Methods in Ecology and Evolution

DR THOMAS J. MATTHEWS (Orcid ID : 0000-0002-7624-244X)
DR MICHAEL KRABBE BORREGAARD (Orcid ID : 0000-0002-8146-8435)

Article type : Application

Handling editor: Dr Timothée Poisot

Submission to: Methods in Ecology & Evolution

Article Type: APPLICATION

Extension of the gambin model to multimodal species abundance distributions

Thomas J. Matthews[1, 2, 3], Michael K. Borregaard[4], Colin S. Gillespie[5], François Rigal[6], Karl I. Ugland[7], Rodrigo Ferreira Krüger[8,9], Roberta Marques[9], Jon P. Sadler[1], Paulo A.V. Borges[2], Yasuhiro Kubota[10, 11], Robert J. Whittaker[4,8]

[1]GEES (School of Geography, Earth and Environmental Sciences), The University of Birmingham, Birmingham, B15 2TT

[2]CE3C – Centre for Ecology, Evolution and Environmental Changes/Azorean Biodiversity Group and Univ. dos Açores – Depto de Ciências e Engenharia do Ambiente, PT-9700-042, Angra do Heroísmo, Açores, Portugal.

[3]Birmingham Institute of Forest Research, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK

[4]Center for Macroecology, Evolution and Climate, Natural History Museum of Denmark, University of Copenhagen, Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark

[5]School of Mathematics & Statistics, Newcastle University, Newcastle upon Tyne, NE1 7RU UK

[6]CNRS-Université de Pau et des Pays de l'Adour, Institut des Sciences Analytiques et de Physico-Chimie pour l'Environnement et les Materiaux, MIRA, Environment and Microbiology Team, UMR 5254, BP 1155, 64013 Pau Cedex, France

[7]Department of Marine Biology, Institute of Biosciences, University of Oslo, P.O. Box 1066, Blindern, 0316 Oslo, Norway

[8]Conservation Biogeography and Macroecology Group, School of Geography and the Environment, University of Oxford, South Parks Road, Oxford, OX1 3QY, UK

[9]Parasite and Vector Ecology Group, Depto de Microbiologia e Parasitologia, Campus Universitário Capão do Leão, s/nº CEP 96010-900, Pelotas, Rio Grande do Sul, Brasil

[10]Faculty of Science, University of the Ryukyus, Nishihara, Okinawa, Japan

[11]Marine and Terrestrial Field Ecology, Tropical Biosphere Research Center, University of the Ryukyus, Nishihara, Okinawa 903-0213, Japan.

*Correspondence: Thomas J. Matthews, School of Geography, Earth and Environmental Sciences, University of Birmingham, Birmingham, B15 2TT, UK

Email: txm676@gmail.com

Running header: Multimodal gambin distributions

Word count: abstract = 231 words; main text = 3149 words; 0 Tables; 2 Figures; 4 appendices

Keywords

Compound distributions, gambin, horse flies, multimodal species abundance distributions

ABSTRACT

1. Species abundance distributions (SADs) are one of the most widely used tools in macroecology, and it has become increasingly apparent that many empirical SADs can best be described as multimodal. However, only a few SAD models have been extended to incorporate multiple modes and no software packages are available to fit multimodal SAD models. In this study, we present an extension of the gambin SAD model to multimodal SADs.

2. We derive the maximum likelihood equations for fitting the bimodal gambin distribution and generalise this approach to fit gambin models with any number of modes. We present these new functions, along with additional functions to aid in the analysis of multimodal SADs, within an updated R package ('gambin'; version 2.4.0) that enables the fitting, plotting and evaluating of gambin models with any number of modes.

3. We use a mixture of simulations and empirical datasets to test our new models, including tests of the sensitivity of the model parameters to the number of individuals and the number of species in a sample. We show that the new multimodal gambin models perform well under a variety of circumstances, and that the application of these new models to empirical SAD and other macroecological (e.g. species range size distributions) datasets can provide interesting insights. The updated software package is simple to use and provides straightforward yet flexible statistical analyses of multimodality in SAD-type datasets.

INTRODUCTION

The species abundance distribution (SAD) has been a core focus of macroecology for over eighty years (e.g. Fisher, Corbet & Williams 1943), and is currently the subject of widespread renewed interest (McGill et al., 2007; Alonso, Etienne & Ostling 2008; Arellano et al., 2017). Recently, it has been argued that a gamma-binomial (herein 'gambin') distribution represents a useful SAD model (Ugland et al., 2007). Gambin is a stochastic unimodal model that combines the gamma distribution, in which the scale parameter is fixed at 1, with a binomial sampling method (see Ugland et al., 2007 for a full description of the model). The use of the gamma distribution as the basis of the model provides gambin with substantial flexibility and tests of the gambin model have found that it generally provides a good fit to a wide range of empirical SAD data, typically out-performing other candidate SAD models (Ugland et al., 2007; Matthews et al., 2014), such as the Poisson lognormal (PLN; Bulmer, 1974) and logseries models (Fisher et al., 1943). The model can also be used with continuous data, and thus extend the analysis of SADs to different measures of abundances (e.g. biomass). The unimodal gambin model has a free parameter (α) that determines the shape of the distribution. Low values of α indicate logseries curve shapes, whilst higher α values indicate more lognormal-like curve shapes. Thus, α is an intuitive parameter that has been found to be of use in comparing the SAD of different ecological communities, e.g. disturbed and undisturbed communities, and for testing what variables drive changes in the shape of the SAD along ecological gradients (Dornelas, Soykan & Ugland 2011; Matthews & Whittaker, 2015; Arellano et al., 2017; Matthews, Borges, de Azevedo & Whittaker 2017). Due to the way the statistical model is defined, gambin can only be fitted to data binned into octaves e.g. classes of log2 transformed abundance data, with octave 0 containing the number of species with 1 individual, octave 1 the number of species with 2 or 3 individuals, and so on.

It has become increasingly apparent that many empirical SADs can best be described as multimodal (Dornelas & Connolly, 2008; Vergnon, van Nes & Scheffer 2012; Antão, Connolly, Magurran, Soares & Dornelas 2017). For example, Antão et al. (2017) found that between 15% and 22% of the 117 empirical SAD datasets they evaluated showed evidence of multimodality, depending on the model selection tools used. Multimodality may be indicative of particular process regimes (Matthews, Whittaker & Borges 2014) or be due to a combination of different types of species (e.g. trophic groups) in a sample, and its detection may also be relevant to, for example, tests of the theory of emergent neutrality (Vergnon et al., 2012). Hence, describing and testing for multimodality is a priority in SAD research (Antão et al., 2017). To date, few SAD models have been extended to incorporate multiple modes (for the PLN see Dornelas & Connolly, 2008), in part because compound probability distribution models are mathematically and computationally complex. Hence the need for an easy-to-use software package permitting straightforward statistical analysis of multimodality in SAD datasets. We set out to provide a multimodal extension of gambin because the gambin model is relatively simple and it would allow comparison of the fit of unimodal and multimodal models analytically using standard statistical methods.

First, we derive the maximum likelihood equations for fitting gambin models with any number (g) of components and incorporate these new functions, along with additional functions to aid in the analysis of multimodal SADs, within an updated version of the R package gambin (version 2.4.0). Second, we use a mixture of simulations and empirical datasets to test the new models, providing examples of the updated package in operation.

MULTIMODAL GAMBIN DISTRIBUTIONS AND THE GAMBIN R PACKAGE (VERSION 2.4.0)

The full derivation of the likelihood functions for multimodal gambin models is provided in Appendix S1 (Supplementary Information). In version 2.4.0 of the gambin R package, the one-component gambin model is taken to have two parameters: the shape parameter (α) and the max octave. It should be noted that this differs from previous implementations of the model (e.g. Matthews et al., 2014) that only considered there to be a single parameter (α). The two-component gambin model is simply the mixture of two gambin distributions. To allow for the subdivision of all of the observed objects (*species* in the context of SADs) (y_{obs}), a parameter (w_1) is needed that describes the fraction of objects belonging to the first distribution (w_i is analogous to the ρ parameter in the multimodal PLN context). The fraction of objects belonging to the second component (w_2) is 1 - w_1. Thus, the expected number of observed objects is split into two components, consisting of $w_1 * y_{obs}$ and $w_2 * y_{obs}$ objects, respectively. Thus, $y_{obs} = (w_1 * y_{obs}) + (w_2 * y_{obs})$. With no extra information, we may therefore assume that the number of objects in the *k*-th interval (k = 1, 2, ..., *i*) are $w_1 * y_k$ and $w_2 * y_k$. Thus, the likelihood function for a bimodal gambin model contains five parameters: the shape parameters for the first and second component (α_1 & α_2), the max octaves for the first (n_{oct1}) and second (n_{oct2}) components, and one splitting parameter (w_1) representing the fraction of objects in the first component. Note that this is the same number of parameters in the bimodal PLN model; it is simply that the parameters represent different aspects of the distribution in each case. It is relatively straightforward to extend the above approach for fitting the two-component gambin model by maximum likelihood, to fitting gambin models with *g* components (where components correspond to the number of modes; see Appendix S1). However, whilst it is possible to use the equations given in Appendix S1 to fit gambin distributions with any number of components, in practice fitting SAD models with more than three (possibly even two depending on sample size) components will likely result in overfitting the data. Sample sizes in ecological studies are generally relatively small, and the number of parameters becomes large with increasing *g* (Dornelas & Connolly, 2008). Thus, optimising the likelihood functions becomes increasingly problematic at larger *g*; ecological interpretation of model fits with large numbers of components is also problematic. Accordingly, we do not advise fitting gambin models with more than three components.

In addition to providing functions to fit multimodal gambin distributions (described below), the gambin R package (version 2.4.0; available on CRAN) has been updated to bring it more in line with other distribution functions within the R base 'stats' package. For example, the updated gambin package now provides *dgambin* (probability density function), *rgambin* (generate random values from a gambin distribution; the returned values relate to a given octave), *qgambin* (quantile function) and *pgambin* (cumulative distribution function) functions. Likelihood optimisation is undertaken using the Nelder–Mead algorithm. As the likelihood optimisation procedure for multimodal gambin models can be time consuming, the updated package provides the option of using parallel processing to speed up optimisation. The gambin R package documentation and associated vignette provide additional information.

The main function within the package is 'fit_abundances':

```
#this fits a gambin distribution with g modes to a vector of abundances,
#with the option of subsampling z individuals. If g is set to 1, the
#standard unimodal gambin distribution is fitted, g = 2 fits the bimodal
#gambin distribution, and so on. When the no_of_components argument is
#greater than 1, the 'cores' argument can be used to enable parallel
#processing using d cores.

fit_abundances(data, subsample = z, no_of_components = g, cores = d)
```

A primary argument for the prevalence of multimodal SADs in nature is the idea that the different modes represent different categories of species (e.g. native and invasive species, or core and satellite species; Magurran & Henderson 2003; Matthews & Whittaker 2015). A natural next step then is to deconstruct the SAD by visualizing and analysing how different categories of species are distributed across the various modes / modal octaves. This is performed with the new function 'deconstruct_modes'. If species category information is provided (e.g. native or invasive), the function returns the number and proportion of the various categories in the different modal octaves (a split barplot where the bar for each octave is split according to the number of species in each category can also be returned). Subsequent statistical test (e.g. χ^2 or G-test) and/or null model tests can then be undertaken to determine whether the number of species representing the different categories significantly differs between octaves. If species category information is not available, the function will simply identify the modal octaves (i.e. the modal octave of each component distribution) in a multimodal gambin model fit (user-specified modal octaves can instead be provided), and also lists the names of the species within each octave (a plot of the model fit with the modal octaves highlighted can also be returned).

```
#Fit the bimodal gambin model to SAD data

fit <- fit_abundances(data, no_of_components = 2)

#Deconstruct the model fit and calculate the number of species of
#different categories (categ) in each of the modal octaves (peak_val is
#set to 'NULL' and thus the modal octaves are calculated from the model
#fit). Return a plot of the model fit with the modal octaves highlighted
#(plot_modes = TRUE) and run the null model bootstrap sampling with 100 (n
#= 100) random draws.

deconstruct_modes(fit, dat = data, peak_val = NULL, categ = "status",
                  plot_modes = TRUE, n = 100)
```

One of the main applications of the gambin model has been to fit gambin to SADs from different sites (e.g. along a disturbance gradient) and then to compare the resultant alpha values (e.g. Dornelas, Soykan & Ugland 2011; Arellano et al., 2017). Thus, we have also added a function that fits the unimodal gambin model to the SADs from multiple sites and returns the standardised and unstandardised alpha values.

```
#Fit the unimodal gambin model to the SADs from multiple sites (mult) and
#return the standardised (based on N subsamples of size 'subsample'; NULL
# = the number of individuals in the site with the fewest individuals) and
#unstandardised alpha values

mult_abundances(mult, N = 100, subsample = NULL)
```

EXAMPLES USING EMPIRICAL DATASETS

A Brazilian horse fly dataset

To illustrate the new functionality, we used an empirical dataset comprising abundance records of horse flies (Diptera, Tabanidae) from a variety of sampling locations in Brazil (see Appendix S2 in the Supplementary Information). As outlined above, multimodal SADs may hypothetically arise from the intersection in nature of samples from different habitat types or of different ecological species groups (Magurran & Henderson, 2003; Antão et al., 2017) within a dataset. To test this proposition, we first fitted the unimodal, bimodal and trimodal versions of gambin to the whole Brazilian dataset. We then took a subset of the dataset relating to one individual locality within Brazil and one type of sampling (see Appendix S2) and again fitted the three models. In both cases the three models were compared using the Bayesian information criterion (BIC):

```
#load the fly datasets

data(fly)

Brazil <- fly[[1]]#select the data for all of Brazil

Site <- fly[[2]]#select the data for a single site within Brazil

#Fit the multimodal gambin models to a given dataset (Brazil or site)

res1 <- lapply(c(1, 2, 3), fit_abundances, abundances = Brazil, subsample
= 0, cores = 3)

#calculate and compare the BIC value of the fitted models

vapply(res1, BIC, FUN.VALUE = numeric(1))

#plot the empirical SADs

barplot.gambin(res1[[1]])
points.gambin (res1[[1]], pch = 17, col = "black") #add the fitted values
points.gambin (res1[[2]], pch = 16, col = "blue")
points.gambin (res1[[3]], pch = 18, col = "green")
```

When the three models were fitted to the whole Brazilian horse fly dataset (Fig. 1a), the bimodal gambin model provided the best fit to the data (BIC = 830.4), followed by the unimodal model (BIC = 832.5) and the trimodal model (BIC = 837.6). When the three models were fitted to the subset of data from a single site (Fig. 1b), the unimodal model provided the best fit (BIC = 236.2), followed by the bimodal model (BIC = 239.9) and the trimodal model (BIC = 246.5). Thus, whilst the data from a single site are characterised by a classical unimodal SAD, when pooling records from different localities across Brazil, the bimodal model was favoured. These findings provide additional support for the claim that multimodal SADs are more prevalent with increasing taxonomic breadth, sampling variation, spatial extent (i.e. increasing ecological heterogeneity; Antão et al., 2017), and heterogeneity in species detectability (Alonso et al. 2008).

A set of 275 woody plant SADs

We took the set of 843 angiosperm woody plant datasets sourced from the literature by Kubota et al. (2018). Each dataset represents an abundance vector of plant species sampled in a forest plot and the datasets have a global distribution. We filtered out datasets with <10 species and <500 individuals. We then fitted the unimodal and bimodal gambin models to the resulting 275 datasets and compared the fits using BIC. The bimodal model was considered as the best fitting model if it had the lowest BIC value and the unimodal model had a ΔBIC value of >2.0 (a lower value indicates the models have similar support, in which case the unimodal model should be preferred on grounds of parsimony).

The bimodal model provided the best fit to 51 of the 275 datasets (19%; see Appendix S3 for the full model comparison results).

Application to other macroecological phenomena

Whilst gambin models have so far only been used to analyse SADs, it is possible to fit them to any other type of ecological or general distribution. For example, there is evidence that some species-range size distributions may exhibit multimodality (e.g. see Gaston, 2003, p. 80). As an illustration, we fitted a selection of gambin models to the global range size distribution of 167 marine mammal species, and the occupancy distribution of intestinal helminths in three species of grebe; we observe evidence of multimodality in both distributions. The full methods and resultant model fits are provided in Appendix S3.

SIMULATION ANALYSES

The results of our simulations indicated that in general the α parameter estimates of the bimodal gambin model were relatively insensitive to the number of species in the sample (Figure S2, Appendix S4).

In contrast, it was found that the α parameter estimates of the bimodal gambin model were sensitive to the number of individuals in a sample (Figs S3 and S4, Appendix S4). The latter is true of most SAD models (see McGill, 2011) and is worrying given that SAD analyses typically involve small datasets. While this sensitivity is problematic for unimodal gambin, it is less of an issue for applications of the bimodal model. With the unimodal gambin model, the α value can be used as a type of diversity metric to compare SAD shape across communities (e.g. Arellano et al., 2017). However, for multimodal gambin models the meaning of the α values is not as clear, and as such, when fitting multimodal gambin models we do not advise using the α parameter estimates as diversity metrics or as response variables in regression-type comparative analyses. Rather, the benefit of multimodal gambin models is to provide a simple, quick and easy to use test for determining whether empirical SADs are multimodal, and to provide a basis for subsequent deconstruction analysis to examine the identities of species within the octaves.

To test the error rate of our models, we simulated unimodal and bimodal gambin SADs using multiple simulations varying the numbers of individuals and species (Appendix S4), fitting both unimodal and bimodal gambin models to the simulated data. We compared models using BIC and calculated the proportion of times that a bimodal model provides a better fit than a unimodal model to a unimodal dataset (i.e. false positive) and the proportion of times a unimodal model provides a better fit than a bimodal model to a multimodal dataset (i.e. false negative). When a unimodal gambin distribution was simulated, the error rate (false positive) was roughly 7.0% (see Appendix S4). When a bimodal gambin distribution was simulated, the mean error rate depended on the sample size and the difference between the $\alpha 1$ and $\alpha 2$ values in the simulated data (Fig. 2). When the difference between $\alpha 1$ and $\alpha 2$ was relatively large, the error rate was very low (e.g. 0%) regardless of the number of species in the sample. In contrast, when the difference between the $\alpha 1$ and $\alpha 2$ values was very small, the error rate was high (e.g. 81%) regardless of the number of species. The fact that the error rate increases as the components become closer together (Fig. 2) is to be expected, as the underlying sample distribution starts to resemble a unimodal distribution. As most empirical multimodal SADs have distinct rarer and more common species modes, this is not a substantive issue. The approach can be considered conservative in that the model comparison test is slightly biased towards selecting the unimodal model over the multimodal model.

A full outline of the methodology, results and discussion for each of the simulations, along with a more detailed discussion, is provided in Appendix S4 in the Supplementary Information. All analyses were undertaken in R (version 3.4.3; R Core Team, 2017).

CONCLUDING REMARKS

In this paper, we have derived the maximum likelihood equations for gambin models with multiple components and integrated these functions into an updated version of the 'gambin' R package available on CRAN. Due to the relatively simple underlying mathematics and binning procedure, the models are easy to fit and the maximum likelihood estimation procedure does not require the user to vary the starting parameter values or the optimisation algorithm employed. Hence, multimodal gambin models represent a novel, easily applied test for determining whether SADs or certain other macroecological datasets exhibit multimodality. We have also provided a number of additional functions to aid in the analysis of multimodal SADs.

As Antão et al. (2017, p. 203) state, "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." The development of multimodal gambin models provides one tool to undertake these types of analyses. Application of these new models to additional datasets will likely be revealing and will help in improving our understanding of multimodality in SADs and possibly in other macroecological data forms.

ACKNOWLEDGEMENTS

Pedro Cardoso and three anonymous reviewers kindly provided comments on an earlier version of the manuscript. RFK was supported by grants from CNPq (Process numbers 202236/2015-3 and 308908/2016-3).

AUTHORS' CONTRIBUTIONS

T.J.M designed the study and led the drafting with input from R.J.W. T.J.M designed and ran the analyses with the help of M.K.B and F.R. K.I.U, C.S.G and T.J.M derived the likelihood functions, and C.S.G and T.J.M built the new version of the R package. R.F.K, R.M and Y.K contributed datasets. All authors contributed to the final manuscript.

DATA ACCESSIBILITY

gambin is freely available from CRAN (https://CRAN.R-project.org/package=gambin), whilst the development version is hosted on GitHub (https://github.com/txm676/gambin), where feature requests and bug reports can be posted. The Brazilian fly data are freely available with the gambin R package.

REFERENCES

Alonso, D., Ostling, A. & Etienne, R.S. (2008) The implicit assumption of symmetry and the species abundance distribution. *Ecology Letters,* **11**, 93-105.

Antão, L.H., Connolly, S.R., Magurran, A.E., Soares, A. & Dornelas, M. (2017) Prevalence of multimodal species abundance distributions is linked to spatial and taxonomic breadth. *Global Ecology and Biogeography,* **26,** 203-215.

Arellano, G., Umaña, M.N., Macía, M.J., Loza, M.I., Fuentes, A., Cala, V. & Jørgensen, P.M. (2017) The role of niche overlap, environmental heterogeneity, landscape roughness and productivity in shaping species abundance distributions along the Amazon–Andes gradient. *Global Ecology and Biogeography,* **26,** 191-202.

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

Burnham, K.P. & Anderson, D.R. (2002) *Model selection and multi-model inference: a practical information-theoretic approach,* 2nd edn. Springer, New-York.

Dornelas, M. & Connolly, S.R. (2008) Multiple modes in a coral species abundance distribution. *Ecology Letters,* **11,** 1008-1016.

Dornelas, M., Soykan, C.U. & Ugland, K.I. (2011) Biodiversity and disturbance. *Biological diversity: frontiers in measurement and assessment* (eds A.E. Magurran & B.J. McGill), pp. 237-251. Oxford University Press, Oxford.

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

Gaston, K.J. (2003) *The structure and dynamics of geographic ranges*. Oxford University Press, Oxford.

Kubota, Y., Kusumoto, B., Shiono, T. & Ulrich, W. (2018) Environmental filters shaping angiosperm trees assembly along climatic and geographical gradients. *Journal of Vegetation Science*. doi:10.1111/jvs.12648

Magurran, A.E. & Henderson, P.A. (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature,* **422,** 714-716.

Matthews, T.J., Borges, P.A.V., de Azevedo, E.B. & Whittaker, R.J. (2017) A biogeographical perspective on species abundance distributions: recent advances and opportunities for future research. *Journal of Biogeography,* **44,** 1705–1710.

Matthews, T.J., Borregaard, M.K., Ugland, K.I., Borges, P.A.V., Rigal, F., Cardoso, P. & Whittaker, R.J. (2014) The gambin model provides a superior fit to species abundance distributions with a single free parameter: evidence, implementation and interpretation. *Ecography,* **37,** 1002-1011.

Matthews, T.J. & Whittaker, R.J. (2015) On the species abundance distribution in applied ecology and biodiversity management. *Journal of Applied Ecology,* **52,** 443-454.

McGill, B.J. (2011) Species abundance distributions. *Biological diversity: frontiers in measurement and assessment* (eds A.E. Magurran & B.J. McGill), pp. 105-122. Oxford University Press, Oxford.

McGill, B.J., Etienne, R.S., Gray, J.S., Alonso, D., Anderson, M.J., Benecha, H.K., … White, E.P. (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters,* **10,** 995-1015.

R Core Team (2017) R: A language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.

Ugland, K.I., Lambshead, P.J.D., McGill, B., Gray, J.S., O'Dea, N., Ladle, R.J. & Whittaker, R.J. (2007) Modelling dimensionality in species abundance distributions: description and evaluation of the Gambin model. *Evolutionary Ecology Research,* **9,** 313-324.

Vergnon, R., van Nes, E.H. & Scheffer, M. (2012) Emergent neutrality leads to multimodal species abundance distributions. *Nature Communications,* **3,** 663.

SUPPORTING INFORMATION

Please see the online Supporting information tab for this article.

FIGURES

FIGURE 1 The fit of the unimodal (blue circles), bimodal (red triangles) and trimodal (black diamonds) gambin models to two horse fly species abundance distribution datasets (black bars) from Brazil. (a) horse fly data from 33 localities across Brazil (number of unique species = 164; total number of individuals = 78,755), and (b) data from one individual locality and one type of sampling (number of unique species = 58; total number of individuals = 1943; see Appendix S3). In (a) the bimodal model provides the best fit according to BIC, whilst the unimodal model provided the best to (b).

FIGURE 2 The multimodal SAD error rate (expressed as a percentage) for an information theoretic model comparison test. For the test, a bimodal SAD was simulated, with one α parameter fixed at 0.5 and the second ($\alpha 2$) set to vary between 2 and 10 in units of 1. The number of species (sample size) was set to: 50, 100, 200, 500. The unimodal and bimodal gambin models were then fitted to this simulated SAD and the best model fit determined using BIC. The error rate percentage relates to the proportion of times the unimodal model provided a better fit than the bimodal model (i.e. a higher error rate percentage indicates that the unimodal model erroneously provided a better fit to the bimodal SAD).