Ecological Archives **E093-155-A1**

Ethan P. White, Katherine M. Thibault, and Xiao Xiao. 2012. Characterizing species abundance distributions across taxa and ecosystems using a simple maximum entropy model. *Ecology* 93:1772–1778. http://dx.doi.org/10.1890/11-2177.1

Appendix A. Additional details on data collection, statistical methods, and maximum entropy.

Data

We compiled species abundance data from six continental- or global-scale databases spanning four major taxonomic groups and encompassing 15,848 sites distributed across all continents except Antarctica. We used only one year of sampling for each site to capture only those individuals partitioning resources in a community at one time in the same place. We used only data for communities comprised of a minimum of 10 species, based on the assumption that abundance distributions will not be well characterized with fewer than 10 data points. In the descriptions of datasets below, averages are reported as the mean ± 1 standard deviation.

Birds

We used community data collected in 2009 from 2,769 routes of the Breeding Bird Survey (Sauer et al. 2011) (BBS) and 1,999 counts of the Christmas Bird Count (National Audubon Society 2002) (CBC). BBS routes are 40 km long, each consisting of 50 three minute point counts, 800 m apart, sampled annually in June. The 2009 data include a total of 1,819,908 individuals representing 347 species of diurnal landbirds, with individual routes averaging 657.2 ± 323.9 individuals (range = 53 – 3,504) and 46 ± 13 species (range = 10 – 81). CBC count circles are 24.1 km in diameter and are censused by multiple observers over the course of a day.

This intensive sampling effort yielded a total of 45,294,162 individuals representing 397 species of diurnal landbirds in 2009 (count year 109), with individual circles averaging 22,658.4 ± 254,604.2 individuals (range = 36 – 10,280,057) and 45 ± 17 species (range = 10 – 116). Nocturnal species, water birds, and raptors were excluded from analyses, because they are poorly sampled by these survey methods.

Trees

We also used two existing data sets of species abundance for communities of trees, the USFS Forest Inventory Analysis program (U.S. Department of Agriculture 2010, Woudenberg et al. 2010, http://apps.fs.fed.us/fiadb-downloads/datamart.html) (FIA), and the Alwyn H. Gentry Forest Transect Data Set (Phillips and Miller 2002) (referred to herein as ‘Gentry’). We used one year of data (calendar year of sampling varies among plots) for FIA phase 2 plots that were sampled using the standardized methodology implemented in 1999 [see the FIA National Core Field Guide for more information (U.S. Department of Agriculture 2010)]. The standard plot consists of four 24.0-foot (7.32 m) radius subplots, on which trees 5.0 inches (12.7 cm) and greater in diameter are identified to species and measured. We used species abundance data for 10,355 FIA plots, encompassing a total of 380,581 individuals and 236 species, with plots averaging 36.8 ± 12.5 individuals (range = 11 – 118) and 11.4 ± 1.6 species (range = 10 – 21). The Gentry data were collected from 226-0.1 hectare sites throughout the world, with each site sampled once over the course of a 22 year period. At each site, all plants with stem diameters of 2.5 cm or greater were identified and measured along ten 2 × 50 m transects. It should be noted that, due to difficulties in the taxonomy and identification of tropical trees, some species in the Gentry dataset are identified only as morpho-species (unique within sites), and species’ names vary among sites due to both typographical errors and synonymy problems. Since we only analyzed data within a site, these issues do not affect our analyses, but they artificially elevate the count of species in the Gentry dataset and therefore the number of species included in the overall analysis. We used data from 222 sites, including 67,405 individuals representing approximately 7,300 species, with individual sites averaging 303.6 ± 115.6 individuals (range = 44 – 779) and approximately 91.4 ± 59.7 species (range = 10 – 250).

Mammals

We used species abundance data for the 103 sites included in the Mammal Community Database (Thibault et al. 2011) (MCDB) that included at least 10 species (mean richness = 13.6 ± 4.0 species; range = 10 – 34). These data have been compiled from various published sources and therefore have not been collected using a standardized protocol across sites. As a result, these data are species-level abundances of small mammals that were captured using various levels of sampling effort spread across varying amounts of time and space. Despite these limitations, these data represent, to our knowledge, the largest collection of mammal community data ever analyzed in one study. The data encompass a total of 380 mammal species and 94,866 individuals (mean abundance per site = 921.0 ± 1,434.9; range = 19 – 10,085).

Butterflies

Butterfly community abundance data were collected as part of the continent-wide North American Butterfly Count program of the North American Butterfly Association (North American Butterfly Association 2009) (NABA). This program includes the 4th of July Butterfly Counts, 1st of July Butterfly Counts, 16th of September Butterfly Counts, and Seasonal Butterfly Counts. We used only one date of sampling per site from the 2009 data, with most of the data collected in July. These counts are censuses of all butterflies seen within a 25-km (15-mile) diameter circle over the course of a day (minimum of 6 hours) by various numbers of observers.

We used species abundance data from 400 count circles, including 453 species (mean richness per site = 34.9 ± 13; range = 10 – 95) and 405,354 individuals (mean per-site abundance = 1,013.4 ± 3,808; range = 23 – 73,435).

Taxonomic and Geographic Limitations

While this is, to our knowledge, the most extensive collection of ecological communities ever assembled, it does contain taxonomic and geographic biases. In particular, while there are communities on six continents, the vast majority of the data are from North America. In addition, we lack data on several major taxonomic groups including both aquatic and microbial communities. Therefore, the inference regarding the ability to capture the form of the species-abundance distribution using information on only richness and abundance is strongest for terrestrial systems in North America.

Statistical Methods

A Brief Introduction to the Method of Maximum Entropy

The maximum entropy principle (MaxEnt) is a method from information theory that is used to make inference related to probability distributions (Jaynes 2003). MaxEnt identifies the probability distribution that is consistent only with a limited amount of prior information regarding the underlying processes or the state of the system. The fundamental idea is that, among all potential configurations concordant with some constraints, the least biased (i.e., most likely) configuration is the one that relies on the fewest assumptions, i.e., the one that includes the least additional information beyond the specified constraints. Intuitively, a peaked distribution is more precise and informative in inference compared to a flat one; thus the application of MaxEnt is equivalent to identifying the flattest (i.e., closest to uniform) distribution possible that also matches the prior information.

The most common measure of information (or rather, lack of information) is Shannon entropy

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log(p(x_i)),$$

where p denotes the probability mass function of a random variable X. This measure is maximized when the distribution is flattest with respect to the prior information. Prior information can often be expressed as constraints on the expected value of some function of the variable:

$$\sum_i f(x_i) p(x_i) = \langle f(x_i) \rangle$$

such as the mean or the variance. This type of constraint is known as a "soft constraint" (*sensu* Haegeman and Etienne 2010). It is also possible to place "hard constraints" on the distribution that fix specific values rather than expectations, which can lead to different predictions (Haegeman and Etienne 2010). Once the constraints are defined, the determination of the least biased distribution is a constrained optimization problem, which can be solved with the method of Lagrange multipliers (Jaynes 2003, Harte 2011).

Harte et al.'s Maximum Entropy Based Theory

In the theory developed by Harte and colleagues (Harte et al. 2008, 2009, Harte 2011), macroecological patterns are inferred based on a joint distribution, $R(n, \varepsilon \mid S_0, N_0, E_0)$, which specifies the probability that a randomly selected species from a given community has abundance *n*, and a randomly selected individual from that species has a metabolic requirement in the interval ($\varepsilon, \varepsilon + d\varepsilon$). The constraints on this distribution are soft constraints on the average abundance per species, N_0/S_0, and the average energy flux per species, E_0/S_0, and hard constraints on the upper bound of the distribution equal to N_0 for n and E_0 for ε. By applying the method of Lagrange multipliers, integrating out energy, and dropping terms that are very small (this approximation has been empirically justified for reasonably large values of S_0 and N_0; see Box 7.4, and Table 7.2 in Harte 2011), it has been shown that the species-abundance distribution (SAD) predicted by MaxEnt follows Fisher's log-series distribution:

$\Phi(n \mid S_0, N_0) = \frac{1}{log(\beta^{-1})} \frac{e^{-\beta n}}{n}$ [Eq. 7.32 from (Harte 2011)]

where $\Phi(n)$ is the probability that a species has abundance n, and β is the sum of the two Lagrange multipliers and is defined by: $\frac{N_0}{S_0} = \frac{\sum_{n=1}^{N_0} e^{-\beta n}}{\sum_{n=1}^{N_0} e^{-\beta n}/n}$ (Harte et al. 2008, Harte 2011). In the traditional parameterization of the log-series $\beta = -\log(p)$.

While E_0 does not appear in the resulting form of the abundance distribution, its presence in the model does influence the general shape of the distribution. If the only constraint included in this application of MaxEnt was the average abundance (N_0/S_0), and the uniform prior is maintained (Harte et al. 2008, 2009), then the resulting prediction is a geometric distribution.

Other applications of MaxEnt to this problem have also predicted log-series distributions but have used priors different from the uniform (Pueyo et al. 2007, Dewar and Porté 2008). Using MaxEnt with different detailed assumptions can also produce distributions that differ from the log-series (Haegeman and Etienne 2010), but alternative MaxEnt approaches to modeling the SAD tend to produce predictions that are generally similar in shape (Banavar et al. 2010, Frank 2011). Given these similarities, our results are likely relatively robust to the application of different MaxEnt approaches. A detailed exploration of the different maximum entropy models for the SAD would represent a valuable contribution to the field, but is beyond the scope of this manuscript. Some preliminary comparisons based on six datasets suggest that alternative formulations may perform even better than Harte et al.'s models (Banavar et al. 2010).

Fitting the predicted log-series distribution to empirical datasets

Harte and colleagues have developed several levels of approximation in calculating the log-series parameter *p* (Harte et al. 2008, 2009, Harte 2011). In our analysis we adopted Eq. 7.27 in (Harte 2011):

$$\frac{N_0}{S_0} = \frac{\sum_{n=1}^{N_0} e^{-\beta n}}{\sum_{n=1}^{N_0} e^{-\beta n}/n} \quad \text{(A.1)}$$

where $\beta = -\log(p)$. This formula is computationally more intensive but makes fewer approximation assumptions compared to the other formulas. Given S_0 and N_0, β was solved numerically for each community, and *p* was obtained by the back transformation $p = \exp(-\beta)$. This is equivalent to the maximum likelihood estimate of the log-series when the log-series is right truncated at N_0 (see below).

One way of assessing the predictive power of a theory for empirical SADs is by binning. However the choice of bin width is arbitrary, and power is lost by grouping data together. In our study we converted theoretical log-series distributions to rank-abundance distributions (RAD), i.e., the abundance of each species within the community from the most abundant to the least abundant, which were then compared to the empirical RADs. Log-series distributions were right truncated at N_0 to accommodate the implicit restriction that no species can have more individuals than the total number of individuals within the community. To obtain RADs, we utilized the concept of the cumulative density function (CDF), F*(n)*, i.e., the fraction of species with abundance no more than n. The predicted abundance for species with rank i (ranging from 1 to S_0) n_i is then calculated by solving $(S_0 - i + 0.5)/S_0 = F(n_i)$ (A.2) (Harte 2011). We present an example to illustrate how the procedure works (Appendix B, Fig. B1). In 2009, the observed species richness of BBS Route 1 in British Columbia is 42, with a total abundance across all species of 436 individuals. For this community, parameter p for the predicted log-series can be obtained by solving Eq. A1 to give 0.974. Fig. B1 (a) shows the CDF of the corresponding truncated distribution. To obtain abundance for each species, we solve Eq. A.2, which is equivalent to drawing a horizontal line for each rank and identifying the corresponding x value of the intersection (Harte 2011). The predicted RAD is shown in Fig. B1 (b).

Predictive power across sites and taxonomic groups

The ability of the maximum entropy model to predict observed abundance distributions is evaluated using observed-predicted plots and associated coefficients of determination (R^2). Instead of regressing the observed abundances with the predicted abundances, which can yield high R^2 values even with consistent deviations from the predictions, we calculated R^2 with respect to the one-to-one line, i.e., the proportion of variance among the observed abundances that is explained by the predicted abundances alone:

$$R^2 = 1 - \sum_{i=1}^{S_0}(n_{i-obs} - n_{i-pred})^2 \Big/ \sum_{i=1}^{S_0}(n_{i-obs} - \overline{n_{i-obs}})^2 \quad \text{(A.3)}$$

where n_{i-obs} and $n_{i\text{-pred}}$ are the log-transformed observed and predicted abundances of the i-th ranked species, respectively, and $\overline{n_{i-obs}}$ is the mean log-transformed observed abundance across all species and sites. All species from different sites within a given dataset (i.e., NABC, BBS, CBC, MCDB, FIA and Gentry) were pooled to calculate a single R^2, such that each point in an observed-predicted plot represents a single rank (i.e., a single species) at one site. As such, the associated R^2 value characterizes the proportion of variance in the abundance across ranks and across sites that can be determined using only information on the richness and abundance at each site. We also calculated R^2 values for each individual site. The results of the individual site analyses are presented as kernel density estimates of the distribution of R^2 values in the insets of Fig. 2.

Comparison between the MaxEnt prediction and the log-normal distribution

A variety of different forms of the log-normal distribution have been used in the literature. Here we chose the Poisson log-normal distribution, a log-normal distribution discretized by a Poisson sampling process (Bulmer 1974), which is generally considered the most appropriate form of the log-normal for fitting to abundance distributions (Connolly et al. 2005, McGill et al. 2007), even though it is not the most commonly used form. We compared the performance of the two distributions using Akaike Information Criterion (AIC) (Burnham and Anderson 2002, Connolly et al. 2005), which is a measure of likelihood while penalizing for complexity. We used $\mathrm{AIC_C}$, a second order variant of AIC that corrects for small sample size:

$$AIC_C = 2k - 2\log(L) + \frac{2k(k+1)}{n-k-1} \quad \text{(A.4)}$$

where L is the likelihood of the corresponding distribution, n is sample size, and k is the number of parameters.

Because right truncated likelihoods for the Poisson log-normal are not readily available, this analysis was conducted by treating both distributions as having no upper bound. For the log-series, the single parameter p was estimated using a form of the MaxEnt solution that is equivalent to the maximum likelihood estimate for the untruncated log-series (Eq. B.4 from (Harte et al. 2008) (see below)

$$\frac{N}{S} = \frac{p}{1-p} \cdot \frac{1}{\log\left(\frac{1}{1-p}\right)}$$

For the Poisson log-normal, the two parameters μ and σ were estimated with maximum likelihood using the approach described in Bulmer (1974), with code adapted from the Palamedes software package developed by the NCEAS working group on "Tools and fresh approaches for species abundance distributions." The $\mathrm{AIC_C}$ values of the two distributions were then converted to Akaike weights which indicate the relative probability that one distribution is superior compared to the other (Burnham and Anderson 2002). The results of this analysis are nearly identical if we use the MaxEnt parameter calculation that is used for the predictive power analyses (which assumes that the log-series distribution is truncated at N_0).

Deriving parameter estimate for log-series using maximum likelihood

In this section we derive the maximum likelihood estimator for the log-series parameter p, and demonstrate that the MaxEnt model developed by Harte and colleagues (Harte et al. 2008, 2009, Harte 2011) is compatible with the maximum likelihood method without making the explicit assumption.

1. Untruncated log-series

If the SAD is assumed to take the form of an untruncated log-series distribution, it follows that

$$\Pr\,(n|p) = \frac{-1}{\log\,(1-p)}\frac{p^n}{n}$$

where $\Pr(n|p)$ denotes the probability that a species has abundance n (i.e., SAD) given fitted parameter p. For a community of S species with abundances (n_1, n_2, …, n_S), the joint probability is

$$\Pr(n_1, n_2, \ldots, n_S|p) = \prod_{i=1}^{S} \Pr(n_i|p) = (\frac{1}{\log\left(\frac{1}{1-p}\right)})^S \cdot \frac{p^{\sum n_i}}{\prod n_i}$$

We then seek the maximum likelihood estimator $\hat{p}$ that maximizes the joint probability by setting its partial derivative to zero:

$$\frac{\partial \Pr(n_1, n_2, \ldots, n_S|p)}{\partial p}$$

$$= (\frac{1}{\log\left(\frac{1}{1-p}\right)})^S \cdot p^{\sum n_i - 1} \cdot \sum n_i + (\frac{1}{\log\left(\frac{1}{1-p}\right)})^{S+1} \cdot S \cdot p^{\sum n_i} \cdot (-\frac{1}{1-p}) = 0$$

$$\Rightarrow \sum_{i=1}^{S} n_i = S \cdot \frac{p}{1-p} \cdot \frac{1}{\log\left(\frac{1}{1-p}\right)}$$

Letting N denote $\sum_{i=1}^{S} n_i$, i.e., the total abundance for all species within the community, we have

$$\frac{N}{S} = \frac{p}{1-p} \cdot \frac{1}{\log\left(\frac{1}{1-p}\right)}$$

which is identical to Eq. B.4 derived by Harte and colleagues (Harte et al. 2008) using MaxEnt .

2. Truncated log-series

Alternatively, if the SAD is assumed to follow a log-series distribution truncated at $N = \sum_{i=1}^{S} n_i$, i.e., if no species can have an abundance higher than the total number of individuals within the community, we have

$$\Pr(n|p) = \frac{p^n}{n} / \sum_{i=1}^{N} \frac{p^i}{i}$$

The joint probability in this case is

$$\Pr(n_1, n_2, \ldots, n_S|p) = \prod_{i=1}^{S} \Pr(n_i|p) = \frac{p^{\sum n_i}}{\prod n_i} \cdot (\sum_{i=1}^{N} \frac{p^i}{i})^{-S}$$

Setting the partial derivative of the log-transformed joint probability to zero,

$$\frac{\partial \log\left(\Pr(n_1, n_2, \ldots, n_S|p)\right)}{\partial p} = \frac{N}{p} - \frac{S}{\sum_{i=1}^{N} \frac{p^i}{i}} \cdot \sum_{i=1}^{N} \frac{i \cdot p^{i-1}}{i} = 0$$

$$\Rightarrow \frac{N}{p} - \frac{S}{\sum_{i=1}^{N} \frac{p^i}{i}} \cdot \frac{1 - p^N}{1 - p} = 0$$

$$\Longrightarrow \frac{S}{N} \cdot \frac{p - p^{N+1}}{1 - p} = \sum_{i=1}^{N} \frac{p^i}{i}$$

which is identical to Eq. 3 derived by Harte and colleagues (Harte et al. 2008) using MaxEnt . Because the primary constraints in the Maximum Entropy model are “soft constraints” (*sensu* Haegeman and Etienne 2010), it is valid to treat the probabilities as independent for each species and therefore to treat the likelihoods in the usual way as done here and in the statistical analysis.

LITERATURE CITED

Banavar, J. R., A. Maritan, and I. Volkov. 2010. Applications of the principle of maximum entropy: from physics to ecology. Journal of Physics: Condensed Matter 22:063101.

Bulmer, M. G. 1974. On Fitting the Poisson Lognormal Distribution to Species-Abundance Data. Biometrics 30:101–110.

Burnham, K. P., and D. Anderson. 2002. Model Selection and Multi-Model Inference, Second edition. Springer.

Connolly, S. R., Hughes, T. P., Bellwood, D. R., and Karlson, R. H. 2005. Community Structure of Corals and Reef Fishes at Multiple Scales. Science 309:1363–1365.

Dewar, R. C., and A. Porté. 2008. Statistical mechanics unifies different ecological patterns. Journal of Theoretical Biology 251:389–403.

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

Haegeman, B., and R. S. Etienne. Entropy maximization and the spatial distribution of species. American Naturalist 175:E74-E90.

Harte, J. 2011. Maximum entropy and ecology. Oxford University Press, Oxford, UK.

Harte, J., A. B. Smith, and D. Storch. 2009. Biodiversity scales from plots to biomes with a universal species–area curve. Ecology Letters 12:789–797.

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

Jaynes, E. T.. 2003. Probability theory: the logic of science. Cambridge University Press, Cambridge, UK.

Marks, C. O., and H. C. Muller-Landau. 2007. Comment on “From Plant Traits to Plant Communities: A Statistical Mechanistic Approach to Biodiversity”. Science 316:1425c.

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, A. H. Hurlbert, A. E. Magurran, P. A. Marquet, B. A. Maurer, A. Ostling, C. U. Soykan, K. I. Ugland, and E. P. White. 2007. Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecology Letters 10:995–1015.

National Audubon Society. 2002. The Christmas Bird Count historical results. Retrieved from http://www.audubon.org/bird/cbc.

North American Butterfly Association. 2009. NABA Butterfly Counts: 2009 Report, http://www.naba.org.

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, USA.

Preston, F. W. 1948. The Commonness, And Rarity, of Species. Ecology 29:254–283.

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

Sauer, J. R., J. E. Hines, J. E. Fallon, D. J. Pardieck, D. J. Ziolkowski, Jr., and W. A. Link. 2011. The North American Breeding Bird Survey 1966-2009. Version 3.23.2011. USGS Patuxent Wildlife Research Center, Laurel, Maryland, USA.

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

U.S. Department of Agriculture, F. S. 2010. Forest inventory and analysis national core field guide (Phase 2 and 3), version 4.0. Washington, DC: U.S. Department of Agriculture Forest Service, Forest Inventory and Analysis.

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

Woudenberg, S. W., B. L. Conkling, B. M. O'Connell, E. B. LaPoint, J. A. Turner, K. L. Waddell. 2010. The Forest Inventory and Analysis Database: Database description and users manual version 4.0 for Phase 2. Gen. Tech. Rep. RMRS-GTR- 245. Fort Collins, CO: U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station. 336 p.