Supplementary materials for this article are available at 10.1007/s13253-013-0146-x.

Finite Mixture of Regression Modeling for High-Dimensional Count and Biomass Data in Ecology

Piers K. DUNSTAN, Scott D. FOSTER, Francis K.C. HUI, and David I. WARTON

Understanding how species distributions respond as a function of environmental gradients is a key question in ecology, and will benefit from a multi-species approach. Multi-species data are often high dimensional, in that the number of species sampled is often large relative to the number of sites, and are commonly quantified as either presence–absence, counts of individuals, or biomass of each species. In this paper, we propose a novel approach to the analysis of multi-species data when the goal is to understand how each species responds to their environment. We use a finite mixture of regression models, grouping species into "Archetypes" according to their environmental response, thereby significantly reducing the dimension of the regression model. Previous research introduced such Species Archetype Models (SAMs), but only for binary assemblage data. Here, we extend this basic framework with three key innovations: (1) the method is expanded to handle count and biomass data, (2) we propose grouping on the slope coefficients only, whilst the intercept terms and nuisance parameters remain species-specific, and (3) we develop model diagnostic tools for SAMs. By grouping on environmental responses only, the model allows for inter-species variation in terms of overall prevalence and abundance. The application of our expanded SAM framework data is illustrated on marine survey data and through simulation.

Supplementary materials accompanying this paper appear on-line.

Key Words: Community-level model; Mixture model; Multi-species; Species archetype model; Species distribution model; Tweedie.

Piers K. Dunstan is Quantitative Ecologist, CSIRO Marine and Atmospheric Research, GPO Box 1538, Hobart, TAS 7001, Australia. Scott D. Foster is Statistician, CSIRO Mathematics, Informatics and Statistics, GPO Box 1538, Hobart, TAS 7001, Australia. Francis K.C. Hui is PhD Candidate, and David I. Warton (✉) is Associate Professor (E-mail: *david.warton@unsw.edu.au*), School of Mathematics and Statistics, The University of New South Wales, Sydney, NSW 2052, Australia.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 18, Number 3, Pages 357–375
DOI: 10.1007/s13253-013-0146-x

1. INTRODUCTION

Understanding how species respond as a function of environmental variables (hereafter "environmental response") is a key question in ecology. Ecosystems, by definition, consist of many species, and ecological survey data are frequently stored as an $N \times S$ matrix of data for S species at N sites. A corresponding matrix of environmental variables e.g., temperature and altitude/depth, is also observed, and the task is to relate the species data to these variables. Species data are usually recorded as a presence–absence, counts (number of individuals of each species) or biomass (total mass of each species). Count data are common in terrestrial systems (e.g., Novotny et al. 2007), whereas biomass are common in marine systems (e.g., Bax and Williams 2000).

Two key properties often present in the species matrix are high-dimensionality (often S is similar in size, or bigger, than N), and sparsity (species are often encountered infrequently), both of which complicate analysis. For example, the marine fish survey data analyzed in Section 4 consist of 70 species which were sampled at 180 locations. Only 28 species were found at more than 30 % of sites. Datasets with similar properties are found not only in marine systems but also in terrestrial assemblages of insects (e.g., Novotny et al. 2007), plants (Ross et al. 2012) and mammals (Thibault et al. 2011).

Multi-species analysis of such data is an important but challenging task. Modeling each species separately using univariate regression tools like Generalized Linear Models (GLMs; McCullagh and Nelder 1989) is impractical and can be difficult to interpret in a multi-species context. Instead, a common approach to multi-species analysis in ecology uses an algorithmic, site-based approach, where differences in species composition and abundance between sites are analyzed (Anderson et al. 2011; Li, Ban, and Santiago 2011). Such an approach does not consider that species may not vary as a "community" but rather as a set of independently varying entities. Furthermore, failure to explicitly account for important statistical properties in the data such as the mean-variance relationships of each species can lead to undesirable and unexpected properties (Warton, Wright, and Wang 2012). Hence there has been recent interest in model-based approaches to multi-species analysis (Yee 2010; Ovaskainen and Soininen 2011; Ives and Helmus 2011). A challenge however is constructing a model which characterizes multi-species environmental response, but does so in a parsimonious and interpretable way.

Recently, Dunstan, Foster, and Darnell (2011) proposed a new approach to the analysis of multi-species data, using finite mixture of regression models (McLachlan and Peel 2000) to simultaneously model and group species based on presence–absence data. Specifically, the environmental response of each species is modeled as a finite mixture of K logistic regressions, with each component characterizing a different type of "archetypal" response. This can be understood as dimension reduction, and is attractive from both statistical and ecological perspectives: it reduces the number of regression parameters to only $Kp \ll Sp$ where p is the number of regression parameters per archetype, and it provides simplified interpretation via the K archetypal responses. We refer to these models as "Species Archetype Models" (SAMs). These models represent a significant departure from the traditional types of analyses for high dimensional datasets in ecology. Each archetypal response

represents a group of species that respond to the environment in a statistically similar way. The groups should not be confused with traditional community concepts, rather we believe that communities observed at a site are comprised of multiple overlapping species distributions, and hence archetypes, that generate a unique assemblage at that one location.

An important limitation of Dunstan, Foster, and Darnell (2011) was that it considered the analysis of presence–absence data only, yet multi-species survey data often arise as counts or biomass. Moreover count and biomass data typically have a very strong mean-variance relationship, and variances of replicate observations may differ across species by a factor of a hundred thousand or more (Warton, Wright, and Wang 2012). Hence it is crucial to carefully model this mean-variance relation and the potentially differing patterns of overdispersion between species. One must also consider how to estimate any nuisance parameters not directly involved in describing the archetypal responses, a problem which did not arise in Dunstan, Foster, and Darnell (2011) with presence–absence data.

A second issue with the model defined in Dunstan, Foster, and Darnell (2011) was that species were grouped by mixing on *all* the parameters including the intercept. This implies the archetypes could in part be defined by species prevalence rather than environmental response. The problem is exacerbated for count and biomass data, where species have inherently different social habits, physiological characteristics, and overall levels of abundance. This could lead to different mean counts and biomass across a group of species, even when they share similar forms of environmental response.

In this work, we propose a novel approach to the analysis of count and biomass data when the aim is to understand how each species responds to their environment. Like in Dunstan, Foster, and Darnell (2011), our approach uses a finite mixture of regression models to group species based on environmental response, but we present three key advances to address the limitations discussed above: (1) the method is extended to count and biomass data via a generalized linear modeling framework, (2) we propose mixing on slope coefficients only, and not intercept terms nor nuisance parameters, and (3) we discuss residual analysis for SAMs to help assess a model's adequacy. In a companion paper (Hui et al. 2013), we show that this extended model, with species-specific intercepts, predicts rare species better than univariate models based on a single species. This is due to the borrowing of information (borrowing strength) from species where information is available.

2. SAMS FOR COUNT AND BIOMASS DATA

Let $\mathbf{y}_j = (y_{1j}, \ldots, y_{Nj})^\top$ be the vector of data (presence–absence, count, or biomass) for species $j \in \{1, \ldots, S\}$, observed at each of N sites indexed by $i \in \{1, \ldots, N\}$. At each site we also have a set of environmental variables $\mathbf{x}_i$ used to model the mean of y_{ij}. The $\mathbf{y}_j$ are assumed to be distributed as a type of finite mixture of regression models (McLachlan and Peel 2000), with the jth species being classified into one of $K \leq S$ species archetypes with probability function and mean model as follows:

$$\sum_{k=1}^{K} \pi_k \prod_{i=1}^{N} f(y_{ij}; \mu_{ijk}, \phi_j); \qquad h(\mu_{ijk}) = \beta_{0j} + \boldsymbol{x}_i^\top \boldsymbol{\beta}_k \tag{2.1}$$

where $f(\cdot)$ is a probability distribution function from the exponential family (McCullagh and Nelder 1989), $h(\cdot)$ is a known link function relating the mean for the kth archetype (μ_{ijk}) to $\mathbf{x}_i$, and π_k is the kth mixing proportion with $\pi_k \in (0, 1)$ and $\sum_{k=1}^{K} \pi_k = 1$.

Equation (2.1) may be regarded as a mixture of GLMs (Wedel and DeSarbo 1995), but with some important distinctions: we are mixing on vectors of N observations (to classify species into archetypes), and we are mixing on a subset of regression parameters. In particular, note the intercept β_{0j} is indexed by species and not archetype, meaning we allow each species to have different levels of abundance irrespective of their archetype. The nuisance parameters ϕ_j are also chosen to be species-specific, allowing each species to have varying amounts of overdispersion. Only the slope parameters $\boldsymbol{\beta}_k$ are indexed by archetype, implying that we cluster species solely in terms of their environmental response.

The SAM defined here assumes data to come from the exponential family of distributions, rather than being limited to Bernoulli data as in Dunstan, Foster, and Darnell (2011). This extension allows the analysis of count or biomass data (although not jointly) in addition to presence–absence data. On the other hand, extending the distributional assumption brings additional difficulties in the estimation of species-specific nuisance parameters and in the assessment of model adequacy. This problem is particularly challenging for models with overdispersion present, given its misspecification can have substantial impacts on model outcomes (Warton, Wright, and Wang 2012).

For count data, we model each archetype using a negative binomial distribution via the NB-2 parameterization (Hilbe 2007), which assumes the mean-variance function $V(\mu_{ijk}) = \mu_{ijk} + \mu_{ijk}^2/\phi_j$. The nuisance parameter $\phi_j > 0$ controls the degree of overdispersion relative to the Poisson, and can be interpreted as a measure of spatial clustering (Hilbe 2007). In having ϕ_j as species-specific, we permit each species to have its own degree of overdispersion.

For biomass data, we model the random variation via a Tweedie distribution (see Jørgenson 1997), with mean-variance function $V(\mu_{ijg}) = \phi_j \mu_{ijg}^{\nu_j}$. The power parameter controls the shape of the distribution, whilst $\phi_j > 0$ is a scale parameter. The Tweedie model should appeal to many quantitative ecologists as the mean-variance relationship is exactly that defined by Taylor's power law (Taylor 1961). In this work, we also exploit the fact that, for $1 < \nu_j < 2$, a Tweedie random variable is a compound Poisson distribution—the sum of a Poisson number of individuals, each of which has a gamma mass (see Jørgenson 1997; Foster and Bravington 2013). A practical advantage of using the Tweedie distribution in our context is that it quite naturally models quantitative data which are scale invariant and have a point mass at zero. The scale-invariance attribute means that the measurement units are inconsequential, up to non-linear transformations of the data. Tweedie densities can be evaluated using the method described in Dunn and Smyth (2005). We chose to pre-specify the power parameter to be $\nu_j = 1.6$ for all species. In principle, species' power parameters could be estimated from the data through maximization of the likelihood, although this process is slow and we believe unnecessary here. Previous studies of fisheries data from the same geographic region as our example dataset have shown $\nu_j = 1.6$ to be a reasonable value (Peel et al. 2013; Foster and Bravington 2013). In addition, we performed a small sensitivity study for 15 individual species examined in Foster and Bravington (2013); there

was only negligible difference between model residuals from estimated ν_j and specified $\nu_j = 1.6$.

2.1. ESTIMATION

We estimated parameters by maximizing the likelihood

$$\ell(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\pi}; \mathbf{y}) = \sum_{j=1}^{S} \log\left(\sum_{k=1}^{K} \pi_k \prod_{i=1}^{N} f(y_{ij} | \beta_{0j}, \boldsymbol{\beta}_k, \boldsymbol{\phi}_j) \right), \tag{2.2}$$

where $\boldsymbol{\beta}$, $\boldsymbol{\phi}$ and $\boldsymbol{\pi}$ store all of the regression, nuisance, and mixing parameters, respectively. Individual species' responses are assumed to be independent between sites conditional on species archetype. This assumption allows the specification of a GLM within each archetype. A common approach to maximizing Equation (2.2) is to apply the Expectation–Maximization algorithm (EM; Dempster, Laird, and Rubin 1977). Although quite robust, the EM algorithm tends to be relatively slow and is not guaranteed to find a global maximum from arbitrary starting values (McLachlan and Peel 2000). To overcome this, we adopted a hybrid approach for maximization consisting of three parts: (1) find good starting values, (2) perform some initial EM steps, and (3) use a descent-based maximizer until convergence. This hybrid algorithm is based on the idea of Aitkin and Aitkin (1996).

Starting values were obtained by fitting a separate GLM to each species and clustering the slope coefficients, then averaging slope coefficients across the species in each cluster. Clustering was performed using the K-means method (e.g., Venables and Ripley 1999). Intercepts and nuisance parameters were retained from the separate GLMs fitted to each species, for use in the second stage of our hybrid algorithm. Next, the starting values were refined using a small number of EM iterations. The EM-algorithm for mixture models (see McLachlan and Peel 2000) iterates between calculating the probabilities for each species belonging to each group and maximizing an augmented likelihood. The group membership probabilities are

$$\tau_{jk} = \frac{\pi_k f(\boldsymbol{y}_j | \beta_{0j}, \boldsymbol{\beta}_k, \phi_j)}{\sum_{k'=1}^{K} \pi_{k'} f(\boldsymbol{y}_j | \beta_{0j}, \boldsymbol{\beta}_{k'}, \phi_j)}$$

and can be arranged into a $S \times K$ matrix whose jth row is denoted by $\{\boldsymbol{\tau}_j\}$. These probabilities are commonly referred to as "posterior probabilities", despite the non-Bayesian context.

A problem that arose in the initial E-step was that it tended to produce starting values of $\{\boldsymbol{\tau}_j\}$ which were zero or one to within machine error. This is unwanted, as it would then not allow for movement away from this initial model and would not provide an adequate exploration of the parameter space. We therefore shrank the $\{\boldsymbol{\tau}_j\}$'s towards 0.5 using

$$\tau^*_{jk} = \frac{2\alpha\tau_{jk} - \alpha + 1}{2\alpha - \alpha K + K} \quad \text{where } \alpha = \frac{1 - 0.8K}{0.8(2 - K) - 1}. \tag{2.3}$$

The above transformation prevented any τ_{jk} from being bigger than 0.8 or less than $(1 - 0.8)/(K - 1)$, whilst maintaining the sum-to-one constraint, $\sum_{k=1}^{K} \tau_{jk} = 1$. An M-step was next applied, which maximized the complete log likelihood with respect to the

slope coefficients. The EM-algorithm was run for four cycles before moving to the next phase of optimization. Subsequent E-steps did not include shrinkage for any τ_{jk}. Note species-specific parameters were not updated during the M-steps, but instead were fixed at the values obtained from the separate species models. The final part of maximization, applied across all model parameters, used a quasi-Newton optimization routine (see Nash and Sofer 1996), which provides super-linear convergence without the need to calculate computationally expensive second derivatives. However, it can also be sensitive to starting values—hence the reason we persisted with the first two steps in the estimation routine.

There is a small amount of stochasticity in the estimation process, through randomly drawing the initial $\boldsymbol{\pi}$ for the initial E-step. This was introduced as we feel that there is little information about these parameters in the initial (species specific) analyses. To safeguard against unfortunate choices the estimation process can be repeated. This is advisable, especially when there is little data and/or a complex model.

The number of archetypes K was chosen to minimize a BIC-style criterion, defined as

$$BIC = -2\ell(\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\pi}; \mathbf{y}) + \ln S \times (\#\text{ of parameters}), \tag{2.4}$$

with the number of parameters set to $(K-1) + (q+1)S + Gp$, q being the number of nuisance parameters per species. The use of S as the sample size rather than N or NS was based on the fact the fundamental unit being grouped on was species and not site. Choosing the order K of a mixture model via BIC minimization has been shown to be consistent for K (Keribin 2000).

2.2. PREDICTION FOR MAPS

We chose to summarize the models through a series of predicted maps, one map for each archetype. The maps consisted of point predictions made on a high density grid of spatial locations throughout the prediction area. As they were based on the spatially varying environmental gradient, these predicted maps allowed the interpretation of results in terms of the processes structuring the distributions of species and assemblages. Since the intercepts in our SAM are species-specific, then the archetype predictions are a relative measure—with an arbitrary intercept. We used $\bar{\beta}_0 = \sum_{j=1}^{S} \hat{\beta}_{0j}$, the average of all the estimated species intercepts, as our arbitrary intercept. Other choices are possible, but $\bar{\beta}_0$ puts the archetypical response on a scale that is likely to be indicative of species responses. Formally, the predictions for the kth archetype are made via

$$\hat{y}_{i\cdot k} = h^{-1}(\hat{\eta}_{i\cdot k}); \quad \text{where } \hat{\eta}_{i\cdot k} = \bar{\beta}_0 + \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_k.$$

Based on the above equation, the variance of the linear predictor $\hat{\eta}_{i\cdot k}$ is then given by $\text{var}(\hat{\eta}_{i\cdot k}) = \boldsymbol{x}_i^\top \hat{\Sigma}_k \boldsymbol{x}_i$, where $\hat{\Sigma}_k = \text{var}(\hat{\boldsymbol{\beta}}_k)$. The variance of the point prediction $\hat{y}_{i\cdot k}$ can be approximated from this via the delta method (see Oehlert 1992).

2.3. MODEL DIAGNOSTICS

We assessed goodness-of-fit using residual plots, having first addressed the non-trivial question of how to define residuals for a mixture of non-Gaussian variables. We followed

Dunn and Smyth (1996), who proposed exploiting the probability integral transform to construct residuals for parametric regression models as follows:

$$r_{ij} = \Phi^{-1}\{u_{ij}F(y_{ij};\mu_{ij},\phi_j) + (1-u_{ij})F(y_{ij}^{-};\mu_{ij},\phi_j)\}, \tag{2.5}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal variable, u_{ij} are independent draws from the standard uniform distribution, $F(y_{ij};\mu_{ij},\phi_j)$ is the cumulative distribution function of the observed y_{ij}, and $F(y_{ij}^{-};\mu_{ij},\phi_j)$ is its limiting value as y_{ij} is approached from the negative direction. A key advantage of this definition of residuals is that, provided the regression model is correct, the r_{ij} are identically distributed as standard normal even though y_{ij} are not. Therefore, standard diagnostic tools for linear regression (e.g., Weisberg 2005) become applicable. The use of uniform random numbers u_{ij} in the transformation ensures this result remains true for discrete data—the u_{ij} can be understood as removing the discreteness inherent in count and presence–absence data, which is an extra but essential step enabling easy-to-interpret residual plots. In our experience, different draws of u_{ij} tend to produce similar plots.

While not previously extended to the mixture modeling context, Dunn–Smyth residuals (Dunn and Smyth 1996) can be readily applied, since integrating Equation (2.1) returns the following cumulative probabilities:

$$F(y_{ij};\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\pi}) = \sum_{k=1}^{K}\pi_k G(y_{ij};\mu_{ijk},\phi_j)$$

where $G(y_{ij};\mu_{ijk},\phi_j)$ is the kth component cumulative distribution function of y_{ij}. Residuals can be readily computed on any software that returns cumulative probabilities for exponential families, by using Equation (2.5) and substituting estimated parameter values for $\boldsymbol{\beta},\boldsymbol{\phi},\boldsymbol{\pi}$. We constructed such residuals and plotted them against fitted values and normal quantiles to diagnose various features of our SAMs.

Dunn and Smyth (1996) emphasized the assumption of independence of observations when using their residuals. However, independence is not required in order for the r_{ij} to be standard normal in distribution—the probability integral transform holds irrespective of dependence. Independence does however simplify interpretation of residual plots—otherwise, if a pattern were observed, dependence would be a possible explanation as well as misspecification of $F(y_{ij};\mu_{ij},\phi_j)$.

3. SIMULATION STUDY

In this section, we use simulation to assess whether the extensions proposed in this article can lead to improved prediction of species distribution. We consider whether analysis of count or biomass data, using the proposed SAMs, can better predict species presence–absence than fitting a Bernoulli SAM directly to presence–absence data (Dunstan, Foster, and Darnell 2011). Results have obvious practical ramifications for ecologists—advantages of collecting or analyzing one type of data over another provide an incentive to design studies ensuring the preferred data type is collected.

We jointly generated biomass (y_{ij}, for site i and species j), count (n_{ij}), and presence–absence (x_{ij}) data by exploiting the fact that biomass data can be generated as a compound sum across the n_{ij} samples:

$$y_{ij} = \sum_{k=1}^{n_{ij}} w_{ijk} \tag{3.1}$$

where w_{ijk} is a positive random variable representing the weight of the kth individual of species j at site i. Presence–absence data can be computed directly from the counts as $x_{ij} = \mathscr{I}(n_{ij} > 0)$, where $\mathscr{I}(\cdot)$ is the indicator function. Hence each of the count, biomass, and presence–absence datasets have exactly the same pattern of presence. We are interested in how to best model the probability of presence i.e., using presence–absence data directly, or whether the additional information in count or biomass data can improve predicted probabilities.

A total of $B = 1,000$ datasets were generated, each with $N = 200$ sites and $S = 50$ species. $K = 4$ archetypal responses were created for the count data, with each varying quadratically (on the log scale) in a different manner with a single environmental covariate, as shown in Figure 5. The covariate was simulated as independent standard normal variates. Species-specific intercepts were drawn from a normal distribution with mean and variance approximately matching that of the fish data in Section 4. Species-specific dispersion parameters were likewise drawn from a gamma distribution.

To simulate biomass data, the count data generated above were combined with weights w_{ijk} using Equation (3.1). The weights w_{ijk} were simulated from a gamma distribution with shape parameter equal to $2/3$ and a species mean taken randomly from a uniform distribution on $[0.1, 2]$. This data-generating model is similar to one which generates a Tweedie distribution, with the only difference being that the number of individuals for a Tweedie distribution is given by a Poisson rather than a negative binomial distribution. The choice of shape parameter reflects this similarity: it corresponds to a Tweedie power parameter of 1.6.

To each of the 1,000 simulated datasets, we fitted three models: (1) a Bernoulli SAM defined in Dunstan, Foster, and Darnell (2011), (2) a negative binomial SAM, and (3) a Tweedie SAM, as proposed in Section 2. Note only the count model matches exactly with the simulated dataset—the Tweedie model and the Bernoulli model are approximations to the data generating process. The three models were compared in terms of their predictive performance on a large test dataset (consisting of $N_t = 10,000$ observations), drawn from the same data generating mechanism as used to model the sample data. The test dataset was large to reduce the randomness in our measures of predictive performance, described below. The different models were compared in terms of posterior predictions of presence in the test data, which in the case of a model for count data, were computed as:

$$\hat{p}_{ij} = \widehat{\Pr}(n_{ij} > 0) = \sum_{k=1}^{K} \hat{\tau}_{jk} \widehat{\Pr}(n_{ij} > 0 | \text{archetype } k),$$

with similar definitions for biomass (y_{ij}) and presence–absence data (x_{ij}). Posterior predictions of presence for a given model ($\hat{p}_{ij}$) were compared to the true probability of presence (p_{ij}) using two summary statistics:

RMSE The average root mean squared error between the model's prediction of $\Pr(n_{ij} > 0)$ and the test data's true values:

$$\text{RMSE} = \frac{1}{B}\sum_{b=1}^{B}\sqrt{\frac{1}{N_t S}\sum_{i=1}^{N_t}\sum_{j=1}^{S}\left(\hat{p}_{ij}^{(b)} - p_{ij}\right)^2}.$$

Cross Entropy The negative of the sum of the expected Bernoulli log-likelihood:

$$\text{CE} = \sum_{b=1}^{B}\sum_{i=1}^{N_t}\sum_{j=1}^{S} -\left(p_{ij}\log\left(\hat{p}_{ij}^{(b)}\right) + (1-p_{ij})\log\left(1-\hat{p}_{ij}^{(b)}\right)\right).$$

For both measures, smaller values indicate better fit, with zero being a lower bound.

Simulation results (Table 1) indicated a clear ordering for the different models types. The best was the negative binomial model, the second best was the Tweedie model and the worst was the Bernoulli model. There are two possible causes for the poor performance of the Bernoulli model: either the lack of species-specific intercepts in the Bernoulli SAM caused a significant lack-of-fit, or the additional information in the quantitative data led to improved predictive performance.

The superior performance of the negative binomial with respect to the Tweedie model may, at first, seem to be at odds with intuition. In most applications, outside ecology, one would expect that continuous data have more information than integer data. However, in this simulation the data generating mechanism creates, in a natural manner, the biomass data as a noisy version of the count data—the extra variation is due to the variation of individual fish masses. Further, the Tweedie model is an approximation to the truth as count data, used in the generation of a Tweedie variate, are negative binomial and not Poisson.

There was one situation where the Tweedie model performed very similarly to the negative binomial model in terms of predictive performance—when the true probability of presence was very high (Figure 1). This remained true when using either summary statistic, although the magnitude of the differences between model types varied with choice of summary statistic (Table 1, Figure 1). This is due to the squared-error statistic placing less emphasis on the extreme probabilities, precisely where there was little difference between the models.

4. APPLICATION—SOUTH EAST FISHERIES SURVEY DATA

We applied the SAM defined in (2.1) to count and biomass data from an ecosystem research program carried out on the South Eastern continental shelf of Australia (Bax and Williams 2000), designed to determine the distribution and abundance of demersal fish species. The survey covered an area from 36 °S to 39.3 °S on the eastern continental shelf in depths from 25 to 200 meters. Samples were taken at depth-stratified stations along seven

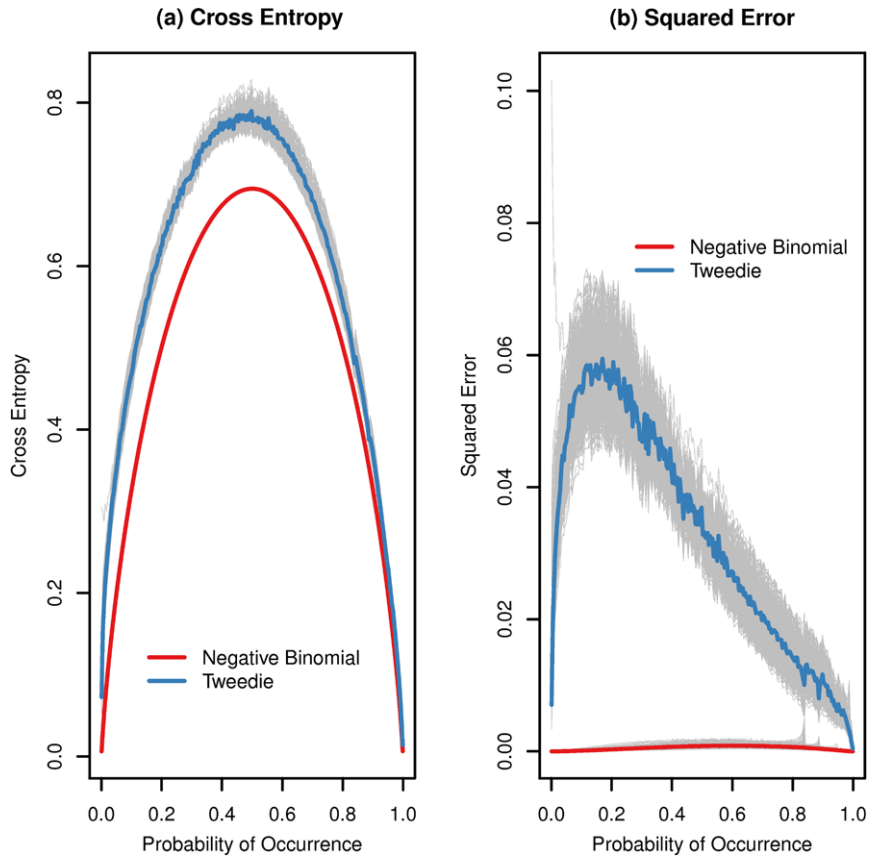

Figure 1. Smoothed version of cross-entropy and squared-error contributions for negative binomial and Tweedie models. The gray lines are realizations for each simulation and the colored lines are their averages. If the models had equivalent predictive power across probabilities, then the lines would be similar. Smoothing was performed by binning the probability into 250 classes.

Table 1. Results from simulation study. $B = 1,000$ datasets were simulated, with the test data containing $N_t = 10,000$ observations. $S = 50$ species were considered. See text for definition of summary statistics. Parenthetic values are standard errors (across simulations).

	RMSE	Cross Entropy ($\times 10^5$)
Bernoulli	0.232 (0.006)	0.266 (0.006)
Negative Binomial	0.018 (0.002)	0.174 (<0.001)
Tweedie	0.150 (0.004)	0.208 (0.002)

cross shelf transects using a commercial trawl net (Bax and Williams 2000, page 191), and for each of the species in the sample, the number of individuals and their total mass in kilograms were recorded. A total of 70 species were used in the analysis.

We used the presence–absence model outlined in Dunstan, Foster, and Darnell (2011) to help to identify a set of nine environmental covariates for modeling: latitude, depth,

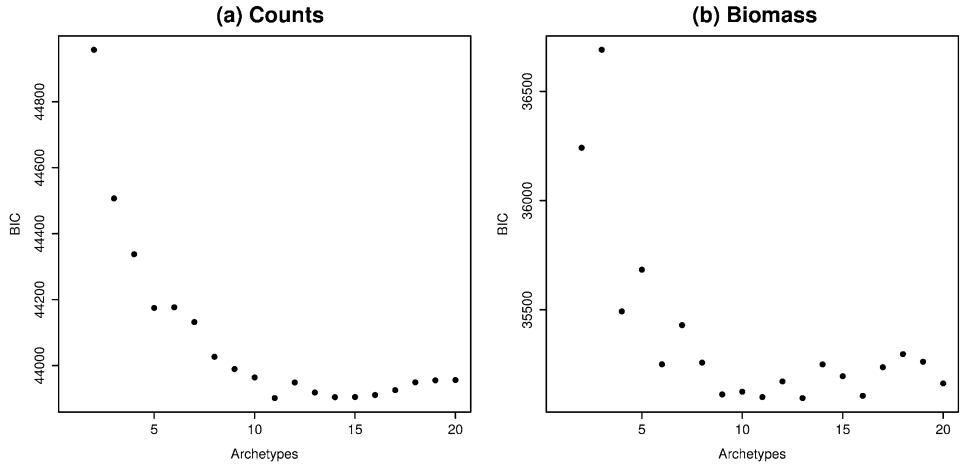

Figure 2. Plots of BIC vs. number of species archetypes using (a) count data, (b) biomass data. Note that a minimum is achieved at 11 archetypes for counts, and the 11-archetype biomass model is close to the minimum BIC.

percent carbonate, percent gravel, percent sand (derived from Geoscience 2009), and the intra-annual standard deviation in the concentrations of nitrate, phosphate, oxygen, and temperature derived from the CSIRO Atlas of Regional Seas (Dunn and Ridgway 2002; Ridgway, Dunn, and Wilkin 2002). The mean values of nitrate, phosphate, oxygen, and temperature were strongly correlated with depth and provided no additional explanatory power, whereas the standard deviations reflect seasonal variation in these variables. Trawl area was estimated from the length of each tow and the known width of the trawl mouth and (*area*) used as an offset in the count and biomass models. The number of components was chosen using BIC in (2.4) for counts and biomass separately, see Figure 2. To safeguard against convergence to local maxima, we re-fitted each model 20 times, keeping the model which achieved the highest likelihood value as defined in Equation (2.2).

The BIC dropped steeply from to $K = 1$ to 5 archetypes, and was minimized at $K = 11$ ($K = 13$) for count (biomass) data. Since the 11-archetype biomass model produced a BIC value close to the minimum achieved (Figure 2), then we opted to use $K = 11$ archetypes for both count and biomass SAMs.

Residual plots exhibited little pattern (Figure 3), suggesting the 11-archetype as a potential model for both counts and biomass. In particular, the lack of any funnel-shaped trend in the residuals suggests our choice of distribution types and model for the mean and variance relationship are reasonable. However, one noteworthy departure from the assumed model was a small cluster of unusually large residuals ($r_{ij} > 5$), suggesting a group of observations with y_{ij} larger than expected were not adequately captured by the fitted model. Future analyses could address this apparent departure, perhaps using different predictor variables or a different functional form for the mean structure, e.g., fitting mixtures of additive models.

Prediction of archetypes was conducted using broad scale environmental gradient datasets for south east Australia, using mean area trawled as an offset. Plots of selected

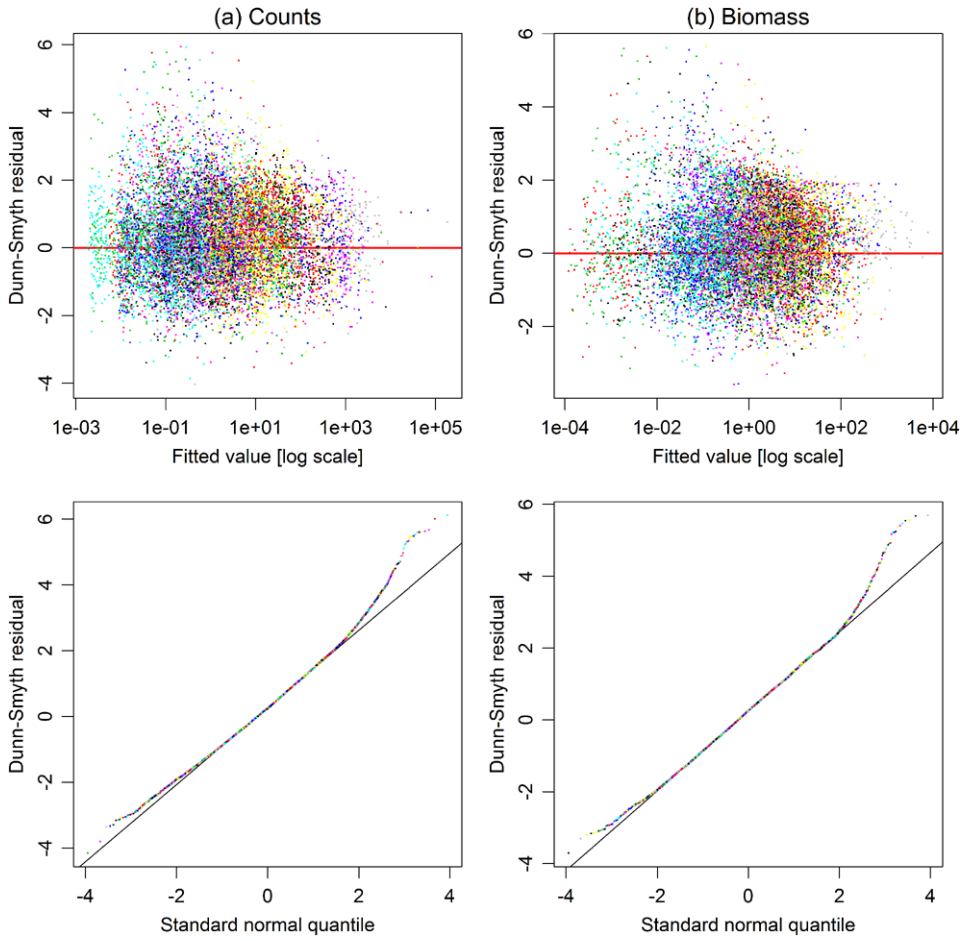

Figure 3. Residual vs. fits plots (top row) and normal quantile plots of residuals (bottom row) for species archetype models of (a) count data, using a negative binomial model, and (b) biomass data, using a Tweedie model. Different species are plotted with different colors. The lack of pattern in the residual plots in conjunction with the quantile plots suggest a reasonable model fit, with the exception of a small cluster of observations with small fitted and large residual. Quantile plots suggest this cluster is the main exception to normality of residuals.

archetypes are shown for counts and biomass in Figure 4. The predictions exhibit complex patterns for both biomass and counts. There were a suite of archetypes associated with inner shelf areas, with the mid shelf and with the shelf edge and upper slope. The scale of changes in abundance across these regions was typically large with respect to the standard errors (see supplementary material). A few of the predictions for both biomass and counts were very large. These may be due to prediction in extreme and/or unsampled locations in covariate space. However, it is worth noting that fish catches in the South Eastern Shelf can be extreme in this region and these values are not that unreasonable.

There were some obvious similarities in some of the spatial patterns from archetypes based on count data and archetypes based on biomass data. The right-most two columns in Figure 4 highlight this; that is Figure 4 panels (b) and (e), and panels (c) and (f). However, it is not guaranteed that the archetypes from the different data types will match up.

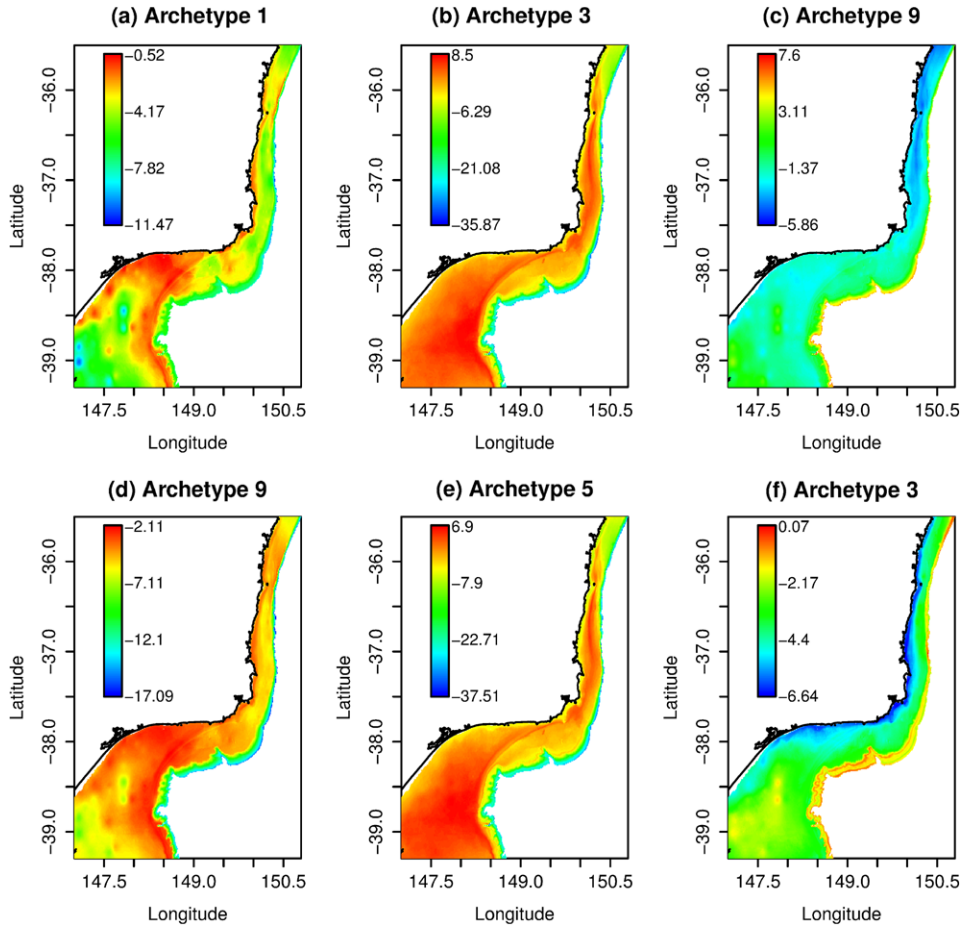

Figure 4. Predictions (log scale) for selected archetypes from the 11-archetype model in south east Australia for counts (top row) and biomass (bottom row).

We can summarize the connectivity between the two sets of archetype groups by studying their $S \times K$ matrices of posterior archetype probabilities. Full tables are presented in supplementary material (Supplementary Tables 1 and 3), but a species list is presented for each of the archetypes in Figure 4, listing any species whose posterior probability was highest for the given archetype. In most cases the maximum posterior probability for a species was very close to one, meaning that archetypal responses can usually be interpreted synonymously with species responses, considerably simplifying interpretation. Now compare the species lists (Table 2) across count and biomass models for the three example archetypes in Figure 4.

- Count Archetype 1 shared three of its seven species with biomass Archetype 9 (Table 2). This explains the difference in the maps in Figure 4(a) and (d). Two additional species were present in the corresponding biomass archetype.

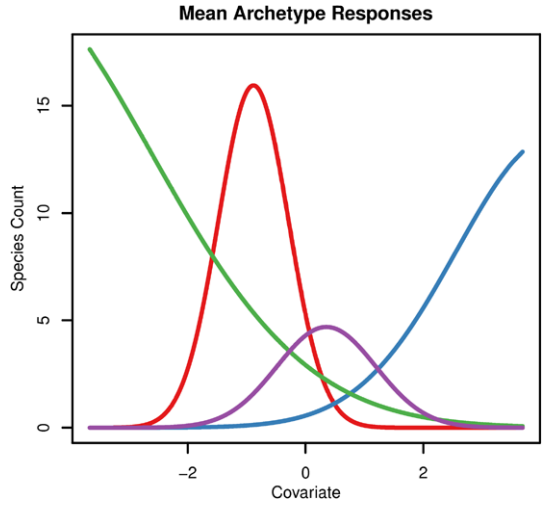

Figure 5. Archetype responses to the covariate for the simulation study. Each curve represents the archetypal response of one of the four different archetypes, from which species data were generated.

- Count Archetype 3 had identical composition to biomass Archetype 5, sharing the same seven species (Table 2). The archetype maps were almost identical, see Figure 4(b) and (e).
- Count Archetype 9 shared species with biomass Archetype 3 only, see Table 2. Biomass Archetype 3 also shared species with other count archetypes. This explains some of the differences in Figure 4(c) and (f). In some respects biomass Archetype 3 is a 'super set' of count Archetype 9.

Overall, there was good agreement between the count and biomass archetype models, with similar estimated archetypes and similar species compositions across the two models. This level of general agreement should be expected as the two measurement types are correlated—the simulation model in Section 3 provides one mechanism for inducing this correlation. Some archetypes matched exactly in species composition (e.g., count Archetype 3 and biomass Archetype 5), while other archetypes only partially matched (e.g., count Archetype 1 only shared half the species of biomass Archetype 9). Full details of the fitted values and standard errors for count and biomass models, the matrices $\{\boldsymbol{\tau}_j\}$ for count and biomass models and predictions of each archetype and the standard error of predictions for both models can be found in the supplemental material.

5. DISCUSSION

In this paper, we proposed a general method for analyzing multi-species data in ecology, and applied it to two commonly encountered types of multi-species data: count and biomass data. The proposed method addresses a clear need in ecology to deal with high-dimensional, multi-species data. The SAM considerably simplifies the problem of understanding the diverse ways in which different species respond to their environment. By

Table 2. The species belonging to Archetypes 1, 3, and 9 for abundance and 9, 5, 3 for biomass. Values for$\{\boldsymbol{\tau}_j\}$ can be found in supplemental Tables 1 and 3—the relevant posterior probabilities are one in most cases, to two decimal places.

Abundance		
Archetype 1	Archetype 3	Archetype 9
Urolophus paucimaculatus	Asymbolus analis	Squalus megalops
Neosebastes scorpaenoides	Scorpaena papillosa	Cyttus novaezelandiae
Chelidonichthys kumu	Caesioperca lepidoptera	Seriolella punctata
Lepidotrigla vanessa	Pagrus auratus	
Kathetostoma laeve	Nemadactylus douglasi	
Synchiropus calauropomus	Eubalichthys mosaicus	
Diodon nicthemerus	Thamnaconus degeni	

Biomass		
Archetype 9	Archetype 5	Archetype 3
Trygonoptera sp B	Asymbolus analis	Cephaloscyllium sp A
Myliobatis australis	Scorpaena papillosa	Squalus megalops
Neosebastes scorpaenoides	Caesioperca lepidoptera	Zenopsis nebulosus
Lepidotrigla vanessa	Pagrus auratus	Cyttus novaezelandiae
Diodon nicthemerus	Nemadactylus douglasi	Hoplichthys haswelli
	Eubalichthys mosaicus	Lepidoperca pulchella
	Thamnaconus degeni	Apogonops anomalus
		Emmelichthys nitidus nitidus
		Nemadactylus macropterus
		Kathetostoma canaster
		Rexea solandri
		Lepidopus caudatus
		Seriolella punctata

classifying S species into K different archetypal environmental response types, the analyst can focus on understanding the nature of environmental response in the K archetypes and, together with group membership probabilities, build up a picture of the assemblage as a whole.

In our example analysis of Section 4, we found significant overlap between the spatial distributions of the archetypes, a pattern which has broader ecological implications. This pattern has been observed over a number of different analyses (e.g., Dunstan, Foster, and Darnell 2011). It suggests that the archetypes we are modeling can not be understood as communities, and more broadly, our data are not consistent with the local concept of community assembly (Ricklefs 2008). If archetypes behaved as communities, they would be aggregated in coherent patches with little overlap. The lack of patchiness suggests that perhaps communities *per se* are not the natural unit to study—which questions the value of studying the spatial distribution of communities directly as in Anderson et al. (2011) or Li, Ban, and Santiago (2011). Rather, the communities we have encountered are better thought of as assortments (or "assemblages") of co-occurring species which have their own patterns of distribution. This is classically known as the individualistic concept of commu-

nity assembly (Gleason 1926), and its implication is that species are the natural unit to be studied, as in a SAM. A SAM puts a slight twist on this individualistic concept, positing that many species share similar "archetypal" environmental responses. This reduction in dimensionality of the fitted model considerably simplifies not just model interpretation, but potentially, conservation management decisions—the problem of managing hundreds of species can be simplified to that of managing a small number of archetypes.

Species Archetype Models as proposed here have been demonstrated to provide adequate fits to multi-species data (Figure 3). Additionally, when applied to count or biomass data, SAMs provide a more nuanced view of the multi-species assemblage than would have been available in an analysis of presence–absence data only (Table 1). Related work used the methods proposed in this paper to show that SAMs have enhanced predictive performance as compared to using separate regression models for each species (Hui et al. 2013). Hence SAMs are not only desirable from the perspective of simplifying interpretation, but also in terms of predictive accuracy.

One important advance in the SAM defined here over the method of Dunstan, Foster, and Darnell (2011) is the use of species-specific intercepts. Such an extension is essential here to account for the fact that different fish species have different sizes and social characteristics. Furthermore, it is this extension that allows species within an archetype to share the same form of environmental response, but have differing absolute count or biomass levels. For counts and presence–absence, it could be argued that while different species may share a form of environmental response, they need not share the same level of abundance or prevalence. For this reason, species-specific intercepts could be justified in the analysis of all types of multi-species data, not just biomass.

There remain considerable opportunities for improvement of the SAM approach used in this paper. The species-specific intercept parameters introduced S parameters into the model, and it would have been quite natural to model these using random effects in place of the fixed effects employed here. This is however a challenging problem from a computational perspective—essentially, it requires the fitting of finite mixtures of generalized linear mixed models. Also, as is usual in mixture modeling (McLachlan and Peel 2000), estimation was complicated by the presence of multiple local maxima on the likelihood surface. One potential solution to this problem is to penalize coefficients in order to regularize the estimation problem (Khalili and Chen 2007).

An important issue not addressed in this paper, but which would be a valuable extension of our approach, is extending the model to better handle correlation between species or between sites. Correlation between species can be anticipated via species interactions. Correlation between sites can occur due to spatial autocorrelation. The usual consequences of omitting important sources of correlation from a model are loss of efficiency and biased estimation of model precision, although unbiased estimates of target model parameters (as in Figure 4) can usually still be obtained. While it would be desirable to incorporate correlation between species or between sites into a SAM, it would be challenging from a computational perspective. Inter-species correlation in particular would be very difficult to model because S is not small compared to N. Ovaskainen, Hottola, and Siitonen (2010) proposed addressing this issue using a pair-wise likelihood, Ives and Helmus (2011) assumed species

correlations operate via phylogeny, whilst others (e.g. Warton 2011) assumed working independence and then resampled clusters of observations to make multivariate inferences robust to potential failure of correlation assumptions. Constructing plausible models for inter-species correlation which can be implemented in multi-species analysis is an important challenge that remains to be addressed adequately.

A final issue not discussed in this paper is the question of model selection—how to select the environmental variables to include in a SAM, and how to select the form of model for the mean of the y_{ij}. Diagnostic tools such as Figure 3 can be of some assistance. But a key issue in the mixture modeling literature for some time (McLachlan and Peel 2000) has been the question of how to formally derive an information criterion with good properties that can be used to inform concerning model choice for mixture models and FMRs, including generalizations of FMRs as considered in this paper. Addressing this deficiency is the subject of current research.

SUPPLEMENTARY MATERIALS

Tables of posterior probabilities, parameter estimates and standard errors, and maps of predictions and standard errors for all archetypes (pdf file).

ACKNOWLEDGEMENTS

PKD and SDF are supported by the Marine Biodiversity Hub through the Australian Government's National Environmental Research Program (NERP). DIW is supported by Australian Research Council Discovery Projects and Future Fellow funding schemes (project number DP130102131 and FT120100501). FKCH is supported by a Research Excellence Award from The University of New South Wales and a CSIRO CMIS PhD Scholarship.

[Published Online May 2013.]

REFERENCES

Aitkin, M., and Aitkin, I. (1996), "A Hybrid EM/Gauss–Newton Algorithm for Maximum Likelihood in Mixture Distributions," *Statistics and Computing*, 6, 127–130.

Anderson, M. J., Crist, T. O., Chase, J. M., Vellend, M., Inouye, B. D., Freestone, A. L., Sanders, N. J., Cornell, H. V., Comita, L. S., Davies, K. F., Harrison, S. P., Kraft, N. J. B., Stegen, J. C., and Swenson, N. G. (2011), "Navigating the Multiple Meanings of β Diversity: A Roadmap for the Practicing Ecologist," *Ecology Letters*, 14, 19–28.

Bax, N., and Williams, A. (2000), "Habitat and Fisheries Production in the South East Fishery Ecosystem," Final Report to the Fisheries Research and Development Corporation, Project No. 94/040.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, 39, 1–38.

Dunn, J. R., and Ridgway, K. R. (2002), "Mapping Ocean Properties in Regions of Complex Topography," *Deep-Sea Research. Part 1. Oceanographic Research Papers*, 49, 591–604.

Dunn, P., and Smyth, G. (2005), "Series Evaluation of Tweedie Exponential Dispersion Model Densities," *Statistics and Computing*, 15, 267–280.

Dunn, P. K., and Smyth, G. K. (1996), "Randomized Quantile Residuals," *Journal of Computational and Graphical Statistics*, 5, 236–244.

Dunstan, P., Foster, S., and Darnell, R. (2011), "Model Based Grouping of Species Across Environmental Gradients," *Ecological Modelling*, 222, 955–963.

Foster, S., and Bravington, M. (2013), "A Poisson-Gamma Model for Analysis of Ecological Non-negative Continuous Data," *Journal of Environmental and Ecological Statistics*, in press.

Geoscience Australia (2009), "GA Australian Bathymetry and Topography Grid, ANZLIC Metadata ANZCW0703013116." Tech. rep., Australian Government Geoscience Australia.

Gleason, H. A. (1926), "The Individualistic Concept of the Plant Association," *Bulletin of the Torrey Botanical Club*, 53, 7–26.

Hilbe, J. M. (2007), *Negative Binomial Regression*, Cambridge: Cambridge University Press.

Hui, F. K. C., Warton, D. J., Foster, S., and Dunstan, P. (2013), "To Mix or Not to Mix: Comparing the Predictive Performance of Mixture Models Versus Separate SDMs," *Ecology*, in press.

Ives, A. R., and Helmus, M. R. (2011), "Generalized Linear Mixed Models for Phylogenetic Analyses of Community Structure," *Ecological Monographs*, 81, 511–525.

Jørgenson, B. (1997), *The Theory of Dispersion Models*, London: Chapman and Hall.

Keribin, C. (2000), "Consistent Estimation of the Order of Mixture Models," *Sankhya. The Indian Journal of Statistics*, 62, 49–66.

Khalili, A., and Chen, J. (2007), "Variable Selection in Finite Mixture of Regression Models," *Journal of the American Statistical Association*, 102, 1025–1038.

Li, J., Ban, J., and Santiago, L. (2011), "Nonparametric Tests for Homogeneity of Species Assemblages: A Data Depth Approach," *Biometrics*, 67, 1481–1488.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models* (2nd ed.), London: Chapman & Hall.

McLachlan, G., and Peel, D. (2000), *Finite Mixture Models*, New York: Wiley.

Nash, S. G., and Sofer, A. (1996), *Linear and Nonlinear Programming* (1st ed.), *McGraw-Hill Series in Industrial Engineering and Management Science*, New York: McGraw-Hill Inc.

Novotny, V., Miller, S., Hulcr, J., Drew, R., Basset, Y., Janda, M., Setliff, G., Darrow, K., Stewart, A., Auga, J., Isua, B., Molem, K., Manumbor, M., Tamtiai, E., Mogia, M., and Weiblen, G. (2007), "Low Beta Diversity of Herbivorous Insects in Tropical Forests," *Nature*, 448, 692–695.

Oehlert, G. (1992), "A Note on the Delta Method," *American Statistician*, 46, 27–29.

Ovaskainen, O., Hottola, J., and Siitonen, J. (2010), "Modeling Species Co-occurrence by Multivariate Logistic Regression Generates New Hypotheses on Fungal Interactions," *Ecology*, 91, 2514–2521.

Ovaskainen, O., and Soininen, J. (2011), "Making More Out of Sparse Data: Hierarchical Modeling of Species Communities," *Ecology*, 92, 289–295.

Peel, D., Bravington, M. V., Kelly, N., Wood, S. N., and Knuckey, I. (2013), "A Model-Based Approach to Designing a Fishery Independent Survey," *Journal of Agricultural, Biological and Environmental Statistics*, 18, 1–21.

Ricklefs, R. E. (2008), "Disintegration of the Ecological Community," *The American Naturalist*, 172, 741–750.

Ridgway, K. R., Dunn, J. R., and Wilkin, J. L. (2002), "Ocean Interpolation by Four-Dimensional Weighted Least Squares—Application to the Waters Around Australia," *Journal of Atmospheric and Oceanic Technology*, 19, 1357–1375.

Ross, L., Woodin, S., Hester, A., Thompson, D., and Birks, H. (2012), "Biotic Homogenization of Upland Vegetation: Patterns and Drivers at Multiple Spatial Scales Over Five Decades," *Journal of Vegetation Science*.

Taylor, L. (1961), "Aggregation, Variance and the Mean," *Nature*, 189, 732–735.

Thibault, K., Supp, S., Giffin, M., White, E., and Ernest, S. (2011), "Species Composition and Abundance of Mammalian Communities," *Ecology*, 92, 2316.

Venables, W. N., and Ripley, B. D. (1999), *Modern Applied Statistics With S* (4th ed.), New York: Springer.

Warton, D. I. (2011), "Regularized Sandwich Estimators for Analysis of High Dimensional Data Using Generalized Estimating Equations," *Biometrics*, 67, 116–123.

Warton, D. I., Wright, S. T., and Wang, Y. (2012), "Distance-Based Multivariate Analyses Confound Location and Dispersion Effects," *Methods in Ecology and Evolution*, 3, 89–101.

Wedel, M., and DeSarbo, W. (1995), "A Mixture Likelihood Approach for Generalized Linear Models," *Journal of Classification*, 12, 21–55.

Weisberg, S. (2005), *Applied Linear Regression* (3rd ed.), Hoboken: Wiley.

Yee, T. W. (2010), "The VGAM Package for Categorical Data Analysis," *Journal of Statistical Software*, 32, 1–34.