Research

Ecography 37: 1002–1011, 2014
doi: 10.1111/ecog.00861

Subject Editor: Joaquin Hortal. Accepted 3 April 2014

The gambin model provides a superior fit to species abundance distributions with a single free parameter: evidence, implementation and interpretation

Thomas J. Matthews, Michael K. Borregaard, Karl I. Ugland, Paulo A. V. Borges, François Rigal, Pedro Cardoso and Robert J. Whittaker

T. J. Matthews (thomas.matthews@ouce.ox.ac.uk), M. K. Borregaard and R. J. Whittaker, Conservation Biogeography and Macroecology Group, School of Geography and the Environment, Univ. of Oxford, South Parks Road, Oxford, OX1 3QY, UK. RJW also at: Center for Macroecology, Evolution and Climate, Dept of Biology, Univ. of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark. – P. A. V. Borges, F. Rigal, P. Cardoso and TJM, Azorean Biodiversity Group (ABG, CITA-A) and Portuguese Platform for Enhancing Ecological Research and Sustainability (PEERS), Depto de Ciências Agrárias, Univ. of the Azores, Rua Capitão João d´Ávila, Pico da Urze, PT-9700-042, Angra do Heroísmo, Portugal. PC also at: Finnish Museum of Natural History, Univ. of Helsinki, PO Box 17, FI-00014 Helsinki, Finland. – K. I. Ugland, Dept of Marine Biology, Inst. of Biosciences, Univ. of Oslo, PO Box 1066, Blindern, NO-0316 Oslo, Norway.

The species abundance distribution (SAD) has been a central focus of community ecology for over fifty years, and is currently the subject of widespread renewed interest. The gambin model has recently been proposed as a model that provides a superior fit to commonly preferred SAD models. It has also been argued that the model's single parameter (α) presents a potentially informative ecological diversity metric, because it summarises the shape of the SAD in a single number. Despite this potential, few empirical tests of the model have been undertaken, perhaps because the necessary methods and software for fitting the model have not existed. Here, we derive a maximum likelihood method to fit the model, and use it to undertake a comprehensive comparative analysis of the fit of the gambin model. The functions and computational code to fit the model are incorporated in a newly developed free-to-download R package (gambin). We test the gambin model using a variety of datasets and compare the fit of the gambin model to fits obtained using the Poisson lognormal, logseries and zero-sum multinomial distributions. We found that gambin almost universally provided a better fit to the data and that the fit was consistent for a variety of sample grain sizes. We demonstrate how α can be used to differentiate intelligibly between community structures of Azorean arthropods sampled in different land use types. We conclude that gambin presents a flexible model capable of fitting a wide variety of observed SAD data, while providing a useful index of SAD form in its single fitted parameter. As such, gambin has wide potential applicability in the study of SADs, and ecology more generally.

A species abundance distribution (herein 'SAD') describes the abundances of all species sampled within a given community (Ulrich et al. 2010). Because SADs characterise the structure of ecological communities, they constitute the foundation of much (macro-) ecological and biogeographical theory (Preston 1948). Thus, considerable effort has been made to characterise empirical SADs in a statistically tractable framework. To date around 30 different SAD models have been published (McGill et al. 2007). The two most commonly used are the logseries (Fisher et al. 1943) and the lognormal (Preston 1948). The Poisson lognormal ('PLN'; Bulmer 1974) is usually preferred to the continuous lognormal as the PLN incorporates a sampling theory and the continuous lognormal allows fractional abundances. The neutral model of Hubbell (2001) also predicts a SAD, termed the zero-sum multinomial distribution ('ZSM').

The logseries has often been found to provide a poor fit to empirical SADs (Ulrich et al. 2010; but see White et al. 2012), and the use of the PLN has been criticised, both in general terms (Williamson and Gaston 2005), and particularly in comparative analysis, as the parameters are not as intuitively interpretable as those of the logseries (but see Engen 2001, Sæther et al. 2013). Thus, there is a need for a model that provides a good fit to a variety of empirical data, is analytically tractable and possesses an easily interpreted parameter (i.e. one that simply describes the shape of the SAD), which can then be used as a system descriptor in further analyses. In a recent paper, Ugland et al. (2007)

suggested that a mixed gamma binomial distribution ('gambin') meets these criteria. Gambin is a stochastic model that combines the gamma distribution with a binomial sampling method. The distribution has a single parameter (alpha: α), which describes the shape of the distribution.

A common context in which the form of SADs is of interest is in analyses of community responses to disturbance and land use change (Ugland and Gray 1982, Mac Nally 2007, Dornelas et al. 2011). For such purposes, a model must provide a good fit to a variety of SAD shapes, and must possess a parameter which summarises, and tracks change in, the shape of the distribution and thus allows a comparison between the SADs of undisturbed and disturbed communities. The limited application of gambin thus far has produced promising results, with good fits to empirical data (Ugland et al. 2007). However, the lack of easily accessible methods for fitting the gambin distribution to empirical datasets, and the absence of maximum likelihood methods for estimating α has limited the number of applications of the model. Here, we address these issues by presenting a maximum likelihood derivation of the gambin distribution in conjunction with an R package (gambin). We then use these methods to undertake a comprehensive evaluation of gambin. We present a rigorous test of the fit of the gambin distribution using a variety of datasets. First, we use an extensive and well-specified arthropod dataset from the Azorean archipelago: a dataset large enough (over 90 000 individuals) to permit a statistically powerful comparison of alternative SAD models. The dataset includes samples from a range of land use types, which allows us to test the applicability of α as an ecological indicator. As the form of the SAD may change with the spatial grain of sampling and the number of individuals in a sample (Preston 1948, McGill 2011, Borda-de-Água et al. 2012) a SAD model must perform well across a variety of spatial scales, and be relatively independent of sample size, to be of general utility. Thus, we also test the utility of α across a range of spatial grain sizes. In addition, we sourced ten species abundance datasets from the literature, representing a range of taxa and ecological contexts. We use these datasets in combination with the Azorean data to address the following questions: 1) does the gambin distribution provide an adequate fit to the data? 2) Are there are any parts of the empirical distribution in which gambin does or does not provide a good fit (cf. Connolly and Dornelas 2011)? 3) How does the fit of the gambin distribution compare to three popular competitor SAD models (Poisson lognormal, logseries and zero-sum multinomial)? 4) Is α sensitive to the proportion of individuals sampled from a community (cf. McGill 2011, Locey and White 2013)? 5) Is the performance of gambin constant across sampling grain sizes (Šizling et al. 2009, Borda-de-Água et al. 2012)? 6) Can α differentiate between community structures in different land use types (cf. Ugland et al. 2007, Dornelas et al. 2011)?

Material and methods

Gambin distribution

Here we present a brief overview of the model, noting that a mathematical description of the gambin distribution is provided by Ugland et al. (2007). The basic idea underpinning gambin is that although the abundance of a species is determined by a mixture of many deterministic and stochastic factors, it is possible to obtain a good approximation to the observed species abundance curves by modelling the population sizes of the species in a community in two steps. First, we define the fitness of a species as the probability of achieving a large population size, and represent the frequencies of fitness values (i.e. the probability density) across species by a gamma distribution with scale parameter fixed at the value 1. The gamma distribution was chosen as the basis of the model as it is known to be a flexible distribution and this flexibility is preserved when the number of parameters is reduced by fixing the scale parameter. The scale parameter of gambin is set to 1 as the scaling of the distribution is achieved by fixing the max octave (see below). This is a pragmatic choice as it allows for all octaves to be fitted whilst simultaneously reducing the number of model parameters to one. Thus, the shape parameter (α), which determines the form of the distribution, is the only free parameter. A small α induces a distribution skewed to the left, i.e. a high density at small abundance values. This is what is observed in logseries-like distributions. A high α induces a distribution closer to normal on a log scale of abundances.

The second step is to link the gamma distribution (i.e. the fitness frequencies) to the actual abundance values. As empirical SADs are highly right skewed, we log transform the number of individuals into 'octaves' (Gray et al. 2006), and then estimate the number of species in each octave (Ugland et al. 2007). While binning has been criticised for resulting in the loss of information, it is unlikely to bias parameter estimates (Connolly and Dornelas 2011). Furthermore, as empirical abundance data often contain a significant amount of variation due to sampling effects, the added precision of models that are not based on binned data is questionable. We created abundance octaves by a simple log2 transform that doubles the number of abundance classes within each octave (a sensitivity analysis demonstrated the robustness of the results to this binning procedure; see below). Thus, octave 0 contains the number of species with 1 individual, octave 1 the number of species with 2 or 3 individuals, octave 2 the number of species with 4 to 7 individuals, and so forth (see Supplementary material Appendix 1 for a more detailed description). This binning procedure is method 3 of Gray et al. (2006), and is listed by that study as the most appropriate binning method from several reviewed. The assignment of a species abundance into octave x is then regarded as the result of a binomial process with x trials (Supplementary material Appendix 1 and Ugland et al. 2007).

A description of the gambin R package

The full derivation of the maximum likelihood estimation of gambin, which is novel to this paper, is presented in Supplementary material Appendix 1 and is incorporated in an R package (gambin; ver. 1.0 accepted by CRAN in September 2013). Examples of use of the gambin package are also presented in Supplementary material Appendix 1. The gambin package contains functions to calculate the gambin

distribution (dgambin) and to fit the gambin distribution to empirical data using maximum likelihood (fitGambin), along with a set of utility functions. The dgambin function returns the density function of the gambin distribution, given α and the octave number of the most abundant species. The syntax is similar to e.g. the basic density functions in the base package stats (e.g. dnorm). To get the gambin distribution in units of species, the gambin_exp function is used.

The fitGambin function accepts a vector of abundances (optionally using a subsample of the individuals), which it bins into octaves using the utility function create_octaves. It then uses optimisation algorithms in R to identify the value of α that maximises the likelihood function. The return value of fitGambin is an S3 object of class 'gambin'. This class provides print, confint, predict and plot functions: confint gives the 95% confidence interval around the estimated α value; predict gives the predicted number of species in each octave; and plot creates a bar graph comparing empirical abundances (shown as grey bars) to the predicted values from the gambin distribution (shown as black dots).

Azorean arthropod data

Our main test dataset forms part of a long-term (1999–2012) biological study, the BALA ('Biodiversity of Arthropods from the Laurisilva of the Azores') project (Borges et al. 2005, Ribeiro et al. 2005). Eighteen fragments of protected native Laurisilva forests were sampled for arthropods across seven islands using 100 randomly located transects (150×5 m). Datasets were created by pooling the fragment samples present on each individual island. One island (Santa Maria) only has a single fragment and was thus not used as an island dataset (see Supplementary material Appendix 2, Table A1 and A2 for site and sample details). Thus, we used 18 fragment samples and six island samples in this study. Along each transect, ground surface arthropods (largely epigean) were surveyed using 30 pitfall traps (Borges et al. 2005), while canopy species were surveyed using a beating tray methodology (Ribeiro et al. 2005) focused on three primary tree species (see Supplementary material Appendix 2, Table A3 for all species information). In all, 6770 samples (3420 pitfall traps and 3350 beating samples) were used for the current study (for further details see Gaspar et al. 2008). In addition to the native Laurisilva fragments we used data from a related study on several of the islands, sampling epigean soil arthropods by the same pitfall methods for two additional land-uses – exotic plantation forest, and pasture (Cardoso et al. 2009, 2013).

Model comparison

Both the 18 native forest fragment samples and the six island samples were used in our first model comparison, resulting in 24 different samples. Our first step was to fit the gambin distribution to each sample using the gambin R Package. We then fit the two most commonly used statistical distributions (PLN and logseries). To fit the PLN we used the poilog R package (Grøtan and Engen 2009). The zero-sum multinomial distribution (ZSM) predicted by Hubbell's (2001, see Matthews and Whittaker 2014 for a review) spatially implicit neutral theory was also fit to the 24 samples using the analytical form and likelihood function derived from Etienne (2005). We used two approaches to compare the goodness of fit of the various models. First, we compared the models using a Pearson's chi-square (χ^2) goodness of fit test. As the choice of goodness of fit test has been found to influence results (McGill 2003) we also used a Kolmogorov–Smirnov test. Results were found to be the same using either test and so we present only the χ^2 test results herein. Sole reliance on traditional goodness of metrics such as χ^2 has been criticised as unreliable (McGill 2003). Thus, we also used an information theoretic approach (Burnham and Anderson 2002) to compare the fit of gambin to that of the three other models (i.e. the PLN, logseries and ZSM). Model performance was compared using the Bayesian information criterion (BIC) and Akaike's information criterion corrected for small sample size (AIC_c). The smallest BIC and AIC_c values were taken to represent the single best model for a given sample providing that ΔBIC or ΔAIC_c to the next best model was > 2 (Burnham and Anderson 2002). The PLN was considered a two parameter model, the ZSM a three parameter model, and gambin and the logseries one parameter models.

Unlike the other distributions in our multi-model comparison, the maximum likelihood estimate of the ZSM is very sensitive to the initial parameters used in the optimisation process. As such, multiple initial parameter values were used in the model fitting process along with different optimisation algorithms in R ('optim' and 'mle2'). As the gambin distribution is calculated using octaves, the other distributions in the model selection framework were also fitted to binned data using the same binning process as described above. The ZSM analysis was restricted to the complete assemblage (i.e. not the sample grain subsets and additional datasets; see below) as the model fitting process was very time intensive. This omission is unlikely to affect our estimation of gambin's performance as initial observations found the ZSM only rarely provided a superior fit to gambin.

While statistical approaches to assessing goodness of fit are evidently more objective, visually analysing the fitted distribution can provide valuable information (Connolly and Dornelas 2011). Thus, for each of the 24 samples we plotted the SAD with the fitted gambin model and visually inspected the fit to determine if there are particular parts of the SAD, or types of SAD patterns, where gambin does or does not provide a good fit.

Sourced datasets

We also compiled ten additional datasets widely used in SAD studies, representing a variety of systems and taxa (Table 1; see Supplementary material Appendix 2 for full acknowledgments to the data providers). Although a number of these datasets were used in the original gambin paper by Ugland et al. (2007), both our software and the determination of goodness of fit are novel to the present study and the inclusion of these datasets thus permits direct comparison with Ugland et al. (2007) and with other comparative SAD analyses that have used these datasets. We again deployed the multi-model comparison for each dataset (excluding the ZSM); examining the fit using the aforementioned graphical methods at each stage.

Table 1. Additional datasets used in the multi-model comparison and the associated dataset characteristics: location, taxon, number of species (S) and number of individuals (N). The full related acknowledgements are presented in Supplementary material Appendix 2. The best model was determined by comparing gambin with the Poisson lognormal distribution (PLN) and the logseries distribution using both a Pearson's χ^2 goodness of fit test and an information theoretic approach. PLN has two parameters and gambin and the logseries are single parameter models. A model was selected as best by means of the information theoretic approaches if it had the lowest BIC and AIC_c using a minimum difference in both criteria of two. + relates to cases where none of the distributions provided a satisfactory fit.

					Best model	
Dataset location	Taxon	S	N	Source	χ^2	AIC_c and BIC
Barro Colorado Island, Panama	Trees (>10 cm)	229	20 852	Hubbell et al. (2005)	Gambin	Gambin
UK	Birds	503	151 781 104	Baker et al. (2006)	+	+
Firth of Clyde, Scotland	Marine nematodes	113	8896	Lambshead (1986)	Gambin	Gambin
Australia	Corals	154	44 225	Dornelas and Connolly (2008)	Gambin	Gambin
English Channel	Marine nematodes	321	1200	unpubl. data	Gambin	Gambin
Irish Sea	Marine nematodes	178	58 372	Lambshead and Boucher (2003)	Gambin	Gambin
Rothamsted, UK	Lepidoptera	195	6813	Williams (1964)	Gambin	Gambin
Pasoh Forest Plot, Malaysia	Trees (>1 cm)	808	295 133	FRIM (2013)	Gambin	Gambin
Hinkley Point, UK	Fish (1981–2003)	82	3891	Magurran and Henderson (2003)	Gambin	Gambin
Sherman Forest Plot, Panama	Trees	129	3363	Condit (1998)	Gambin	Gambin

Sample size and grain size

To determine the sensitivity of model parameters (gambin's α, the mean of the PLN, and the alpha of the logseries) to the proportion of individuals in a sample, we simulated a large number of different metacommunities of two types, lognormal (i.e. the SAD of the metacommunity was lognormal) and logseries. The set of lognormal metacommunities comprised subsets of metacommunities with 50, 100, 500 and 1000 species. For each category of species richness, we simulated metacommunities with the number of individuals varying from 1000 to 1 000 000. There were 9 lognormal metacommunities simulated (see Supplementary material Appendix 3, Table A4 for their exact properties). In addition, six logseries metacommunities were simulated, comprising subsets with 50, 100 and 500 species (each with 1000 and 10 000 individuals, details in Supplementary material Appendix 3, Table A5). This range of metacommunities covered the two extremes of observed empirical distributions (i.e. lognormal and logseries) and a wide range in number of individuals (N) and number of species (S). For each of the described metacommunities we randomly sampled 1, 10, 25 and 50% of individuals using functionality in the gambin R package. We then fit the three distributions to each sample and to the full community (i.e. 100% of individuals), recording the parameters of each distribution in addition to the maximum and modal octaves (Locey and White 2013) of the observed distribution. This procedure was repeated 100 times in each case.

In addition, for a subset of four of the simulated lognormal metacommunities representing a range of S and N, 19 samples were taken, ranging from 5% of individuals in each metacommunity, to 95% of individuals, at intervals of 5%. The three models (gambin, PLN, and logseries) were fitted to the samples and the relevant parameters recorded. This process was repeated 100 times and the mean of the parameter values calculated. We then calculated the sample size needed in order for the parameter estimates of the sample to be within 10, 25 and 50% of the parameter estimates for the whole community.

To test whether the fit of gambin was consistent across sampling grain sizes, three native Azorean forest fragments that contained large numbers of transects were selected. For each of these fragments the multi-model comparison (excluding the ZSM) was calculated for a sample consisting of a single transect, then two transects and so on iteratively up to the maximum number of transects in a given fragment. Species abundances were averaged across all possible combinations of transects in each case. So, for example, when creating a sample from two transects, we averaged species abundances across all possible combinations of two transects within that fragment. All samples from each island were combined into island-level datasets as examples of larger grain sizes. It was not necessary to keep N constant to calculate α in this particular analysis as here we were testing whether the fit of gambin is consistent across a range of sample sizes (i.e. a range of N).

Comparison of alpha values between land use types

To evaluate the performance of gambin's α as an ecological indicator, we ranked the different habitat types in the Azorean dataset from untransformed (native forest) through moderately transformed (non-native forest plantations) to highly transformed (agricultural pasture). Three islands, Terceira, Flores and Faial, had suitable sample sizes for each habitat type and for these purposes each island was treated as a distinct system. Within each habitat type the gambin distribution was fitted to each transect and the α value recorded. This was done using a re-sampled set of samples using a fixed N value determined by the number of individuals in the least populated transect in order to remove any bias due to differences in sample size between land uses. A Wilcoxon rank sum test was used to determine whether α varied significantly as a function of land use. This method was repeated for the alpha parameter of the logseries and the sigma parameter of the PLN, again after accounting for differences in sample size.

Testing the sensitivity of alpha to the binning method

While our binning method (log2) is frequently employed in SAD studies, the choice of log2 is somewhat arbitrary and theoretically any base could be used to bin the data into octaves. Thus, to test the robustness of the decision to

use log2, we conducted two analyses. First, we calculated gambin's α for the 18 fragments using our standard log2 binning method and compared these values to α derived using a variety of other bases (base e, base 3, base 4 and base 5). In order to compare alpha values derived from these different binning methods we plotted pairwise comparisons of each α set and tested the degree of correlation in each instance using Pearson's product-moment correlation. Second, we undertook a simulation approach. For each run of the analysis, the aforementioned metacommunity simulation method was used to simulate communities with a) 10 000 individuals, and b) 100 000 individuals. For the first iteration of the run, the number of species was set to 50. At each subsequent iteration, species number was increased by 50, up to the maximum of 1000 species (i.e. 20 iterations). At each stage the gambin model was fit to the data using the five different binning methods, and the α values recorded. At the end of each run, a Pearson's product moment correlation was calculated for each pairwise comparison of α sets (i.e. different binning methods) and the values stored. The number of runs was set to 100 and the mean of the correlation values, along with the corresponding standard deviation, was calculated for each pairwise comparison. We were unable to test bases higher than 5 due to constraints on the number of individuals in our samples. For example, use of base ten resulted in only three octaves for certain samples. All analyses were conducted in R (R Development Core Team).

Results

Model comparison and inspection of the fit

For the Azorean fragment and island model comparison analyses, gambin provided a better fit to the data according to the Pearson's χ^2 test for all but six of the fragments (Table 2), and for all but one of the six islands (Supplementary material Appendix 3, Table A6). For three fragments (16–18), the PLN provided a better fit, and for three fragments (7, 10 and 15) and one island (3: Pico) the ZSM provided a better fit, according to χ^2. Gambin outperformed the other three distributions according to both BIC and AIC_c for all 24 samples (Table 2 and Supplementary material Appendix 3, Table A6). Gambin also performed best for all of the non-Azorean datasets according to χ^2 tests (Table 1, see Supplementary material Appendix 3, Fig. A1 for examples), and for nine of the ten datasets according to AIC_c and BIC. None of the models provided an adequate fit for the British breeding bird data.

Visual examination suggested that gambin provides a good fit to a variety of empirically observed distribution shapes (e.g. Fig. 1a, see also Supplementary material Appendix 3, Fig. A1), ranging from logseries-like to lognormal-like patterns. Equally, the fit is consistent across the different parts of the SAD, i.e. for both the rarer species and more common species. Nonetheless, visually the fit does not appear to be as good for those samples which exhibited a degree of multi-modality within the SAD (e.g. Fig. 1b).

The effect of sample size

Our analyses showed that when simulating lognormal metacommunities, the shape of the observed SAD, and thus the value of gambin's α, varied as a function of the proportion of individuals sampled (Supplementary material Appendix 3, Table A4 and A5). Figure 2 illustrates this for one particular lognormal metacommunity. When basing analysis on a small sample, the observed SAD exhibits a logseries shape (Fig. 2a), but as the sample size increases, the shape of the

Table 2. Goodness of fit and model selection results for arthropod SADs of 18 (No.) native Laurisilva forest fragments in the Azores. Arthropods were sampled using a standardised pitfall trap and canopy beating methodology between 1999 and 2004. For each fragment the Pearson's χ^2 statistic and associated p value (in parentheses) are presented for the gambin, logseries, Poisson lognormal (PLN) and zero-sum multinomial (ZSM) distributions. The Bayesian information criterion (BIC) and Akaike's information criterion corrected for small sample size (AIC_c) are also given for all four distributions. PLN has two parameters and the ZSM has three parameters. Gambin and the logseries are single parameter models. The best model according to each criterion, using a minimum difference of two, is highlighted in bold for each fragment. Fragment information can be found in Supplementary material Appendix 2, Table A1.

	No.			Logseries			PLN			ZSM		
No.	χ^2 (p)	BIC	AIC_c	χ^2 (p)	BIC	AIC_c	χ^2 (p)	BIC	AIC_c	χ^2 (p)	BIC	AIC_c
1	**19.6 (0.03)**	**395.6**	**395.7**	52.8 (0.03)	421.3	421.3	25.4 (0.01)	403.1	403.8	20.0 (0.04)	414.0	417.1
2	**3.8 (0.92)**	**339.0**	**339.2**	42.2 (0.16)	377.1	377.3	8.1 (0.53)	344.2	345.3	3.9 (0.98)	367.9	371.0
3	**11.3 (0.34)**	**491.2**	**491.3**	36.5 (0.4)	515.4	515.4	18.1 (0.05)	500.8	501.5	19.0 (0.06)	508.1	510.3
4	**8.2 (0.51)**	**365.9**	**366.1**	37.3 (0.32)	393.6	393.8	13.4 (0.15)	372.1	373.2	10.0 (0.63)	386.6	389.7
5	**10.9 (0.36)**	**413.1**	**413.1**	38.1 (0.33)	439.3	439.4	16.0 (0.1)	420.7	421.4	11.3 (0.42)	431.9	434.2
6	**23.7 (0.01)**	**412.1**	**412.2**	68.4 (0.01)	446.9	447.0	29.8 (0)	418.8	419.5	33.3 (0.11)	436.5	439.6
7	25.5 (0.01)	**526.5**	**526.5**	67.2 (0.01)	563.4	563.5	26.3 (0.01)	532.6	533.3	**10.5 (0.49)**	550.4	552.6
8	**19.0 (0.06)**	**458.4**	**458.3**	45.1 (0.14)	480.4	480.3	27.8 (0.01)	468.3	468.6	44.2 (0)	473.6	476.7
9	**9.2 (0.51)**	**418.3**	**418.3**	35.6 (0.44)	445.8	445.8	14.2 (0.16)	425.0	425.7	31.7 (0)	442.1	445.2
10	19.5 (0.01)	**531.8**	**532.2**	45.3 (0.07)	559.2	559.5	20.3 (0.01)	539.3	540.9	**7.8 (0.56)**	549.9	554.1
11	**7.5 (0.67)**	**331.2**	**331.2**	24.8 (0.9)	346.8	346.9	13.3 (0.21)	337.7	338.4	70.8 (0)	343.7	346.7
12	**23.6 (0.01)**	**510.7**	**510.9**	50.9 (0.03)	538.7	538.9	24.0 (0.01)	517.9	519.0	27.4 (0.01)	529.0	532.1
13	**25.2 (0.01)**	**582.9**	**582.9**	58.5 (0.01)	615.0	615.0	34.8 (0.01)	593.1	593.8	28.8 (0)	604.5	607.6
14	**15.7 (0.07)**	**485.7**	**485.9**	45.0 (0.1)	514.2	514.4	19.7 (0.02)	493.7	494.8	20.2 (0.31)	505.3	508.4
15	24.9 (0.01)	**387.2**	**387.4**	33.0 (0.52)	404.4	404.6	17.4 (0.07)	390.8	391.9	**8.5 (0.58)**	398.3	401.3
16	28.6 (0.01)	**467.1**	**467.5**	43.4 (0.11)	487.1	487.4	**23.3 (0.01)**	471.9	473.5	32.8 (0.04)	479.7	483.9
17	30.3 (0.01)	**244.4**	**245.2**	55.8 (0.01)	261.9	262.8	**30.0 (0.01)**	247.1	250.2	34.9 (0.04)	258.1	266.3
18	29.9 (0.01)	**440.2**	**440.6**	50.3 (0.03)	460.5	460.9	**26.1 (0.01)**	444.6	446.2	31.3 (0.14)	453.1	457.3

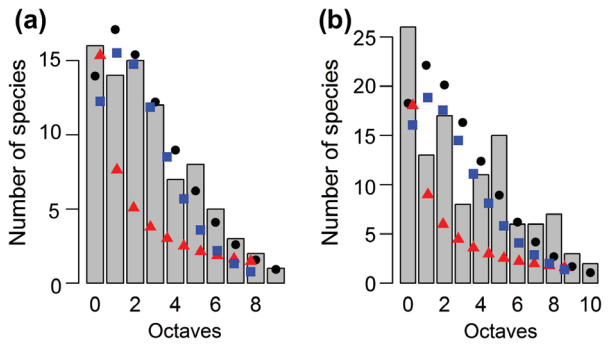

Figure 1. Examples of the fit of the gambin distribution (black dots), the logseries (red triangles), and Poisson lognormal (blue squares) to observed data (bars). (a) Data are for arthropods from a fragment (No. 2, Supplementary material Appendix 2, Table A1) of native Laurisilva forest on the island of Faial, Azores. The α parameter of the gambin distribution is 1.9. (b) Data are for arthropods from a fragment (7, Supplementary material Appendix 2, Table A1) of native Laurisilva forest on the island of Pico, Azores. The α parameter of the gambin distribution is 1.8. In both plots, gambin provides the best fit according to both BIC and AIC_c.

SAD becomes more lognormal and the modal and maximum octaves both shift to the right (Fig. 2c, d), as predicted by Preston's (1948) veil line concept. The mean and sigma of the PLN, and the alpha parameter of the logseries distribution, also change notably with the proportion of individuals in the sample (Fig. 2). For the logseries metacommunities, the change in gambin's α with proportion of individuals sampled was not as pronounced as for the lognormal metacommunities.

Gambin and the PLN generally required similar sample sizes in order for the parameter estimates of the sample to asymptote towards the parameter estimates of the whole community; the logseries generally required slightly smaller sample sizes (see Supplementary material Appendix 3, Table A7 for the full results). For example, for a sample estimate to be within 10% of the whole metacommunity parameter estimate, the proportion of individuals sampled from the metacommunity needed to be on average: 68% (range: 55–90%) for gambin's α, 65% (range: 60–70%) for the mean of the PLN, and 53% for the logseries' alpha (range: 45%–60%; see Supplementary material Appendix 3, Table A7 for the equivalent results based on 25 and 50% accuracy).

The performance of gambin was consistent across sample grain sizes. For each of the three Azorean fragments (total of 24 samples) used in the sample size analysis, gambin outperformed the PLN and logseries distribution for each combination of transect number according to BIC and AIC_c (see Supplementary material Appendix 3, Table A8 for results for all three fragments).

Land use gradient

Considering Terceira Island, gambin's α was found to differ significantly between habitat types, with the highest mean α belonging to the native forest samples, and the lowest mean α belonging to the pasture samples (Fig. 3a). The results were

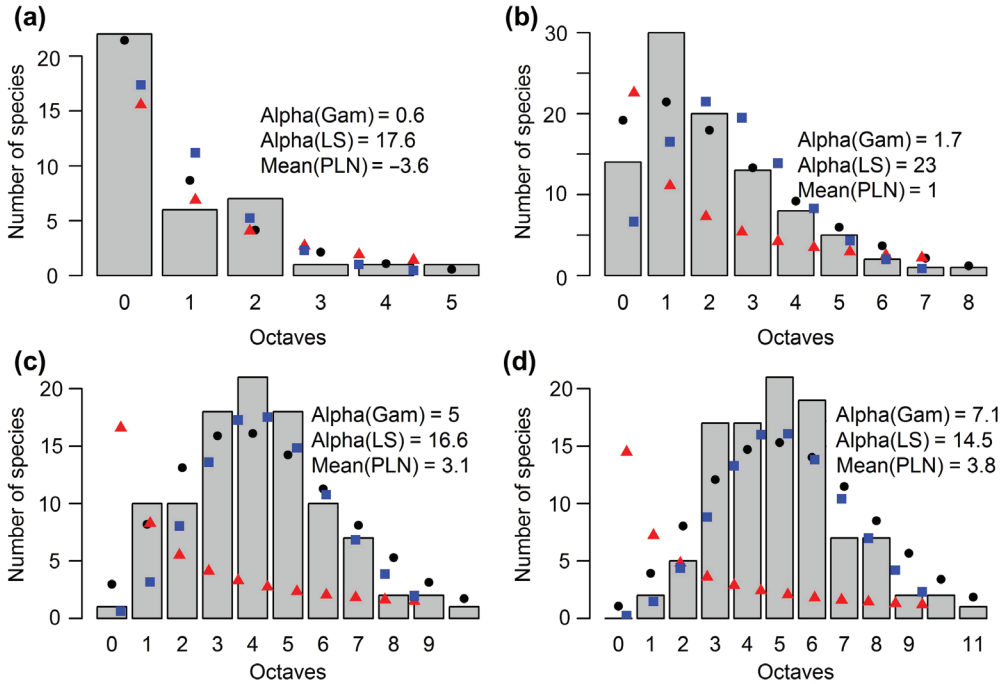

Figure 2. Changing shape of the observed species abundance distribution with sample size. Data were sampled from a simulated lognormal metacommunity (number of individuals (N) = 10 000, number of species = 100). The four plots correspond to different levels of sampling from the metacommunity: (a) 1% of individuals (N = 100) sampled from the metacommunity, (b) 10% of individuals (1000), (c) 50% of individuals (5000), and (d) 100% of individuals, i.e. the full metacommunity (N = 10 000). On each plot the α parameter of the gambin distribution (Gam), the alpha of the logseries distribution (LS), and the mean of the Poisson lognormal distribution (PLN) calculated using the sampled data are given. The fit of the gambin distribution (black dots), PLN (blue squares) and logseries (red triangles) to each sample is also presented.

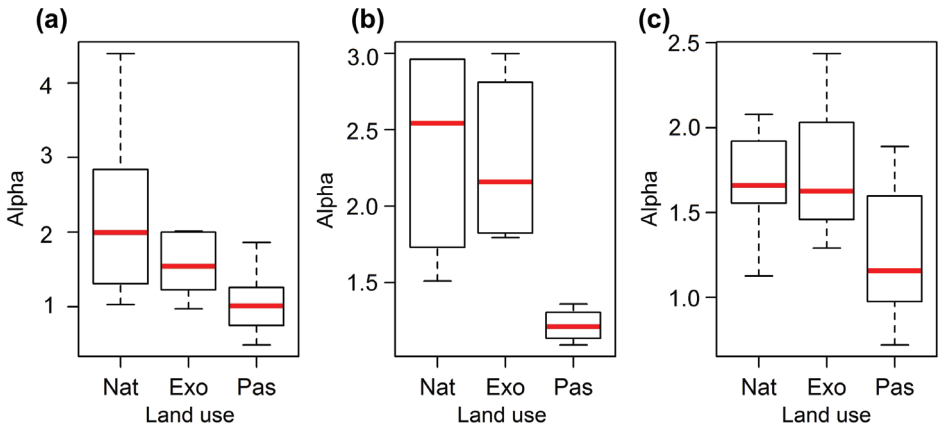

Figure 3. Difference in the gambin distribution α parameter values (standardised to keep sample size constant) between transects of the three land use types (native forest, exotic plantation forest, and pasture) on three Azorean islands: (a) Terceira, (b) Faial, (c) Flores. The number of individuals used to standardise the calculations in each instance is given in Supplementary material Appendix 2, Table A2. The box plots display the median (red line), the first and third quartiles (black box), and the minimum and maximum values (whiskers). The significance of differences between land use types according to Wilcoxon rank sum tests are presented in Table 3.

similar for Faial and Flores, except that for these islands the differences between native forest and exotic plantation forest were not significant at the 0.05 level (Table 3; see also Fig. 3b, c). The alpha of the logseries significantly differentiated between communities in the different land uses (i.e. $p < 0.05$) in two instances (native forest and pasture on Terceira, and native and exotic forest on Flores; Supplementary material Appendix 3, Fig. A2). The sigma of the PLN did not significantly differentiate between any of the communities (Supplementary material Appendix 3, Fig. A2).

Sensitivity to the binning method

The use of log2 to bin the data into octaves appears to be a robust choice. For the 18 Azorean fragments the pairwise comparisons of the alpha values revealed them all to be highly significantly positively correlated ($p < 0.001$ in each instance, Supplementary material Appendix 3, Fig. A3). The comparisons were also highly correlated according to the simulations (Supplementary material Appendix 3, Table A9).

Table 3. Wilcoxon rank sum test results for the significance of difference of α values of the gambin distribution between transects of different land use groups, for three Azorean islands: Faial, Flores and Terceira. For each island the analysis was run using the standardised re-sampled island samples (i.e. all transects were re-sampled in order to keep sample size constant). The number of individuals used to standardise the calculations in each instance is given in Supplementary material Appendix 2, Table A2. Significance was set at the 0.05 level and all significant p values are highlighted in bold. The three land-uses were native Laurisilva forest, exotic plantation forest and pasture. The sample size (i.e. number of transects) for each grouping was as follows. Faial: native forest (8), exotic forest (6), and pastures (8). Flores: native forest (12), exotic forest (4), and pastures (8). Terceira: native forest (33), exotic forest (13), and pastures (35).

Island	Test statistics	Native/exotic	Native/pastures	Exotic/pastures
Faial	W	22	15	1
	p value	0.53	**0.036**	**0.016**
Flores	W	12	32	29
	p value	0.864	**0.034**	**0.028**
Terceira	W	195	1210	80
	p value	**0.05**	**<0.001**	**0.001**

Discussion

We derived a maximum likelihood estimation of the gambin distribution and used it to undertake a comprehensive test of the model. We found that gambin is a flexible model, which provides a superior fit to empirical SADs compared to three popular alternatives. This flexibility, with only a single free parameter, is advantageous as it enables gambin to provide a good and parsimonious fit to a variety of empirical data. Also, the α parameter, which summarises the shape of the model, is of analytical utility as it can potentially be used, for example, as an explanatory variable in regression models exploring how environmental properties influence the SAD. Multi-parameter models, such as the PLN and ZSM, do not have this capability, and while the logseries also possesses a single parameter, the logseries is not a very flexible distribution and often provides a poor fit to empirical data (see e.g. Table 2). The flexibility of gambin is also beneficial because the shape of empirical SADs change as a result of many processes; for instance, as a function of spatial grain size and as the result of ecological disturbance. Gambin is able to track these changes in SAD shape, and as this information is captured in the single parameter, α can be used to compare community SADs analytically.

Goodness of fit and multi-model comparison

Despite only having a single shape parameter (α), gambin provided a better fit than the other models in the majority of instances, including both Azorean and non-Azorean datasets. This included distribution patterns where the modal class was the singleton octave (i.e. logseries form), and patterns where the modal class represented an octave of more common species (i.e. lognormal form). Additionally, the fit

provided a more accurate representation of the data than the other distributions for the various sample grain subsets (Supplementary material Appendix 3, Table A8). It can be highly informative to examine how fit changes with scale. Our analyses show that, relative to the other distributions, both the fit and performance of gambin remained consistently good at spatial scales ranging from a single transect up to fragment- and island-scales of analysis. The different scales represent contrasting points along a continuum of community structure; the good fit of gambin to all subsets illustrates the potential of the model for SAD research.

The gambin distribution is restricted to be unimodal, and thus the poor fit to those SADs that visually exhibit multimodality is unsurprising (e.g. Fig. 1b). Recent work has suggested multimodal SADs may be more prevalent than previously assumed (Dornelas and Connolly 2008, Vergnon et al. 2012, Matthews et al. 2014). However, functions capable of generating multimodal distributions tend to be complex. For instance, the tri-modal Poisson lognormal (PLN3) distribution (Dornelas and Connolly 2008, Matthews et al. 2014) possesses eight parameters compared to gambin's one. Moreover, complex distributions such as the ZSM and PLN3 can be problematic to fit, with an increased chance of locating local maxima when deriving the log-likelihood. While gambin cannot capture multimodality, to the extent that distributions analysed herein are in fact multimodal in character, the gambin model nonetheless provided a better statistical fit than the logseries, PLN, and the ZSM.

Model performance and sample size

That the shape of the SAD is not a constant property of any given community is well known (Magurran 2007). However, the separate issue of sample SADs being poor representations of community SADs when sample size is low is often overlooked (but see Preston 1948, Pielou 1975, Green and Plotkin 2007, McGill 2011, Locey and White 2013). Our simulations indicate that this issue is particularly problematic when the SAD of the community being sampled is lognormal, bringing into question the utility of SADs when sample sizes are low. In particular, it appears that if the community SAD is lognormal, a sample size of 10% of the community's individuals will result in a sample parameter estimate that may differ from the community α estimate by up to 50%. This issue arises because the shape of the sample SAD changes considerably with varying sample size in relation to the community being sampled (Fig. 2), and as outlined above, the flexibility of gambin means the model accurately tracks this change in SAD shape. This small sample problem appears to be general, as it affects the parameters of the PLN and logseries distributions as well as gambin (Supplementary material Appendix 3, Table A7; see also Pielou 1975, McGill 2011). The relative stability in gambin's α for samples from the logseries metacommunities makes intuitive sense, as the modal octave of each sample always corresponded to the singletons' octave. Providing specific recommendations regarding a minimum value of N necessary to obtain accurate estimates of gambin's α is difficult as the accuracy depends in part on the size of the community being sampled, which in empirical systems is often impractical to calculate. However, as an indicative guideline, we agree with McGill (2011), who suggests a minimum N of 1000 individuals.

As gambin's α varies with sample size, any rigorous comparative analysis of α should be based on keeping N constant across samples. A recent study by Locey and White (2013) drew similar conclusions. Focusing on SADs and the feasible set concept, they reported that the specific form of the SAD is constrained by N, and that differences in the form of the SAD between communities may result purely from differences in sample size. Our methodology of comparing gambin α values using repeated re-sampling to a common N (using functionality provided in the gambin R package) provides a way of circumventing the sample size problem, and our comparative analysis of variations in α between land use types (Table 3) indicates that gambin α can have discriminatory power as an ecological indicator when used in this way. We advocate this method more generally in SAD research, as our analyses indicate the parameters of various SAD models (i.e. not just gambin) are related to the number of individuals in the sample and thus using the re-sampling method provides an unbiased way of using model parameters in comparative analysis.

We are also confident that our choice of log2 to bin the data into octaves is appropriate. While using different bases did result in small differences between alpha values, each set of alpha values was highly correlated with every other set (Supplementary material Appendix 3, Fig. A2 and Table A9). Thus, the behaviour of alpha was consistent irrespective of the binning method employed and we are confident our study findings are not dependent on the log base used.

The alpha parameter as an ecological metric

It has long been appreciated that the SAD of undisturbed communities (i.e. those in equilibrium) resemble the lognormal, while disturbed communities more closely follow a logseries SAD (Gray and Mirza 1979, Ugland and Gray 1982). We have shown that the gambin distribution is flexible, meaning it can fit a variety of empirical SAD shapes (including lognormal and logseries-like shapes), and that the distribution shape is adequately characterised by the model's single parameter (α): low values of alpha indicate logseries-like SADs, and high alpha values indicate lognormal-like SADs. Again, the only other distribution with a single parameter which can be used in a similar way is the logseries, and this is much less flexible than gambin and fits far fewer datasets.

Our Azorean analyses support the claim (Ugland et al. 2007) that gambin's α is a useful metric with which to compare communities affected by different types of disturbance or habitat transformation, as they showed that α generally decreased from the relatively untransformed native forest transects, to exotic plantation forest, to the highly transformed pastures (cf. Meijer et al. 2011, Cardoso et al. 2013). These differences were significant in seven out of nine cases, indicating that this metric provides a reasonably sensitive diagnostic tool. Land use change is frequently cited

as the main driver of current biodiversity loss (Fischer and Lindenmayer 2007), and thus any ecological metric which can effectively characterise the impact of land use change on the SAD, and community structure more generally, is a useful addition to the ecologist's statistical toolbox. Furthermore, this finding is of particular interest considering that the logseries and PLN distributions were generally unable to differentiate between the different land uses to the same degree (Supplementary material Appendix 3, Fig. A3).

That α fails to differentiate between Azorean native and exotic forest in Faial and Flores simply indicates that the SADs of these two land uses are similar in form, for these islands. This is an ecologically interesting result, which likely reflects the large number of introduced arthropod species on the Azores. The majority of these introduced species are well adapted to exotic forest habitats in the archipelago, and are often sampled in relatively high abundances (Cardoso et al. 2013). Thus, it is likely that the inclusion of such species obscures the impact of the loss of native Laurisilva forest on community structure, resulting in similar SAD shapes between forest types. However, removing the introduced species from the analysis is not an option in this case as it often results in sample sizes too low to permit accurate estimation of α (above; see also McGill 2011).

The lack of easily accessible methods to fit the gambin distribution hitherto, coupled with the absence of a maximum likelihood derivation, has restricted gambin's dissemination among the ecological community. The new methods and R package presented in this paper alleviate this issue and allow for both easy computation of α, and the incorporation of gambin within information theoretic model comparisons. Characterised by a single parameter that is analytically practical, can be easily interpreted, and provides flexibility, and taking into account the results of our tests, gambin presents a promising tool for future SAD research.

Acknowledgements – We thank Zhe Sha for help in deriving the maximum likelihood function, Alison Pool for help with data entry, and Luís Borda-de-Água for advice which greatly improved the manuscript. We are grateful to all the researchers who collaborated in the field and laboratory work, and to the Azorean Forest Services and Environment Services for providing local support on each island. Azorean data used herein were obtained from the projects funded by Direcção Regional dos Recursos Florestais (Project: 17.01-080203, 1999–2004) and Direcção Regional da Ciência e Tecnologia (Project: 'Consequences of land-use change on Azorean fauna and flora – the 2010 Target', M.2.1.2/I/003/2008). TJM's work in the Azores was funded through a Santander Academic Travel Grant and he also acknowledges funding from the Royal Geographical Society, the Sidney Perry Foundation, the Sir Richard Stapley Trust, and the EPA Cephalosporin Fund. MKB is supported by an individual postdoctoral grant from the Danish Council for Independent Research. FR and PAVB are supported by the project PTDC/BIA-BIC/119255/2010 – 'Biodiversity on oceanic islands: towards a unified theory'.

References

Baker, H. et al. 2006. Population estimates of birds in Great Britain and the United Kingdom. – Br. Birds 99: 25–44.

Borda-de-Água, L. et al. 2012. Spatial scaling of species abundance distributions. – Ecography 35: 549–556.

Borges, P. A. V. et al. 2005. Ranking protected areas in the Azores using standardized sampling of soil epigean arthropods. – Biodivers. Conserv. 14: 2029–2060.

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

Burnham, K. P. and Anderson, D. R. 2002. Model selection and multi-model inference: a practical information-theoretic approach, 2nd ed. – Springer.

Cardoso, P. et al. 2009. A spatial scale assessment of habitat effects on arthropod communities of an oceanic island. – Acta Oecol. 30: 590–597.

Cardoso, P. et al. 2013. Integrating landscape disturbance and indicator species in conservation studies. – PLoS One 8: e63294.

Condit, R. 1998. Tropical forest census plots. – Springer and R. G. Landes.

Connolly, S. R. and Dornelas, M. 2011. Fitting and empirical evaluation of models for species abundance distributions. – In: Magurran, A. E. and McGill, B. J. (eds), Biological diversity: frontiers in measurement and assessment. Oxford Univ. Press, pp. 123–140.

Dornelas, M. and Connolly, S. R. 2008. Multiple modes in a coral species abundance distribution. – Ecol. Lett. 11: 1008–1016.

Dornelas, M. et al. 2011. Biodiversity and disturbance. – In: Magurran, A. E. and McGill, B. J. (eds), Biological diversity: frontiers in measurement and assessment. Oxford Univ. Press, pp. 237–251.

Engen, S. 2001. A dynamic and spatial model with migration generating the log-Gaussian field of population densities. – Math. Biosci. 173: 85–102.

Etienne, R. S. 2005. A new sampling formula for neutral biodiversity. – Ecol. Lett. 8: 253–260.

Fischer, J. and Lindenmayer, D. B. 2007. Landscape modification and habitat fragmentation: a synthesis. – Global Ecol. Biogeogr. 16: 265–280.

Fisher, R. A. et al. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. – J. Anim. Ecol. 12: 42–58.

FRIM 2013. Pasoh species abundance. – Forest Research Inst. Malaysia, <www.ctfs.si.edu/site/Pasoh/abundance>.

Gaspar, C. et al. 2008. Diversity and distribution of arthropods in native forests of the Azores archipelago. – Arquipél. Life Mar. Sci. 25: 1–30.

Gray, J. S. and Mirza, F. B. 1979. A possible method for the detection of pollution-induced disturbance on marine benthic communities. – Mar. Pollut. Bull. 10: 142–146.

Gray, J. S. et al. 2006. On plotting species abundance distributions. – J. Anim. Ecol. 75: 752–756.

Green, J. L. and Plotkin, J. B. 2007. A statistical theory for sampling species abundances. – Ecol. Lett. 10: 1037–1045.

Grøtan, V. and Engen, S. 2009. poilog: Poisson lognormal and bivariate Poisson lognormal distribution. – R package ver. 0.4.

Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography. – Princeton Univ. Press.

Hubbell, S. P. et al. 2005. Barro Colorado Forest census plot data. – <https://ctfs.arnarb.harvard.edu/webatlas/datasets/bci>, accessed 3 April 2013.

Lambshead, P. J. D. 1986. Sub-catastrophic sewage and industrial waste contamination as revealed by marine nematode faunal analysis. – Mar. Ecol. Prog. Ser. 29: 247–260.

Lambshead, P. J. D. and Boucher, G. 2003. Marine nematode deep-sea biodiversity – hyperdiverse or hype? – J. Biogeogr. 30: 475–485.

Locey, K. J. and White, E. P. 2013. How species richness and total abundance constrain the distribution of abundance. – Ecol. Lett. 16: 1177–1185.

Mac Nally, R. 2007. Use of the abundance spectrum and relative abundance distributions to analyze assemblage change in massively altered landscapes. – Am. Nat. 170: 319–330.
Magurran, A. E. 2007. Species abundance distributions over time. – Ecol. Lett. 10: 347–354.
Magurran, A. E. and Henderson, P. A. 2003. Explaining the excess of rare species in natural species abundance distributions. – Nature 422: 714–716.
Matthews, T. J. and Whittaker, R. J. 2014. Neutral theory and the species abundance distribution: recent developments and prospects for unifying niche and neutral perspectives. – Ecol. Evol. doi: 10.1002/ece3.1092
Matthews, T. J. et al. 2014. Multimodal species abundance distributions: a deconstruction approach reveals the process behind the pattern. – Oikos doi: 10.1111/j.1600-0706.2013.00829.x
McGill, B. 2003. Strong and weak tests of macroecological theory. – Oikos 102: 679–685.
McGill, B. J. 2011. Species abundance distributions. – In: Magurran, A. E. and McGill, B. J. (eds), Biological diversity: frontiers in measurement and assessment. Oxford Univ. Press, pp. 105–122.
McGill, B. J. et al. 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. – Ecol. Lett. 10: 995–1015.
Meijer, S. S. et al. 2011. The effects of land-use change on arthropod richness and abundance on Santa Maria Island (Azores): unmanaged plantations favour endemic beetles. – J. Insect Conserv. 15: 505–522.
Pielou, E. C. 1975. Ecological diversity. – Wiley.
Preston, F. W. 1948. The commonness, and rarity, of species. – Ecology 29: 254–283.
Ribeiro, S. P. et al. 2005. Canopy insect herbivores in the Azorean Laurisilva forests: key host plant species in a highly generalist insect community. – Ecography 28: 315–330.
Sæther, B.-E. et al. 2013. Species diversity and community similarity in fluctuating environments: parametric approaches using species abundance distributions. – J. Anim. Ecol. 82: 721–738.
Šizling, A. et al. 2009. Invariance in species abundance distributions. – Theor. Ecol. 2: 89–103.
Ugland, K. I. and Gray, J. S. 1982. Lognormal distributions and the concept of community equilibrium. – Oikos 39: 171–178.
Ugland, K. I. et al. 2007. Modelling dimensionality in species abundance distributions: description and evaluation of the Gambin model. – Evol. Ecol. Res. 9: 313–324.
Ulrich, W. et al. 2010. A meta-analysis of species abundance distributions. – Oikos 119: 1149–1155.
Vergnon, R. et al. 2012. Emergent neutrality leads to multimodal species abundance distributions. – Nat. Commun. 3: 663.
White, E. P. et al. 2012. Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. – Ecology 93: 1772–1778.
Williams, C. B. 1964. Patterns in the balance of nature. – Academic Press.
Williamson, M. and Gaston, K. J. 2005. The lognormal distribution is not an appropriate null hypothesis for the species abundance distribution. – J. Anim. Ecol. 74: 409–422.

Supplementary material (Appendix ECOG-00861 at <www.ecography.org/readers/appendix>). Appendix 1–3.

Correction added on 10 June 2014, after first online publication: Following publication, two issues with our analyses were brought to our attention: 1) we were inadvertently fitting a truncated form of the Poisson lognormal distribution (PLN), and 2) there was a problem with how we derived the predicted values from the PLN and logseries models. Correcting these issues does not change our qualitative results. Corrected versions of Table 2, Fig. 1 and Fig. 2 are provided in Appendix 3.