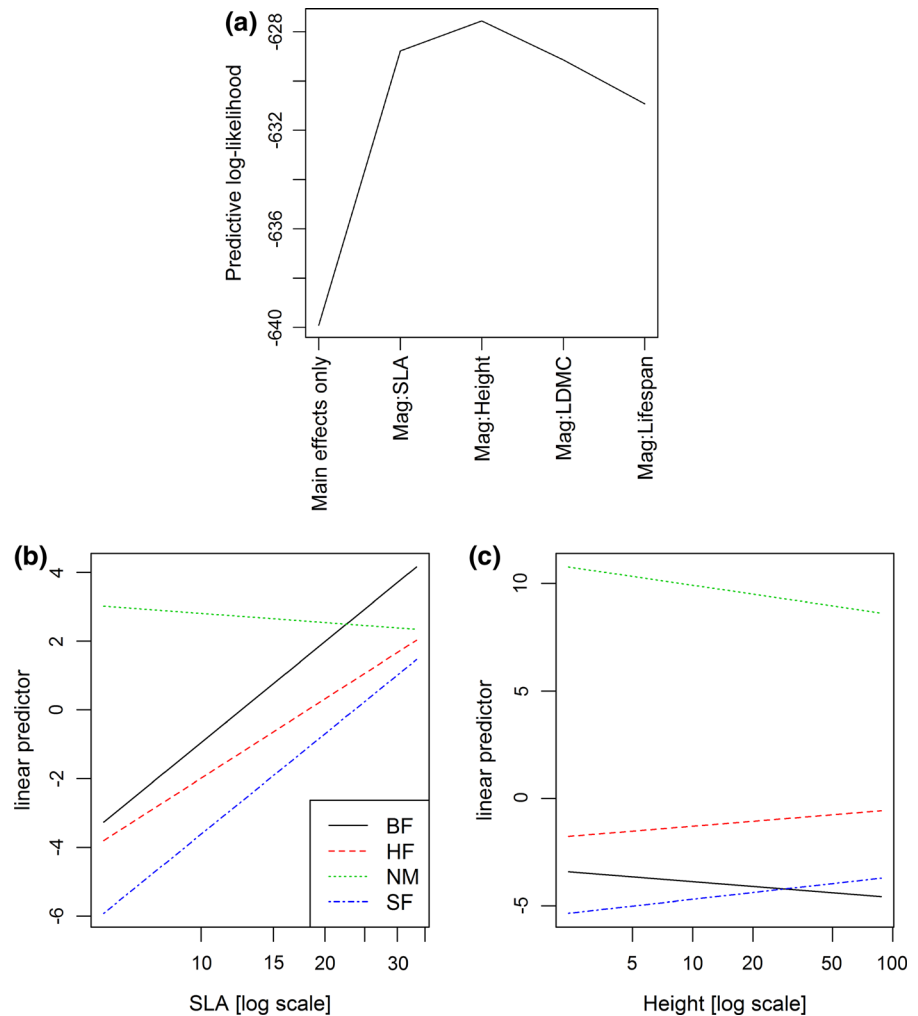
No trend was found in residual plots (not presented).

Interpretation

The maximum value of the log-likelihood for test data was achieved by a model with just two interaction terms, representing the interactions between management type and each of specific leaf area (SLA) and height (Fig. 3a). That is, the variables specific leaf area and plant height appear to be useful in explaining interspecific differences in response to the land management. Plots of these interactions suggest that a flat, slightly negative relationship between abundance and each of SLA and height for sites in conservation reserves, but steeper increasing relationships in most other situations (Fig. 3b, c). That is, taller plants with “thinner” leaves (larger SLA) tended to be found on the farmed sites.

The model with maximum log-likelihood for test data was a mixed effects model with a random intercept term for species. This model had a substantially higher predictive log-likelihood than its fixed effect only counterpart (−627.6 vs −655.6), but was not appreciably improved on by additionally including random species terms for the management effect. This suggests that there was some interspecific variation in abundance that was not adequately explained by traits. However, inclusion or exclusion of this species effect did not affect outcomes of variable selection, models with and without a species effect both selecting interaction terms with SLA and height, and the fixed

Fig. 3 Analysis of the dune meadow ordinal data using mixed effects proportional odds logistic regression with a random intercept term for species. **a** Predictive performance under cross-validation of models of different complexity; and interaction plot for **b** the SLA:Management interaction and **c** the Height:Management interaction. Order in which interaction terms were added was decided by forward selection, and a two-term model was optimal, with interactions as in **b–c**. “Main effects only” indicates a model with main effects for all five of the predictors used in this analysis. *BF* biological farming, *HF* hobby farming, *NM* natural conservation management, *SF* standard farming



effects model producing an interaction plot similar to Fig. 3b, c.

Jamil et al. (2013) analysed the same dataset but as presence–absence, and using a different model and procedure for variable screening. This led them to characterise environmental response not using management type, but instead looking at the effects of two soil properties (moisture and manure content). These two variables (and especially manure content) are correlated with management type and thus are likely to be picking up the same trend as our analyses. They did, however, find that their soil variables were related to the same species traits (height and specific leaf area) as in our analyses. Interestingly, the variables we found to be important corresponded well with the results of their ordination (Jamil et al. 2013, Fig. 4).

One difference in the methodology we used as compared to Jamil et al. (2013) is that we analysed ordinal data rather than reducing the data to presence–absence, and it is natural to ask whether there was any additional benefit from doing so. We have compared the coefficients and their estimated standard errors between a logistic regression of the presence/absence data and a proportional odds model of the ordinal response. (Table 2). Estimated coefficients were broadly similar. While the standard errors should be interpreted with caution, since they did not take into account species interactions, they were indicative of a slight improvement in efficiency from using the ordinal data in analyses. Most interaction terms have a smaller standard error by 10–20 % in the ordinal analysis, because the inclusion of additional

Table 2 Estimated coefficients (and standard errors) for environment-trait interaction terms in each of a mixed effects proportional odds model fitted to ordinal data, and a mixed

effects logistic regression fitted to the presence/absence data, with a random intercept term for species

Interaction term	Ordinal	Presence/absence
MagHF × SLA	−0.91502 (1.27797)	−0.61612 (1.49153)
MagNM × SLA	−4.62628 (1.24162)	−4.78519 (1.41440)
MagSF × SLA	−0.02506 (1.33266)	−0.42832 (1.49756)
MagHF × Height	0.65523 (0.36257)	0.92039 (0.43448)
MagNM × Height	−0.26979 (0.36418)	0.07702 (0.42826)
MagSF × Height	0.77657 (0.37041)	0.98282 (0.43903)

Notice the coefficient estimates are broadly similar, but the standard errors are consistently smaller, when analysing ordinal data (because more information is available for use in modelling)

information on relative abundance improved the accuracy of model estimates. This kind of comparison would have been more difficult without the use of models, as it is not obvious how to compare algorithmic analyses of the presence/absence vs ordinal data.

Discussion

While historically, a model-based approach to multivariate analysis has not been feasible in ecology that time has passed and there are now a range of tools available that can be used for model-based analysis. Specifying and estimating a statistical model for multivariate data have a number of advantages, most strikingly their interpretability and the flexibility to handle a range of data types, study design features and research objectives. They have also been demonstrated to have better properties than methods currently used in ecology, with a mismatch between the analysis method and the statistical properties of the data being analysed often having dramatic consequences for algorithmic approaches to multivariate analysis (Warton et al. 2012).

A key current challenge in community-level analysis is the development of realistic models for interspecies correlation that are sufficiently parsimonious to be estimable from multivariate abundance data under the dual challenges of sparsity (many zeros) and high dimensionality (many taxa). Until such methods are developed and demonstrated to be capable of accurate inference for high-dimensional data a design-based approach to inference is encouraged, e.g. resampling rows for hypothesis testing or cross-validation of sites for predictive modelling. We

do not need to throw out the model-based paradigm, when we do not have a good model for species correlation. Rather, we can use independent units in the study design as a robust basis for inference, despite possible failure of assumptions about interspecies correlation.

There is a vast array of different types of questions that can be asked about the community–environment association. Some of these are new, while others are novel adaptations of historical questions. In both the cases, an explicit and direct method to answer them, using available data, is to utilise advances in statistical modelling (e.g. Gelfand et al. 2005; Dunstan et al. 2011; Ovaskainen and Soininen 2011; Foster et al. 2013; Pledger and Arnold 2014). The model-based approach has many scientific advantages, as we have outlined. However, it also has the advantage that it enables ecologists to leverage the main thrust of statistical science that has developed substantially in recent years. Model-based approaches offer exciting potential by providing the capacity to answer ecological questions more directly, potentially opening up the field of community ecology to a deeper focus on underlying processes.

Acknowledgments DIW is supported by the Australian Research Council Future Fellow scheme (project number FT120100501). SDF, GD and PKD were supported by the Marine Biodiversity Hub, a collaborative partnership supported through funding from the Australian Government's National Environmental Research Program (NERP). NERP Marine Biodiversity Hub partners include the Institute for Marine and Antarctic Studies, University of Tasmania; CSIRO Wealth from Oceans National Flagship, Geoscience Australia, Australian Institute of Marine Science, Museum Victoria, Charles Darwin University and the University of Western Australia.

References

- Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecol* 26:32–46
- Anderson TW (2003) An introduction to multivariate statistical analysis, 3rd edn. Wiley, New York
- Bolker BM, Brooks ME, Clark CJ, Geange SW, Poulsen JR, Stevens MHH, White J-SS (2009) Generalized linear mixed models: a practical guide for ecology and evolution. *Trends Ecol Evol* 24:127–135
- Brown AM, Warton DI, Andrew NR, Binns M, Cassis G, Gibb H (2014) The fourth-corner solution using predictive models to understand how species traits interact with the environment. *Methods Ecol Evol* 5(4):344–352
- Burnham KP, Anderson DR (1998) Model selection and inference: a practical information-theoretic approach. Springer, New York
- Christensen RHB (2013) Ordinal–regression models for ordinal data. R package version 2013.9-30. <http://www.cran.r-project.org/package=ordinal/>
- Clark J (2007) Models for ecological data. Princeton University Press, Princeton
- Cressie N, Calder CA, Clark JS, Hoef JMV, Wikle CK (2009) Accounting for uncertainty in ecological analysis: the strengths and limitations of hierarchical statistical modeling. *Ecol Appl* 19:553–570
- Draper NR, Smith H (1998) Applied regression analysis, 3rd edn. Wiley, New York
- Dunn P, Smyth G (1996) Randomized quantile residuals. *J Comput Graph Stat* 5:236–244
- Dunstan PK, Foster SD, Darnell R (2011) Model based grouping of species across environmental gradients. *Ecol Model* 222:955–963
- Dunstan PK, Foster SD, Hui FK, Warton DI (2013) Finite mixture of regression modelling for high-dimensional count and biomass data in ecology. *J Agric Biol Environ Stat* 18:357–375
- Elith J, Leathwick J (2007) Predicting species distributions from museum and herbarium records using multiresponse models fitted with multivariate adaptive regression splines. *Divers Distrib* 13:265–275
- Elith J, Leathwick J (2009) Species distribution models: ecological explanation and prediction across space and time. *Ann Rev Ecol Syst* 40:677–697
- Etienne RS (2007) A neutral sampling formula for multiple samples and an ‘exact’ test of neutrality. *Ecol Lett* 10:608–618
- Ferrier S, Guisan A (2006) Spatial modelling of biodiversity at the community level. *J Appl Ecol* 43:393–404
- Foster S, Bravington M (2011) Graphical diagnostics for markov models for categorical data. *J Comput Graph Stat* 20:355–374
- Foster S, Givens G, Dorman G, Dunstan P, Darnell R (2013) Modelling biological regions from multi-species and environmental data. *Environmetrics* 24:489–499
- Gauch H, Chase GB, Whittaker RH (1974) Ordination of vegetation samples by Gaussian species distributions. *Ecology* 55:1382–1390
- Gelfand AE, Schmidt AM, Wu S, Latimer A (2005) Modelling species diversity through species level hierarchical modelling. *J R Stat Soc* 54:1–20
- Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB (2013) Bayesian data analysis, 3rd edn. CRC Press, Boca Raton
- Goodall D, Johnson R (1982) Non-linear ordination in several dimensions. *Vegetatio* 48:197–208
- Hui FK, Warton DI, Foster S, Dunstan P (2013) To mix or not to mix: comparing the predictive performance of mixture models versus separate species distribution models. *Ecology* 94:1913–1919
- Ives AR, Helmus MR (2011) Generalized linear mixed models for phylogenetic analyses of community structure. *Ecol Monogr* 81:511–525
- Jamil T, Ozinga WA, Kleyer M, ter Braak CJF (2013) Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *J Veg Sci* 24:988–1000
- Jongman RHG, ter Braak CJF, van Tongeren OFR (1987) Data analysis in community and landscape ecology. Pudoc, Wageningen
- Lahiri SN (2003) Resampling methods for dependent data. Springer, New York
- Legendre P, Legendre L (2012) Numerical ecology. Elsevier, Amsterdam
- Manly BFJ (2007) Randomization, bootstrap and Monte Carlo methods in biology, 3rd edn. Chapman & Hall, London
- Neter J, Kutner M, Natchtsheim C, Wasserman W (1996) Applied linear statistical models, 4th edn. Irwin, Chicago
- O’Hara RB, Kotze DJ (2010) Do not log-transform count data. *Methods Ecol Evol* 1:118–122
- Ovaskainen O, Soininen J (2011) Making more out of sparse data: hierarchical modeling of species communities. *Ecology* 92:289–295
- Pledger S, Arnold R (2014) Multivariate methods using mixtures: correspondence analysis, scaling and pattern-detection. *Comput Stat Data Anal* 71:241–261
- Pollock LJ, Morris WK, Vesik PA (2012) The role of functional traits in species distributions revealed through a hierarchical model. *Ecography* 35:716–725
- Shuster JJ (2005) Diagnostics for assumptions in moderate to large simple clinical trials: do they really help? *Stat Med* 24:2431–2438
- Sousa P, Azevedo M, Gomes M (2006) Species-richness patterns in space, depth, and time (1989–1999) of the Portuguese fauna sampled by bottom trawl. *Aquat Living Resour* 19:93–103
- Steel E, Kennedy M, Cunningham P, Stanovick J (2013) Applied statistics in ecology: common pitfalls and simple solutions. *Ecosphere* 4:115
- ter Braak CJ, Hooijtkink H, Akkermans W, Verdonschot PF (2003) Bayesian model-based cluster analysis for predicting macrofaunal communities. *Ecol Model* 160:235–248
- ter Braak CJF (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67:1167–1179
- Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading
- Wang Y, Naumann U, Wright ST, Warton DI (2012) mvabund: an R package for model-based analysis of multivariate abundance data. *Methods Ecol Evol* 3:471–474

- Warton DI (2008) Raw data graphing: an informative but underutilized tool for the analysis of multivariate abundances. *Austral Ecol* 33:290–300
- Warton DI (2011) Regularized sandwich estimators for analysis of high dimensional data using generalized estimating equations. *Biometrics* 67:116–123
- Warton DI, Hudson HM (2004) A MANOVA statistic is just as powerful as distance-based statistics, for multivariate abundances. *Ecology* 85:858–874
- Warton DI, Hui FKC (2011) The arcsine is asinine: the analysis of proportions in ecology. *Ecology* 92:3–10
- Warton DI, Wang YA (in review) The PIT-trap: a general bootstrap procedure for inference about regression models with non-normal response
- Warton DI, Wright ST, Wang Y (2012) Distance-based multivariate analyses confound location and dispersion effects. *Methods Ecol Evol* 3:89–101
- Wenger SJ, Olden JD (2012) Assessing transferability of ecological models: an underappreciated aspect of statistical validation. *Methods Ecol Evol* 3:260–267. doi:[10.1111/j.2041-210X.2011.00170.x](https://doi.org/10.1111/j.2041-210X.2011.00170.x)
- Yee TW (2010) The VGAM package for categorical data analysis. *J Stat Softw* 32:1–34
- Zuur AF, Ieno EN, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. *Methods Ecol Evol* 1:3–14