Biometrics 66, 186–195
March 2010
DOI: 10.1111/j.1541-0420.2009.01263.x

The Analysis of Biodiversity Using Rank Abundance Distributions

Scott D. Foster[1,*] and Piers K. Dunstan[2,]**

[1]CSIRO Mathematical and Information Sciences, GPO Box 1538, Hobart 7001, Tasmania, Australia
[2]CSIRO Marine and Atmospheric Research, GPO Box 1538, Hobart 7001, Tasmania, Australia
[*]*email:* scott.foster@csiro.au
[**]*email:* piers.dunstan@csiro.au

Summary. Biodiversity is an important topic of ecological research. A common form of data collected to investigate patterns of biodiversity is the number of individuals of each species at a series of locations. These data contain information on the number of individuals (abundance), the number of species (richness), and the relative proportion of each species within the sampled assemblage (evenness). If there are enough sampled locations across an environmental gradient then the data should contain information on how these three attributes of biodiversity change over gradients. We show that the rank abundance distribution (RAD) representation of the data provides a convenient method for quantifying these three attributes constituting biodiversity. We present a statistical framework for modeling RADs and allow their multivariate distribution to vary according to environmental gradients. The method relies on three models: a negative binomial model, a truncated negative binomial model, and a novel model based on a modified Dirichlet-multinomial that allows for a particular type of heterogeneity observed in RAD data. The method is motivated by, and applied to, a large-scale marine survey off the coast of Western Australia, Australia. It provides a rich description of biodiversity and how it changes with environmental conditions.

Key words: Abundance data; Biodiversity; Dirichlet-multinomial; Evenness; Negative binomial; Overdispersion; Rank abundance distribution; Species abundance distribution; Species richness; Truncated negative binomial.

1. Introduction

Understanding patterns of biodiversity is a key goal from a scientific and natural resource management perspective. Consistent definitions of biodiversity that target important aspects of ecological processes are necessary for understanding how biodiversity changes over environmental gradients.

In this article, we consider a common form of ecological data collected to study biodiversity: species counts in a sample from a community. Biodiversity from these data can be usefully quantified by considering the following attributes of biodiversity:

(1) A measure of the total number of organisms (abundance),
(2) A measure of the number of species within the sample (richness or α-diversity), and
(3) A measure of the relative proportional abundances of the species (evenness).

There are other definitions of biodiversity and they may be more important in specific contexts. However, the three attributes listed above have an almost universal appeal.

Our interest in biodiversity stems from modeling these attributes as functions of environmental gradients in the marine environment. The motivating data for the methods in this article are the species enumerations from the Voyage of Discovery (VoD; Section 2). The survey spanned a large area of the marine environment off the coast of Western Australia, Australia.

Researchers often use diversity indices as a measure of biodiversity and they have been used for parametric modeling (e.g., Gutiérrez-Estrada, Vasconcelos, and Costa, 2008). These indices are likely to be an oversimplification of the complexities of biodiversity. Another approach to investigating biodiversity is to use multivariate methods where species identities are preserved (e.g., Anderson, Ellingsen, and McArdle, 2006). This approach is powerful but is not suitable for the modeling task addressed in this article as these methods test changes in biodiversity between sites (β-diversity) and do not formally model biodiversity on the environment. A full multivariate analogy to the methods presented in this article that retains species identity would require parametric modeling of each species' abundance as functions of environmental gradients as well as specification of species covariances. Typically, the amount of data required to perform this analysis is not available and a much simpler solution is needed.

The method proposed in this article is based on the rank abundance distribution (RAD, see McGill et al., 2007). A RAD is a representation of species enumeration data that is applicable to all environments. RADs allow comparisons of samples taken from geographically separated locations that have few or no species in common. Many researchers, e.g., Wilson (1991), Beck and Vun Khen (2006), and McGill et al. (2007), choose to represent data as RADs and consider that the RAD is a fundamental quantity of the community. However, cohesive statistical studies of the sampling properties of RADs have been, until this article, absent from the literature (McGill et al., 2007).

The proposed models are demonstrated using the VoD example throughout the article. For comparison we also analyze these data using biodiversity metrics. The interpretation of

the data presented here is intended to be illustrative only. For a thorough ecological interpretation of the analysis we refer to a companion article, which is in preparation.

2. Voyage of Discovery Data

The primary goal of the VoD was to characterize benthic ecosystems on the continental shelf and slope at water depths from 100 to 1500 m. Data on species abundances were collected using samples from benthic tows. The length of tow for each sample was variable (range of 96–3265 m). The data consist of all benthic species from six phyla (Mollusca, Echinodermata, Pycnogonida, Ascidiacea, Cnidaria, and Crustacea). There were 120 locations sampled successfully, with one sample at each location. See Figure 1 for the spatial locations of the samples.

There were a total of 1548 species identified over all the locations. Of these 55.5% were found at only 1 location, 89.7% were found at 5 locations or less, and the most common species was found at only 25 locations.

The samples were accurately geolocated which enabled cross-referencing with the CSIRO Atlas of Regional Seas oceanography data. The oceanographic data are the time averaged means and intraannual standard deviations of temperature, salinity, and oxygen. They are interpolations of observed data to the locations of the samples (Ridgway, Dunn, and Wilkin, 2002) and are based on approximately 3700 vertical profiles for temperature and salinity, and 2600 vertical profiles for oxygen. These profiles have been collected over a period of 60 years with reliable data density for the last 40 years. The particular oceanographic quantities used in this study were chosen, in conjunction with depth and latitude, as they are known to affect individual species' distributions (e.g., Jennings et al., 1999; Callaway et al., 2002).

3. Rank Abundance Distributions

The RAD for S species at a single site is $\boldsymbol{n} = (n_1, n_2, \ldots, n_S)$, where $n_k, 1 \leq k \leq S$, is the abundance of the kth most abundant species at this site. RADs are applicable to any type of environment because there is no dependence on the presence of a particular set of species. This is particularly attractive for data such as the VoD data as many species are present at a very small number of sites.

Representing community data with RADs makes specific assumptions about ecological process. It assumes that there is some degree of symmetry between the species; one species can be replaced by another without affecting the community function (see McGill et al., 2007). This is a strong assumption and may not hold due to species differences (e.g., demography, biomass, trophic level). RADs also ignore any correlations between abundances of species, which is likely to result in a loss of information obtainable from the data. In spite of these issues we, like many others (e.g., Wilson, 1991; Beck and Vun Khen, 2006; McGill et al., 2007), consider RADs to be a useful tool for describing these types of data.

RADs potentially differ between environments in three ways: the total number of individuals, the total number of species, and/or the relative abundance of the species. These are precisely the attributes of biodiversity that we wish to investigate. Hence, modeling RAD data is synonymous to modeling attributes of biodiversity listed in the Introduction.

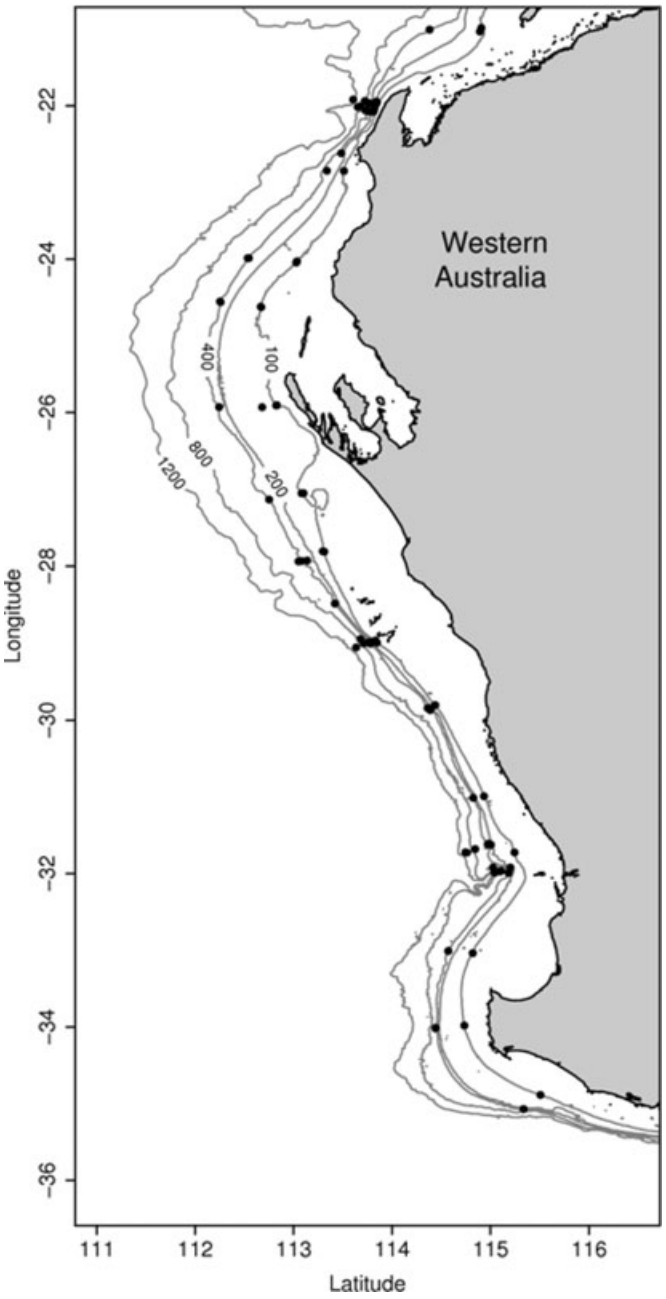

Figure 1. Locations of the 120 samples off the Western Australian coast with selected depth contours.

A closely related quantity to the RAD is the species abundance distribution (SAD). It is the distribution of the number of species with different levels of abundance in the sample. SADs and RADs are different, but similar, representations of the observed data (Pielou, 1975; McGill et al., 2007). We prefer the RAD representation as the three elements of biodiversity emerge as natural attributes. Also, many ecological models, such as niche preemption models (Pielou, 1975), attempt to determine community composition using RADs

as the fundamental quantity (e.g., the broken stick model of Whittaker, 1972).

The literature on data analytic approaches for SADs/RADs predominantly centers around SADs (e.g., Fisher, Corbet, and Williams, 1943; Bulmer, 1974) with only limited studies on RADs (Wilson, 1991; Mac Nally, 2007). Both these RAD studies assume that the ranked abundances are normally distributed about their means. This assumption is unlikely to hold for data such as the VoD data, as they are counts whose mean is low for most species.

Investigation of the way that RADs change with environmental gradients has typically been limited to separately plotting RADs at different sites and visually comparing them (e.g., Whittaker, 1965), or by categorizing the different sites by what type of model fits best (Wilson, 1991; Beck and Vun Khen, 2006; Mac Nally, 2007). The visual inspection approach is clearly inefficient, especially when there are multiple environmental gradients. The categorization approach fails to model the relationship of the RAD with the environmental variables.

4. Modeling Elements of a RAD

In its rawest form the data constituting a RAD for the ith sample is given by the $S_i \times 1$ vector $\boldsymbol{n}_i$ and is the sampled species' ranked abundances. The vector $\boldsymbol{n}_i$ contains information on the total number of individuals N_i, the relative proportions of each of the rankings, and its length gives the number of species.

Modeling the first three attributes of biodiversity using RADs is equivalent to modeling the joint distribution of $(S_i,\ \boldsymbol{n}_i)$ or equivalently $(S_i,\ N_i,\ \boldsymbol{n}_i)$, where $\boldsymbol{n}_i$ in the second representation is constrained to sum to N_i. The joint distribution can be factorized as

$$\begin{aligned}\Pr(S_i, N_i, \boldsymbol{n}_i) &= \Pr(N_i)\Pr(S_i, \boldsymbol{n}_i \mid N_i) \\ &= \Pr(N_i)\Pr(S_i \mid N_i)\Pr(\boldsymbol{n}_i \mid N_i, S_i). \qquad (1)\end{aligned}$$

We condition on abundance, N_i, first as we consider the fundamental quantity in the data is an individual; individuals occupy space within the sampling area and their species is secondary. We then condition on the number of species, S_i, as relative abundance is meaningless without knowing how many species. We acknowledge that factorization (1) could be performed in a number of ways and that none of the factorizations are statistically preferable. However, we choose factorization (1) for the above ecological reasons.

The task of modeling the joint distribution is reduced to modeling three separate quantities. Details of the models will be given in the following three sections. Developmental `R` code is available from the authors upon request.

5. Total Number of Individuals, N_i

A common method for modeling ecological abundances (counts) is to use a negative binomial (NB) model. It allows for extra variation with respect to a Poisson model, a phenomenon commonly encountered in ecological data for various reasons (e.g., McArdle and Andersen, 2004). The mean for the ith sample is given by the relation $\log(\mu_i) = \boldsymbol{x}_i^\top \boldsymbol{\tau}_N + \log(d_i)$, where $\boldsymbol{x}_i^\top$ is the ith row of a suitably chosen design matrix containing the environmental covariates, $\boldsymbol{\tau}_N$ is the mean parameter vector, d_i is the length of tow in meters, and $\log(d_i)$ is an offset term for standardizing for the varying length of tow. This model has a single parameter for overdispersion, θ_N.

The model for the mean must be specified, and to do this parsimoniously some sort of model selection is required. We use a forward selection procedure that considers inclusion of first-, second-, and third-order polynomials of covariates not already included in the model, as well as all linear two-way interactions of terms already in the model. The selection process was terminated when the addition of all terms not in the model increased the Akaike information criteria (AIC). We diagnose the model using the randomized quantile residuals of Dunn and Smyth (1996) but do not include the randomization step. We feel that the randomization step does not aid diagnosis as it injects an element of randomness into the procedure which is unnecessary for these models.

5.1 N_i *in the VoD Data*

There were three environmental covariates identified by the model selection process. These are given in order of selection in Table 1 along with the estimates from the selected model. The AIC value for the final model is substantially smaller than the initial model, indicating that the environmental variables are useful for modeling purposes.

The selected model contains the covariates: depth, variation in temperature, and latitude. Deeper waters tend to have lower abundance than shallower waters, agreeing with previous studies (e.g., Maldonado and Young, 1996; Koslow, Williams, and Paxton, 1997; Labropoulou and Papaconstantinou, 2004). Environments with high and low variation in temperature tend to have less abundance than sites with a more moderate amount of temperature variation (temperature standard deviation about 0.6°C). The southern waters tended to have less abundance than the northern waters.

6. Conditional Species Richness, $S_i \mid N_i$

The model for $S_i \mid N_i$ should reflect the relationship $S_i \leq N_i$. To account for this we define a truncated negative binomial (TNB) model. The model is parameterized by a vector of location parameters $\boldsymbol{\tau}_S$ and an overdispersion parameter θ_S. Define $\log(\kappa_i) = \boldsymbol{x}_i^\top \boldsymbol{\tau}_S + \log(d_i)$, where $\boldsymbol{x}_i^\top$ is the ith row of a design matrix and d_i is the ith sample's length of tow. The log likelihood for the TNB model is defined to be

$$\begin{aligned}\ell_S(\boldsymbol{\tau}_S, \theta_S; \boldsymbol{S} \mid \boldsymbol{N}) = \sum_{i=1}^{T} [&\log\{\Pr(Y = S_i)\} \\ &- \log\{\Pr(Y \leq N_i)\}], \qquad (2)\end{aligned}$$

where $\boldsymbol{S}$ is the vector containing all the sample species richness values, $\boldsymbol{N}$ is the corresponding vector for abundances, T is the number of samples, $\Pr(Y = S_i)$ is the probability from a NB model with mean κ_i, overdispersion parameter θ_S, and N_i is the observed abundance for the ith sample.

We again use the forward selection method already introduced in Section 5. Scaled abundance N_i/d_i is added to the set of potential explanatory variables. This addition completes the conditioning of S_i on N_i and is used in preference to unscaled abundance to ensure that units of the outcome and this covariate agree.

Table 1
Summary of selected model and the model-fitting processes for abundance, richness, and evenness

Model	Polynomial term	Number of parameters[a]	AIC[a]	Estimate[b]	SE[a*]
Total abundance N_i					
dispersion (θ_N)	–			1.220	0.144
mean	–	2	1411.88	−2.274	0.747
+ depth	1	3	1386.29	−0.002	4.781×10^{-4}
+ temperature SD[**]	1			0.992	0.860
	2	5	1380.32	−0.803	0.334
+ latitude	1	6	1375.41	−0.046	0.017
Richness $S_i \mid N_i$					
dispersion (θ_S)	–			11.668	2.431
mean	–	2	953.61	−21.383	7.032
+ depth	1			−0.006	0.001
	2	5	883.99	9.528×10^{-6}	1.560×10^{-6}
	3			-4.444×10^{-9}	8.016×10^{-10}
+ scaled abundance	1			6.474	0.690
	2	7	837.31	−6.828	0.914
+ salinity	1	8	832.71	0.521	0.197
Evenness $\boldsymbol{n}_i \mid S_i,\ N_i$					
dispersion (θ)	–			16.567	2.600
dispersion (ν)	–	3	9831.56	2.799	0.167
mean	–			0.124	1.172
+ temperature	1			0.127	0.069
	2	6	9602.02	−0.009	0.005
	3			2.081×10^{-4}	1.093×10^{-4}
+ scaled abundance	1			7.797	0.966
	2	9	9458.08	−10.084	1.992
	3			5.419	1.690
+ scaled richness	1			−16.868	2.159
	2	11	9371.68	56.988	10.638
+ salinity SD[**]	1			−8.045	2.229
	2	13	9356.60	30.772	7.651
+ oxygen	1	14	9350.36	0.146	0.039
+ abundance : oxygen	1 : 1	15	9345.81	−0.337	0.132

[*]Standard error.
[**]Standard deviation.
[a]From model with terms higher in the table.
[b]From final model.

Expected values, for the TNB model are given by

$$\mathrm{E}\,(S_i \mid N_i) = \frac{1}{\sum_{j=0}^{N_i} \Pr(Y=j)} \sum_{j=0}^{N_i} j \Pr(Y=j), \qquad (3)$$

where $\Pr(Y=j)$ is the NB distribution in equation (2). Standard errors for this expectation can be found using parametric bootstrap methods (Davison and Hinkley, 1997).

The TNB model for richness is conditional on, not marginal to, abundance. In Section 5, a model was specified for abundance and hence marginalization can occur, via bootstrap methods. A bootstrap sample of $(\hat{\boldsymbol{\tau}}_N, \hat{\theta}_N)$ and $(\hat{\boldsymbol{\tau}}_S, \hat{\theta}_S)$ is generated from their asymptotic distributions, and a bootstrap sample N^* is drawn from the NB model using the sampled $(\hat{\boldsymbol{\tau}}_N, \hat{\theta}_N)$ as plug-in values. Confidence intervals are obtained using the expectation (3) calculated for each of the bootstrap samples. Prediction intervals are generated by further sampling from the TNB model with $(\hat{\boldsymbol{\tau}}_S, \hat{\theta}_S)$ as plug-in values.

6.1 *S_i in the VoD Data*

The variables included in the model for $S_i \mid N_i$, their order of inclusion, their associated reduction in AIC, and the parameter estimates are given in Table 1. The model shows that the environmental variables explain a substantial amount of variation. The estimate of the dispersion parameter suggests that the simpler truncated Poisson model is inadequate.

Consider the relationship with depth and richness, either conditional on, or marginal to, abundance (Figure 2). The model shows a decrease in species richness as depth increases and this decrease is nonlinear. This relationship is in agreement with other studies in this area (Koslow et al., 1997) and in other areas (e.g., Maldonado and Young, 1996). However, studies in shallower European marine biomes show that the relationship is different (Labropoulou and Papaconstantinou,

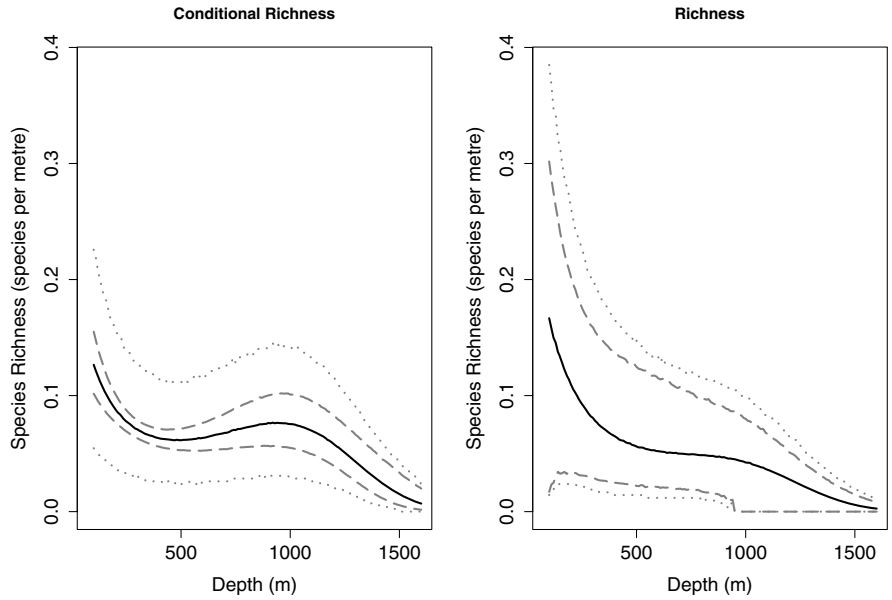

Figure 2. Modeled partial effects of depth on species richness. Left panel are predictions conditional on total abundance ($N_i = \bar{N} = 135.18$). Right panel are predictions marginal to total abundance. Both panels are predictions with salinity held constant at its mean. Conditioning on total abundance and salinity were performed to isolate the effect of depth. The solid line is the mean, the dashed lines are the 95% confidence intervals, and the dotted lines are the 95% prediction intervals. A total of 20,000 bootstrap samples were used to generate the predictions and intervals.

2004; Sousa, Azevedo, and Gomes, 2006). This discrepancy may highlight that such relationships are reflective of particular study areas such as oceanographic conditions and community composition. The confidence intervals for the marginal relationship is wider than that for the conditional relationship (Figure 2) and is due to uncertainty in abundance.

7. Conditional Evenness, $\boldsymbol{n_i} \mid \boldsymbol{S_i}, \boldsymbol{N_i}$

An obvious modeling framework for $\boldsymbol{n}_i \mid N_i, S_i$ is multinomial (Mn) regression, as the $\boldsymbol{n}_i$ is a partition of N_i individuals into S_i categories. The variance structure of the Mn model may not adequately reflect the observed data; ranking will affect variance and the species are likely to be heterogeneous (e.g., McArdle and Andersen, 2004).

We describe the modeling approach for $\boldsymbol{n}_i \mid N_i,\ S_i$ by describing a series of nested models. The first and simplest model is the Mn model. The second is the Dirichlet-multinomial (DMn), which is overdispersed with respect to the Mn model. The final model is a modification of the DMn model that accommodates particular patterns of overdispersion.

7.1 *Multinomial Model*

The probabilities relating to each of the species rankings are, by definition of a RAD, a decreasing function of species rank, and they should be parameterized to be so. With p_{ij} being the probability of the jth species at the ith site a possibility is

$$p_{ij} = \frac{1}{K}\exp\left(-\beta_i \log j\right), \tag{4}$$

where $j = 1, \ldots, S_i$ indexes species rank, $\beta_i = \boldsymbol{x}_i^\top \boldsymbol{\tau}_n > 0, \boldsymbol{x}_i$ is the vector of covariates for sample $i, \boldsymbol{\tau}_n$ is the parameter vector, and $K = \sum_{k=1}^{S_i} \exp(-\beta_i \log k)$ is a standardizing term.

The functional form (4) was chosen empirically after observing that the plots of $\log p_{ij}$ versus log species rank for each site in the VoD data are surprisingly linear. We note that this function does have ecological motivation as it is closely related to the continuous form for the broken stick model (Whittaker, 1972; Pielou, 1975). It is not the only choice, and any decreasing function that sums to one could be used.

Quantile residuals are used for diagnostics of the mean model and are generated for each species rank, marginal to all the other species ranks. For the Mn model, the quantile residuals are generated from a binomial (e.g., Johnson, Kotz, and Balakrishnan, 1997, p. 32). Raw residuals, $r_{ij} = n_{ij} - \hat{n}_{ij}$, are used for diagnosing the variance model, as they are not standardized and their expected absolute values should represent the standard deviation of the model.

Covariates and their associated estimates for the Mn model are given in Web Table 1. To enable direct comparison we use the terms selected from the selection procedure from the modified DMn model (Section 7.3). Residuals for the Mn model are presented in the upper panels of Figure 3. If the Mn model's variance structure is adequate then the raw residuals (upper right panel of Figure 3) should scatter around the expected standard deviation (the solid line). This is not observed; the raw residuals are generally greater than the expected standard deviation and hence the data are overdispersed. The diagnostic plot for the mean model (upper left panel of Figure 3) suggests that there may be some further inadequacies but the proof is inconclusive due to the small number of large observations and the more obvious deficiencies in the variance model.

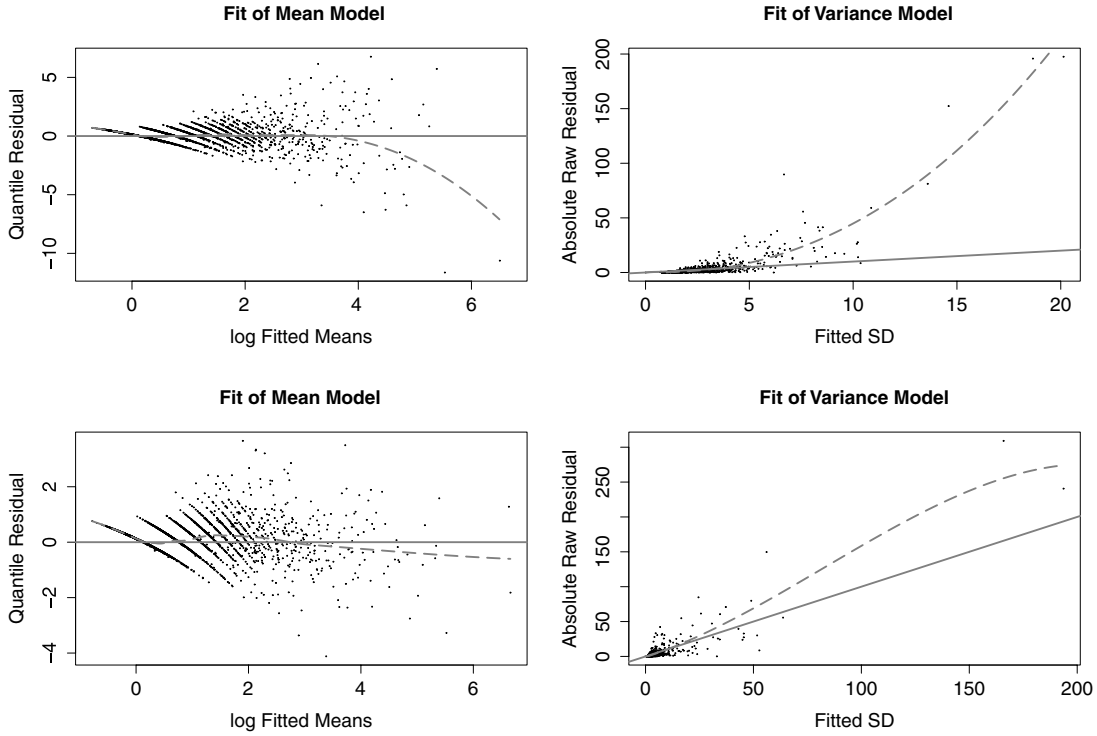

Figure 3. Residuals from models for $\boldsymbol{n}_{ij} \mid S_i, N_i$. Top panels: Mn model. Bottom panels: modified DMn model. Left panels: quantile residuals versus log fitted value. The solid grey line is $y = 0$ and the dashed grey line is a loess smooth. Right panels: raw residuals versus fitted standard deviation. The solid grey line is $y = x$ and the dashed line is a loess smooth. If the model represented the variation in the data well, then the raw residuals should be scattered around the $y = x$ line.

7.2 *Dirichlet-Multinomial Model*

A commonly used model that allows for overdispersion with respect to a Mn model is the DMn (Mosimann, 1962; Johnson et al., 1997). The S_i dimensional DMn distribution for the ranked abundances $\boldsymbol{n}_i$ is parameterized by the vector $(\theta_{i1}, \theta_{i2}, \ldots, \theta_{iS_i})$. The DMn distribution is overdispersed if $\theta = \sum_{j=1}^{S_i} \theta_{ij}$ is small and converges to the Mn when $\theta \to \infty$ (see Figure 1 of Stedinger, Shoemaker, and Tenga, 1985). Stedinger et al. (1985) suggested an alternative parameterization that is based on the probabilities of each species rank; the parameters $(\theta_{i1}, \theta_{i2}, \ldots, \theta_{iS_i})$ are replaced by the probabilities $(p_{i1}, p_{i2}, \ldots, p_{iS_i})$ and a single parameter θ. The two parameterizations are equivalent as $p_{ij} = \theta_{ij}/\theta$. We choose to use the parameterization based on the probabilities, as the functional form from equation (4) can easily be preserved.

The DMn model's first two moments are

$$\mathrm{E}(n_{ij}) = N_i p_{ij}, \ \mathrm{Var}(n_{ij}) = \frac{\theta + N_i}{\theta + 1} N_i p_{ij}(1 - p_{ij})$$

$$\text{and } \mathrm{Cov}(n_{ij}, n_{ij'}) = -\frac{\theta + N_i}{\theta + 1} N_i p_{ij} p_{ij'}.$$

A comparison of these moments with those from a Mn shows that the mean is identical and the (co-)variances are inflated by a constant overdispersion factor.

The estimated mean parameters for the DMn were similar to those from the Mn model. The overdispersion parameter was estimated to be $\hat{\theta} = 3408.89$ (± 600.63). This indicates a very modest increase in variance with respect to the Mn. The DMn's constant overdispersion with respect to the Mn is inconsistent with the increasing overdispersion observed in the data (see upper right panel of Figure 3).

7.3 *A Modified Dirichlet-Multinomial (M-DMn) Model*

The empirical heterogeneity pattern in the VoD data is quite distinctive (Figure 3, upper right panel). In comparison to the Mn model, the abundant species at a site, with high modeled abundance and standard deviation, are more overdispersed than the rarer species. Hence, with respect to the Mn model, the overdispersion is decreasing with species rank and this observation can be exploited by modeling the overdispersion as a function of species rank. The proposed model is based on the DMn model but the resulting model is not DMn.

There are two important attributes of the DMn that are pivotal to the derivation (Johnson et al., 1997).

(i) The distribution of the jth ranked species marginal to all the other species ranks is a beta-binomial (BBn) with parameters $(N_i, \theta_{ij}, \theta - \theta_{ij})$, and
(ii) The distribution of a subset of the ranked species conditional on the other ranked species is a DMn.

These results allow the factorization of the DMn distribution via

$$\Pr(\boldsymbol{n}_i) = \Pr(n_{i1}) \Pr\left(n_{i2} \mid \boldsymbol{n}_i^{(1)}\right) \times \Pr\left(n_{i3} \mid \boldsymbol{n}_i^{(2)}\right) \ldots \Pr\left(n_{iS_i} \mid \boldsymbol{n}_i^{(S_i - 1)}\right), \quad (5)$$

where the notation $\boldsymbol{n}_i^{(j)}$ is a j-dimensional vector consisting of the j highest abundances. Explicit notation for conditioning on S_i and N_i is dropped for clarity and is not used for the remainder of this section.

All the component distributions in equation (5) are BBn distributions. This stems from a sequential factorization. First, the full DMn is factorized into a BBn and a DMn of dimension $(S_i - 1)$. Next, the remaining DMn, of lower dimension, is factorized further into another BBn and another DMn. The process is repeated until the final element is reached. The BBn distributions are

$$n_{ij} \,|\, \boldsymbol{n}_i^{(j-1)} \sim \text{BBn}\left(N_i - \sum_{k=1}^{j-1} n_{ik}, \theta_{ij}, \theta - \sum_{k=1}^{j} \theta_{ik}\right), \quad (6)$$

where $\Pr(n_{i1}) = \Pr(n_{i1} \,|\, \boldsymbol{n}_i^{(0)})$. The expected value and variance of each of these BBn distributions are

$$\text{E}\big(n_{ij} \,|\, \boldsymbol{n}_i^{(j-1)}\big) = \frac{\theta}{\theta - \sum_{k=1}^{j} \theta_{ik}} \left(N_i - \sum_{k=1}^{j-1} n_{ik}\right) p_{ij} \quad \text{and,} \quad (7)$$

$$\text{Var}\big(n_{ij} \,|\, \boldsymbol{n}_i^{(j-1)}\big) = \left\{1 + \frac{\left(N_i - \sum_{k=1}^{j-1} n_{ik}\right) - 1}{1 + \theta_{ij} + \theta - \sum_{k=1}^{j} \theta_{ik}}\right\} \times \left(N_i - \sum_{k=1}^{j-1} n_{ik}\right) p_{ij}(1 - p_{ij}). \quad (8)$$

Our modification of the DMn alters each of the BBn distributions (6). For the jth highest abundance the modified distribution is

$$n_{ij} \,|\, \boldsymbol{n}_i^{(j-1)} \sim \text{BBn}\left\{N_i - \sum_{k=1}^{j-1} n_{ik}, \phi_j \theta_{ij}, \phi_j \left(\theta - \sum_{k=1}^{j} \theta_{ik}\right)\right\},$$

which has an additional parameter ϕ_j to (6). The expected value for the modified BBn is given by equation (7) and the variance is

$$\text{Var}\big(n_{ij} \,|\, \boldsymbol{n}_i^{(j-1)}\big) = \left\{1 + \frac{\left(N_i - \sum_{k=1}^{j-1} n_{ik}\right) - 1}{1 + \phi_j \left(\theta_{ij} + \theta - \sum_{k=1}^{j} \theta_{ik}\right)}\right\} \times \left(N_i - \sum_{k=1}^{j-1} n_{ik}\right) p_{ij}(1 - p_{ij}).$$

The M-DMn distribution is overdispersed with respect to the DMn for the jth ranked species if $\phi_j < 1$, it is equivalent if $\phi_j = 1$, and it is underdispersed if $\phi_j > 1$. We restrict $\phi_j > 0$ and so the M-DMn will always be overdispersed with respect to the corresponding Mn. It is possible to specify ϕ_j as a function of species rank, such as $\phi_j = j^\nu$ where ν is positive if the overdispersion is decreasing over species rank.

Estimation of the parameters $\boldsymbol{\tau}_n, \theta$ and ν is performed by maximizing the likelihood using numerical methods. Details are given in Web Appendix A.

Model selection was performed using the forward selection method described previously (Sections 5 and 6) with scaled richness, S_i/d_i, included to the set of covariates. The models are diagnosed using quantile residuals (Dunn and Smyth, 1996) and raw residuals. For the M-DMn model it is not obvious on which distribution the quantile residuals should be based. We use the seemingly natural choice of a BBn distribution with parameters $N_i, \phi_j\theta_{ij}$, and $\phi_j(\theta - \theta_{ij})$ for the jth ranked species from the ith site, but we can offer no formal justification.

7.4 *Model Interpretation and Prediction*

The functional form for the probability of observing each ranked species (4) suggests a natural measure of evenness; the derivative of the unstandardized probability function evaluated at the first ranked species. For the ith sample it is

$$\eta_i = -\beta_i \exp\{-\beta_i \log(1)\} = -\beta_i. \quad (9)$$

Small negative values of η indicate that the assemblage is even. Large negative values of η suggest assemblages are uneven.

Formal inference on $\boldsymbol{n}_i \,|\, N_i$, S_i can be performed using standard maximum likelihood methods. Dependence of evenness on the environmental explanatory variables marginal to, and conditional on, abundance and richness will aid interpretation and utility. A bootstrap routine, similar to that proposed for the marginal richness (Section 6) is used to obtain the marginal evenness. The marginalization is now over two random variables and bootstrap samples are needed from both in a sequential manner. Details directly follow those described in Section 6.

7.5 $\boldsymbol{n}_i$ *in the VoD Data*

The M-DMn model appears to fit the data substantially better than either the Mn or the DMn models. The AIC values for each of the models were 10,255.93, 10,221.46, and 9345.81 for the Mn, DMn, and the M-DMn models respectively. The AIC values were calculated using the model identified by the model-selection process for the M-DMn model.

The residual plots for the M-DMn model (lower panels of Figure 3) indicate that this model represents the data better than the Mn or DMn models (upper panels of Figure 3). The model selection process highlighted that there is a substantial amount of variation explained by the environmental variables. These variables are given in Table 1, along with the AIC at the time of inclusion, and parameter estimates. The parameter estimates from the M-DMn model are substantially different to those obtained from the Mn model (Web Table 1). This would suggest that accounting for the correct pattern of overdispersion in the observed RADs is extremely important for inference and prediction.

The estimated overdispersion parameters for the M-DMn model (Table 1) imply that there is a substantial amount of overdispersion with respect to the Mn model and the overdispersion decreases substantially over the ranks. To illustrate the relationship of overdispersion with rank we assume BBn marginal distributions for n_{ij} (Section 7.3). Then

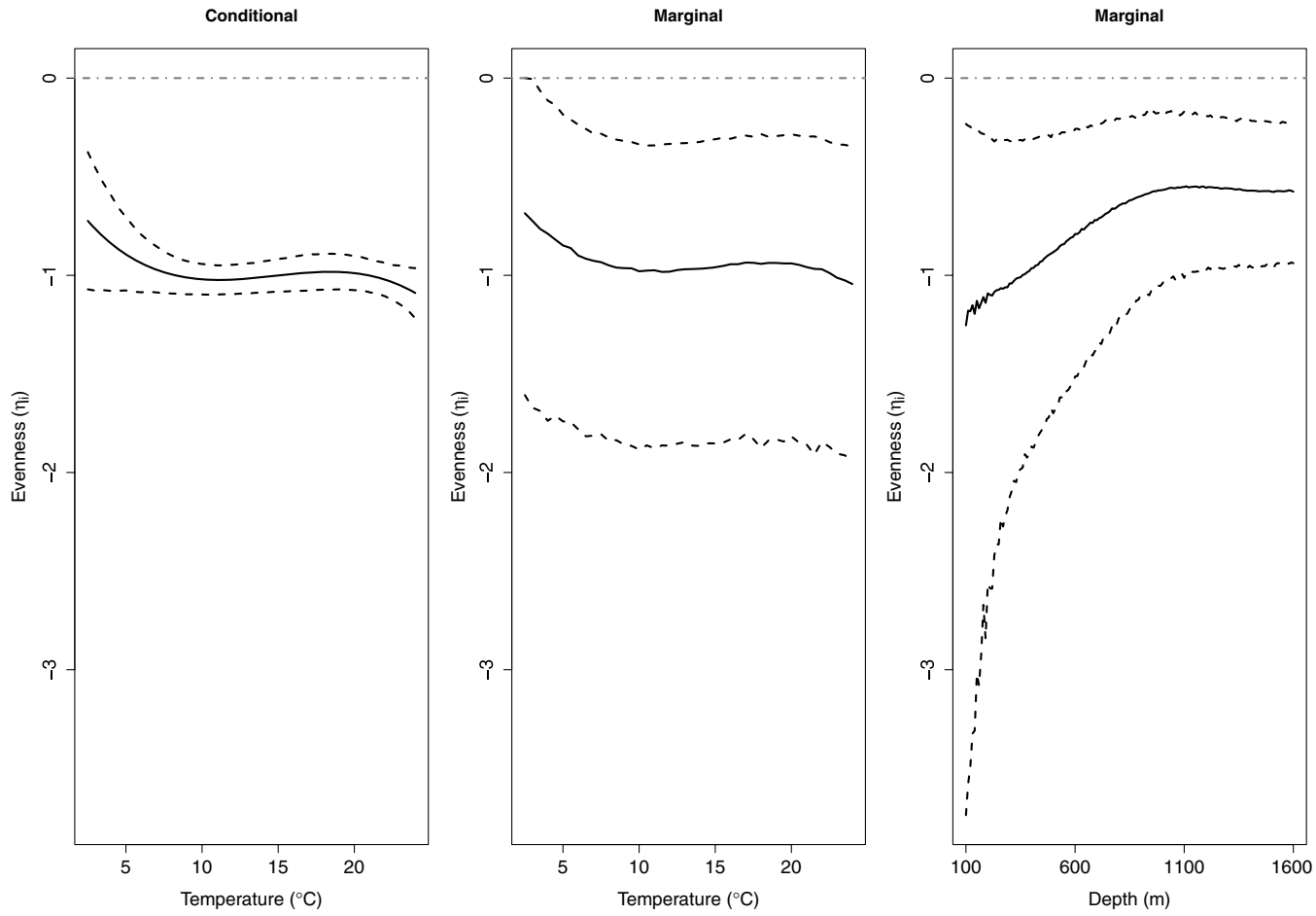

Figure 4. Relationships of evenness (η) with temperature and depth. Left panel conditions on the mean scaled total abundance and mean scaled species richness while the center and right panels are marginal to these variables. All other covariates are held constant at the mean of the observations. In all panels the solid line is the mean, dashed lines are the 95% confidence intervals for that mean, and the horizontal grey line is $y = 0$.

the overdispersion factor with respect to a Mn model is the decreasing function of rank

$$\hat{\xi}_{ij} = 1 + \frac{N_i - 1}{j^{\hat{\nu}}\hat{\theta} + 1}.$$

The marginal relationship of evenness, η_i, with temperature and depth is given in Figure 4. Evenness decreases with temperature and increases with depth. Depth is not included as an explanatory variable in the model for $\boldsymbol{n}_i \,|\, N_i, S_i$, but it is important for explaining variation in N_i and $S_i \,|\, N_i$. The evenness–depth relationship is a product of these lower level relationships.

Assemblages in shallower and/or warmer environments are more dominated by a small number of species than deeper and/or colder environments (Figure 4). We suspect that this is related to the supply of particulate organic matter (POM). Shallower environments have more food input, in the form of POM, allowing for greater abundances (Smith et al., 2008, also observed for N_i in these data). Where food is less limiting, individual species may attain abundances not possible in more restrictive environments, resulting in greater evenness as depth increases.

These data highlight that the observed appearance of a RAD is dependent on a number of different, related processes. The marginal relationships in the center and right panels of Figure 4 have wide confidence intervals, reflecting the high variability in all three levels of the factorization.

7.6 *Modeling Classic Evenness Measures*

To compare the M-DMn model's results to existing methods we model well-established measures of evenness, namely, a transformation of Simpson's D and Pielou's J (e.g., Magurran, 2004). The metrics were calculated for the ith sample as

$$D_i = -\log\left(\sum_{k=1}^{S_i} q_{ik}^2\right) \quad \text{and} \quad J_i = \frac{\sum_{k=1}^{S_i} q_{ik}\log\left(q_{ik}\right)}{\log\left(S_i\right)}, \tag{10}$$

where $q_{ik} = n_{ik}/N_i$ and is an empirical probability. We use the transformed version of Simpson's metric to address variance heterogeneity issues (Pielou, 1975). Multiple regression

was used to relate the two sets of metrics to the environmental covariates. A suitable model was found using forward selection, with details matching those used previously (see Section 5). The covariates considered for inclusion in the model were identical to those considered for modeling $\boldsymbol{n}_i \mid S_i$, N_i and the results should be comparable. This approach is not novel as biodiversity metrics have been modeled on covariates previously (e.g., Josefson and Hansen, 2004; Sousa et al., 2006; Gutiérrez-Estrada et al., 2008).

The chosen model for Simpson's index depended only on scaled total abundance, scaled species richness, and the length of tow (estimates in Web Table 2). It contained no term relating evenness to the physical environment. The selected model for Pielou's J contained terms for depth and variation in oxygen, in addition to covariates in the model for Simpson's index (estimates in Web Table 2). Neither of these physical covariates were included in the RAD model for evenness but the marginal relationship of η with depth showed an increase (Figure 4) due to the dependence of abundance and richness. The relationship with depth and J loosely agrees between the right panel of Figure 4. The RAD model suggests that the dependence on depth for J may be driven by the dependence of richness and/or abundance on depth and not an inherent relationship of evenness and depth. This confounding with evenness and richness in Pielou's J has been shown in previous studies (Smith and Wilson, 1996).

8. Discussion

Understanding how biodiversity varies with respect to environmental gradients is of central importance to scientists and resource managers alike. These models enable predictions of ecological patterns and enhance sustainable biodiversity conservation. Species abundance data are commonly recorded from many surveys and can be used to describe important biodiversity attributes. In this article, we present a statistical method for the analysis of three attributes of biodiversity. The method is based on the RAD representation of the data. McGill et al. (2007) identified that the statistical properties of RADs have not been fully explored and this article addresses this. Summaries of the models provide focused information on three universally important attributes of biodiversity and shows how these change with environmental gradients.

The cause of the overdispersion with respect to the Mn model for the ranked species abundances is unclear. If the unranked species distribution was truly Mn then intuitively one would expect that the resulting RADs should be underdispersed with respect to a Mn on the ranked data as the ranking process groups similar observations. Ranked species classes will have greater underdispersion for species whose Mn probabilities are similar. In species enumeration data these species are typically less common and so, it is expected to see decreasing variance with species rank: a pattern observed in the VoD RAD data. However, the observed data is overdispersed with respect to a Mn model and not underdispersed. This implies that the single species distributions, prior to ranking, must have more dispersion than the Mn prescribes. Also, we expect that the level of overdispersion will vary between species and this will affect the pattern of dispersion in the corresponding RAD. It is difficult to conclude anything about the effect of this heterogeneity by inspecting the RADs alone.

We believe that the proposed modeling framework provides a substantial advancement in the analysis of biodiversity with species composition data. It certainly provides a richer suite of modeled biodiversity relationships than a disparate set of univariate measures with unclear links between them. We realize that there may be other model assumptions for any of the three component models that are ecologically more realistic. For example, the model for the mean probabilities in the modified DMn model could be taken to be one of the large number of theoretical models from the ecological literature (e.g., Pielou, 1975; Wilson, 1991; Magurran, 2004). The framework presented here provides an objective likelihood-based method to compare these models using real data.

9. Supplementary Materials

Web Appendices and Tables referenced in Section 7 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

We thank A. Williams, F. Althaus, F. McEnnulty, R. Kloser, G. Keith, M. Fuller, J. Dunn, G. Poore, and the Captain and crew of the RV Southern Surveyor. The Australian Government's Department of Environment, Water, Heritage and the Arts contributed to the VoD survey. N. Ellis, R. Darnell, the associate editor, and the referees provided constructive criticisms of the article that led to substantial improvements.

References

Anderson, M. J., Ellingsen, K. E., and McArdle, B. H. (2006). Multivariate dispersion as a measure of beta diversity. *Ecology Letters* **9,** 683–693.

Beck, J. and Vun Khen, C. (2006). Are rank-abundance distributions a useful tool of assemblage discrimination in tropical moths? *Acta Zoologica Sinica* **52,** 1148–1154.

Bulmer, M. G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics* **30,** 101–110.

Callaway, R., Alsvåg, J., de Boois, I., Cotter, J., Ford, A., Hinz, H., Jennings, S., Kroncke, I., Lancaster, J., Piet, G., Prince, P., and Ehrich, S. (2002). Diversity and community structure of epibenthic invertebrates and fish in the North Sea. *ICES Journal of Marine Science* **59,** 1199–1214.

Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Applications*. New York: Cambridge University Press.

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

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* **12,** 42–58.

Gutiérrez-Estrada, J. C., Vasconcelos, R., and Costa, M. J. (2008). estimating fish community diversity from environmental features in the Tagus estuary (Portugal): Multiple linear regression and artificial neural network approaches. *Journal of Applied Ichthyology* **24,** 150–162.

Jennings, S., Lancaster, J., Woolmer, A., and Cotter, J. (1999). Distribution, diversity and abundance of epibenthic fauna in the North Sea. *Journal of the Marine Biological Association of the United Kingdom* **79,** 385–399.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. New York: John Wiley & Sons.

Josefson, A. B. and Hansen, J. L. S. (2004). Species richness of benthic macrofauna in Danish estuaries and coastal areas. *Global Ecology and Biogeography* **13,** 273–288.

Koslow, J. A., Williams, A., and Paxton, J. R. (1997). How many demersal fish species in the deep sea? A test of a method to extrapolate from local to global diversity. *Biodiversity and Conservation* **6,** 1523–1532.

Labropoulou, M. and Papaconstantinou, C. (2004). Community structure and diversity of demersal fish assemblages: The role of fishery. *Scientia Marina* **68,** 215–226.

Mac Nally, R. (2007). Use of the abundance spectrum and relative-abundance distribution to analyze assemblage change in massively altered landscapes. *The American Naturalist* **170,** 319–330.

Magurran, A. E. (2004). *Measuring Biological Diversity*. Malden, Massachusetts: Blackwell Publishing

Maldonado, M. and Young, C. M. (1996). Bathymetric patterns of sponge distribution on the Bahamian slope. *Deep Sea Research I* **43,** 897–915.

McArdle, B. H. and Andersen, M. J. (2004). Variance heterogeneity, transformations, and models of species abundance: A cautionary tale. *Canadian Journal of Fisheries and Aquatic Science* **61,** 1294–1302.

McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Benecha, H. K., Dornelas, M., Enquist, B. J., Green, J. L., He, F., Hurlbert, A. H., Magurran, A. E., Marquet, P. A., Maruer, B. A., Ostling, A., Soykan, C. U., Ugland, K. I., and White, E. P. (2007). Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters* **10,** 995–1015.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution and correlations among proportions. *Biometrika* **49,** 65–82.

Pielou, E. C. (1975). *Ecological Diversity*. New York: John Wiley & Sons.

Ridgway, K. R., Dunn, J. R., and Wilkin, J. L. (2002). Ocean interpolation by four-dimensional least squares—application to the waters around Australia. *Journal of Atmospheric and Oceanic Technology* **19,** 1357–1375.

Smith, B. and Wilson, J. B. (1996). A consumer's guide to evenness measures. *Oikos* **76,** 70–82.

Smith, C. R., De Leo, F. C., Bernardino, A. F., Sweetman, A. K., and Martinez Arbizu, P. (2008). Abyssal food limitation, ecosystem structure and climate change. *Trends in Ecology and Evolution* **23,** 518–528.

Sousa, P., Azevedo, M., and Gomes, M. C. (2006). Species-richness patterns in space, depth, and time (1989-1999) of the Portuguese fauna sampled by bottom trawl. *Aquatic Living Resources* **19,** 93–103.

Stedinger, J. R., Shoemaker, C. A., and Tenga, R. F. (1985). A stochastic model of insect phenology for a population with spatially variable development rates. *Biometrics* **41,** 691–701.

Whittaker, R. H. (1965). Dominance and diversity in land plant communities. *Science* **147,** 250–260.

Whittaker, R. H. (1972). Evolution and measurement of species diversity. *Taxon* **21,** 213–251.

Wilson, J. B. (1991). Methods for fitting dominance/diversity curves. *Journal of Vegetation Science* **2,** 35–46.

Received August 2008. *Revised January* 2009.
Accepted January 2009.