Prevalence of multimodal species abundance distributions is linked to spatial and taxonomic breadth

Laura Henriques Antão¹,²,*, Sean R. Connolly³, Anne E. Magurran¹, Amadeu Soares² and Maria Dornelas¹

¹Centre for Biological Diversity, University of St Andrews, Sir Harold Mitchell Building, St Andrews KY16 9TF, Scotland, UK,
²Department of Biology and CESAM, Universidade de Aveiro, Campus Universitário de Santiago, Aveiro, 3810-193, Portugal, ³College of Science and Engineering, and ARC Centre of Excellence for Coral Reef Studies, James Cook University, Townsville, QLD 4810, Australia

ABSTRACT

Aim Species abundance distributions (SADs) are a synthetic measure of biodiversity and community structure. Although typically described by unimodal logseries or lognormal distributions, empirical SADs can also exhibit multiple modes. However, we do not know how prevalent multimodality is, nor do we have an understanding of the factors leading to this pattern. Here we quantify the prevalence of multimodality in SADs across a wide range of taxa, habitats and spatial extents.

Location Global.

Methods We used the second-order Akaike information criterion for small sample sizes (AICc) and likelihood ratio tests (LRTs) to test whether models with more than one mode accurately describe the empirical abundance frequency distributions of the underlying communities. We analysed 117 empirical datasets from intensely sampled communities, including taxa ranging from birds, plants, fish and invertebrates, from terrestrial, marine and freshwater habitats.

Results We find evidence for multimodality in 14.5% of the SADs when using AICc and LRT. This is a conservative estimate, as AICc alone estimates a prevalence of multimodality of 22%. We additionally show that the pattern is more common in data encompassing broader spatial scales and greater taxonomic breadth, suggesting that multimodality increases with ecological heterogeneity.

Main conclusions We suggest that higher levels of ecological heterogeneity, underpinned by larger spatial extent and higher taxonomic breadth, can yield multimodal SADs. Our analysis shows that multimodality occurs with a prevalence that warrants its systematic consideration when assessing SAD shape and emphasizes the need for macroecological theories to include multimodality in the range of SADs they predict.

Keywords Community structure, diversity patterns, lognormal, logseries, Poisson lognormal mixture, spatial scale, taxonomic breadth.

INTRODUCTION

Species abundance distributions (SADs) depict the relative abundance of the species present in a community and describe one of the most fundamental patterns of species diversity – most communities contain many rare and only a few common species (McGill et al., 2007). Empirical datasets consistently produce SADs that are quasi-hyperbolic on an arithmetic scale – the ubiquitous ‘hollow curve’. However, on
a logarithmic scale of abundance, SADs exhibit more variability, with species abundance distributions alternately exhibiting no internal mode (most species occur at the lowest abundance class, i.e. as singletons), one internal mode or multiple internal modes. Despite seven decades of study and dozens of different proposed models (McGill et al., 2007), there is still no consensus about what drives variation in the shape of SADs, nor how SAD shape might be connected to factors structuring ecological communities (Fisher et al., 1943; Preston, 1948; Magurran & Henderson, 2003; McGill, 2003b; Green & Plotkin, 2007; Dornelas et al., 2009). The extent to which current theories of biodiversity are able to accommodate and explain such variation is a critical criterion for their evaluation and application (McGill et al., 2007).

The two distributions recurrently proposed to describe SADs are the logseries (Fisher et al., 1943) and the lognormal (Preston, 1948) distributions (Fig. 1). While many intensely sampled communities seem to follow a lognormal distribution (Magurran, 2004), it has become increasingly clear that empirical SADs often deviate from a lognormal by having more than one internal mode (Ugland & Gray, 1982; Gray et al., 2005; Dornelas & Connolly, 2008). Multimodality is seldom reported, and its implications little explored (McGill et al., 2007), with a few notable, but dispersed, exceptions. Ugland & Gray (1982) proposed three lognormal distributions, corresponding to rare, intermediate abundant and common species, to describe non-equilibrium marine benthic communities. Magurran & Henderson (2003) ‘deconstructed’ an estuarine fish community into two groups – ‘core’ and ‘occasional’ – based on species persistence and habitat preferences, where the ‘core’ group was better fitted by a lognormal while the ‘occasional’ group of rare species followed a logseries distribution. Gray et al. (2005) showed that a mixture of two lognormal distributions provided a good fit to a marine benthos and a tropical tree data, again separating the species into ‘abundant’ and ‘rare’.

In the first statistical analysis comparing the fit of distributions with varying numbers of modes, Dornelas & Connolly (2008) showed that the SAD of an intensely sampled coral community was multimodal. However, the different modes could not be explained by mixture of species associated with different habitats, and were only partially explained by different spatial aggregations. Recently, Matthews et al. (2014), using the same methodology for an arthropod community, showed that multimodal distributions performed better for many of the samples analysed, and that grouping ecologically different species leads to multimodality, with the rarest species mode containing a higher proportion of satellite and introduced species and species better adapted to other habitats. However, the effect of dispersal ability was unclear, and a body size niche axis was unrelated to the multimodal patterns. What these studies have in common is that they

**Figure 1** Examples of random sampled communities for a logseries, a single lognormal Poisson (1PLN), and mixtures of two and three lognormal Poisson distributions (2PLN and 3PLN, respectively). For the logseries distribution, the single parameter is Fisher’s alpha. For the PLN models, $\mu$ and $\sigma$ are the mean and standard deviation of log-abundance for one of the underlying lognormal community abundance distributions (one pair of parameters for each mode), and $p_n$ is the probability that a species comes from distribution $n$. The parameters used to generate the random sampled abundance data for each model are shown, and the species pool size was $S = 100$ (the code to generate the 2PLN example can be found in Appendix S2).
indicate that multimodality is linked to ecological heterogeneity, broadly defined as groups of species with different ecological or functional characteristics. This suggests that multimodality should have higher prevalence among communities with higher ecological heterogeneity. Our concept of ecological heterogeneity is intended to encompass the spatial, environmental, taxonomic and functional aspects of ecological systems, rather than simply the number of species or functional groups.

The prevalence of multimodality in empirical SADs is as yet unknown. In a recent theoretical study, Barabás et al. (2013) reported that stochastic versions of both resource partitioning and neutral models can produce multimodal SADs with a 50% prevalence. The authors argue that, in nature, individual realizations are likely to differ from the mean predicted pattern due to stochastic processes, while disputing that the emergent neutrality model proposed by Vergnon et al. (2012) is the only one able to produce multimodal SADs. Thus, assessing the prevalence of multimodality in empirical datasets is warranted to establish the generality of the pattern, as well as help elucidate how it can be related to different ecological explanations.

Here, we undertake a comprehensive global assessment of the prevalence of multimodality for a wide range of communities. This is, to our knowledge, the first assessment of the prevalence of multiple modes in SADs. We improve the method for detecting multimodality and show with high confidence that multimodality occurs in 17 out of 117 assemblages. Additionally, we test the hypothesis that more heterogeneous communities are more likely to exhibit multimodality. We show that multimodality has higher prevalence for large-scale or taxonomically heterogeneous communities. Broader spatial extent and higher taxonomic breadth (as measured by family diversity) underpin higher ecological heterogeneity, and hence we suggest these as potential explanations for multimodality in SADs.

**MATERIALS AND METHODS**

**Model selection**

To test whether models with more than one mode accurately reflect the abundance distributions of the underlying communities we used maximum likelihood methods to explicitly compare the fit of mixtures of one, two and three Poisson lognormal distributions (1PLN, 2PLN and 3PLN, respectively) (Pielou, 1969; Bulmer, 1974); a logseries distribution was also included (Fig. 1). All the calculations were performed in the software R (R Core Team, 2013). Functions to fit the PLN mixtures and to calculate maximum likelihood estimates (MLE) were adapted from Dornelas & Connolly (2008), but using the dpolilog() function from the poilog package (Grotan & Engen, 2008); the log-likelihood functions are otherwise similar and best fit parameters were found by minimizing the negative log-likelihood (functions are available in Appendix S1 in the Supporting Information). Parameter estimation was performed using the R optimization routine nlmnb, and parameter searches were initialized from multiple starting points due to the possibility of several local maxima for more complex distributions (Dornelas & Connolly, 2008; Connolly & Dornelas, 2011).

Model comparison was performed under a multi-model information-theoretic framework (Burnham & Anderson, 2002), using the second-order Akaike information criterion for small sample sizes (AICc; Burnham & Anderson, 2002) and the Bayesian information criterion (BIC; Schwarz 1978). AICc was used throughout as it converges to AIC when the sample size is large (Burnham & Anderson, 2002, 2004). AIC and BIC are model selection tools that provide quantitative relative support for alternative hypotheses, while finding a compromise between goodness of fit and model complexity. AIC tends to overestimate the number of distributions in mixture models, while BIC tends to underestimate them (McLachlan & Peel, 2000; Henson et al., 2007). Hence, we evaluated the performance of these two model selection criteria with a simulation study.

We evaluated model performance in slightly different ways in the empirical and simulation studies. For the analysis of the empirical data, relative support for the models was calculated as AAIC, which is the difference between the AICc of each model and the lowest AICc in the model set. Differences larger than two indicate substantial evidence against the model with the higher AICc (or BIC) (Burnham & Anderson, 2002). However, for the simulation study, the ‘true model’ (the model used to generate the simulated data) is known. Therefore, we calculated AIC differences relative to this true model, a quantity we term ‘AICdiff’. Specifically, AICdiff is the AICc of the true model, minus the smallest AICc of the remaining models. This quantity is negative whenever the true model is the best fitting model (the one with the lowest AIC score). Conversely, if one or more of the alternative models actually fits better than the true model does, then AICdiff will be positive (note that AICdiff = 0 does not indicate the best fitting model). An analogous quantity was calculated for BIC.

**Simulation study**

Because the PLN-mixture method has only been applied to specific datasets (Dornelas & Connolly, 2008; Vergnon et al., 2012; Matthews et al., 2014), we conducted a simulation study to assess how it performed under a broad range of parameter combinations. We were specifically interested in determining which conditions lead to selection of a model with the wrong number of internal modes. We defined a false positive as simulated samples where a multimodal model was selected with high confidence when the true model generating the sample was not multimodal; and a false negative as simulated samples where the true model was multimodal but for which a ‘non-multimodal’ model was selected. A range of species richness and parameter values for the four alternative abundance distributions models was used to generate simulated count data. The spectrum of parameters used was
designed to cover a realistic range for species abundance data (Connolly & Thibaut, 2012), and to provide a quantitative picture of whether and when the method fails to select the true number of underlying modes. A total of 162 parameter combinations were examined; for each parameter combination, 100 simulated SAD samples were generated and the alternative log-likelihood functions were fitted (see Appendix S2 for more details and code).

**Parametric bootstrap**

Following the results of the simulation study, we identified some 1PLN parameter combinations where AICc strongly selected a more complex model than the one generating the data, with a frequency of up to c. 25% of the simulated samples (Fig. S2.2 in Appendix S2). We additionally calculated likelihood ratio tests (LRTs) to minimize the chance of a multimodal model being selected due to overfitting. LRTs assess if the improvement in goodness of fit of a more complex model is greater than would be expected by chance if the simpler model were true. LRTs are only applicable to nested models, so the logseries was not included in this analysis. Because the null distribution of LRTs is known to occasionally deviate from a $\chi^2$ distribution (McLachlan, 1987; McLachlan & Peel, 2000), we generated null LRT frequency distributions from 1PLN-simulated communities. This allows us to calculate the equivalent of a P-value for the null hypothesis that the sampled data are consistent with a 1PLN distribution, thus providing an alternative assessment of whether a multimodal model provided the best fit for that parameter combination. For the simulation study, we illustrate this by comparing LRT distributions for two parameter combinations, one from the parameter space where AICc successfully selected 1PLN, and the other from the space where AICc has a higher probability of selecting a more complex model.

For the empirical data, we conducted a parametric bootstrap likelihood ratio test (PBLRT; see Knape & de Valpine, 2012, for an example) for all the SADs selected as multimodal by AICc. The parametric bootstrap procedure consisted of randomly generating species abundance values from a 1PLN density function parameterized using the model’s maximum likelihood estimates for that empirical dataset (Connolly et al., 2009). As these analyses are very computationally intensive (Dornelas & Connolly, 2008; Connolly & Dornelas, 2011), 100 parametric bootstrap samples were generated for each dataset, using $\mu$ and $\sigma$ (the estimated mean and standard deviation of log-abundances) and sample size as the observed number of species, and the log-likelihood functions were fitted (details and code in Appendix S2). This procedure allowed us to compare the empirical likelihood ratio, calculated from the empirical SAD fitting, with the frequency distribution expected under the null hypothesis that the data are actually a single PLN.

**Empirical data**

One hundred and seventeen datasets from intensely sampled communities were collected from three online repositories: OBIS (Ocean Biogeographic Information System; http://www.iobis.org/), Ecological Data Wiki (http://ecologicaldata.org/) and GBIF (Global Biodiversity Information Facility; http://www.gbif.org/) (Fig. 2). These data are freely available and a complete list of the data sources can be found in Appendix I, and additional information in Appendix S3. For each dataset a simplified vector of abundances was obtained, corresponding to only 1 year of sampling (the most recent year with at least 10,000 individuals where multiple years were sampled).
Multimodal SADs linked to spatial and taxonomic breadth

Global Ecology and Biogeography, © 2016 John Wiley & Sons Ltd
This was intended to prevent interannual variability from inducing multimodality sensu Magurran & Henderson (2003), as we were interested in assessing the prevalence of multimodal SADs independent of any temporal effect of fluctuations in species abundances among years.

Datasets were classified according to spatial extent and taxonomic breadth. These two variables were intended to represent different axes of ecological heterogeneity. Regarding spatial extent, as explicit estimates of extent were not available for all datasets, datasets were classified as Local when data originated from plots or sampling stations within less than 1° latitude/longitude, as Regional when data comprised larger areas (e.g. countrywide or larger biome patches) and as Continental when data spanned broader areas such as the whole eastern North American coast or Antarctica. Regarding taxonomic breadth, we used the number of families to quantify this variable. We analysed whether the prevalence of multimodality was influenced by spatial extent and taxonomic breadth (and their interaction) using two models: first, we used a binomial generalised linear model (GLM), aggregating 1PLN and logseries as ‘non-multimodal’, using the R function glm() with the logit link function. Additionally, we used a multinomial Bayesian generalised linear model to assess the prevalence of multimodality, 1PLN and logseries separately. We performed the Markov chain Monte Carlo (MCMC) estimation using the R package MCMCglmm (Hadfield, 2010). We fitted a model with a random intercept to obtain improved parameter estimates for each level of the fixed effects (see the MCMCglmm vignette; Hadfield, 2010; Gelman & Hill, 2007), running 5,000,000 iterations with a burn-in of 100,000 and a thinning interval of 25.

RESULTS

Simulation study

Overall, the PLN-mixture method was robust to large variation in the parameters used to perform the simulations. The false positive frequency was very low, particularly for BIC where in only 1% of the cases was a multimodal model selected with high confidence as the best fit model when the true model was not multimodal; for AICc, it was 6%. Species richness ($S$) had strong and disparate effects on the 1PLN simulations. For AICc, the percentage of false positives increased with $S$, while for BIC the percentage of failures decreased. When inspecting the 1PLN simulation results in more detail, some particular parameter combinations led AICc to consistently and strongly favour more complex models than the one generating the data (Fig. S2.2 in Appendix S2).

The overall false negative frequency, i.e. simulations where the model generating the sampled communities was multimodal but for which a ‘non-multimodal’ model was selected as the best fit, was 25% for AICc and 39% for BIC. For the 2PLN and 3PLN simulations, the true model was selected when the modes were clearly separated, for smaller $\sigma$ values and for higher species richness. BIC started to select a simpler model as the distance between the modes decreased ‘earlier’ than AICc, which was still able to select the true model for closer modes. Overall, the position of the modes, species richness and particularly $\sigma$ values showed strong effects in the best-fit model selection for both AICc and BIC, sometimes with different directions (see Appendix S2 for more detailed results and figures for each set of simulations).

When LRTs were used in addition to AICc, the chance of selecting a more complex model decreased compared with when using AICc alone (Fig. S2.5 in Appendix S2). For the parameter space where AICc very rarely selected a multimodal model, the LRT distribution overlapped with the AICc selection pattern (Fig. S2.5a,b). When AICc had a higher false positive frequency, using the LRT reduced the chance of erroneously selecting a multimodal model. Furthermore, the parametric bootstrap P-value is more conservative than the critical value from a $\chi^2$ distribution for the latter case (Fig. S2.5c,d). Hence, both AICc and PBLRT were used to analyse the empirical SADs, as the high false negative frequency for BIC suggests that it might not effectively detect multimodality.

Empirical data

Of the 117 SADs analysed, AICc selected a multimodal model for 47 (26 with high confidence). For many SADs, the estimated 1PLN parameters fell within the parameter space for which AICc often selects a multimodal model with high confidence when the true distribution is unimodal (specifically with an estimated standard deviation of log abundance, $\sigma$, of about 2). On the other hand, all the SADs selected as logseries had also estimated $\sigma \geq 2$ for the 1PLN model. This suggests that the method is not overfitting generally, but can occasionally select a more complex model. On visual inspection, none of the fitted curves seemed to be odd-looking or out of phase with the empirical SAD (Fig. 3 and Fig. S3.1 in Appendix S3), although it is possible that SADs that appear unimodal are better fitted by multimodal models, and vice versa (Matthews et al., 2014).

For the SADs selected as multimodal, PBLRT supported AICc model selection for 17 SADs (empirical likelihood ratio values were higher than the bootstrap P-value from the PBLRT distribution; Fig. 3, Table S2.2). For the cases where...
the PBLRT results did not support multimodality, the second best model was assumed to be the best model (either logseries or 1PLN). Overall, 17 SADs are multimodal with high confidence, 1PLN was the best model for 54 SADs and for 46 it was logseries (Table S3.1). None of the datasets selected as logseries had continental spatial scale.

Both spatial extent and taxonomic breadth have a positive effect on the prevalence of multimodality (Table 1). For the binomial GLM, SADs with local spatial extent were significantly less likely to be multimodal ($P = 0.0073$) than continental and regional scales, and there is a positive effect of the interaction between number of families and the local scale ($P = 0.00407$). When using the multinomial GLM, SADs with local spatial extent were again significantly less likely to be multimodal versus 1PLN (Fig. 4; $pMCMC = 0.01943$, where $pMCMC$ is the MCMC $P$-value), but not at continental and regional scales. There is a positive effect of the interaction between number of families and the local scale, with the proportion of multimodality versus 1PLN increasing as the number of families increases ($pMCMC = 0.00106$). In other words, relative to 1PLN, multimodality is significantly less prevalent at local scales and low family richness than when family richness is higher or spatial extent is regional or continental. Conversely, logseries is less prevalent versus 1PLN at continental scales ($pMCMC = 0.01636$), and more prevalent at regional and local scales ($pMCMC = 0.00923$ and $pMCMC = 0.01578$, respectively; Fig. 4, Table 1). These effects are independent of the number of families, which does not significantly influence the proportion of logseries versus 1PLN.

**DISCUSSION**

Here we show that 17 out of 117 SADs are multimodal with high confidence (c. 15%). Further, there is a higher prevalence of multimodality for communities with broader spatial scale or higher taxonomic breadth, suggesting that multimodality increases with ecological heterogeneity. This warrants systematic consideration of multimodality in the quantification of SAD shape.

Our analysis across different taxa, biomes and species richness indicates that multimodality is not an artefact of particular SADs. The only particularity of the SADs analysed is that they were intensely sampled, and there is no reason to suspect that this holds any influence as to whether the underlying ecological community is multimodal. Furthermore, because each empirical SAD analysed corresponds to only one year of sampling, multimodality reflects the structure of the community at a particular point in time. Additionally, we infer multimodality only when it is supported by both AICc and PBLRT. Given that in our simulation study false negatives were more prevalent than false positives, this renders our conclusions highly conservative. A caveat of our study is that the SADs analysed here do not fully represent the spectrum of community variability in terms of spatial and taxonomic coverage; furthermore, our sample of SADs was not intended to be representative of taxa, habitat, climatic regions or even realm. Nevertheless, our results show a positive effect of both spatial scale and taxonomic breadth on the prevalence of multimodality, regardless of taxa and realm.

The prevalence of multimodality that we found differs from that suggested by Barabás et al. (2013). Our simulation study showed that, depending on the parameter combination, sampled communities from a single PLN can indeed produce apparently multimodal SADs, as the authors suggested. However, we believe that the method developed here improves our ability to test for multimodality. Despite there being no direct correspondence between Barabás et al.’s parameterization and ours, their Fig. 4 suggests that the mode of the average unimodal distributions is located around octave 6 of the SAD, with the distributions spanning 11 octaves. This could be compared with our 1PLN simulations for larger $\mu$, $\sigma$ and species richness values, which fall in the parameter space for
Table 1 Results of binomial (a) and multinomial (b) generalized linear model (GLM) fitting, showing a positive effect of spatial scale or higher taxonomic breadth on the prevalence of multimodality.

(a) Binomial GLM

| Reference: 1PLN-SpatialExtent.Continental | Estimate | SE  | z-value | Pr(>|z|) |
|-------------------------------------------|----------|-----|---------|---------|
| SpatialExtent.Continental                | -0.2207  | 1.0135 | -0.2180 | 0.8277  |
| SpatialExtent.Regional                   | -1.5511  | 1.2830 | -1.2090 | 0.2267  |
| SpatialExtent.Local                      | -3.7396  | 1.3940 | -2.6830 | 0.0073  |
| NumberFamilies                           | -0.0127  | 0.0154 | -0.8230 | 0.4105  |
| SpatialExtent.Regional : NumberFamilies  | 0.0060   | 0.0247 | 0.2430  | 0.8084  |
| SpatialExtent.Local : NumberFamilies     | 0.0747   | 0.0260 | 2.8730  | 0.0041  |

(b) Multinomial GLM

<table>
<thead>
<tr>
<th>Reference: SpatialExtent.Continental</th>
<th>Posterior mean</th>
<th>Lower 95% CI</th>
<th>Upper 95% CI</th>
<th>pMCMC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multimodality : SpatialExtent.Continental</td>
<td>-0.0013</td>
<td>-2.4390</td>
<td>2.3690</td>
<td>0.9930</td>
</tr>
<tr>
<td>Multimodality : SpatialExtent.Regional</td>
<td>0.0183</td>
<td>-3.1700</td>
<td>3.2370</td>
<td>0.9944</td>
</tr>
<tr>
<td>Multimodality : SpatialExtent.Local</td>
<td>-3.7030</td>
<td>-6.9170</td>
<td>-0.5603</td>
<td>0.0194</td>
</tr>
<tr>
<td>Multimodality : SpatialExtent.Regional : NumberFamilies</td>
<td>-0.0248</td>
<td>-0.0669</td>
<td>0.0098</td>
<td>0.1597</td>
</tr>
<tr>
<td>Multimodality : SpatialExtent.Regional : NumberFamilies</td>
<td>-0.0124</td>
<td>-0.0805</td>
<td>0.0541</td>
<td>0.7079</td>
</tr>
<tr>
<td>Multimodality : SpatialExtent.Local : NumberFamilies</td>
<td>0.0897</td>
<td>0.0301</td>
<td>0.1519</td>
<td>0.0011</td>
</tr>
<tr>
<td>Logser : SpatialExtent.Continental</td>
<td>-186.0000</td>
<td>-336.7000</td>
<td>-0.4936</td>
<td>0.0164</td>
</tr>
<tr>
<td>Logser : SpatialExtent.Regional</td>
<td>188.1000</td>
<td>1.9480</td>
<td>338.0000</td>
<td>0.0092</td>
</tr>
<tr>
<td>Logser : SpatialExtent.Local</td>
<td>186.2000</td>
<td>1.4960</td>
<td>337.8000</td>
<td>0.0041</td>
</tr>
<tr>
<td>Logser : SpatialExtent.Regional : NumberFamilies</td>
<td>1.7340</td>
<td>-0.8160</td>
<td>4.5390</td>
<td>0.3565</td>
</tr>
<tr>
<td>Logser : SpatialExtent.Local : NumberFamilies</td>
<td>1.7650</td>
<td>-0.8011</td>
<td>4.5560</td>
<td>0.3454</td>
</tr>
</tbody>
</table>

Pr, two-tailed p-value; pMCMC, Markov chain Monte Carlo P-values; PLN, Poisson lognormal distribution.

For the Bayesian GLM, the posterior mean estimates, the 95% credible intervals and the pMCMC values are shown. The parameter estimates were considered statistically significant when pMCMC < 0.05, and the 95% credible intervals (CIs) did not include 0. The term ‘Multimodality : SpatialExtent.Local : NumberFamilies’ refers to the estimation of multimodality versus 1PLN at the continental scale with the interaction of number of families.

which AICc has a higher chance of erroneously selecting multimodality. Thus it would be interesting to investigate whether performing the additional LRT on the SADs generated using Barabási et al.’s parameterization would still yield similar multimodality frequencies.

General explanations for multimodality

Scale is fundamental to understanding patterns of biodiversity (Levin, 1992; McGill, 2010). Our results indicate that multimodality is more likely to occur for regional- to continental-scale SADs, albeit not exclusively. Some SADs selected as multimodal consist of local samples or plots, but all of these are taxonomically diverse (between 12 and 76 families): ID3 consists of macrobenthos samples from the Belgian Continental Shelf; IDs 95 and 96 consist of tropical forest plots in Malaysia; ID99 in Thailand and IDs 101 and 102 in Brazil and Colombia, respectively; and IDs 45, 92 and 108 consist of vegetation plots in the USA (desert, shortgrass steppe and dune vegetation, respectively). This matches our regression analysis, for which local SADs with low family richness exhibited lower prevalence of multimodality than it did at high family richness or broad spatial scales.

The explanatory variables we analysed here mirror the spatial and organizational scales suggested by Levin (1992) as underpinning the variability of ecological patterns, and they support previous explanations for multimodality. Multimodality has been proposed to arise as consequence of species differences in ecological or functional characteristics (e.g. Magurran & Henderson 2003; Alonso et al., 2008) and environmental heterogeneity (Dornelas et al., 2009). Both of these explanations are consistent with a greater prevalence of multimodality in communities with greater spatial extent or taxonomic diversity. Our goal was not to develop a predictive model for multimodality, but to quantify its prevalence and test its association with relevant ecological variables. We believe that exploring in more detail the effects of environmental heterogeneity, functional diversity and core–transient species will prove a fruitful avenue for further understanding of what aspects of ecological heterogeneity affect SAD shape and lead to multimodality.

An additional interesting research question is how temporal variability in species abundances might affect the shapes of SADs over time. In the present study, we were interested in removing the potential effect of temporal fluctuations in the relative abundances of species across years, to avoid the
posibility that multimodality could arise as an artefact of a single mode changing position over time. In principle, it is also possible that pooling could reduce multimodality, if changes in the position of modes over time make multiple modes more difficult to detect (for instance, if multimodality arises as a transient feature of communities, as an effect of particular stochastic environmental effects). Because the models we use implicitly account for sampling effects, and require actual counts (the number of individuals sampled), an investigation into the effects of temporal averaging would require the development of an alternative statistical approach.

Rarity and commonness

SAD studies have often focused on the left-hand side of the distribution and on the ability of different theoretical models to accommodate the rarest species mode (e.g. Hubbell 2001; McGill 2003b), and several studies have described the rarer mode as the one leading to a multimodal pattern (Magurran & Henderson, 2003; Borda-de-Água et al., 2012; Matthews et al., 2014). Although a mode was often fitted to the rarest species, some of the empirical SADs also exhibited modes for very abundant species (e.g. IDs 30, 92, 99 and 108 in Fig. 3). This highlights the observation that communities characterized by very high abundances of the most abundant species might not be accommodated within a single lognormal SAD, and a multimodal model provides a better description, similarly to communities with a very high prevalence of rare species. While the majority of species are rare and the universal ‘hollow curve’ SAD is the definitive description of this, the few most common species disproportionately dominate communities in terms of abundance and ecological processes (Gaston, 2010, 2011), and might also have considerable influence on SAD shape (e.g. Connolly et al., 2014).

The logseries was selected as best model relatively frequently, despite all the data coming from intensely sampled communities. This suggests that, even for high sampling intensity, some communities are characterized by a very high proportion of rare species. The logseries was more often selected for communities encompassing smaller spatial scales, a finding consistent with the results of our regression analysis. Additionally, visual inspection suggests that there was a slight tendency for the logseries to be favoured when species richness was lower (not shown), and in our analysis logseries was never the model with the best absolute fit (in terms of negative log-likelihood values only; cf. Baldrige et al., 2015). Interestingly, none of the SADs selected as logseries had the largest spatial extent, contrasting with the predictions of neutral theory with point-mutation speciation (Hubbell, 2001), which predicts a logseries SAD for the metacommunity. On the other hand, the maximum entropy theory of ecology (METE; Harte et al., 2008) predicts a logseries SAD, contrasting with the support for multimodality that we found, and with the effect of spatial scale and taxonomic breadth on model frequency.

Model selection tools

The simulation study showed that the position of the modes, species richness and particularly \( \sigma \) values greatly affected model selection, for both AICc and BIC. Additionally, species richness often had contrary effects on the information criteria; this can be related to the high level of penalization exerted by AICc as sample size decreases (Burnham & Anderson, 2002), while the opposite happens for BIC (by definition), which can be problematic when testing for multimodality in SADs. As expected, BIC was more conservative than AICc, reflected in both the very low false positive frequency and particularly in the relatively high frequency of false negatives. While the former is a highly desirable feature of a selection method, the latter suggests that BIC can be insensitive to deviations in SADs indicative of multimodality.

On the other hand, although AICc overestimated the number of modes for some parameter combinations, for a large number of empirical SADs with estimated parameters within that space, the more parsimonious model was selected. This suggests that AICc does not overestimate the number of modes generally, and that model selection criteria might be affected by parameter values in a non-directional fashion. As noted before for SADs, comparative measures of goodness of fit can often produce conflicting results (McGill, 2003a; McGill et al., 2007). We showed that the additional calculation of LRT frequency distributions further reduces the probability of erroneously selecting multimodality compared with using AICc alone.

CONCLUSIONS

Multimodal SADs occur at a non-negligible frequency. Larger spatial scale or higher taxonomic breadth can yield multimodal SADs. Greater spatial scale and taxonomic breadth of communities imply higher ecological heterogeneity. In turn, this is expressed as different levels of species abundances, thus being reflected in the shape of the SAD and providing information on community structure. Here we show that the dichotomy between logseries and lognormal as the sole adequate descriptors of SAD should be expanded to include multimodal models. This will enhance our ability to use SADs to detect the effects of ecological or functional mechanisms affecting the communities. Furthermore, differences in SAD shape across different scales provide important insights into the current endeavour of biodiversity scaling.

ACKNOWLEDGEMENTS

We thank Maria João Janeiro, Petr Keil, Felix May and another anonymous referee for their thoughtful comments which considerably improved our manuscript. The authors are grateful to the online repositories OBIS, Ecological Data Wiki and GBIF. We are also grateful to all the scientists and participants in data collection, to all the data providers for making data publicly available and to their funders (see Appendix S3 for the complete list of acknowledgements). We thank the University of St
Andrews MHD Cluster and the Bioinformatics Unit (Wellcome Trust ISSF grant 105621/Z/14/Z). L.H.A. was supported by Fundação para a Ciência e Tecnologia, Portugal (POPH/FSE SRH/BD/90469/2012). A.E.M. acknowledges the ERC (BIOTIME 250189). M.D. acknowledges funding from the Marine Alliance for Science and Technology Scotland (MASTS), funded by the Scottish Funding Council (grant reference HR09011) and contributing institutions.

The authors declare no conflict of interest.

REFERENCES


Multimodal SADs linked to spatial and taxonomic breadth

BIOSKETCH

Laura Henriques Antão is interested in general patterns in ecology and macroecology, across taxa and ecosystems, and also in biodiversity and the processes explaining species coexistence. Recently, her research has focused on species abundance distributions.

Author contributions: L.H.A. assembled the datasets and performed all the analyses and modelling work. S.C. and M.D. contributed to the modelling development and analysis of the results. L.H.A. wrote the first draft of the manuscript; all authors have discussed the results and contributed extensively to improved revisions.

Editor: Pet Keil