Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2015

doi: 10.1111/2041-210X.12349

Estimating diversity and entropy profiles via discovery rates of new species

Anne Chao[1]* and Lou Jost[2]

[1]*Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan; and* [2]*EcoMinga Foundation, Via a Runtun, Baños, Tungurahua, Ecuador*

Summary

1. The compositional complexity of an assemblage is not expressible as a single number; standard measures such as diversities (Hill numbers) and entropies (Rényi entropies and Tsallis entropies) vary in their order q which determines the measures' emphasis on rare or common species. Ranking and comparing assemblages depend on the choice of q. Rather than selecting one or a few measures to describe an assemblage, it is preferable to convey the complete story by presenting a continuous profile, a plot of diversity or entropy as a function of $q \geq 0$. This makes it easy to visually compare the compositional complexities of multiple assemblages and to judge the evenness of the relative abundance distributions of the assemblages. In practice, the profile is plotted for all values of q from 0 to $q = 3$ or 4 (beyond which it generally changes little).

2. These profiles are usually generated by substituting species sample proportions into the complexity measures. However, this empirical approach typically underestimates the true profile for low values of q, because samples usually miss some of the assemblage's species due to under-sampling. Although bias-reduction methods exist for individual measures of order $q = 0$, 1 and 2, there has been no analytic method that unifies these bias-corrected estimates into a continuous profile.

3. For incomplete sampling data, this work proposes a novel analytic method to obtain accurate, continuous, low-bias diversity and entropy profiles with focus on low orders of q ($0 \leq q \leq 3$). Our approach is based on reformulating the diversity and entropy of any order q in terms of the successive discovery rates of new species with respect to sample size, that is the successive slopes of the species accumulation curve. A bootstrap method is applied to obtain approximate variances of our proposed profiles and to construct the associated confidence intervals.

4. Extensive simulations from theoretical models and real surveys show that the proposed profiles greatly reduce under-sampling bias and have substantially lower bias and mean-squared error than the empirical profile, especially for $q \leq 1$. Our method is also extended to deal with incidence data.

Key-words: diversity, diversity profile, entropy profile, Hill numbers, Rényi entropy, species accumulation curve, species discovery rate, Tsallis entropy

Introduction

Measures of compositional complexity are key elements of a biologist's theoretical and empirical toolbox. There are many such measures. The most widely used families of these measures include Hill numbers (diversity), Tsallis entropies, Rényi entropies and others. Rényi (1961) integrated several classic complexity measures into a class of indices, now known as Rényi entropies in many disciplines. Hill numbers (or effective number of species) were first used in ecology by MacArthur (1965), developed by Hill (1973) and reintroduced to ecologists by Jost (2006, 2007). Hill numbers were recently extended to incorporate species phylogeny and/or function (Leinster & Cobbold 2012; Scheiner 2012; Chao, Chiu & Jost 2010, 2014). Although Tsallis entropies were introduced in physics by Tsallis (1988), they had been discovered earlier in other fields, so they are also known as HCDT entropies, to recognize all four discovering authors (Havrda & Charvát 1967; Daróczy 1970; Tsallis 1988). Patil & Taillie (1982) independently discovered them in the statistical literature. These three important families are monotonic transformations of each other.

All these measures (Hill numbers, Tsallis entropies and Rényi entropies) are parameterized by the order q, where q is any real number greater than or equal to zero. Throughout the paper, entropy refers to Tsallis entropy and/or Rényi entropy. The parameter q determines the measure's sensitivity to species relative abundances. Measures using $q = 0$ count species equally without regard to their relative abundances, while measures using high values of q discount all but the dominant species. For a given data set, biologists often report just one or a few of these complexity measures, usually those based on $q = 0$, 1 and/or 2. However, conclusions about such things as ranking of assemblages, or changes in a single assemblage over

*Correspondence author. E-mail: chao@stat.nthu.edu.tw

time, depend on the choice of q used in the analysis. For this reason, Hill (1973), Tóthmérész (1995), Gotelli & Chao (2013), Chao *et al.* (2014) and others suggested that biologists should instead use all the information contained in their data, by plotting the complexity measure as a continuous function of q. This profile for $q \geq 0$ contains all the information in the species relative abundance distribution. If profiles of two assemblages do not cross, then one of the assemblages is unambiguously more diverse than the other. If they cross, only statements conditional on q can be made about their ranking. Profiles not only provide a complete account of the rankings of assemblages, they also provide complete information about the evenness of the relative abundance distribution of a single assemblage. Standard evenness measures are functions of profile slopes (Jost 2010). In most applications, the diversity and entropy profiles are plotted for all values (including non-integers) of q from 0 to $q = 3$ or 4, beyond which it generally does not change much and thus contains little information on compositional complexity.

All profiles are meant to characterize the true population values of the chosen complexity measures in an assemblage. In practice, the true values of the chosen measures are unknown and must be estimated from samples. To plot a continuous profile, empirical profiles are obtained by substituting species sample proportions into the complexity measure. Empirical profiles generally work well for large values of order q (say, $q > 3$) because these measures are determined by the dominant species, which always appear in samples, if sample size is not unreasonably small. However, for low orders of q (especially for $0 \leq q \leq 3$), empirical profiles typically underestimate the true population profiles due to the effect of the undetected species in samples; in the region $0 \leq q \leq 1$, the biases are substantial for severely under-sampled cases (Magurran 2004, p. 107). Under-sampling is a major source of bias for the empirical profiles of low orders; the magnitude of this bias increases as the proportion of undetected species increases. In biodiversity studies, under-sampling is a common problem. It is important to reduce or remove the bias inherent in the empirical measures due to incomplete samples (see Gotelli & Colwell (2001) and Beck & Schwanghart (2010) for related examples and statistical inferences).

In this paper, we focus on the estimators of diversity profiles expressed as Hill numbers. Simple transformations of Hill numbers lead to Tsallis entropies and Rényi entropies, if desired. Diversity of order $q = 0$ is the species richness of an assemblage. The empirical diversity of $q = 0$ is the observed species richness in a sample, well known for its negative bias. A wide range of estimators have been proposed to reduce this bias (Chao 2005; O'Hara 2005). Diversity of order $q = 1$ is the exponential of Shannon entropy. The empirical diversity of $q = 1$ in a sample is also well known to exhibit negative bias due to the effect of undetected species in samples. Its estimation is surprisingly non-trivial as shown by Chao, Wang & Jost (2013). For $q = 2$, improved estimators include the usual nearly unbiased estimator (Gotelli & Chao 2013) and a specific estimator (Nielsen, Tarpy & Reeve 2003). Although bias-reduction methods exist for individual measures of orders $q = 0$, 1 and 2, there has been no analytical method that unifies these bias-corrected estimates into a *continuous* profile, but see Chao *et al.* (2015, their appendix E) for a recent method that requires solving intensive numerical computations and estimating the species-rank abundance distribution.

This work proposes a novel analytic method to obtain a continuous, accurate and low-bias profile from incomplete sampling data. We mainly focus on the case $0 \leq q \leq 3$ because this is the range of orders that diversity and entropy profiles are usually plotted and the range which contains nearly all useful information; it is also the range in which the empirical profiles typically underestimate and differ from our proposed estimates. In our approach, we first reformulate Hill numbers and entropies of order q in terms of the successive discovery rates of new species with respect to sample size, which are the successive slopes of the sample size-based species accumulation curve. By applying slope estimators derived from an improved Good–Turing frequency formula (Chao & Jost 2012), we obtain a more accurate continuous diversity and entropy profile. Our estimation takes into account the effect of undetected species in samples. A bootstrap method is applied to obtain approximate variances of our proposed profiles and to construct the associated confidence intervals.

The performance of the proposed profile is examined by data sets generated from theoretical species abundance models and real large surveys. We compare the proposed profile estimator with the empirical profile in terms of bias and accuracy (mean-squared error) for $0 \leq q \leq 3$. The simulations reveal that our proposed profile estimator always improves the empirical profiles in this range of q, and the improvement can be substantial for severely under-sampled data. The negative bias due to undetected species can be greatly reduced. We illustrate the application of our formulas and estimators using real data sets, which demonstrate that the empirical and proposed diversity profile may give qualitatively different answers when comparing biodiversity surveys.

Methods

COMPOSITIONAL COMPLEXITY IN TERMS OF DISCOVERY RATE OF NEW SPECIES

We first link Hill numbers, Rényi entropies and Tsallis entropies to the species accumulation curve (SAC) and then formulate them in terms of the successive slopes of the species accumulation curve. Assume that there are S species in an assemblage, with true species relative abundances $\{p_1, p_2, \ldots, p_S\}$. Let $S(k)$ be the expected number of species in a random sample of k individuals taken (with replacement) from the assemblage. Given species relative abundances, Good (1953) derived the expression for $S(k)$ as a function of sample size k as follows:

$$S(k) = \sum_{i=1}^{S} [1 - (1 - p_i)^k] = S - \sum_{i=1}^{S} (1 - p_i)^k, \; k = 0, 1, 2, \ldots, \qquad \text{eqn 1}$$

with $S(0) = 0$ and $S(1) = 1$. The sample size-based SAC plots the expected species richness $S(k)$ vs. sample size k. The horizontal asymptote of this curve as k tends to infinity is the true species richness. Based on eqn 1, the slope of the line connecting two adjacent points $(k, S(k))$ and $(k+1, S(k+1))$ can be expressed as follows:

$$\Delta(k) = \frac{S(k+1) - S(k)}{(k+1) - k} = \sum_{i=1}^{S} p_i (1 - p_i)^k, \qquad \text{eqn 2}$$

with the initial value $\Delta(0) = 1$. The curve's successive slopes show the rates at which new species are expected to be detected in the sampling process. The slope $\Delta(k)$ is a decreasing function of k, implying the expected rate declines as sample size is increased. Since the slope $\Delta(k)$ is also the complement of the expected sample coverage of a sample of size k, Chao & Jost (2012) referred to the slope as the 'coverage deficit'. Sample coverage, originally developed by Alan Turing (Good 1953, 2000), is the fraction of the population belonging to species represented in the sample; it is an objective measure of sample completeness. The coverage deficit is an aspect of undetected species that can be accurately estimated by sample data (Good 1953; Lande, DeVries & Walla 2000).

We now express Hill numbers and entropies of order q in terms of the slopes of the SAC. By convention, we define $0! \equiv 1$, and define for a real number u and a positive integer k as follows:

$$\binom{u}{0} \equiv 1;$$

$$\binom{u}{k} \equiv u(u-1)(u-2)\ldots(u-k+1)/k!.$$

In the special case that u is a non-negative integer, then

$$\binom{u}{k} = \begin{cases} \frac{u!}{k!(u-k)!} & k \le u; \\ 0 & k > u. \end{cases}$$

We first write the qth order basic sum of the species relative abundances as follows:

$$\sum_{i=1}^{S} p_i^q = \sum_{i=1}^{S} p_i [1 - (1 - p_i)]^{q-1}$$

$$= \sum_{i=1}^{S} \sum_{k=0}^{\infty} p_i \binom{q-1}{k} (-1)^k (1 - p_i)^k = \sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k). \qquad \text{eqn 3a}$$

Thus, Tsallis entropy of order q is a linear function of the slopes of the SAC as follows:

$${}^qH_{\text{Tsallis}} = \frac{1}{1-q}\left(\sum_{i=1}^{S} p_i^q - 1\right) = \frac{1}{1-q}\left[\sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k) - 1\right]. \qquad \text{eqn 3b}$$

Rényi entropy of order q can be expressed as follows:

$${}^qH_{\text{Rényi}} = \frac{1}{1-q} \log\left(\sum_{i=1}^{S} p_i^q\right) = \frac{1}{1-q} \log\left[\sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k)\right]. \qquad \text{eqn 3c}$$

In this paper, we mainly focus on the estimation of Hill numbers of order q, which can be expressed as nonlinear functions of the slopes of the SAC as follows:

$${}^qD = \left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)} = \left(\sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k)\right)^{1/(1-q)}, \; q \ge 0. \qquad \text{eqn 4a}$$

Below we give three special cases of Hill numbers and a general form (see Appendix S1 for proofs):

1. For $q = 0$, eqn 4a reduces to the following equation which implies that species richness is the sum of the slopes. This is another way of expressing that S is the asymptote of the SAC:

$${}^0D = S = \sum_{k=0}^{\infty} \Delta(k). \qquad \text{eqn 4b}$$

2. As q tends to unity, the diversity of order 1 (i.e. exponential of Shannon entropy or Shannon diversity) can be expressed as the exponential of a harmonic infinite sum of the successive slopes as follows:

$${}^1D = \lim_{q \to 1}\left(\sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k)\right)^{1/(1-q)} = \exp\left(\sum_{k=1}^{\infty} \frac{1}{k} \Delta(k)\right). \qquad \text{eqn 4c}$$

3. When q is an integer ≥ 2, eqn 4a becomes a sum of finite terms as follows:

$${}^qD = \left(\sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^k \Delta(k)\right)^{1/(1-q)}. \qquad \text{eqn 4d}$$

For the special case of $q = 2$ and 3, we have ${}^2D = [\Delta(0) - \Delta(1)]^{-1}$ and ${}^3D = [\Delta(0) - 2\Delta(1) + \Delta(2)]^{-1/2}$. These numbers have an elegant link to the finite differences of the slopes (see the expression derived in Appendix S1 for a general order q).

4. For any value of $q \ge 0$, the general formula in eqn 4a is a sum of an infinite number of terms: **(4e)**

$${}^qD = \left(\Delta(0) - (q-1)\Delta(1) + \ldots + (-1)^k \frac{(q-1)(q-2)\ldots(q-k)}{k!} \Delta(k) + \ldots\right)^{1/(1-q)}. \qquad \text{eqn 4e}$$

DIVERSITY PROFILE ESTIMATOR

Our new formulation of diversity (eqn 4a) and entropies (eqns 3b and 3c) opens a novel way to estimate diversity and entropy profiles via discovery rates of new species. Here, we illustrate our method by estimating diversity profiles. Since the expected slopes of a species accumulation curve can be accurately estimated when the sample size is sufficiently large (Chao & Jost 2012), an estimated diversity profile for any $q \ge 0$ can be obtained via estimation of these slopes. Based on a sample of fixed size n, we separate the infinite sum in eqn 4a into two parts: the first part with $k < n$ and the second part with $k \ge n$ as follows:

$${}^qD = \left(\sum_{k=0}^{n-1} \binom{q-1}{k} (-1)^k \Delta(k) + \sum_{k=n}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k)\right)^{1/(1-q)}. \qquad \text{eqn 5}$$

The reason for this separation is because the first sum can be estimated without bias, whereas there exists no unbiased estimator for the second sum. Under the commonly used multinomial model, in which species frequencies $(X_1, X_2, \ldots, X_S)$ follow a multinomial distribution with cell total $\sum_{x_i > 0} X_i = n$ and cell probabilities $\{p_1, p_2, \ldots, p_S\}$, we separately estimate each sum in eqn (5).

For the first sum, it is known from statistical theory that the minimum variance unbiased estimator for the expected slope $\Delta(k)$ exists for any size k less than n. This unbiased estimator for the first part is as follows (Chao & Jost 2012):

$$\hat{\Delta}(k) = \sum_{1 \le X_i \le n-k} \frac{X_i}{n} \frac{\binom{n-X_i}{k}}{\binom{n-1}{k}} = \sum_{1 \le X_i \le n-k} \frac{\binom{n-k-1}{X_i-1}}{\binom{n}{X_i}}, \quad k < n. \qquad \text{eqn 6a}$$

Here, $\hat{\Delta}(0) = 1$. The second sum in eqn (5) involves the expected slopes for sample sizes greater than n, and no unbiased estimator exists. This part is usually dominated by rare undetected species whose effect on diversity cannot be ignored, especially when the sample size is much smaller than the number of species. Thus, the most difficult part of our diversity profile estimation is accurately estimating this second sum.

We follow Chao, Wang & Jost (2013) and use the slope estimators to approximate the second sum. Chao & Jost (2012) derived slope estimators for successive slopes $\Delta(k)$ at size $k \ge n$ based on the wisdom of Turing and Good (Good 1953, 2000). The core idea in Turing and Good's approach is that the singletons and doubletons in samples provide the most essential information about the undetected species. Let f_1 denote the number of singletons and f_2 denote the number of doubletons in the sample. Good–Turing's original frequency formula implies that the estimated mean relative frequency of the singletons in the population is not $1/n$, but $2f_2/(nf_1)$, contrary to most people's intuition. When n is sufficiently large, Chao & Jost (2012) derived a more accurate modification of Good–Turing's frequency formula for singletons; we refer to this estimated mean relative frequency as A with the following expression:

$$A = \begin{cases} 2f_2/[(n-1)f_1 + 2f_2], & \text{if } f_2 > 0; \\ 2/[(n-1)(f_1-1)+2], & \text{if } f_2 = 0,\ f_1 \ne 0; \\ 1, & \text{if } f_2 = f_1 = 0. \end{cases} \qquad \text{eqn 6b}$$

Based on these formulas for A, an estimator for the slope at sample size $n + m$ is as follows (Chao & Jost 2012):

$$\hat{\Delta}(n+m) = \frac{f_1}{n}(1-A)^{m+1},\ m \ge 0. \qquad \text{eqn 6c}$$

Substituting the slope estimators (eqns 6a and 6c) into eqn (5), we obtain the following diversity estimator of order q as follows:

$${}^q\hat{D} = \left(\sum_{k=0}^{n-1} \binom{q-1}{k} (-1)^k \hat{\Delta}(k) + \sum_{k=n}^{\infty} \binom{q-1}{k} (-1)^k \hat{\Delta}(k) \right)^{1/(1-q)}. \qquad \text{eqn 6d}$$

Although the estimator is expressed as an infinite sum, eqn 6d is an estimator with finite terms as described below (All proofs are provided in Appendix S1).

(1) For $q = 0$, eqn 6d gives a species richness estimator ${}^0\hat{D} = \hat{S} = \sum_{k=0}^{\infty} \hat{\Delta}(k) = \sum_{k=0}^{n-1} \hat{\Delta}(k) + \sum_{k=n}^{\infty} \hat{\Delta}(k)$, and we prove that $\sum_{k=0}^{n-1} \hat{\Delta}(k) = S_{obs}$ (the observed species richness in the sample) and $\sum_{k=n}^{\infty} \hat{\Delta}(k) = \hat{f}_0$, where

$$\hat{f}_0 = \begin{cases} \frac{(n-1)}{n} \frac{f_1^2}{(2f_2)}, & \text{if } f_2 > 0; \\ \frac{(n-1)}{n} \frac{f_1(f_1-1)}{2}, & \text{if } f_2 = 0 \end{cases} \qquad \text{eqn 7a}$$

Thus, the species richness estimator becomes the Chao1 estimator (Chao 1984). This estimator is actually a lower bound of species richness. Extensive simulations (Chiu *et al.* 2014) have suggested the Chao1 lower bound is preferable to some widely used estimators.

(2) For $q = 1$, eqn 6d reduces to the exponential of the Shannon entropy estimator derived in Chao, Wang & Jost (2013): **(7b)**

$${}^1\hat{D} = \exp\left(\sum_{1 \le X_i \le n-1} \frac{X_i}{n} \left(\sum_{k=X_i}^{n-1} \frac{1}{k} \right) + \frac{f_1}{n}(1-A)^{-n+1} \left[-\log A - \sum_{r=1}^{n-1} \frac{(1-A)^r}{r} \right] \right). \qquad \text{eqn 7b}$$

(3) If q is an integer between 2 and max X_i, where max X_i denotes the maximum species frequency, then our estimator becomes the nearly unbiased estimator as follows (Gotelli & Chao 2013):

$$\begin{aligned} {}^q\hat{D} &= \left(\sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^k \hat{\Delta}(k) \right)^{1/(1-q)} \\ &= \left(\sum_{X_i \ge q} \frac{X_i(X_i-1)\ldots(X_i-q+1)}{n(n-1)\ldots(n-q+1)} \right)^{1/(1-q)}. \end{aligned} \qquad \text{eqn 7c}$$

(4) For any value of q up to max X_i, the general formula for our estimator is as follows: **(7d)**

$${}^q\hat{D} = \left(\sum_{k=0}^{n-1} \binom{q-1}{k} (-1)^k \hat{\Delta}(k) + \frac{f_1}{n}(1-A)^{-n+1} \left[A^{q-1} - \sum_{r=0}^{n-1} \binom{q-1}{r} (A-1)^r \right] \right)^{1/(1-q)}. \qquad \text{eqn 7d}$$

When $q = 0$ and 1, eqn 7d reduces, respectively, to the formula described above. Note that if q is an integer > 1, then $A^{q-1} - \sum_{r=0}^{n-1} \binom{q-1}{r} (A-1)^r = 0$ and eqn 7d reduces to eqn 7c.

Our proposed diversity profile estimator is the plot of ${}^q\hat{D}$ with respect to q for $0 \le q \le 3$ in most applications. This profile is a continuous function q and its performance is examined in the next section (see Discussion for its behaviour for higher orders). The variance of the proposed estimator can be estimated by a bootstrap method, which is a modified version of the bootstrap method proposed in Chao, Wang & Jost (2013). Details are provided in Appendix S2. The resulting variance estimate can then be used to construct a confidence interval for the diversity profile to reflect sampling uncertainty.

TRANSFORMATIONS TO ENTROPY PROFILES

The diversity profiles based on Hill numbers are the most useful profiles in many applications. Hill numbers obey the replication principle, an important mathematical property that implicitly underlies most biological thinking about diversity (Jost 2006, 2007; Ellison 2010; Chao *et al.* 2014); if K equally large, equally diverse assemblages with no shared species are pooled, then the diversity of the pooled assemblages equals K times the diversity of an individual

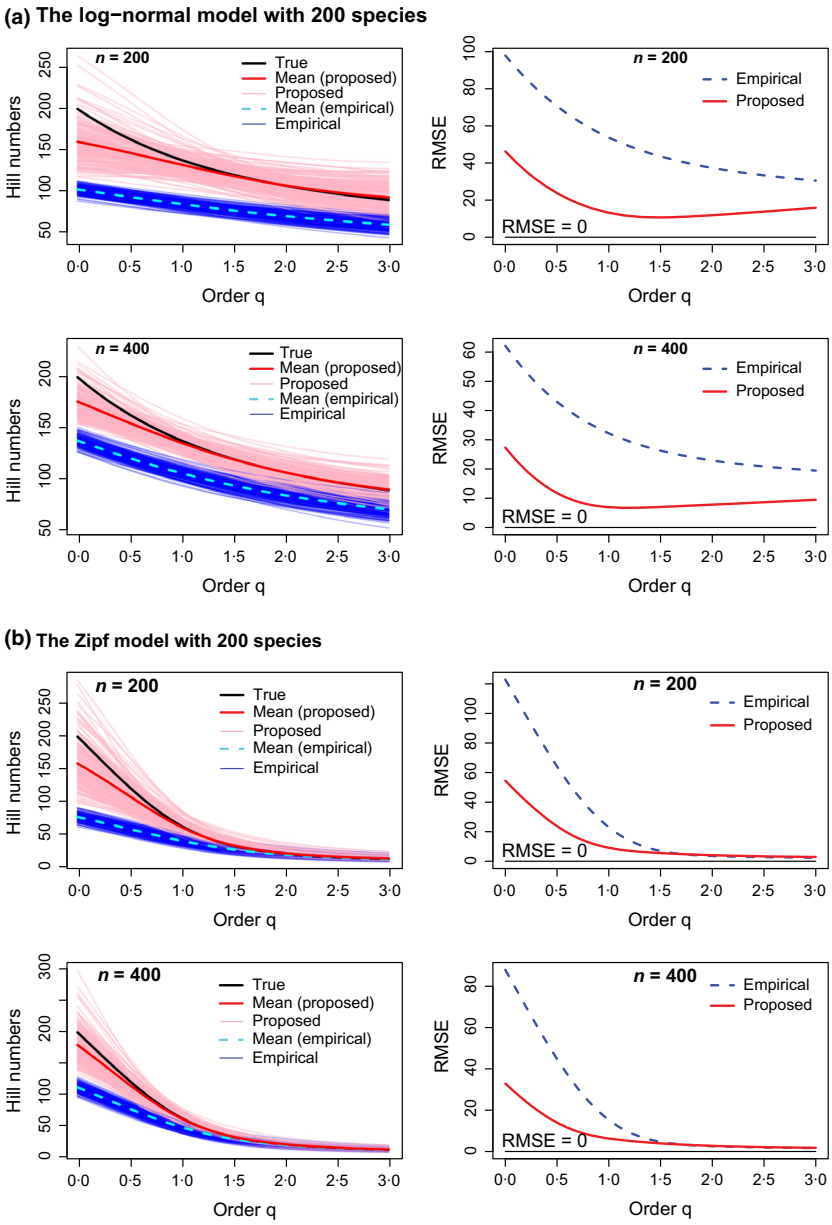

Fig. 1. Simulation comparison of the empirical and proposed diversity profiles for two theoretical abundance models with 200 species: (a) the log-normal model with 200 species, CV = 0·937 and (b) the Zipf model with 200 species, CV = 2·947. For each model, 200 data sets of size 200 and 400 were generated; thus, there are 200 empirical and proposed diversity profiles for each size. Left panels compare the true diversity profile (black line), empirical diversity profiles (superimposed blue lines with 200 replications; light blue line for the mean profile over 200 replications) and the proposed estimated diversity profiles (superimposed pink lines with 200 replications; red line for the mean profile over 200 replications). The right panels compare the root-mean-squared errors (RMSEs) of the empirical and proposed diversity profiles based on 200 replications.

assemblage. This linearity with respect to pooling makes Hill numbers behave properly in direct comparisons of magnitudes and in ratios of within group to total diversity.

Neither Tsallis entropies nor Rényi entropies satisfy the replication principle. The use of them as diversity measures may lead to uninterpretable results (Jost 2006, 2007). Nevertheless, the two classes of measures are useful in other disciplines. It follows from eqn 3b that Tsallis entropy of any order of q is a linear function of slopes, and its estimation is direct by substituting slope estimators. Based on eqn 7d, the general formula for the Tsallis entropy estimator is $({}^{q}\hat{D}^{1-q}-1)/(1-q)$ $(q \neq 1)$ and $\log({}^{1}\hat{D})$ $(q = 1)$. Applying eqns 7a, 7b and 7c, we can obtain some special cases. Note, here an unbiased estimator for Tsallis entropy of integer order $q = 2, 3, \ldots, n$ is given by:

$$ {}^{q}\hat{H}_{Tsallis} = \frac{1}{1-q}\left(\sum_{X_i \geq q} \frac{X_i(X_i-1)\ldots(X_i-q+1)}{n(n-1)\ldots(n-q+1)} - 1\right). \qquad \text{eqn 7e} $$

Statistical theory implies that no unbiased estimator exists for $q > n$.

Based on eqn 3c, the Rényi entropy profile can be estimated by taking the logarithm of our estimated diversity at each point. The resulting profile is only nearly unbiased, due to the log transformation. The values of the Rényi profile are more difficult to interpret than the values of a diversity profile, because they do not obey the replication principle. However, Rényi entropies do have an interesting related property. Suppose we have K assemblages that share no species but whose species relative abundance distributions are exact replicates of each other. Pool any number of these assemblages. The Rényi entropy profile of the pooled assemblage will have exactly the same shape as the Rényi profile of each of the single assemblages, but will be shifted upward.

SIMULATION

We simulated data from several theoretical species abundance models and large surveys which are treated as the 'true' entire assemblages. We report here the representative results for two models and two surveys (more simulation scenarios are described in Appendix S3 and the simulation results are provided in Appendix S4). In each case, we also give the CV (coefficient of variation, which is the

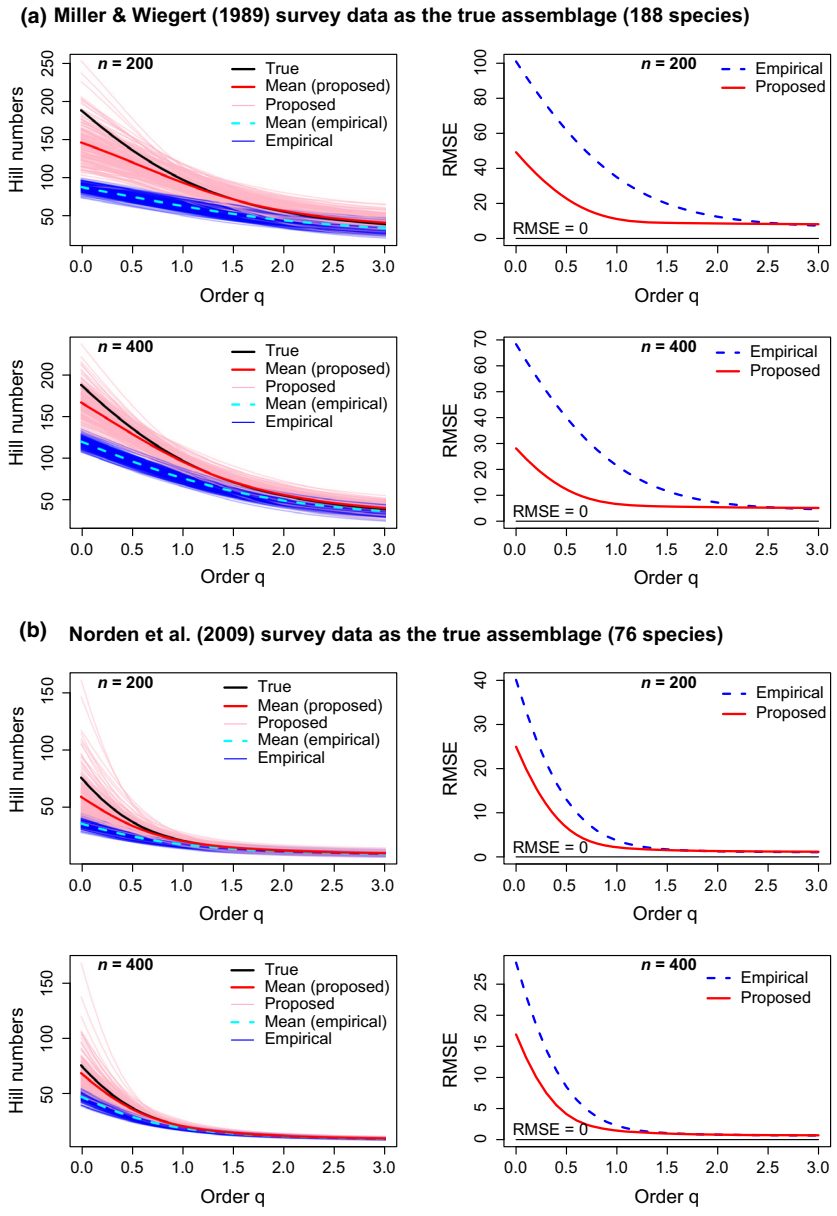

Fig. 2. Simulation comparison of the empirical and proposed diversity profiles for two surveys which are used to define the complete true assemblages: (a) Miller & Wiegert (1989) plant species survey with 188 species, CV = 1·563. (b) Norden *et al.* (2009) tree species survey with 76 species, CV = 2·305. For each assemblage, 200 data sets of size 200 and 400 were generated (see captions of Fig. 1 for explanation of all plots).

ratio of standard deviation and mean) of $(p_1, p_2, \ldots, p_S)$. The CV value quantifies the degree of heterogeneity of the species relative abundances $(p_1, p_2, \ldots, p_S)$. When all species are equally abundant, CV = 0. A larger value of CV signifies higher degree of heterogeneity among species relative probabilities. We here use the following two theoretical abundance models:

1. The log-normal model with 200 species.
   We first generated 200 random variables $(a_1, a_2, \ldots, a_{200})$ from a log-normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. The species relative abundances take the form $p_i = ca_i$, where c is a normalizing constant such that the sum of the relative abundances is 1. Then, the species relative abundances $(p_1, p_2, \ldots, p_{200})$ represent the true species relative abundances of the complete assemblage; CV = 0·937 for the set we used in our simulations.
2. The Zipf model with 200 species (Zipf 1965).
   The relative abundances of the complete assemblage take the general form $p_i = c/i$, $i = 1, 2, \ldots, 200$, where c is a normalizing constant. CV = 2·947.

We also report the results for the following two surveys which are used to define the complete true assemblages in our simulations:

1. Miller & Wiegert (1989) plant species survey with 188 species.
   We considered the 150-year field observations (Miller & Wiegert 1989) for endangered and rare vascular plant species in the central portion of the southern Appalachian region as the true entire assemblage. The species abundance frequency counts for this survey are reproduced in Table S3.1 (Appendix S3); a total of 188 species were represented by 1008 individuals with CV = 1·563.
2. Norden *et al.* (2009) tree species survey with 76 species.
   We considered the tree assemblage in LSUR younger (21 years) second-growth forest site in north-eastern Costa Rica (Norden *et al.* 2009; Colwell *et al.* 2012). The species abundance frequency counts for this survey are reproduced in Table S3.2 (Appendix S3); a total of 76 species were represented by 1020 individuals with CV = 2·305.

For each model or assemblage, we generated 200 data sets for sample sizes of 200 and 400. The smaller size was chosen so that a large portion

Table 1. Beetle species frequency counts in two sites (Janzen 1973). Osa second-growth site (140 species, $n = 976$); Osa old-growth site (112 species, $n = 237$). The species abundance frequency count f_i denotes the number of species represented by exactly i individuals in the survey

Osa second-growth site						Osa old-growth site	
i	f_i	i	f_i	i	f_i	i	f_i
1	70	11	3	57	2	1	84
2	17	12	2	60	1	2	10
3	4	14	2	64	1	3	4
4	5	17	1	71	1	4	3
5	5	19	2	77	1	5	5
6	5	20	3			6	1
7	5	21	1			7	2
8	3	24	1			8	1
9	1	26	1			14	1
10	2	40	1			42	1

of species is undetected; this provides a severe challenge for any estimator and shows that ours can handle the challenge. For each generated data set, we obtained the true diversity profile, the empirical diversity profile and our new proposed diversity profile for q between 0 and 3. See Fig. 1 for the simulation results using the two models and Fig. 2 for the two real surveys. For each model/assemblage and each sample size, there are 200 superimposed empirical diversity profiles (blue lines, with each line corresponding to an empirical diversity profile for each generated data set) and their mean profile (by averaging over the 200 values for each q, light blue line). There are also 200 superimposed diversity profiles using our proposed estimator (pink lines), and their mean profile (by averaging over 200 values for each q, red line). All these are compared to the true diversity profile (black line). For each q between 0 and 3, we also calculated the root-mean-squared error based on the 200 diversity estimates; the plots of the RMSE as a function of q are also shown in Figs 1 and 2.

All empirical diversity profiles (blue lines) for $0 \le q \le 3$ are typically below the true curves and thus exhibit negative biases; the biases for $q \le 1$ are substantial for $n = 200$. In such cases, our estimator (pink lines) greatly reduces the under-sampling bias and always outperforms the empirical profile estimator. When q is not close to 0, our method produces nearly unbiased diversity estimates (which fluctuate below and above the true profile). Note that when q is equal to 0, our estimate is the Chao1 index, which is an estimated lower bound of true species richness. The plots of RMSE reveal that our proposed estimator is more accurate than the empirical diversity estimator for nearly all values of q considered here. Consistent findings are also revealed by other simulation results provided in Appendix S4 for $0 \le q \le 3$ (see Discussion about comparison for higher orders).

Applications

We illustrate the use of our diversity profile estimator by analysing two beetle data sets (Osa second-growth site and Osa old-growth site) obtained from sweep samples in Costa Rica (Janzen 1973). The abundance frequency counts for the two sites are given in Table 1. In the Osa second-growth site, Janzen found 140 species among 976 individuals; the number of singletons was $f_1 = 70$, and the number of doubletons was $f_2 = 17$. The sample coverage estimate (Chao & Jost 2012) for the sample is 93% (SE 0·74%). In the Osa old-growth site, there were 112 species, 237 individuals, and $f_1 = 84$ and $f_2 = 10$, yielding a coverage estimate of 65% (SE 3·5%). This implies that the sample completeness for the old-growth site is much lower than that for the second-growth site.

In Table 2, we give for each site the empirical and proposed diversity estimates and their standard errors (SE) for $q = 0$ to 3 in increments of 0·5. An expanded table covering the case $0 \le q \le 10$ is provided in Appendix S5. The corresponding continuous profile plots along with 95% confidence intervals are shown in Fig. 3. All SE in Table 2 and the confidence intervals in Fig. 3 were calculated based on a bootstrap method of 1000 replications (Appendix S2). The confidence intervals are quite wide for the old-growth site due to relatively sparse data. Nevertheless, our estimated profile paints a very different picture than the empirical profile, and the differences are statistically significant.

According to the empirical diversity profiles, the second-growth site is more diverse than the old-growth site for $q = 0$ (140 vs. 112 species). For $q > 0$, the two empirical profiles are not distinguishable because the two 95% confidence intervals overlap. Our proposed diversity profiles reverse this assessment, implying instead that the true population values of the diversities in old-growth forest are much higher than those of second-growth forest when $q < 1{\cdot}3$. For q between 0·2 and 1·3, the differences are statistically significant, as reflected by the two non-overlapped confidence intervals. The curves only

Table 2. The empirical and proposed diversity estimates for $q = 0$ to 3 in increments of 0·5 (with SE based on a bootstrap method of 1000 replications) for beetle species data in the old-growth and second-growth sites (Janzen 1973)

	Old-growth site		Second-growth site	
Diversity order	Empirical diversity (SE)	Proposed diversity (SE)	Empirical diversity (SE)	Proposed diversity (SE)
$q = 0$	112·0 (6·4)	463·3 (106·1)	140·0 (6·6)	284·0 (42·1)
$q = 0{\cdot}5$	86·3 (6·6)	290·5 (56·7)	83·2 (3·8)	126·0 (11·0)
$q = 1$	58·3 (6·1)	123·5 (18·9)	51·2 (2·3)	59·2 (3·0)
$q = 1{\cdot}5$	35·6 (5·1)	47·2 (8·2)	36·2 (1·7)	38·0 (1·9)
$q = 2$	22·5 (4·2)	24·8 (5·2)	28·9 (1·4)	29·7 (1·5)
$q = 2{\cdot}5$	16·2 (3·2)	17·0 (3·6)	24·9 (1·3)	25·6 (1·3)
$q = 3$	13·0 (2·5)	13·5 (2·7)	22·5 (1·2)	23·1 (1·3)

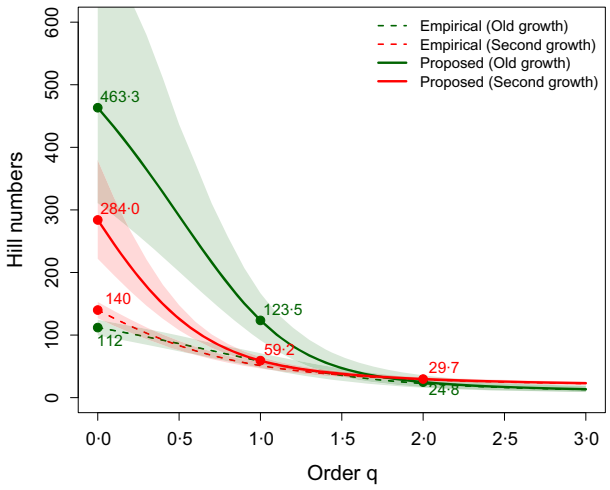

Fig. 3. The empirical (dotted lines) and proposed (solid lines) diversity profiles for beetle species data in the second-growth (red dotted and solid lines) and old-growth (green dotted and solid lines) sites for q between 0 and 3 with 95% confidence interval (shaded areas based on a bootstrap method of 1000 replications). The numbers show the empirical and estimated diversities for $q = 0$, 1 and 2.

become indistinguishable again (at the 95% confidence level) for $q \geq 1{\cdot}3$. Thus, our new method better distinguishes the two profiles and also changes their ranking. Biologists would reach different conclusions using the new method, and our simulations of the previous section show that the new conclusions would be more likely to describe the relationships between the true population profiles. The R scripts for computing the empirical and proposed diversity profiles are available in Appendix S8 with examples. All the proposed estimators are also featured in the online freeware application SpadeR (Species Prediction And Diversity Estimation in R) at the first author's website (http://chao.stat.nthu.edu.tw/).

Conclusion and Discussion

In this paper, we have formulated Hill numbers and entropies of order q as functions of the slopes of the species accumulation curve (see eqns 3b, 3c and 4a). Substituting the slope estimators derived in Chao & Jost (2012), we obtain estimated continuous diversity profiles as a function of q. Simple transformations lead to profiles of the Rényi entropies and Tsallis entropies. Extensive simulations have shown that our proposed diversity profiles improve the empirical diversity profiles for all values of q from 0 to $q = 3$ (the range of orders that diversity and entropy profiles are usually plotted); the improvement may be substantial when there are many undetected species in samples. Our real data example shows that comparisons based on the empirical and proposed diversity profiles may lead to contrasting answers, and simulations and theory demonstrate that the more statistically valid and more accurate answer will be given by our new method rather than by the empirical method.

Note that the proposed estimation method produces the estimated asymptotes of diversities and entropies so that we can make fair comparison across multiple assemblages. However, when q is close to zero our estimators based on severely under-sampling data may exhibit negative biases as shown by our simulations. For $q = 0$, our theory (near eqn 7a) implies that the asymptotic species richness estimator is a lower bound for incomplete data taken from highly heterogeneous assemblages. In this case, fair comparison of species richness across multiple assemblages can be made by standardizing sample completeness (i.e. comparing species richness for a standardized fraction of population) based on integrated rarefaction and extrapolation sampling as shown by Chao & Jost (2012) and Chao *et al.* (2014).

COMPARISON WITH THE EMPIRICAL PROFILES FOR HIGH ORDERS

Our diversity and entropy profile estimators are mainly designed to take into account the effect of the undetected species in samples. Thus, this paper has been focused on the low orders of q ($0 \leq q \leq 3$) in which the proposed method can remove most of the biases associated with the empirical profiles. One may wonder about the relative performance of the two profiles when $q > 3$. In Appendix S5, we theoretically show and numerically demonstrate that the performance of our estimated profile differs little from the empirical profile if $3 < q \leq R$, where R denotes the square root of the maximum species frequency (provided $R > 3$, which is valid for nearly all ecological data). Thus, if desired, our proposed bias-corrected diversity and entropy profiles can be extended to the order R because our proposed profile is superior to the empirical profile for $0 \leq q \leq 3$, and both work equally well for $3 < q \leq R$. Beyond the order R, our bias-corrected estimator either does not exist or may be subject to larger variation (due to its complicated form) and thus may result in larger RMSE than the empirical diversity. Nevertheless, such high orders of diversities and entropies are rarely used in most practical applications.

INDIVIDUAL-BASED SAMPLING ASSUMPTION

In our model formulation for individual-based data, we assume that individuals are randomly selected with replacement, and our profile estimators are derived under this assumption. When the goal of an investigation was to measure the diversity of a finite set of objects and sampling is done without replacement, the model for species frequencies should be formulated for species absolute abundances rather than relative abundances; thus, the multinomial distribution for species frequencies used in this paper for sampling with replacement should be replaced by the generalized hypergeometric model as shown by Chao & Lin (2012). They derived a species richness estimator under sampling without replacement based on this model; they also concluded that if sampling is done without replacement, the traditional estimators derived under sampling with replacement tend to overestimate richness for relatively high sampling fractions (ratio of the sample size to the population

size) and do not converge to the true species richness when the sampling fraction approaches one. In Appendix S6, we present some simulations which confirm that the conclusions of Chao & Lin (2012) are also valid for diversity profiles when $q > 0$. How to derive an analytic diversity profile estimator under sampling without replacement is a worthwhile topic of future research. However, in most real cases where under-sampling bias is important, the sample constitutes a tiny fraction of the total population, so that the two sampling strategies give nearly identical results.

INCIDENCE DATA

Our derivation in this paper is based on samples in which individuals are taken randomly from assemblages. In many ecological field surveys, the sampling unit is not an individual, but a trap, net, quadrat, plot or timed survey. It is these sampling units, not the individuals, that are sampled randomly and independently. Often, it is not possible to count individuals within a sampling unit (e.g. in plant communities or microbial assemblages), so estimation is usually based on a set of sampling units in which only the incidence (detection or non-detection) of each species is recorded. This type of data is referred to as (multiple) incidence data. The sampling model and estimation are provided in Colwell *et al*. (2012) and Chao *et al*. (2014). Our diversity profile and its sampling framework as we discussed for abundance data needs proper modification to treat this kind of data. In Appendix S7, we provide a detailed model formulation and show that our derivation for individual-based data can be extended to obtain an accurate estimator of the diversity profile for multiple incidence data. A real data example with interpretation is also provided in Appendix S7.

PHYLOGENETIC AND FUNCTIONAL DIVERSITY PROFILES

In abundance-based species diversity, all species are treated as equally different from each other; no attempt is made to quantify how different they are. Many researchers have recognized the importance of incorporating species differences into biodiversity studies. Differences among species can be based on their evolutionary histories, as estimated by taxonomic classification or well-supported phylogenetic trees, or can be based on their different trophic guilds or functional traits. Many of the most pressing and fundamental questions in ecology and evolution require robust and meaningful measures of this expanded concept of diversity. Hill numbers have been extended to phylogenetic diversity, which incorporates species' evolutionary history or phylogenetic distance between species, as well as to distance-based functional diversity based on species traits (see Chao, Chiu & Jost (2014) for a review). As with Hill numbers, we can similarly formulate the phylogenetic diversity profile and functional diversity profile. We are currently working on the generalization of the proposed method in this paper to phylogenetic diversity and functional diversity.

Acknowledgements

The authors thank the Editor (Robert O'Hara), the Associate Editor (Ryan Chisholm), Samuel Scheiner and one anonymous reviewer for providing thoughtful suggestions and comments. We also thank Y. H. Chen and T. C. Hsieh for discussion and computational help. AC is supported by Taiwan Ministry of Science and Technology under Contract 103-2628-M007-007. LJ acknowledges support from John V. Moore through a grant to the Population Biology Foundation.

Data accessibility

All data used in this manuscript are presented in the manuscript and its supporting information. The R scripts for obtaining the empirical and proposed diversity profiles based on abundance data or incidence data are available in Appendix S8 with illustrative examples.

References

Beck, J. & Schwanghart, W. (2010) Comparing measures of species diversity from incomplete inventories: an update. *Methods in Ecology and Evolution*, **1**, 38–44.

Chao, A. (1984) Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics*, **11**, 265–270.

Chao, A. (2005) Species estimation and applications. *Encyclopedia of Statistical Sciences*, 12, (eds S. Kotz, N. Balakrishnan, C.B. Read & B. Vidakovic), pp. 7907–7916. Wiley, New York.

Chao, A., Chiu, C.-H. & Jost, L. (2010) Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B*, **365**, 3599–3609.

Chao, A., Chiu, C.-H. & Jost, L. (2014) Unifying species diversity, phylogenetic diversity, functional diversity and related similarity and differentiation measures through Hill numbers. *Annual Reviews of Ecology, Evolution, and Systematics*, **45**, 297–324.

Chao, A. & Jost, L. (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, **93**, 2533–2547.

Chao, A. & Lin, C.W. (2012) Nonparametric lower bounds for species richness and shared species richness under sampling without replacement. *Biometrics*, **68**, 912–921.

Chao, A., Wang, Y.T. & Jost, L. (2013) Entropy and the species accumulation curve: a novel estimator of entropy via discovery rates of new species. *Methods in Ecology and Evolution*, **4**, 1091–1110.

Chao, A., Gotelli, N.G., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K. & Ellison, A.M. (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species biodiversity studies. *Ecological Monographs*, **84**, 45–67.

Chao, A., Hsieh, T.C., Chazdon, R.L., Colwell, R.K. & Gotelli, N.J. (2015) Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. *Ecology*, doi: 10.1890/14-0550.1.

Chiu, C.-H., Wang, Y.T., Walther, B.A. & Chao, A. (2014) An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula. *Biometrics*, **70**, 671–682.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.Y., Mao, C.X., Chazdon, R.L. & Longino, J.T. (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation, and comparison of assemblage. *Journal of Plant Ecology*, **5**, 3–21.

Daróczy, Z. (1970) Generalized information functions. *Information and Control*, **16**, 36–51.

Ellison, A.M. (2010) Partitioning diversity. *Ecology*, **91**, 1962–1963.

Good, I.J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Good, I.J. (2000) Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation*, **66**, 101–111.

Gotelli, N.J. & Chao, A. (2013) Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. *Encyclopedia of Biodiversity*, 2nd edn. Vol. 5 (ed. S.A. Levin), pp. 195–211. Academic Press, Waltham, MA.

Gotelli, N.J. & Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Havrda, J. & Charvát, F. (1967) Quantification method of classification processes. Concept of structural α-entropy. *Kybernetika*, **3**, 30–35.

Hill, M.O. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427–432.
Janzen, D.H. (1973) Sweep samples of tropical foliage insects: description of study sites, with data on species abundances and size distributions. *Ecology*, **54**, 659–686.
Jost, L. (2006) Entropy and diversity. *Oikos*, **113**, 363–375.
Jost, L. (2007) Partitioning diversity into independent alpha and beta components. *Ecology*, **88**, 2427–2439.
Jost, L. (2010) The relation between evenness and diversity. *Diversity*, **2**, 207–232.
Lande, R., DeVries, P.J. & Walla, T.R. (2000) When species accumulation curves intersect: implications for ranking diversity using small samples. *Oikos*, **89**, 601–605.
Leinster, T. & Cobbold, C.A. (2012) Measuring diversity: the importance of species similarity. *Ecology*, **93**, 477–489.
MacArthur, R.H. (1965) Patterns of species diversity. *Biological Review*, **40**, 510–533.
Magurran, A.E. (2004) *Measuring Biological Diversity*. Blackwell, Oxford.
Miller, R.I. & Wiegert, R.G. (1989) Documenting completeness, species-area relations, and the species-abundance distribution of a regional flora. *Ecology*, **70**, 16–22.
Nielsen, R., Tarpy, D.R. & Reeve, H.K. (2003) Estimating effective paternity number in social insects and the effective number of alleles in a population. *Molecular Ecology*, **12**, 3157–3164.
Norden, N., Chazdon, R., Chao, A., Jiang, Y.-H. & Vilchez-Alvarado, B. (2009) Resilience of tropical rain forests: rapid tree community reassembly in secondary forests. *Ecology Letters*, **12**, 385–394.
O'Hara, R.B. (2005) Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology*, **74**, 375–386.
Patil, G.P. & Taillie, C. (1982) Diversity as a concept and its measurement. *Journal of the American Statistical Association*, **77**, 548–561.
Rényi, A. (1961) On measures of entropy and information. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1 (ed J. Neyman), pp. 547–560. University California Press, Berkeley, CA.
Scheiner, S.M. (2012) A metric of biodiversity that integrates abundance, phylogeny, and function. *Oikos*, **121**, 1191–1202.
Tóthmérész, B. (1995) Comparison of different methods for diversity ordering. *Journal of Vegetation Science*, **6**, 283–290.
Tsallis, C. (1988) Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, **52**, 479–487.
Zipf, G.K. (1965) *Human Behavior and Principle of Least Effort*, 2nd edn. Addison-Wesley, New York.

Received 22 October 2014; accepted 15 January 2015
Handling Editor: Ryan Chisholm

Supporting Information

Additional Supporting Information may be found in the online version of this article.

Appendix S1. Proof details.

Appendix S2. A bootstrap method to obtain variance estimators and confidence intervals.

Appendix S3. Biological surveys used in the simulations.

Appendix S4. Additional simulation results.

Appendix S5. More comparisons with the empirical profiles.

Appendix S6. The behavior of our diversity profile estimator under sampling without replacement.

Appendix S7. Diversity profile estimation for multiple incidence data.

Appendix S8. R scripts for obtaining empirical and proposed diversity profiles based on abundance data or incidence data.

