An extensive comparison of species-abundance distribution models

Elita Baldridge [1], **David J Harris** [2], **Xiao Xiao** [3, 4, 5, 6], **Ethan P White** Corresp. [2, 3, 4, 7]

[1] Department of Biology and the Ecology Center, Utah State University, Logan, UT, United States
[2] Department of Wildlife Ecology and Conservation, University of Florida, Gainesville, FL, United States
[3] Department of Biology, Utah State University, Logan, Utah, United States
[4] Ecology Center, Utah State University, Logan, Utah, United States
[5] School of Biology and Ecology, University of Maine, Orono, Maine, United States
[6] Mitchell Center for Sustainability Solutions, University of Maine, Orono, Maine, United States
[7] Informatics Institute, University of Florida, Gainesville, FL, United States

Corresponding Author: Ethan P White
Email address: ethan@weecology.org

A number of different models have been proposed as descriptions of the species-abundance distribution (SAD). Most evaluations of these models use only one or two models, focus only a single ecosystem or taxonomic group, or fail to use appropriate statistical methods. We use likelihood and AIC to compare the fit of four of the most widely used models to data on over 16,000 communities from a diverse array of taxonomic groups and ecosystems. Across all datasets combined the log-series, Poisson lognormal, and negative binomial all yield similar overall fits to the data. Therefore, when correcting for differences in the number of parameters the log-series generally provides the best fit to data. Within individual datasets some other distributions performed nearly as well as the log-series even after correcting for the number of parameters. The Zipf distribution is generally a poor characterization of the SAD.

An extensive comparison of species-abundance distribution models

Elita Baldridge[1,2], David J. Harris[3], Xiao Xiao[1,2,4,5], Ethan P. White[*,1,2,3,6]

[1] Department of Biology, Utah State University, Logan, Utah, USA

[2] Ecology Center, Utah State University, Logan, Utah, USA

[3] Department of Wildlife Ecology & Conservation, University of Florida, Gainesville, Florida, USA

[4] School of Biology and Ecology, University of Maine, Orono, Maine, USA

[5] Mitchell Center for Sustainability Solutions, University of Maine, Orono, Maine, USA

[6] Informatics Institute, University of Florida, Gainesville, Florida, USA

Abstract

A number of different models have been proposed as descriptions of the species-abundance distribution (SAD). Most evaluations of these models use only one or two models, focus only a single ecosystem or taxonomic group, or fail to use appropriate statistical methods. We use likelihood and AIC to compare the fit of four of the most widely used models to data on over 16,000 communities from a diverse array of taxonomic groups and ecosystems. Across all datasets combined the log-series, Poisson lognormal, and negative binomial all yield similar overall fits to the data. Therefore, when correcting for differences in the number of parameters the log-series generally provides the best fit to data. Within individual datasets some other distributions performed nearly as well as the log-series even after correcting for the number of parameters. The Zipf distribution is generally a poor characterization of the SAD.

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. 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 the most likely states of generic random systems (Frank 2011, Harte 2011, Locey and White 2013), to models based more directly on ecological processes (Tokeshi 1993, Hubbell 2001, Volkov et al. 2003, Alroy 2015).

Which model or models provide the best fit to the data, and the resulting implications for the processes structuring ecological systems, is 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, see Matthews and Whittaker 2014 for discussion of best statistical methods for fitting SADs). 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 four of the most widely used models for the species abundance distribution using likelihood-based model selection on data from 16,209 communities and nine major 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 16,209 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; USDA Forest Service 2010), the North American Butterfly Association's North American Butterfly Count (NABC; North American Butterfly Assoc. 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; National Audubon 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 acquired using the EcoData Retriever (http://ecodataretriever.org; Morris and White 2013). Details of the treatment of these datasets can be found in Appendix A of White et al. (2012). We modified the data slightly by removing sites 102 and 179 from the Gentry data due to issues with decimal abundances appearing in raw data due to either data entry or data structure errors. Data on Actinopterygii, Reptilia, Coleoptera, Arachnida, and Amphibia, were mined from literature by Baldridge and are publicly available (Baldridge 2013). All abundances in the compiled datasets were counts of individuals.

Table 1: Details of datasets used to evaluate the form of the species abundance distribution. Datasets marked as Private were obtained through data requests to the providers.

Dataset	Dataset code	Availability	Total sites	Citation
Breeding Bird Survey	BBS	Public	2769	Pardieck et al. (2014)
Christmas Bird Count	CBC	Private	1999	National Audubon Society (2002)
Gentry's Forest Transects	Gentry	Public	220	Phillips and Miller (2002)

Dataset	Dataset code	Availability	Total sites	Citation
Forest Inventory Analysis	FIA	Public	10355	USDA Forest Service (2010)
Mammal Community DB	MCDB	Public	103	Thibault et al. (2011)
NA Butterfly Count	NABA	Private	400	North American Butterfly Assoc. (2009)
Actinopterygii	Actinopterygii	Public	161	Baldridge (2013)
Reptilia	Reptilia	Public	129	Baldridge (2013)
Amphibia	Amphibia	Public	43	Baldridge (2013)
Coleoptera	Coleoptera	Public	5	Baldridge (2013)
Arachnida	Arachnida	Public	25	Baldridge (2013)

Models

The majority of species abundance distributions (SADs) are constructed using counts of individuals (for discussion of alternative approaches see McGill et al. 2007 and @morlon2009). As such, the data are discrete and discrete distributions (i.e., those that only have non-zero probabilities for positive integer values of abundance) are more appropriate.

McGill et al. (2007) classified species abundance distribution models into five different families: purely statistical, branching process, population dynamics, niche partitioning, and spatial distribution of individuals. We evaluated models from each of these families, with some models having been derived from more than one family of processes. We evaluated the log-series, the Poisson lognormal, the negative binomial, and the Zipf distributions. All distributions were defined to be capable of having non-zero probability at integer values from 1 to infinity.

The log-series is one of the first distributions used to describe the SAD, being derived as a purely statistical distribution by Fisher (1943). It has since been derived as the result of 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 Gamma-distributed mixture of Poisson 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 Zipf (or power law) distribution was derived based on both branching processes and as the outcome of the McGill and Collin's (2003) spatial model. It was one of the best fitting distributions in a recent meta-analysis of SADs (Ulrich et al. 2010). We use the discrete form of the distribution which is appropriate for fitting discrete abundance data (White et al. 2008).

Figure 1 shows three example sites with the empirical distribution and associated models fit to the data Zipf distributions tend to predict the most rare species followed by the log-series, the negative binomial, and Poisson lognormal.

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 (Burnham and Anderson 2002, 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). This requires that likelihoods for the models can be solved for and therefore we excluded models that lack probability mass functions and therefore likelihoods. While methods have been proposed for comparing models without probability mass functions in this context (Alroy 2015),

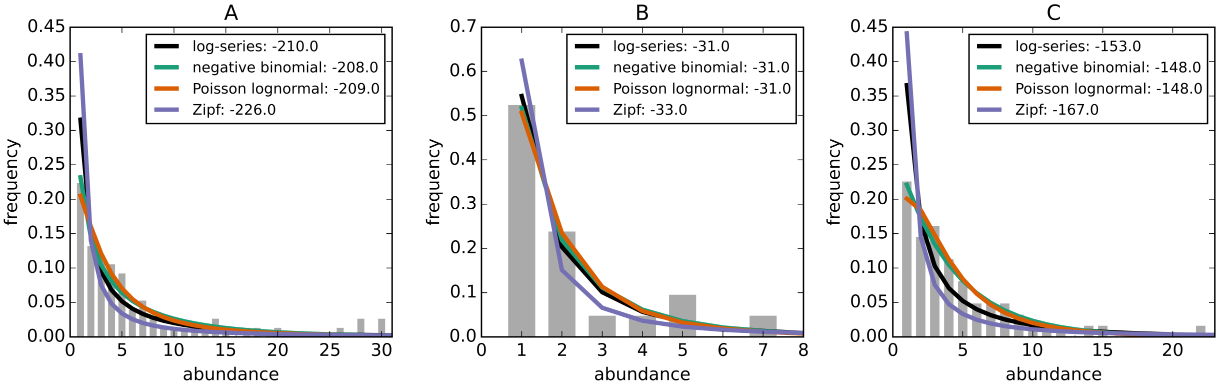

Figure 1: Example species-abundance distributions including the empirical distributions (grey bars) and the best fitting log-series (black line), negative binomial (green line), Poisson lognormal (red line), and Zipf (purple line). Distributions are for (a) Breeding Bird Survey - Route 36 in New York, (b) Forest Inventory and Analysis - Unit 4, County 57, Plot 12 in Alabama, and (c) Gentry - Araracuara High Campina site in Colombia. Log-likelihoods of the models are included in parenthesis in the legend

these methods have not been evaluated to determine how well they perform compared to the widely accepted likelihood-based approaches.

For model comparison we used corrected Akaike 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 lognormal and the negative binomial each have two fitted parameters, while the log-series distribution and the 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. This also allows us to make more direct comparisons to previous analyses that have not corrected for the number of parameters (i.e., Ulrich et al. 2010, Alroy 2015)

Model fitting, log-likelihood, and AICc calculations were performed using Python (Van Rossum and Drake 2011) and R (R Core Team 2015). Python packages used for analysis include numpy (Oliphant 2007, Van Der Walt et al. 2011), matplotlib (Hunter and others 2007), sqlalchemy (Bayer 2014), pandas (McKinney and others 2010), macroecotools https://github.com/weecology/macroecotools, retriever (Morris and White 2013), R packages used for analysis include ggplot2 (Wickham 2009), magrittr(Bache and Wickham 2014), tidyr (Wickham 2016), dplyr (Wickham and Francois 2016). All of the code and all of the publicly available 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.

Results

Across all data sets, the negative binomial and Poisson lognormal distributions had very similar average log-likelihoods (within 0.01 of one another; Figure 2). The log-likelihoods for each of these distributions averaged 0.8 units higher than for the log-series distribution and 5 units higher than for the Zipf distribution (corresponding to likelihoods that were twice as high and 140 times as high, respectively).

Although the negative binomial and Poisson lognormal distributions matched the data most closely, the likelihood provides a biased estimate of these distributions' ability to generalize to unobserved species. AICc approximately removes this bias by penalizing models with more degrees of freedom (e.g. the negative binomial and Poisson lognormal distributions, which have two free parameters instead of one like the log-series and Zipf distributions). After applying this penalty, the log-series distribution would be expected to make the best predictions for 69.2% of the sites. The Poisson lognormal and negative binomial distributions were each preferred in about 12% of the sites, and the Zipf distribution was preferred least often (6.0% of sites; Figure 3).

Across all data sets and taxonomic groups, the log-series distribution had the highest AICc weights

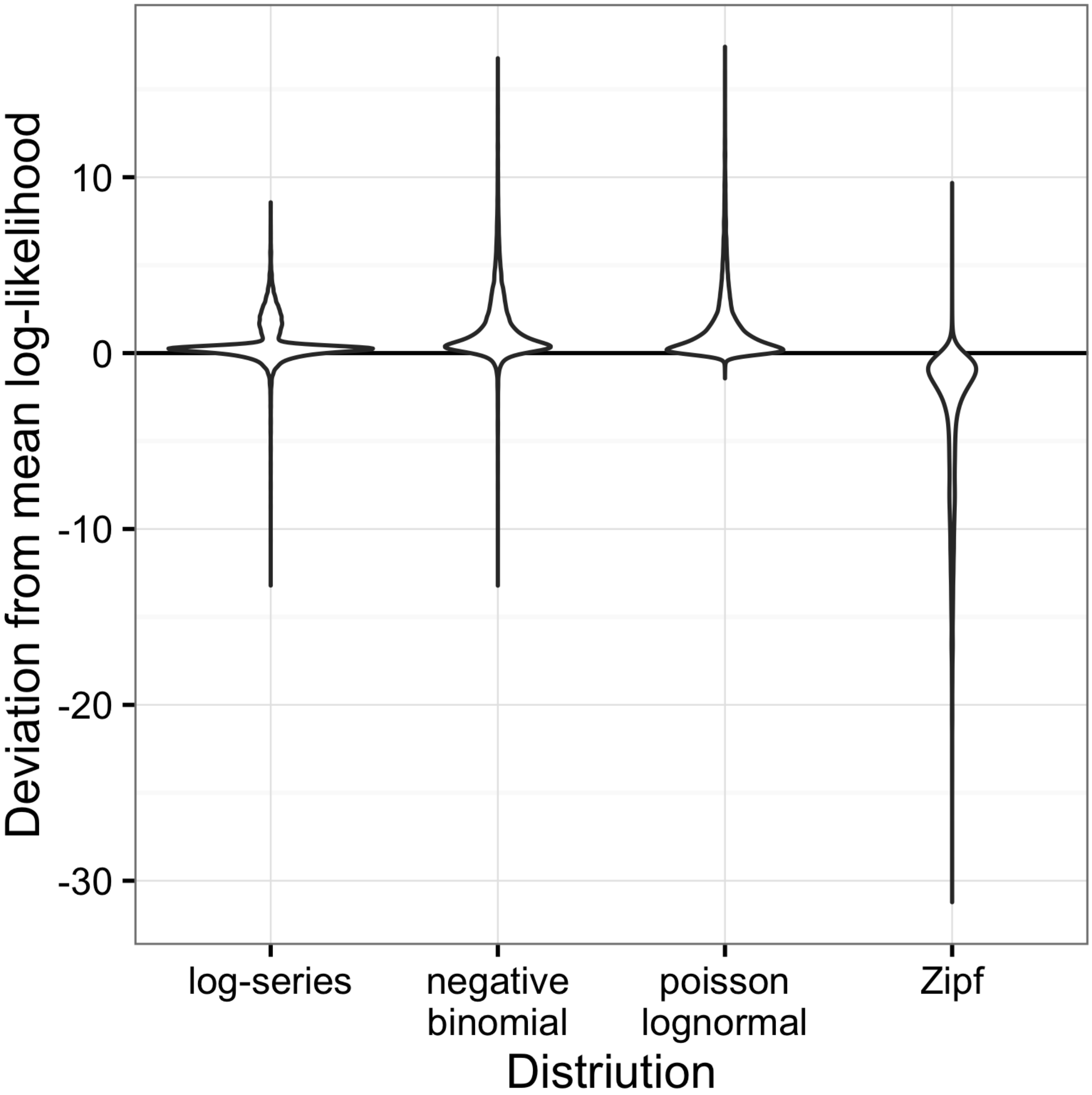

Figure 2: Violin plots of the deviation from the mean log-likelihood for each site for all datasets combined. Positive values indicate that the model fits better than the average fit across the four models.

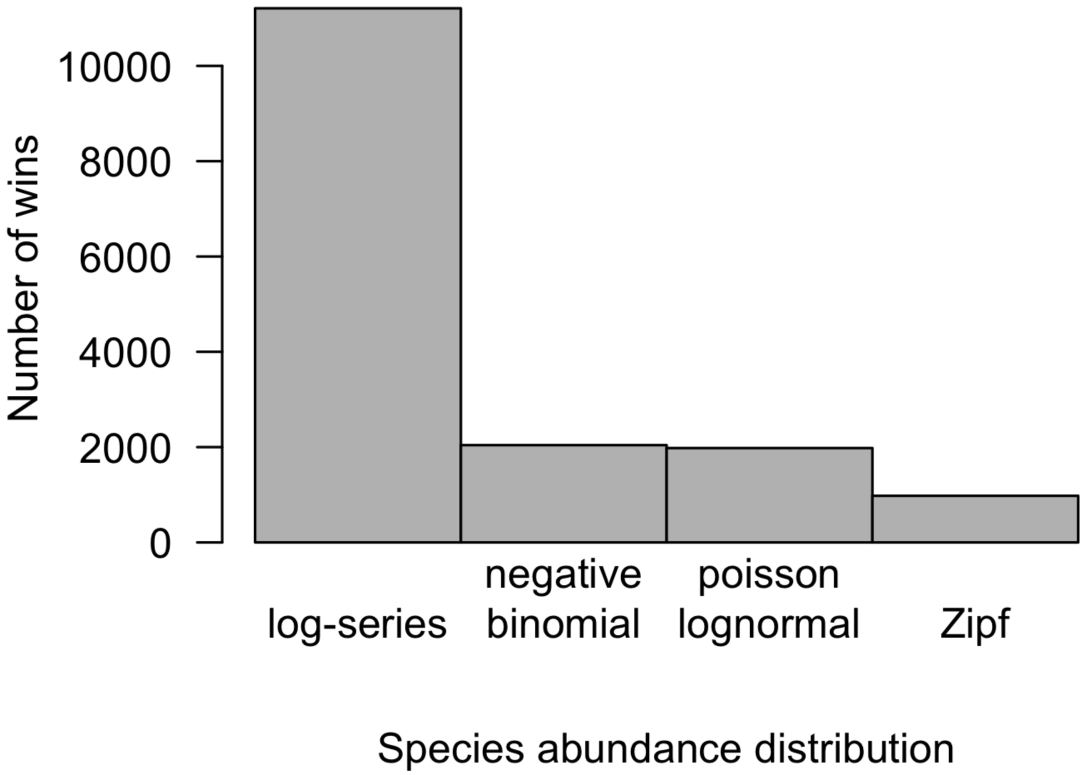

Figure 3: Number of cases where each model provided the best fit to the data based on AICc for all datasets combined.

more often than any other model. The negative binomial performed well for BBS, but was almost never the best fitting model for plants (FIA and Gentry), butterflies (NABA), Acintopterygii, or Coleoptera. The Poisson lognormal performed well for the bird datasets (BBS and CBC) and the Gentry tree data, but was almost never best in the FIA and Coleoptera datasets (Figure 4). The Zipf distribution only performed consistently well for Arachnida.

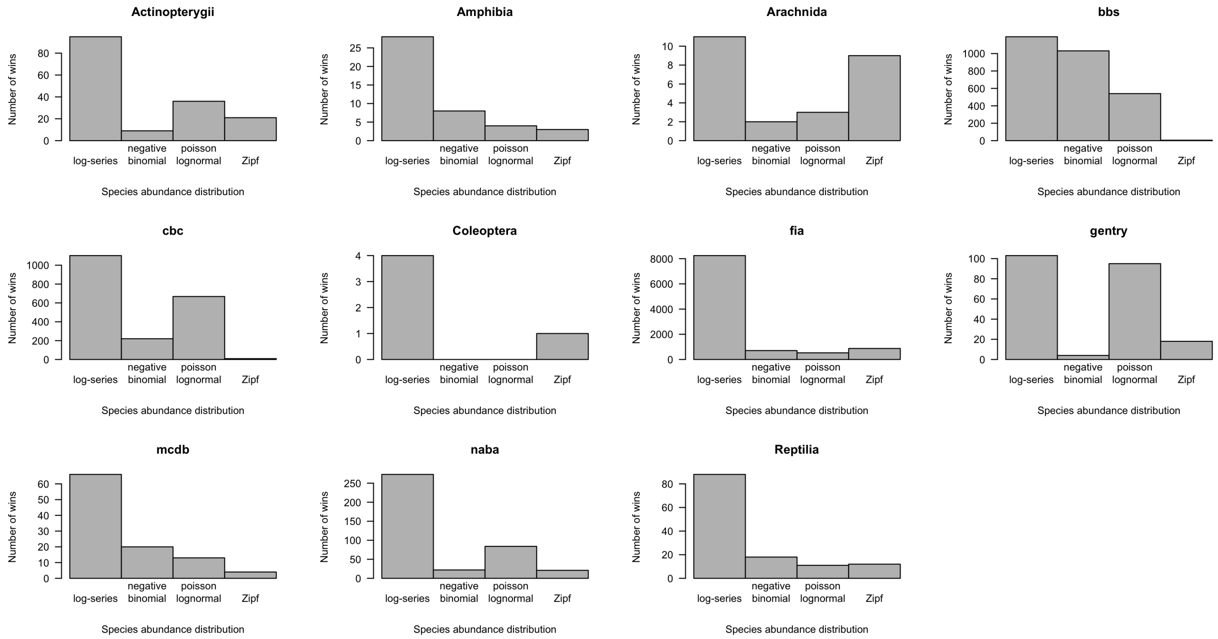

Figure 4: Number of cases where each model provided the best fit to the data based on AICc for each dataset separately.

The full distribution of AICc weights shows separation among models (Figure 5). Although the log-series distribution had the best AICc score much more often than the other models, its lead was never decisive: across all 16,209 sites, it never had more than about 75% of the AICc weight (Figure 5). Most of the remaining weight was assigned to the negative binomial and Poisson lognormal distributions (each of which usually had at least 12-15% of the weight but was occasionally favored very strongly). The Zipf distribution showed a strong mode near zero, and usually had less than 7% of the weight.

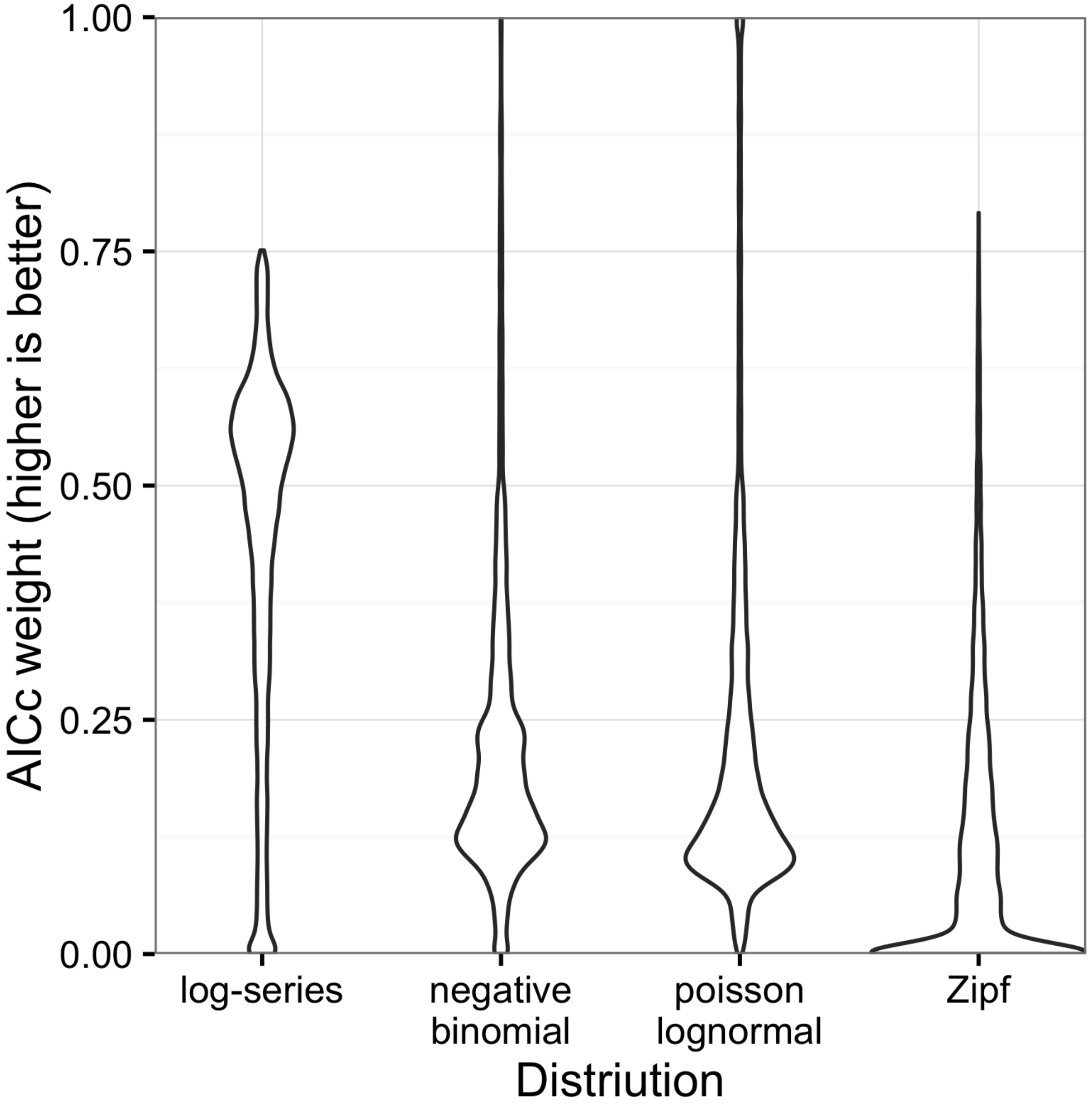

Figure 5: Violin plots of the AICc weights for each model. Weights indicate the probability that the model is the best model for the data

Discussion

Our extensive comparison of different models for the species abundance distribution (SAD) using rigorous statistical methods demonstrates that several of the most popular existing models provide equivalently good absolute fits to empirical data. Log-series, negative binomial, and Poisson lognormal all had model relative likelihoods between 0.25 and 0.5 suggesting that the three distributions provide roughly equivalent fits in most cases, but with the two parameters models doing slightly better on average. Because the log-series has only a single parameter but fits the data almost as well as the two parameter models, the log-series performed better in AICc-based model selection, which penalizes model complexity. These results differ from two other recent analyses of large numbers of species abundance distributions (Ulrich et al. 2010, Connolly et al. 2014) and are generally consistent with a third recent analysis (Alroy 2015).

Ulrich et al. (2010) analyzed ~500 SADs and found support for three major forms of the SAD that changed depending on whether the community had been fully censused or not. They found that "fully censused" communities were best fit by the lognormal, and "incompletely sampled" communities, best fit by the Zipf and log-series (Ulrich et al. 2010). In contrast we find effectively no support for the Zipf across ecosystems and taxonomic groups, including a number of datasets that are incompletely sampled. Our AICc value results also do not support the conclusion that the lognormal outperforms the log-series in fully censused communities. The Gentry and FIA forest inventories both involve large stationary organisms and were collected with the goal of including all trees above a certain stem diameter. Therefore, above the minimum stem diameter, they are as close to fully censused communities as is typically possible. In these communities the log-series provides the best fit to the data most frequently. The discrepancy between our results and those found in (Ulrich et al. 2010) may be due to: 1) their use of binning and fitting curves to rank abundance plots, which deviates from the likelihood-based best practices (Matthews and Whittaker 2014) used in this paper; 2) the statistical methods they use to identify communities as "fully censused", which tend to exclude communities with large numbers of singletons that would be better fit by distributions like the log-series; 3) the use of the continuous lognormal instead of the Poisson lognormal.

Connolly et al. (2014) use likelihood-based methods to compare the the negative binomial distribution (which they call the Poisson gamma) to the Poisson lognormal for a large number of marine communities. They found that they Poisson lognormal provides a substantially better fit than the negative binomial to empirical data and that the negatvie-binomial provides a better fit to communities simulated using neutral models. They conclude that these analyses of the SAD demonstrates that marine communities are structured by non-neutral processes. Our analysis differs from that in Connolly et al. (2014) in that they aggregate communities at larger spatial scales than those sampled and find the strongest results at large spatial scales. This may explain the difference between the two analysis or there may be differences between the terrestrial systems analyzed here and the marine systems analyzed by Connolly et al. (2014). The explanation for these differences is being explored elsewhere (Connolly et al. unpublished data).

Alroy (2015) compared the fits of the lognormal, log-series, Zipf, geometric series, broken stick, and a new model dubbed the “double geometric”, to over 1000 terrestrial community datasets assembled from the literature. To incorporate the geometric series, broken stick, and the double geometric, this research used non-standard methods for evaluating the fits of the models to the data, however the results were generally consistent with those presented here. The central Kullback-Leibler divergence statistics results show that: 1) the Zipf, geometric series, and broken stick all perform consistently worse than the other distributions; 2) the double geometric, log-series, and lognormal all provide the best overall fit for at least one taxonomic group; and 3) the lognormal and double geometric fit the data equivalently well and slightly better than the log-series when not controlling for differences in the number of parameters (Alroy’s tables S1, S2, and S3). Penalizing the two-parameter models (lognormal and double geometric) for their complexity, as we do here with AICc, would likewise improve the relative performance of the log-series distribution.

In combination, the results of these three papers suggest that in general the Zipf is a poor characterization of species-abundance distributions and that both the log-series and lognormal distributions provide reasonable fits in many cases. Differences in the performance of the log-series, lognormal, double geometric, and negative binomial, appear to be more minor. How these differences relate to differences in intensity of sampling, spatial scale, taxonomy, and ecosystem type (marine vs. terrestrial) remain open questions. Our analyses suggest that controlling for the number of parameters makes the log-series a slightly better fitting model, at least in the terrestrial systems we studied. Neither of the other papers that include the log-series (Ulrich et al. 2010, Alroy 2015) make this correction and both show that it is still a reasonably competitive model even against those with more parameters.

The relatively similar fits of several 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, Boswell and Patil 1971, Pielou 1975, McGill et al. 2007). To the extent that SADs are determined by random statistical processes, one might expect the observed distributions to be compatible with a wide variety of different process-based and process-free models (Frank 2009, 2011, Locey and White 2013). Regardless of the underlying reason that the models performed similarly, our results indicate that the SAD usually does not contain sufficient information to distinguish among the possible statistical processes—let alone biological processes—with any degree of certainty (Volkov et al. 2005), though it is possible that this result differs in marine systems (see Connolly et al. 2014). A more promising way to draw inferences about ecological processes is to evaluate each model's ability to simultaneously explain multiple macroecological patterns, rather than relying on a single pattern like the SAD (McGill 2003, McGill et al. 2006, Newman et al. 2014, Xiao et al. 2015). It has also been suggested that examining second-order effects, such as the scale-dependence of macroecological patterns (Blonder et al. 2014) or 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.

Acknowledgments

We thank all of the individuals involved in the collection and provision of the data used in this paper, including the citizen scientists who collect the BBS, CBC, and NABC data, the USGS and CWS scientists and managers, the Audubon Society, the North American Butterfly Association, the USDA Forest Service, the Missouri Botanical Garden, and Alwyn H. Gentry. We also thank all of the scientists who published their raw data allowing it to be combined in Baldridge (2013). 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

Alroy, J. 2015. The shape of terrestrial abundance distributions. Science advances 1:e1500082.

Bache, S. M., and H. Wickham. 2014. Magrittr: A forward-pipe operator for r.

Baldridge, E. 2013. Community abundance data.

Bayer, M. 2014. Sqlalchemy. The Architecture of Open Source Applications: Elegance, Evolution, and a Few More Fearless Hacks 2.

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:8524–8529.

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.

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.

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

Hunter, J. D., and others. 2007. Matplotlib: A 2D graphics environment. Computing in science and engineering 9:90–95.

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

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

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

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

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

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

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

McGill, B., and C. Collins. 2003. A unified theory for macroecology based on spatial patterns of abundance. Evolutionary Ecology Research 5:469–492.

McKinney, W., and others. 2010. Data structures for statistical computing in python. Pages 51–56 *in* Proceedings of the 9th python in science conference.

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

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

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 95:2815–2825.

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

North American Butterfly Assoc. 2009. NABA butterfly counts: 2009 report. NABA, Morristown, New Jersey, USA.

Oliphant, T. E. 2007. Python for scientific computing. Computing in Science & Engineering 9:10–20.

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.

R Core Team. 2015. R: A language and environment for statistical computing. R Foundation for

Statistical Computing, Vienna, Austria.

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.

Thibault, K. M., S. R. Supp, M. Giffin, E. P. White, and S. M. Ernest. 2011. Species composition and abundance of mammalian communities: Ecological archives e092-201. Ecology 92:2316–2316.

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.

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

Van Der Walt, S., S. C. Colbert, and G. Varoquaux. 2011. The numPy array: A structure for efficient numerical computation. Computing in Science & Engineering 13:22–30.

Van Rossum, G., and F. L. Drake. 2011. The python language reference manual. Network Theory Ltd.

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.

Wickham, H. 2009. Ggplot2: Elegant graphics for data analysis. Springer-Verlag New York.

Wickham, H. 2016. Tidyr: Easily tidy data with ‘spread()‘ and ‘gather()‘ functions.

Wickham, H., and R. Francois. 2016. Dplyr: A grammar of data manipulation.

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