Chapter 1 - An extensive comparison of species-abundance distribution models

Introduction

The species abundance distribution (SAD) describes the full distribution of commonness and rarity in ecological systems. It is one of the most fundamental and ubiquitous patterns in ecology, and exhibits a consistent general form with many rare species and few abundant species occurring within a community. This general shape is often referred to as a hollow curve distribution.

The SAD is one of the most widely studied patterns in ecology, leading to a proliferation of models that attempt to characterize the shape of the distribution and identify potential mechanisms for the pattern (see (McGill et al. 2007) for a recent review of SADs). These models range from arbitrary distributions that are chosen based on providing a good fit to the data (Fisher et al. 1943), to distributions chosen based on combinatorics and the most likely state of the system, (Frank 2011, Harte 2011, Locey and White 2013), to models based on ecological process (Tokeshi 1993, Hubbell 2001, Volkov et al. 2003).

Which model or models provide the best fit to the data, and the resulting implications for the processes structuring ecological systems, has been an active area of research (e.g., (McGill 2003, Volkov et al. 2003, Ulrich et al. 2010, White et al. 2012, Connolly et al. 2014)). However, most comparisons of the different models: 1) use only a small subset of available models (typically two; e.g., (McGill 2003, Volkov et al. 2003, White et al. 2012, Connolly et al. 2014)); 2) focus on a single ecosystem or taxonomic group (e.g., (McGill 2003, Volkov et al. 2003)); or 3) fail to use the most appropriate statistical methods (e.g., (Ulrich et al. 2010)). This makes it difficult to draw general conclusions about which, if any, models provide the best empirical fit to species abundance distributions.

Here, we evaluate the performance of five of the most widely used models for the species abundance

distribution. We evaluate their performance using likelihood based model selection on data from 16,218 communities, from nine taxonomic groups. This includes data from terrestrial, aquatic, and marine ecosystems representing roughly 50 million individual organisms in total.

Methods

Data

We compiled data from citizen science projects, government surveys, and literature mining to produce a dataset with 16,218 communities, from nine taxonomic groups, representing nearly 50 million individual terrestrial, aquatic, and marine organisms. Data for trees, birds, butterflies and mammals was compiled by White et al. 2012 from six data sources: the US Forest Service Forest Inventory and Analysis (FIA; (Service 2010)), the North American Butterfly Associations North American Butterfly Count (NABC; American Butterfly Association] (2009)), the Mammal Community Database (MCDB; Thibault et al. 2011), Alwyn Gentry's Forest Transect Data Set (Gentry; Phillips and Miller (2002)), the Audubon Society Christmas Bird Count (CBC; Society (2002)), and the US Geological Survey's North American Breeding Bird Survey (BBS; Pardieck et al. (2014)). The publicly available datasets (FIA, MCDB, Gentry, and BBS) were aquired using the EcoData Retriever (Morris and White 2013). Details of the treatment of these datasets can be found in Appendix A of White et al. (White et al. 2012). Data on Actinopterygii, Reptilia, Coleoptera, Arachnida, and Amphibia, were mined from literature by Baldridge (see details in Chapter 2 of this dissertation).

Table 1: Details of datasets used to evaluate the form of the species-abundance distribution.

Dataset	Dataset code	Availability
North American Breeding Bird Survey	BBS	Publicly available
Christmas Bird Count	CBC	Data request; Memorandum of Understand

Dataset	Dataset code	Availability
Alwyn Gentry's Forest Transects	Gentry	Publicly available
Forest Inventory Analysis	FIA	Publicly available
Mammal Community Database	MCDB	Publicly available
North American Butterfly Count	NABA	Data request with Memorandum of Unders
Actinopterygii; Miscellaneous abundance database	Actinopterygii	Publicly available
Reptilia; Miscellaneous abundance database	Reptilia	Publicly available
Amphibia; Miscellaneous abundance database	Amphibia	Publicly available
Coleoptera; Miscellaneous abundance database	Coleoptera	Publicly available
Arachnida; Miscellaneous abundance database	Arachnida	Publicly available

All abundances in the compiled datasets where counts of individuals.

Models

The majority of species-abundance distributions (SADs) are constructed using counts of individuals (for discussion of alternative approaches see (McGill et al. 2007)). As such, the data are discrete and therefore the most appropriate models are discrete distributions. Therefore we used only abundance data based on individual counts and used only discrete distributions that have been used as models for SADs.

(McGill et al. 2007) classified models into five different families: purely statistical, branching process, population dynamics, niche partitioning, and spatial distribution of individuals. We evaluated models from each of the separate families, excluding the spatial distribution family, which requires spatially explicit data. Specifically, we evaluated the log-series, the Poisson log-normal, the

negative binomial, the geometric series, and the Zipf distributions (Table 2). All distributions were defined to have support defined by the positive integers (i.e., they are capable of having non-zero probability at values from 1 to infinity). We excluded models from analysis that do not have explicit likelihoods (e.g., some niche partitioning models; (Sugihara 1980, Tokeshi 1993)) so that we could use the likelihood based methods for fitting and evaluating distributions (see Analysis).

The log-series is one of the first distributions used to describe the SAD, being derived as a purely statistical distribution by Fisher (Fisher et al. 1943). It has since been derived as the result of both ecological processes, the metacommunity SAD for ecological neutral theory (Hubbell 2001, Volkov et al. 2003), and several different maximum entropy models (Pueyo et al. 2007, Harte et al. 2008).

The lognormal is one of the most commonly used distributions for describing the SAD (McGill 2003) and has been derived as a null form of the distribution resulting from the central limit theorem (May 1975), population dynamics (Engen and Lande 1996), and niche partitioning (Sugihara 1980). We use the Poisson lognormal because it is a discrete form of the distribution appropriate for fitting discrete abundance data (Bulmer 1974).

The negative-binomial (which can be derived as a mixture of the Poisson and Gamma distributions) provides a good characterization of the SAD predictions for several different ecological neutral models for the purposes of model selection (Connolly et al. 2014). We use it to represent neutral models as a class.

The geometric series was one of the first distributions derived as a model of the SAD and was derived based on niche partitioning (Motomura 1932).

The Zipf (or power law) distribution was derived based on branching processes and was one of the best fitting distributions in a recent meta-analysis of SADs (Ulrich et al. 2010)

Table 2: Species abundance distribution models evaluated, their mathematical forms and model classifications.

Species abundance distribution model	Code implementation	Model classification (
Untruncated logseries	https://github.com/weecology/macroecotools.git	Purely statistical
Poisson lognormal	https://github.com/weecology/macroecotools.git	Purely statistical
Negative binomial	https://github.com/weecology/macroecotools.git	Purely statistical and p
Geometric series	https://github.com/weecology/macroecotools.git	Niche partitioning
Zipf distribution (Zipf-Mandelbrot)	https://github.com/weecology/macroecotools.git	Branching process

Analysis

Following current best practices for fitting distributions to data and evaluating their fit, we used maximum likelihood estimation to fit models to the data (Clark et al. 1999, Newman 2005, White et al. 2008) and likelihood based model selection to compare the fits of the different models (burnham2002, Edwards et al. 2007) These general best practices have recently been affirmed as best practices for species abundance distributions (Connolly et al. 2014, Matthews and Whittaker 2014).

For model comparison we used corrected Aikaike Information Criterion (AICc) weights to compare the fits of models while correcting for differences in the number of parameters and appropriately handling the small sample sizes (i.e., numbers of species) in some communities (Burnham and Anderson 2002). The Poisson log-normal and the negative binomial each have two fitted parameters, while the log-series, geometric series, and Zipf distributions have one fitted parameter each. The model with the greatest AICc weight in each community was considered to be the best fitting model for that community. We also assessed the full distribution of AICc weights to evaluate the similarity of the fits of the different models.

In addition to evaluating AICc of each model, we also examined the log-likelihood values of the models directly. We did this to assess the fit of the model while ignoring corrections for the number

of parameters and the influence of similarities to other models in the set of candidate models.

Model fitting, log-likelihood, and AICc calculations were performed using the macroecotools Python package (https://github.com/weecology/macroecotools). All of the code and the majority of the data necessary to replicate these analyses is available at (https://github.com/weecology/sad-comparison). The CBC datasets and NABA datasets are not publicly available and therefore are not included.

The negative-binomial distribution failed to converge for 1444 sites in FIA (13.9%), 5 sites in Gentry (2.3%), 3 sites in Reptilia (2.2%), and 1 site in NABA (0.25%). For these sites likelihoods and AICc weights were calculated for only those models which successfully converged.

Results

Across all datasets, the log-series had the lowest value of AICc, indicating the best fit to the data, in the greatest proportion of datasets (4X.XX%). The geometric series also performed well based on AICc, providing the best fit in 3X.XX% of the datasets. The Poisson lognormal and negative binomial distributions each provided the best fit in XX.XX% of the datasets, and the Zipf distribution had the fewest cases of the lowest AICc with X.XX% of datasets (Figure 1).

Evaluating the best fitting distributions within individual datasets and taxonomic groups, the log-series was the most frequent best fitting model for all datasets except FIA (Figure 2). For the FIA data the geometric series provided the most frequent best fit to the data, and the strong performance of the geometric series in the FIA data is the cause of it's strong performance when all of the data are analyzed together. The relative performance of the other models varies among datasets and taxonomic groups. The negative binomial performed well in the bird datasets (BBS and CBC), but was almost never the best fitting model for plants (FIA and Gentry), Coleoptera, Arachnida, or Reptilia. The Poisson lognormal performed well for the bird datasets and the Gentry tree data, but almost never won in the FIA and Coleoptera datasets (Figure 2). The Zipf distribution performed

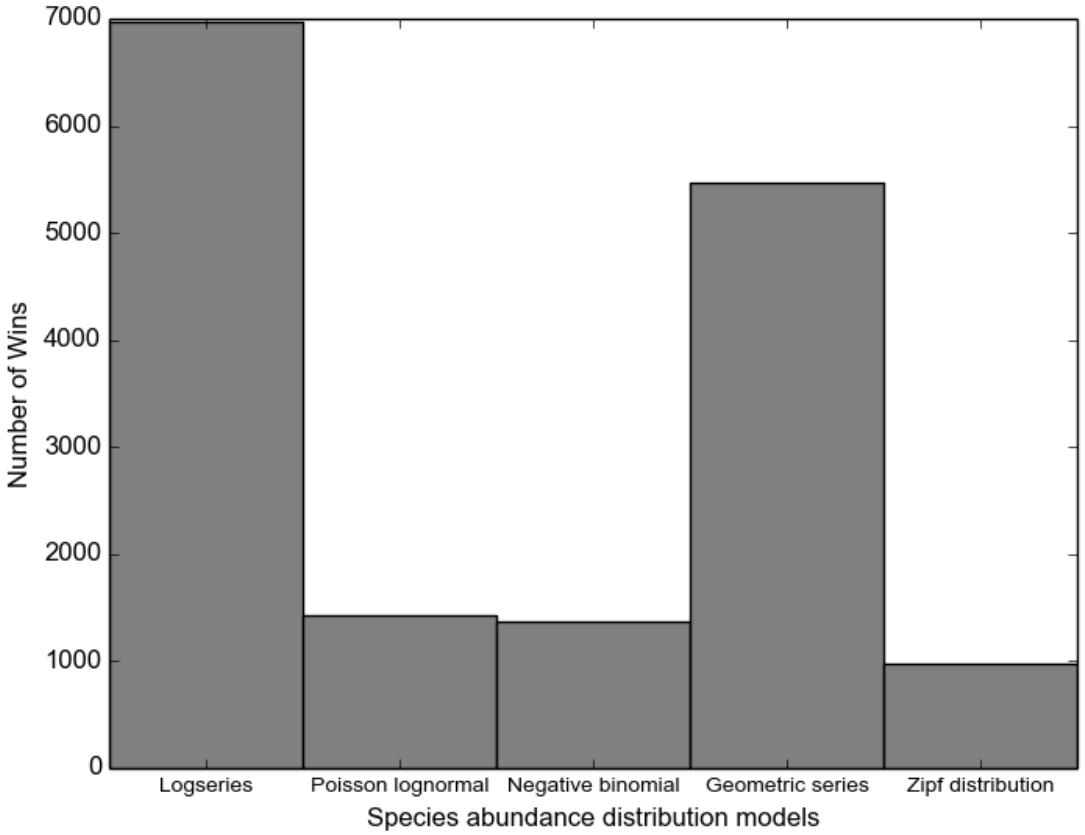

Figure 1: Total wins by model for all datasets combined.

well for Arachnida, but was never the best fitting model for the bird datasets.

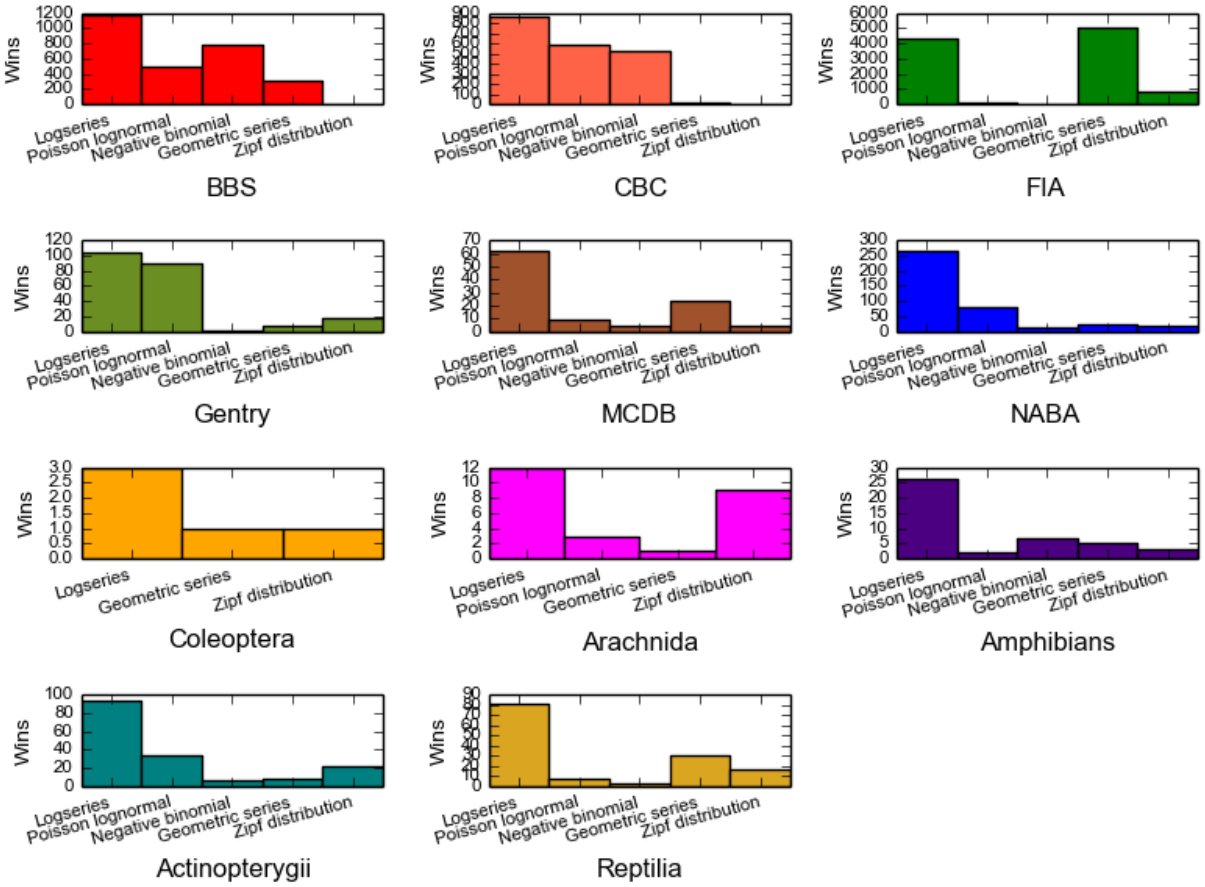

Figure 2: Total wins by model for each dataset individually.

The full distribution of AICc weights shows separation among models (Figure 3). On average, the Zipf and geometric distributions perform poorly, with the primary mode of the weight distribution occurring near 0 (Figure 3). However, the geometric distribution also exhibits better performance for a subset of communities, with a secondary mode near 0.5. This mode is driven by the FIA data. The negative binomial and the Poisson lognormal distributions have peaks around 0.1, with the Poisson lognormal also having a small peak close to 1.0 indicating that in a small number of cases it provides a fit that is clearly superior to that of the other distributions (Figure 3). The logseries performs the best overall, with a large mode spanning AICc values from 0.3 to 0.5, and secondary mode from 0.6-0.7 (Figure 3).

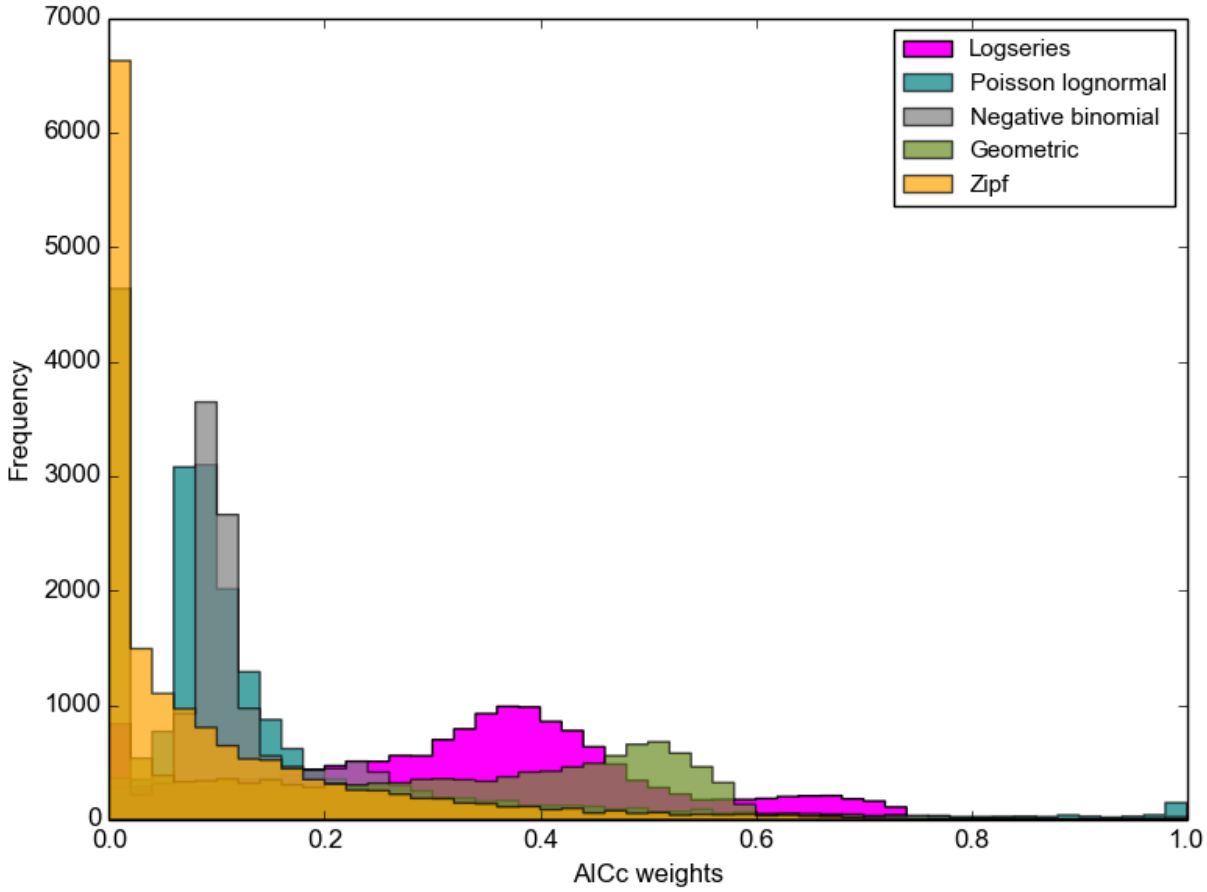

Figure 3: AICc weights by model for all datasets combined.

While the AICc weights show separation among models, these values include a correction for the number of parameters and are also influenced by the similarity between models. Therefore we also compared the negative log-likelihoods of the different models to determine whether or not their absolute fits differed. Frequency distributions of log-likelihoods show almost complete overlap among models (Figure 4) and one-to-one plots of the likelihoods of each model against the likelihood of the log-series show that the likelihoods of the different models correspond almost perfectly for individual distributions (Figure 5). This indicates that all models fit the data equivalently and that differences in AICc weights resulted primarily from differences in the number of parameters and differences in how similar different models in the set of models were (i.e., if three identically fitting models are included in the analysis none of them can have a AICc weight > 0.34).

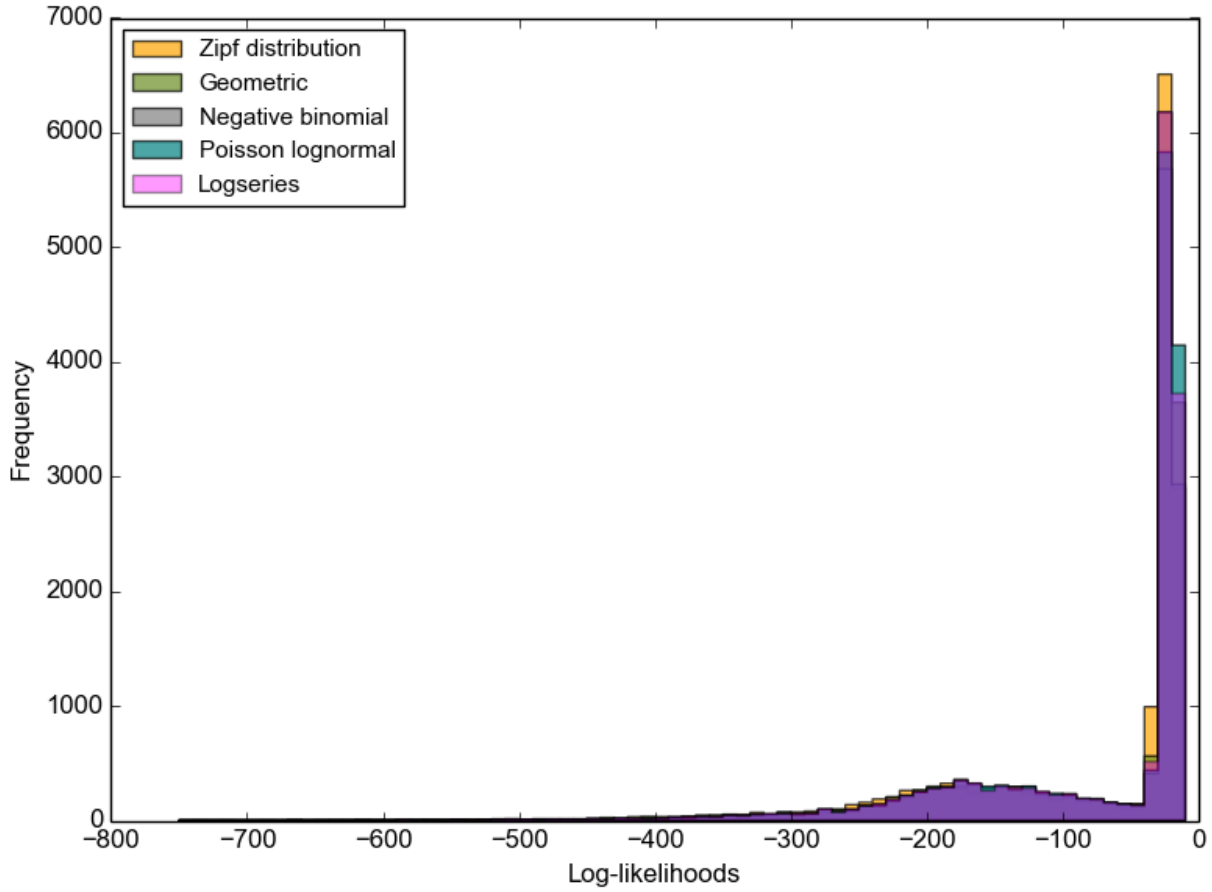

Figure 4: Log-likelihoods by model for all datasets combined.

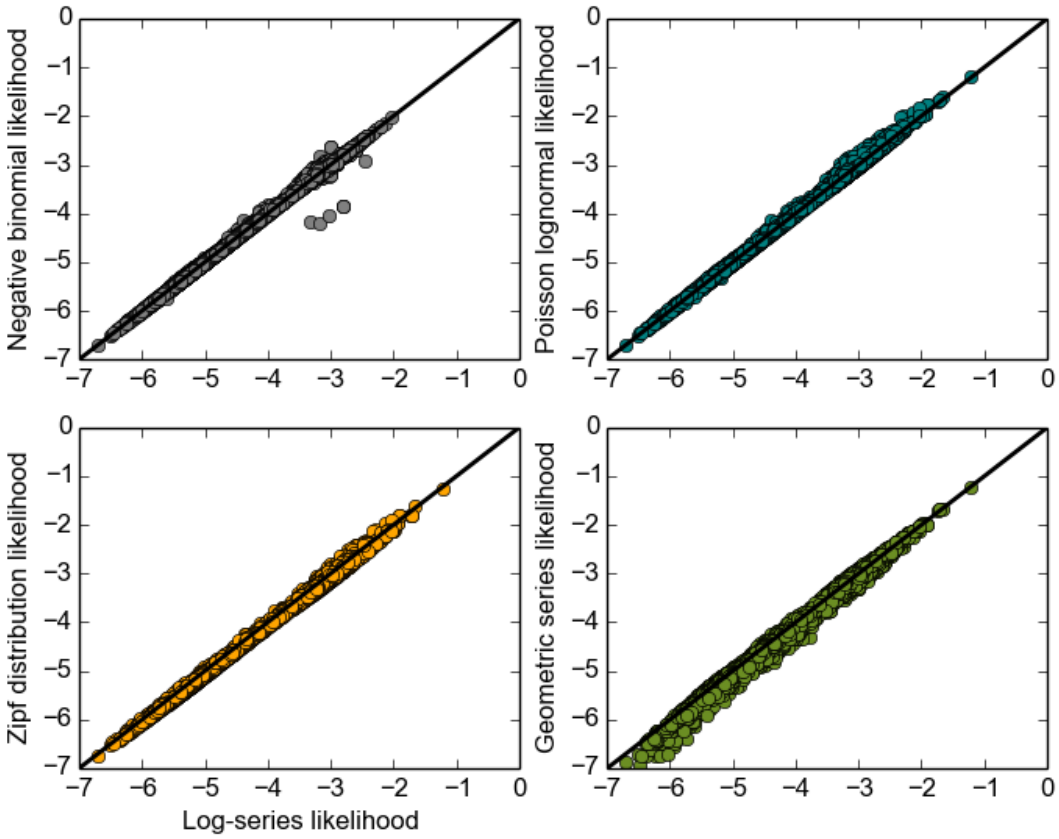

Figure 5: Log-likelihoods by model for all datasets combined.

Discussion

Our extensive comparison of different models for the species abundance distribution (SAD) using rigorous statistical methods demonstrates that most existing models provide equivalently good absolute fits to empirical data. As a result, the models with the fewest parameters perform better in AIC-based model selection since these approaches penalize model complexity. Since the log-series provides equivalent likelihoods to the other distributions, has a single fitted parameter, is easy to fit to empirical data, and is the best overall model using standard model selection, it provides a good naive model for fitting SADs.

The similar absolute fits of these five commonly used distributions emphasizes the challenges of inferring the processes operating in ecological systems from the form of the abundance distribution. It is already well established that models based on different processes can yield equivalent models of the SAD, i.e., they predict distributions of exactly the same form (Cohen 1968). It is also possible for the same biological explanations to result in different forms of the species abundance distribution depending on community conditions (Hughes 1986). Our results support the idea that even when models do differ in their precise mathematical predictions that they are often not distinguishable enough to identify potential mechanisms with any degree of certainty (Volkov et al. 2005). In other words, it is difficult to distinguish among the different distributions used to characterize the SAD, let alone the processes that generate the form of a particular distribution.

In cases where it is desirable to infer process based on macroecological patterns like the SAD, compare the predictions of different models using multiple macroecological patterns simultaneously is likely to be more effective (McGill 2003). It has also been suggested that examining second-order effects, such as the scale-dependence of macroecological patterns (Blonder et al. 2014) or the how the parameters of the distribution change across gradients (Mac Nally et al. 2014), can provide better inference about process from these kinds of pattern.

A previous analysis of ~500 SADs comparing three models, concluded that the form of the distribution varied consistently between fully censused communities, best fit by the lognormal, and

incompletely sampled communities, best fit by the Zipf and logseries (Ulrich et al. 2010). The most completely sampled data in our analysis is arguable the forest inventories (Gentry, FIA), since these inventories count all trees above a certain stem diameter and detection of trees is straightforward so they are unlikely to be missed. The lognormal model is not the best fitting model in either of these datasets. The methods used by Ulrich et al. (Ulrich et al. 2010) involve the use of binning and fitting models to rank abundance plots, which deviates from the best practices (Matthews and Whittaker 2014) used in this paper. A comparison of these two studies with equivalent methods will be necessary to resolve the discrepancies with respect to the influence of sampling on the observed form of the SAD.

In some cases linking ecological patterns to particular sets of processes is not the goal. In particular, ecological patterns can be used for prediction in the absence of any link to process. For example, the species-area relationship, which characterizes how the number of species observed changes with spatial scale, is often used to make predictions for how many species will occur at larger and smaller scales than those observed. This is done without a strong link between biological processes and the empirical pattern. The SAD has been similarly used by White et al. (White et al. 2012) who used the log-series to make predictions for the number of rare species occurring in a community. These predictions are independent of the processes generating the log-series. Given the equivalent fit of the five different distributions observed in this study, it is likely that any choice of distribution would have yielded equivalently strong predictions. In fact, patterns that not strongly contingent on the operation of specific processes can be applied to prediction more broadly, because it is not necessary to understand the detailed biology of the system in order to use them.

It is interesting to consider why so many different models for the SAD yield similar predictions and fits to empirical data. Frank (Frank 2009, Frank (2014)) suggests that general patterns do not result from specific processes, but from the fact that there are many possible ways in which that pattern can be generated. For the SAD it has been shown that of the possible forms of the SAD (the "feasible set") most have similar general shapes (Locey and White 2013). This suggests that most data and most model predictions will have similar forms because most possible forms are similar.

Maximum entropy based predictions for the SAD similarly suggest that the observed SAD should be the most likely possible form based on the random assignment of abundances to species under some basic constraints (Pueyo et al. 2007, Harte et al. 2008, Harte 2011, White et al. 2012). The fact that we observed equivalent log-likelihoods across five different models from a diverse array of ecosystems and taxonomic groups, that are likely being influenced by a diverse array of processes, supports the idea that the detailed processes operating in ecological systems are not having direct and meaningful influences on the SAD (White et al. 2012, but see Mac Nally et al. 2014).

Acknowledgments

This research was supported by the National Science Foundation through a CAREER Grant 0953694 to Ethan White, and by the Gordon and Betty Moore Foundation's Data-Driven Discovery Initiative through Grant GBMF4563 to Ethan White.

References

American Butterfly Association], N. [North. 2009. NABA butterfly counts: 2009 report. NABA, Morristown, New Jersey, USA.

Blonder, B., L. Sloat, B. J. Enquist, and B. McGill. 2014. Separating macroecological pattern and process: Comparing ecological, economic, and geological systems. PloS one 9:e112850.

Boswell, M., and G. Patil. 1971. Chance mechanisms generating the logarithmic series distribution used in the analysis of number of species and individuals. Statistical ecology 1:99–130.

Bulmer, M. 1974. On fitting the poisson lognormal distribution to species-abundance data. Biometrics:101–110.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. Springer.

Clark, R., S. Cox, and G. Laslett. 1999. Generalizations of power-law distributions applicable to sampled fault-trace lengths: model choice, parameter estimation and caveats. Geophysical Journal International 136:357–372.

Cohen, J. E. 1968. Alternate derivations of a species-abundance relation. American naturalist:165–172.

Connolly, S. R., M. A. MacNeil, M. J. Caley, N. Knowlton, E. Cripps, M. Hisano, L. M. Thibaut, B. D. Bhattacharya, L. Benedetti-Cecchi, R. E. Brainard, and others. 2014. Commonness and rarity in the marine biosphere. Proceedings of the National Academy of Sciences:201406664.

Edwards, A. M., R. A. Phillips, N. W. Watkins, M. P. Freeman, E. J. Murphy, V. Afanasyev, S. V. Buldyrev, M. G. da Luz, E. P. Raposo, H. E. Stanley, and others. 2007. Revisiting lévy flight search patterns of wandering albatrosses, bumblebees and deer. Nature 449:1044–1048.

Engen, S., and R. Lande. 1996. Population dynamic models generating species abundance distributions of the gamma type. Journal of Theoretical Biology 178:325–331.

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

Frank, S. A. 2009. The common patterns of nature. Journal of evolutionary biology 22:1563–1585.

Frank, S. A. 2011. Measurement scale in maximum entropy models of species abundance. Journal of evolutionary biology 24:485–496.

Frank, S. A. 2014. Generative models versus underlying symmetries to explain biological pattern. Journal of evolutionary biology 27:1172–1178.

Harte, J. 2011. Maximum entropy and ecology: a theory of abundance, distribution, and energetics. Oxford University Press.

Harte, J., T. Zillio, E. Conlisk, and A. Smith. 2008. Maximum entropy and the state-variable approach to macroecology. Ecology 89:2700–2711.

241 Hubbell, S. P. 2001. The unified neutral theory of biodiversity and biogeography (mPB-32).
242 Princeton University Press.

243 Hughes, R. 1986. Theories and models of species abundance. American Naturalist:879–899.

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

246 Mac Nally, R., C. A. McAlpine, H. P. Possingham, and M. Maron. 2014. The control of rank-
247 abundance distributions by a competitive despotic species. Oecologia 176:849–857.

248 Matthews, T. J., and R. J. Whittaker. 2014. Fitting and comparing competing models of the species
249 abundance distribution: assessment and prospect. Frontiers of Biogeography 6.

250 May, R. M. 1975. Patterns of species abundance and diversity. Ecology and evolution of
251 communities:81–120.

252 McGill, B. J. 2003. A test of the unified neutral theory of biodiversity. Nature 422:881–885.

253 McGill, B. J., R. S. Etienne, J. S. Gray, D. Alonso, M. J. Anderson, H. K. Benecha, M. Dornelas,
254 B. J. Enquist, J. L. Green, F. He, and others. 2007. Species abundance distributions: moving
255 beyond single prediction theories to integration within an ecological framework. Ecology letters
256 10:995–1015.

257 McGill, B. J., B. A. Maurer, and M. D. Weiser. 2006. Empirical evaluation of neutral theory.
258 Ecology 87:1411–1423.

259 Morlon, H., E. P. White, R. S. Etienne, J. L. Green, A. Ostling, D. Alonso, B. J. Enquist, F. He,
260 A. Hurlbert, A. E. Magurran, and others. 2009. Taking species abundance distributions beyond
261 individuals. Ecology Letters 12:488–501.

262 Morris, B. D., and E. P. White. 2013. The ecoData retriever: Improving access to existing ecological
263 data. PloS one 8:e65848.

264 Motomura, I. 1932. On the statistical treatment of communities. Zool. Mag 44:379–383.

Newman, E. A., M. E. Harte, N. Lowell, M. Wilber, and J. Harte. 2014. Empirical tests of within-and across-species energetics in a diverse plant community. Ecology.

Newman, M. E. 2005. Power laws, pareto distributions and zipf's law. Contemporary physics 46:323–351.

Pardieck, K. L., D. J. Ziolkowski Jr, and M.-A. Hudson. 2014. North american breeding bird survey dataset 1966 - 2013, version 2013.0. U.S. Geological Survey, Patuxent Wildlife Research Center.

Phillips, O., and J. S. Miller. 2002. Global patterns of plant diversity: Alwyn h. gentry's forest transect data set. Missouri Botanical Garden Press St., Louis, Missouri.

Pielou, E. 1975. Ecological diversity. Wiley, New York.

Pueyo, S., F. He, and T. Zillio. 2007. The maximum entropy formalism and the idiosyncratic theory of biodiversity. Ecology Letters 10:1017–1028.

Service, U. F. 2010. Forest inventory and analysis national core field guide (phase 2 and 3). version 4.0. USDA Forest Service, Forest Inventory; Analysis.

Society, N. A. 2002. The christmas bird count historical results. National Audobon Society, New York, New York, USA.

Sugihara, G. 1980. Minimal community structure: an explanation of species abundance patterns. American naturalist:770–787.

Tokeshi, M. 1993. Species abundance patterns and community structure. Advances in ecological research 24:111–186.

Ulrich, W., M. Ollik, and K. I. Ugland. 2010. A meta-analysis of species–abundance distributions. Oikos 119:1149–1155.

Volkov, I., J. R. Banavar, F. He, S. P. Hubbell, and A. Maritan. 2005. Density dependence explains tree species abundance and diversity in tropical forests. Nature 438:658–661.

Volkov, I., J. R. Banavar, S. P. Hubbell, and A. Maritan. 2003. Neutral theory and relative species abundance in ecology. Nature 424:1035–1037.

White, E. P., B. J. Enquist, and J. L. Green. 2008. On estimating the exponent of power-law frequency distributions. Ecology 89:905–912.

White, E. P., K. M. Thibault, and X. Xiao. 2012. Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. Ecology 93:1772–1778.

Xiao, X., D. J. McGlinn, and E. P. White. 2015. A strong test of the maximum entropy theory of ecology. American Naturalist.