Estimating diversity and entropy profiles via discovery rates of new species

Anne Chao¹* and Lou Jost²

¹Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan; and ²EcoMinga 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 ≥ 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 ≤ q ≤ 3). Our approach is based on formulating 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 ≤ 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...
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 extensive 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:

\[
S(k) = \sum_{j=1}^{S} [1 - (1 - p_j)^k] = S - \sum_{j=1}^{S} (1 - p_j)^k, \quad k = 0, 1, 2, \ldots, \quad \text{eqn 1}
\]

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:

\[ \Delta(k) = \frac{S(k+1) - S(k)}{(k+1) - k} = \sum_{i=1}^{k} p_i (1 - p_i)^k, \]  
\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 & Wälla 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:

\[ \left( \frac{u}{0} \right) \equiv 1; \]
\[ \left( \frac{u}{k} \right) \equiv u(u - 1)(u - 2) \ldots (u - k + 1)/k!. \]

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

\[ \left( \frac{u}{k} \right) = \begin{cases} \frac{u!}{k!(u-k)!} & k \leq u, \\ 0 & k > u. \end{cases} \]

We first write the \( q \)th order basic sum of the species relative abundances as follows:

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

\[ q D = \left( \Delta(0) - (q - 1) \Delta(1) + \ldots + (-1)^k (q - 1)^k \ldots (q - k) \right) \frac{k!}{\Delta(k) + \ldots} \frac{1}{(1 - q)}. \]  
\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:

\[ q D = S = \sum_{k=0}^{\infty} \Delta(k). \]  
\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:

\[ ^1D = \lim_{q \to -1} \left( \Delta(0) - (q - 1) \frac{1}{\Delta(1)} + \frac{1}{\Delta(2)} \right)^{1/(1-q)}. \]  
\text{eqn 4c} \]

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

\[ q D = \left( \sum_{k=0}^{q-1} \frac{q - 1}{k} \right)^{1/(1-q)}. \]  
\text{eqn 4d} \]

For the special case of \( q = 2 \) and 3, we have \( ^2D = [\Delta(0) - \Delta(1)]^{-1} \) and \( ^3D = [\Delta(0) - 2\Delta(1) + \Delta(2)]^{-1} \). 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 \( q \geq 0 \), the general formula in eqn 4a is a sum of an infinite number of terms: (4e)

\[ \text{DIVERSITY PROFILE ESTIMATOR} \]

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 \( q \geq 0 \) 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 \( k \geq n \) as follows:

\[ q D = \left( \sum_{k=0}^{q-1} \frac{q - 1}{k} \right)^{1/(1-q)} + \sum_{k=q}^{\infty} \frac{1}{\Delta(k)} \right)^{1/(1-q)} \]  
\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 \( \sum_{i>0} x_i = n \) and cell probabilities \( \{p_1, p_2, \ldots, p_S\} \), we separately estimate each sum in eqn (5).
Let $\Delta(k)$ be the singletons and doubletons in samples $T$ of species richness. The core idea in Turing and Good (Good 1953, 2000) is that the estimated mean relative frequency of the singletons 

$$
\hat{\lambda}(k) = \sum_{k=1}^{\infty} \binom{n-k-1}{k} = \sum_{k=0}^{n-1} \binom{n-k-1}{k} \binom{n}{k} \quad k < n.
$$

eqn 6a

Here, $\hat{\lambda}(0) = 1$. 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.

$$
\hat{D} = \exp \left( \sum_{k=1}^{\infty} \binom{X_i}{n} - \binom{X_i-1}{k} \right) + f_i \binom{n}{1} \binom{n-1}{1-A}^{n+1} \left[ -\log A - \sum_{r=1}^{n-1} \frac{1}{A-r} \right].
$$

eqn 7b

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 doubletons in the population is not $1/n$, but $2f_2/(nf_1^2)$, 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:

$$
A = \begin{cases} 
\frac{2f_2}{\binom{n-1}{f_1+2f_2}}, & \text{if } f_2 > 0; \\
\frac{1}{2} \binom{n-1}{f_1+2f_2} + 2, & \text{if } f_2 = 0, f_1 \neq 0; \\
1, & \text{if } f_2 = f_1 = 0.
\end{cases}
$$

eqn 6b

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

$$
\hat{\lambda}(n+m) = \frac{f_1}{n} \binom{n-1}{1-A}^{m+1}, \quad m \ge 0.
$$

eqn 6c

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

$$
q^{\hat{D}} = \left( \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k \hat{\lambda}(k) + f_i \binom{n}{1} \binom{n}{1-A}^{n+1} \left[ A^{q-1} - \sum_{r=0}^{n-1} \binom{n-1}{r} (-1)^q (A-1)^{q-r} \right] \right)^{1/(1-q)}.
$$

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).

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):

$$
\hat{\lambda}(k) = \sum_{n-k-1}^{n} \binom{x_i}{n} = \sum_{k=0}^{n-1} \binom{n-k-1}{k} \binom{n}{k} \cdot k < n.
$$

eqn 6a

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): $\hat{D}$

$$
\left( \sum_{k=0}^{n-1} \binom{n-1}{k} (-1)^k \hat{\lambda}(k) + f_i \binom{n}{1} \binom{n}{1-A}^{n+1} \left[ A^{q-1} - \sum_{r=0}^{n-1} \binom{n-1}{r} (-1)^q (A-1)^{q-r} \right] \right)^{1/(1-q)}.
$$

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

$$
A^{q-1} - \sum_{r=0}^{n-1} \binom{n-1}{r} (-1)^q (A-1)^{q-r} = 0 \text{ and eqn 7d reduces to eqn 7c.}
$$

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 $K$ equals a constant, equally diverse assemblages with no shared species are pooled, then the diversity of the pooled assemblages equals $K$ times the diversity of an individual
Theoretical abundance models with 200 species:

(a) the log-normal model with 200 species,
(b) the Zipf model with 200 species.

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 ratio of the standard deviation to the mean). Pool any number of these 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 ratio of the standard deviation to the mean). Pool any number of these 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.
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 \( p_i = c a_i \), 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:

   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 \( CV = 1.563 \).

   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
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/assembly 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 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 \(0 \leq q \leq 3\) (see Discussion for comparison for higher orders).

### Applications

We illustrate the use of our diversity profile estimator by analyzing 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 \(0 \leq q \leq 10\) 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 \(q < 1.3\). 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 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.

<table>
<thead>
<tr>
<th>Osa second-growth site</th>
<th>Osa old-growth site</th>
</tr>
</thead>
<tbody>
<tr>
<td>(i)</td>
<td>(f_i)</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1</td>
<td>70</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
</tr>
</tbody>
</table>

**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).

<table>
<thead>
<tr>
<th>Diversity order</th>
<th>Old-growth site</th>
<th>Proposed diversity (SE)</th>
<th>Second-growth site</th>
<th>Proposed diversity (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(q = 0)</td>
<td>112.0 (6.4)</td>
<td>463.3 (106.1)</td>
<td>140.0 (6.6)</td>
<td>284.0 (42.1)</td>
</tr>
<tr>
<td>(q = 0.5)</td>
<td>186.3 (6.6)</td>
<td>290.5 (56.7)</td>
<td>83.2 (3.8)</td>
<td>126.0 (11.0)</td>
</tr>
<tr>
<td>(q = 1)</td>
<td>58.6 (6.1)</td>
<td>123.5 (18.9)</td>
<td>51.2 (2.3)</td>
<td>59.2 (3.0)</td>
</tr>
<tr>
<td>(q = 1.5)</td>
<td>35.6 (5.1)</td>
<td>47.2 (8.2)</td>
<td>36.2 (1.7)</td>
<td>38.0 (1.9)</td>
</tr>
<tr>
<td>(q = 2)</td>
<td>22.5 (4.2)</td>
<td>24.8 (5.2)</td>
<td>28.9 (1.4)</td>
<td>29.7 (1.5)</td>
</tr>
<tr>
<td>(q = 2.5)</td>
<td>16.2 (3.2)</td>
<td>17.0 (3.6)</td>
<td>24.9 (1.3)</td>
<td>25.6 (1.3)</td>
</tr>
<tr>
<td>(q = 3)</td>
<td>13.0 (2.5)</td>
<td>13.5 (2.7)</td>
<td>22.5 (1.2)</td>
<td>23.1 (1.3)</td>
</tr>
</tbody>
</table>

become indistinguishable again (at the 95% confidence level) for \( q \geq 1.3 \). 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-M-007-007. 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**


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.
Appendix S1. Proof details

In the following proofs, we define $0! = 1$ by convention, and define for a real number $u$ and a positive integer $k$:

$$
\binom{u}{0} = 1; \quad \binom{u}{k} = u(u-1)(u-2)...(u-k+1)/k!.
$$

In the special case that $u$ is a non-negative integer, the above definition reduces to

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

Theorem S1.1. In eqn 4a of the main text, we have expressed Hill numbers of order $q$ in terms of the slopes of the species accumulation curve (SAC):

$$
^{q}D = \left( \sum_{k=0}^{\infty} \frac{q-1}{k} (-1)^k \Delta(k) \right)^{1/(1-q)},
$$

where $\Delta(k) = S(k+1) - S(k)$ denotes the slope of the line connecting two adjacent points $(k, S(k))$ and $(k+1, S(k+1))$ in the SAC, $S(0) = 0$ and $\Delta(0) = 1$.

(a) We prove that for the special case $q = 0$, $^{0}D = \sum_{k=0}^{\infty} \Delta(k)$.

(b) We prove that as $q$ tends to unity,

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

(c) We prove that if $q$ is an integer $\geq 2$, then we can express $^{q}D$ as a finite sum of slopes and
\[ qD = \left( \sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^k \Delta(k) \right)^{\frac{1}{1-q}} = \left[ (-1)^{q-1} \Phi^{q-1} \Delta(0) \right]^{\frac{1}{1-q}}, \tag{S1.3} \]

where \( \Phi^n \Delta(x) \) denotes the \( n \)-th forward difference of the function \( \Delta(x) \), \( \Phi \) is a forward difference operator defined by \( \Phi \Delta(x) = \Delta(x+1) - \Delta(x) \), and \( \Phi^k \Delta(x) = \Phi[\Phi^{k-1} \Delta(x)] \) for \( k = 2, 3 \ldots \)

Proofs:
(a) For the special case of \( q = 0 \), eqn (S1.1) reduces to
\[ 0D = \sum_{k=0}^{\infty} \binom{-1}{k} (-1)^k \Delta(k). \]

It follows from the definition that we have
\[ \binom{-1}{k} = \frac{(-1)(-2)...(-k)}{k!} = \frac{(-1)^k k!}{k!} = (-1)^k. \]

Then we obtain \( 0D = \sum_{k=0}^{\infty} \Delta(k). \)
(b) For the case of \( q = 1 \), we first find the following limit:
\[ \lim_{q \to 1} \frac{1}{q-1} \log \left( \sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k) \right). \]

Note that the function \( \sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k) \) as \( q \) tends to unity is \( \Delta(0) = 1 \); the derivative of \( \binom{q-1}{k} \) with respect to \( q \) evaluated at \( q = 1 \) is \( (-1)^{k+1} / k \) because
\[ \frac{d}{dq} \binom{q-1}{k} \bigg|_{q=1} = \frac{1}{k!} (q-2)...(q-k) \bigg|_{q=1} = \frac{1}{k!} (1-2)(1-3)...(1-k) = \frac{(-1)^{k+1}}{k}. \]

Thus, we obtain
\[ \lim_{q \to 1} \frac{1}{q-1} \log \left( \sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \Delta(k) \right) = \sum_{k=1}^{\infty} \frac{\Delta(k)}{k}. \]
(c) If \( q \) is an integer \( \geq 2 \), then
\[
\left( \begin{array}{c} q-1 \\ k \end{array} \right) = 0 \text{ if } k \geq q.
\]
Thus the eqn (S1.1) becomes a finite sum where the index \( q \) runs from \( k = 0 \) to \( k = q - 1 \). It follows from the definition of the \((q-1)\)-th forward difference (Johnson & Kotz, 1977, p. 7) that
\[
\phi^{q-1} \Delta(0) = \sum_{k=0}^{q-1} \left( \begin{array}{c} q-1 \\ k \end{array} \right) (-1)^{q+k} \Delta(k).
\]

Then
\[
D^{q} = \left[ \sum_{k=0}^{q-1} \left( \begin{array}{c} q-1 \\ k \end{array} \right) (-1)^{k} \Delta(k) \right]^{1/(1-q)} = \left[ (-1)^{q-1} \sum_{k=0}^{q-1} \left( \begin{array}{c} q-1 \\ k \end{array} \right) (-1)^{q+k} \Delta(k) \right]^{1/(1-q)} = \left[ (-1)^{q-1} \phi^{q-1} \Delta(0) \right]^{1/(1-q)}.
\]

**Theorem S1.2.** Our proposed diversity profile estimator in terms of the estimated slopes of the SAC is expressed as
\[
D^{q} = \left( \sum_{k=0}^{q-1} \left( \begin{array}{c} q-1 \\ k \end{array} \right) (-1)^{k} \hat{A}(k) + \sum_{k=n}^{q-1} \left( \begin{array}{c} q-1 \\ k \end{array} \right) (-1)^{q+k} \hat{A}(k) \right)^{1/(1-q)}, \quad q \geq 0.
\]

We prove the following:

(a) For \( q = 0 \), we have \( D^{0} = \hat{S} = \sum_{k=0}^{q} \hat{A}(k) = \sum_{k=0}^{q-1} \hat{A}(k) + \sum_{k=n}^{q-1} \hat{A}(k), \sum_{k=0}^{q} \hat{A}(k) = \hat{S}_{ob} \) (the observed species richness in the sample), and \( \sum_{k=n}^{q} \hat{A}(k) = \hat{f}_{0} \), where
\[
\hat{f}_{0} = \begin{cases} 
\frac{(n-1) f_{1}^{2}}{n (2 f_{2})}, & \text{if } f_{2} > 0; \\
\frac{(n-1) f_{1} (f_{1} - 1)}{2 n}, & \text{if } f_{2} = 0.
\end{cases} \quad (S1.4)
\]

(b) Letting \( q \) tend to 1, we have
\[
D^{1} = \exp \left( \sum_{k=1, s=A}^{n} \frac{X_{k}}{n} \left( \sum_{k=1}^{\frac{n-1}{2}} \frac{1}{k} + \frac{f_{1}}{n} (1 - A)^{-n} \left[ -\log A - \sum_{r=1}^{\frac{n-1}{2}} \frac{(1-A)^{r}}{r} \right] \right) \right). \quad (S1.5)
\]
(c) For \( q = 2, 3, \ldots, n \), we have
\[
q \tilde{D}^{i-q} = \sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^{k} \hat{A}(k) = \sum_{X_i \in q} X_i(X_i-1) \cdot \frac{X_i-q+1}{n(n-1) \cdots (n-q+1)}.
\]

The right hand side of the above formula is positive only when \( q \leq \max X_i \) under which we have the estimator
\[
q \hat{D} = \left( \sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^{k} \hat{A}(k) \right)^{1/(1-q)} = \left( \sum_{X_i \in q} X_i(X_i-1) \cdot \frac{X_i-q+1}{n(n-1) \cdots (n-q+1)} \right)^{1/(1-q)}.
\] (S1.6)

(d) For any \( q \) between 0 and \( \max X_i \), the general formula for our estimator can be expressed as the following formula with finite terms:
\[
q \hat{D} = \left( \sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^{k} \hat{A}(k) + \frac{f}{n} (1 - A)^{-r+1} \left[ A^{r-1} - \sum_{r=0}^{n-1} \binom{q-1}{r} (A-1)^{r} \right] \right)^{1/(1-q)}.
\] (S1.7)

When \( q = 0, 1 \) and any integer between 2 and \( n \), eqn (S1.7) reduces respectively to the formula in (a), (b) and (c) above.

Proofs:

(a) Based on the slope estimator \( \hat{A}(k) \) for \( k < n \) in eqn 6a of the main text, we obtain
\[
\sum_{k=0}^{n-1} \hat{A}(k) = \sum_{k=0}^{n-1} \sum_{X_i \in k} \frac{n-k-1}{X_i-1} \binom{n}{X_i} = \sum_{X_i \in n} \sum_{k=0}^{n-X_i} \frac{n-k-1}{X_i-1} \binom{n}{X_i}.
\]

By mathematical induction on \( n \), it can be proved that for any fixed \( 1 \leq X_i \leq n \),
\[
\sum_{k=0}^{n-X_i} \binom{n-k-1}{X_i-1} = \binom{n}{X_i}.
\]

Then we obtain
\[
\sum_{k=0}^{n-1} \hat{A}(k) = \sum_{1 \leq X_i \leq n} \binom{n}{X_i} = \sum_{1 \leq X_i \leq n} I(1 \leq X_i \leq n) = S_{\text{obs}}.
\]
Since $\hat{A}(n + m) = \frac{f_i}{n}(1 - A)^{m+1}$ (eqn 6c in the main text), we can further evaluate the sum
\[
\sum_{k=0}^{\infty} \hat{A}(k) : \quad \sum_{k=0}^{\infty} \hat{A}(k) = \frac{f_i}{n} [(1 - A) + (1 - A)^2 + \ldots] = \frac{f_i}{n} \frac{(1 - A)}{A},
\]
where $A$ is defined as (eqn 6b of the main text)
\[
A = \begin{cases} 
2 f_2 / [(n-1)f_1 + 2f_2] & \text{if } f_2 > 0 \\
2 / [(n-1)(f_1 - 1) + 2] & \text{if } f_2 = 0, f_1 \neq 0 \\
1 & \text{if } f_2 = f_1 = 0.
\end{cases}
\]

Substituting $A$ into the formula $f_i(1 - A)/(nA)$, we obtain
\[
\sum_{k=0}^{\infty} \hat{A}(k) = \hat{f}_0,
\]
where $\hat{f}_0$ is defined in eqn (S1.4).

(b) Replacing $A(k)$ in eqn (S1.1) by $\hat{A}(k)$ and following the proof steps of Theorem S1.1 (b), we readily obtain
\[
\lim_{q \to 1} \frac{1}{1 - q} \log \left( \sum_{k=0}^{\infty} \frac{(q - 1)^k \hat{A}(k)}{k} \right) = \sum_{k=0}^{\infty} \frac{\hat{A}(k)}{k}.
\]
Then eqn (S1.5) follows from a proof given in Chao, Wang & Jost (2013).

(c) It is sufficient to prove that for $q = 2, 3, \ldots, n$,
\[
\sum_{k=0}^{q-1} \frac{(q - 1)^k}{k} (-1)^k \hat{A}(k) = \sum_{k \geq q} X_i(X_j - 1)(X_j - q + 1) \frac{X_i(X_j - 1)(X_j - q + 1)}{n(n-1)(n-q+1)}. \quad (S1.8)
\]
The left hand side of (S1.8) can be expressed as
\[
\sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^k \Delta(k) = \sum_{k=0}^{q-1} \binom{q-1}{k} (-1)^k \sum_{i \leq X_i \leq n-k} \frac{X_i}{n} \binom{n-X_i}{k}
\]

\[
= \sum_{i \leq X_i \leq n-q+1} \sum_{k=0}^{q-1} \frac{X_i}{n} (-1)^k \binom{q-1}{k} \binom{n-X_i}{k} + \sum_{n-q+1 < X_i \leq n} \sum_{k=0}^{n-X_i} (-1)^k \binom{q-1}{k} \binom{n-X_i}{k}.
\]

The right hand side of (S1.8) can be rewritten as

\[
\sum_{X_i \leq q} \frac{X_i(X_i-1) \ldots (X_i-q+1)}{n(n-1) \ldots (n-q+1)} = \sum_{X_i \leq q} \frac{X_i}{n} \binom{q-1}{n-1}.
\]  

(S1.9)

Let \( \min = \min \{n-X_i, q-1\} \). We have \( \min = q-1 \) if \( 1 \leq X_i \leq n-q+1 \), and \( \min = n-X_i \) if \( n-q+1 < X_i \leq n \). Then it is sufficient to prove for any \( 1 \leq X_i \leq n \) that

\[
\sum_{k=0}^{\min} (-1)^k \binom{q-1}{k} \binom{n-X_i}{k} = \frac{X_i - 1}{q - 1} \binom{q-1}{n-1}.
\]  

(S1.10)

This can be obtained by considering an urn model: suppose there are \( X_i - 1 \) red balls and \( n - X_i \) white balls in an urn. If we draw \( q-1 \) balls sequentially (one ball at a time) without replacement from the urn, the probability that all \( q-1 \) balls are red balls is the right hand side of (S1.10). Let the event \( E_k \) be the event that the \( k \)-th selected ball is a red ball, \( k = 1, 2, \ldots, q-1 \). It follows from basic probability theory that
\[
\frac{(X_i - 1)}{q - 1} = P(E_1 \cap E_2 \cap \ldots \cap E_{q-1}) = 1 - P(E_1^c \cup E_2^c \cup \ldots \cup E_{q-1}^c) \\
= 1 - \sum_{k=1}^{\min} (-1)^{k-1} \binom{q-1}{k} P(E_1^c \cap E_2^c \cap \ldots \cap E_k^c).
\]

Also, we have
\[
P(E_1^c \cap E_2^c \cap \ldots \cap E_k^c) = \binom{n - X_i}{k} \binom{n-1}{k}.
\]

Then we obtain eqn (S1.10). For the case \(X_i < q\), the left hand side of (S1.10) must be zero because the right hand side is zero. Alternatively, we can use a mathematical induction on \(q\) to directly prove that the left hand side of (S1.10) is zero for all \(X_i < q\).

(d) It is sufficient to prove
\[
\sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \hat{A}(k) = \frac{f_i}{n} (1 - A)^{-q+1} \left[ A^{q-1} - \sum_{r=0}^{\frac{q-1}{r}} \left( \frac{q-1}{r} \right) (A-1)^r \right].
\]

Since \(\hat{A}(n+m) = \frac{f_i}{n} (1 - A)^{-q+1}\) for \(m \geq 0\) (eqn 6c in the main text), we have
\[
\sum_{k=0}^{\infty} \binom{q-1}{k} (-1)^k \hat{A}(k) = \frac{f_i}{n} \sum_{m=0}^{\infty} \left( \frac{q-1}{n+m} \right) (-1)^{q+m} (1 - A)^{q+m} \\
= \frac{f_i}{n} (1 - A)^{-q+1} \sum_{m=0}^{\infty} \left( \frac{q-1}{n+m} \right) (-1)^{q+m} (1 - A)^{q+m} \\
= \frac{f_i}{n} (1 - A)^{-q+1} \left[ 1 - (1 - A)^{q+1} \sum_{r=0}^{q-1} \left( \frac{q-1}{r} \right) (A-1)^r \right] \\
= \frac{f_i}{n} (1 - A)^{-q+1} \left[ A^{q-1} - \sum_{r=0}^{q-1} \left( \frac{q-1}{r} \right) (A-1)^r \right].
\]

Then we obtain the formula in eqn (S1.7). It is readily checked that eqn (S1.7) reduces to the formula in (a) for \(q = 0\). Note that if \(q\) is an integer between 2 and \(n - 1\), then we have
\[ a^{q-1} \sum_{r=0}^{n-1} \frac{q-1}{r} (A-1)r = 0, \] and eqn (S1.7) reduces to eqn (S1.6). To obtain \( \lim_{q \to 1} D_q \), we first evaluate

\[
\lim_{q \to 1} \frac{1}{1-q} \log \left( \sum_{k=0}^{\infty} \frac{q-1}{k} (-1)^k \hat{A}(k) + \frac{f_i}{n} (1-A)^{-n+1} \left[ A^{q-1} - \sum_{r=0}^{n-1} \frac{q-1}{r} (A-1)r \right] \right) \] .

Note that the function \( \sum_{k=0}^{\infty} \frac{q-1}{k} (-1)^k \hat{A}(k) + \frac{f_i}{n} (1-A)^{-n+1} \left[ A^{q-1} - \sum_{r=0}^{n-1} \frac{q-1}{r} (A-1)r \right] \) as \( q \) tends to unity is \( \Delta(0) = 1 \). From Theorem S1.1 (b), the derivative of \( \frac{q-1}{k} \) with respect to \( q \) evaluated at \( q = 1 \), is \( -1 \frac{k+i}{k} \), yielding

\[
\frac{d}{dq} \left( \sum_{k=0}^{\infty} \frac{q-1}{k} (-1)^k \hat{A}(k) + \frac{f_i}{n} (1-A)^{-n+1} \left[ A^{q-1} - \sum_{r=0}^{n-1} \frac{q-1}{r} (A-1)r \right] \right) \bigg|_{q=1}
\]

\[
= \sum_{k=1}^{\infty} \frac{\hat{A}(k)}{k} (-1) + \frac{f_i}{n} (1-A)^{-n+1} \left[ \log A + \sum_{r=1}^{n-1} \frac{(1-A)^r}{r} \right] .
\]

Thus,

\[
\lim_{q \to 1} \frac{1}{1-q} \log \left( \sum_{k=0}^{\infty} \frac{q-1}{k} (-1)^k \hat{A}(k) + \frac{f_i}{n} (1-A)^{-n+1} \left[ A^{q-1} - \sum_{r=0}^{n-1} \frac{q-1}{r} (A-1)r \right] \right) \bigg|_{q=1}
\]

\[
= \sum_{k=1}^{\infty} \frac{\hat{A}(k)}{k} + \frac{f_i}{n} (1-A)^{-n+1} \left[ -\log A - \sum_{r=1}^{n-1} \frac{(1-A)^r}{r} \right]
\]

\[
= \sum_{i \leq X, d \leq n} \frac{X_i}{n} \left( \frac{\sum_{k=1}^{\infty} \frac{1}{k} \hat{A}(k)}{k} + \frac{f_i}{n} (1-A)^{-n+1} \left[ -\log A - \sum_{r=1}^{n-1} \frac{(1-A)^r}{r} \right] \right) .
\]

The last step follows from a proof in Chao, Wang & Jost (2013).
References


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

This appendix is a slightly modified version of Appendix S2 of Chao, Wang & Jost (2013) and Appendix G of Chao *et al.* (2014). Consider a multinomial model in which species sample frequencies \( X_1, X_2, \ldots, X_S \) follow a multinomial distribution with cell total \( \sum_{i>0} X_i = n \) and cell probabilities \( \{p_1, p_2, \ldots, p_S\} \). We suggest the use of a bootstrap method to approximate the variance of our proposed diversity estimator \( \hat{D}^q \) (in eqn 7d of the main text) and to construct the associated confidence intervals.

In the bootstrap procedure, we first need to construct the “bootstrap assemblage” which mimics the “true” entire assemblage. We first determine the “true” species richness in this bootstrap assemblage. Define the abundance frequency count \( f_k \) as the number of species each represented by exactly \( k \) individuals in the sample. Thus, \( f_0 \) denotes the number of species present in the entire assemblage but are not detected in the sample. Also, \( f_1 \) is the number of singletons and \( f_2 \) is the number of doubletons in the sample. Based on the Chao1 estimator (Chao 1984, 1987), an estimator of the number of undetected species is:

\[
\hat{f}_0 = \begin{cases} 
\frac{(n-1)f_i^2}{n(2f_2)}, & \text{if } f_2 > 0; \\
\frac{(n-1)f_i(f_i-1)}{2}, & \text{if } f_2 = 0.
\end{cases}
\]

Since the number of species in the bootstrap assemblage must be an integer, we define \( \hat{f}_0^* \) as the smallest integer which is greater than or equal to \( \hat{f}_0 \). Thus, there are \( S_{\text{obs}} + \hat{f}_0^* \) species in the bootstrap assemblage, where \( S_{\text{obs}} \) denotes the number of observed species in the sample. Although this species richness estimator \( S_{\text{obs}} + \hat{f}_0^* \) is theoretically a lower bound (Chao 1984), simulations have suggested that it can be used to estimate the variance of any estimator. This is
mainly because very rare species that are not inferred in the lower bound have almost negligible effect on variances.

Next we determine the “true” relative abundances for those \( S_{\text{obs}} + \hat{f}_n^* \) species in the bootstrap assemblage. For the \( S_{\text{obs}} \) species that are observed in the actual sample, the sample relative abundance \( X_i/n \) of species \( i \) on average overestimates the true relative abundance \( p_i \). This is seen from the following conditional expectation for \( X_i > 0 \):

\[
E\left( \frac{X_i}{n} \bigg| X_i > 0 \right) = \frac{p_i}{1-(1-p_i)^n} > p_i.
\]

Thus, we need to adjust the sample frequency \( X_i/n \). Based on the above conditional expectation, Chao, Wang & Jost (2013) derived that the following adjusted relative abundance in the bootstrap assemblage for the \( i \)-th species:

\[
\hat{p}_i = \frac{X_i}{n} \left[ 1 - \hat{\lambda} \left( 1 - \frac{X_i}{n} \right)^n \right],
\]

where

\[
\hat{\lambda} = \frac{1 - \hat{C}}{\sum_{i \neq i} \frac{X_i}{n} \left( 1 - \frac{X_i}{n} \right)^n},
\]

and \( \hat{C} \) is an estimate of sample coverage:

\[
\hat{C} = \begin{cases} 
1 - \frac{f_i}{n} \left[ \frac{(n-1)f_i}{(n-1)f_i + 2f_2} \right], & \text{if } f_2 > 0; \\
1 - \frac{f_i}{n} \left[ \frac{(n-1)(f_i - 1)}{(n-1)(f_i - 1) + 2} \right], & \text{if } f_2 = 0.
\end{cases}
\]

Chao et al. (2015) further proposed a refined estimator in terms of two parameters for inferring \( p_i \). However, one-parameter formula used here is sufficient for assessing variances of diversity estimators.
For the remaining \( \hat{f}_0^* \) species in the bootstrap assemblage (i.e., those species that were not detected in the sample but exist in the bootstrap assemblage), we assume they all have the same probability \( (1 - \hat{C}) / \hat{f}_0^* \). This assumption may look to be restrictive, but the effect on the resulting variance estimator is small.

After the bootstrap assemblage is determined, a random sample of \( n \) individuals is then generated with replacement. Then a bootstrap estimate \( \hat{\eta} \) is calculated for the generated sample, i.e., all statistics in our estimators are replaced by those computed from the generated data. Replicate the procedure \( B \) times and obtain \( B \) bootstrap estimates \( \{\hat{\eta}^1, \hat{\eta}^2, \ldots, \hat{\eta}^B\} \). Here \( B = 1000 \) is suggested in constructing confidence intervals. The bootstrap variance estimator of the proposed estimator \( \hat{\eta} \) is the sample variance of these \( B \) estimates \( \{\hat{\eta}^1, \hat{\eta}^2, \ldots, \hat{\eta}^B\} \). For these \( B \) bootstrap estimates, let their average, the 2.5% percentile (lower confidence bound), and the 97.5% percentile (upper confidence bound) be denoted respectively as \( \text{Ave}, \text{LCB} \) and \( \text{UCB} \). The interval \([\text{LCB}, \text{UCB}]\) is then “centralized” by shifting the average towards our estimate \( \hat{\eta} \) and we obtain the proposed 95% confidence interval as \([\hat{\eta} - (\text{Ave} - \text{LCB}), \hat{\eta} + (\text{UCB} - \text{Ave})]\). Our method can be applied to assess the variance of any estimator and its associated confidence interval.

References


**Appendix S3. Biological surveys used in the simulations**

**Table S3.1.** (Data used in the Simulation section of the main text) Abundance frequency counts of the extant rare vascular plant species (188 species, 1008 individuals) in the southern Appalachians (Miller & Wiegert 1989), and CV = 1.563. The count $f_k$ is the number of species represented by exactly $k$ individuals in the survey. Diversities: $S = ^0D = 188$, $^1D = 96.9$, $^2D = 54.8$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>61</td>
<td>35</td>
<td>18</td>
<td>12</td>
<td>15</td>
<td>4</td>
<td>8</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>16</th>
<th>19</th>
<th>20</th>
<th>22</th>
<th>29</th>
<th>32</th>
<th>40</th>
<th>43</th>
<th>48</th>
<th>67</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

**Table S3.2.** (Data used in the Simulation section of the main text) Abundance frequency counts of the tree assemblage in LSUR younger (21 years) second growth forest site in northeastern Costa Rica (Norden et al. 2009; Colwell et al. 2012), 76 species, 1020 individuals, and CV = 2.305. The count $f_k$ is the number of species represented by exactly $k$ individuals in the survey. Diversities: $S = ^0D = 76$, $^1D = 21.0$, $^2D = 12.2$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>7</th>
<th>8</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>15</th>
<th>31</th>
<th>33</th>
<th>34</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>29</td>
<td>13</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>35</th>
<th>66</th>
<th>72</th>
<th>78</th>
<th>127</th>
<th>131</th>
<th>174</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Table S3.3. Abundance frequency counts of the Malayan butterfly survey (Fisher, Corbet & Williams 1943), 620 species, 9031 individuals, and CV = 1.435. The count $f_k$ is the number of species represented by exactly $k$ individuals in the survey. Diversities: $S = ^0D = 620, ^1D = 309.7, ^2D = 202.9$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>118</td>
<td>74</td>
<td>44</td>
<td>24</td>
<td>29</td>
<td>22</td>
<td>20</td>
<td>19</td>
<td>20</td>
<td>15</td>
<td>12</td>
<td>14</td>
<td>6</td>
<td>12</td>
<td>6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>9</td>
<td>9</td>
<td>6</td>
<td>10</td>
<td>10</td>
<td>11</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>8</td>
<td>3</td>
<td>3</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
<th>44</th>
<th>45</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>5</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>5</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>46</th>
<th>48</th>
<th>49</th>
<th>50</th>
<th>51</th>
<th>52</th>
<th>53</th>
<th>54</th>
<th>55</th>
<th>56</th>
<th>58</th>
<th>59</th>
<th>60</th>
<th>64</th>
<th>66</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>68</th>
<th>70</th>
<th>71</th>
<th>76</th>
<th>84</th>
<th>89</th>
<th>92</th>
<th>93</th>
<th>100</th>
<th>105</th>
<th>108</th>
<th>119</th>
<th>141</th>
<th>147</th>
<th>194</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table S3.4. Abundance frequency counts of the Lepidoptera trapping at Rothamsted Station (Fisher, Corbet & Williams 1943), 240 species, 15609 individuals, and CV = 2.915. The count $f_k$ is the number of species represented by exactly $k$ individuals in the survey. Diversities: $S = ^0D = 240, ^1D = 59.2, ^2D = 25.4$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>35</td>
<td>11</td>
<td>15</td>
<td>14</td>
<td>10</td>
<td>11</td>
<td>5</td>
<td>6</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>5</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>32</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Table S3.5. Abundance frequency counts of the tree assemblage in LEP old growth forest site in northeastern Costa Rica (Norden et al. 2009; Colwell et al. 2012), 152 species, 943 individuals, and $CV = 1.545$. The count $f_k$ is the number of species represented by exactly $k$ individuals in the survey. Diversities: $S = ^6D = 152$, $^1D = 74.6$, $^2D = 45.1$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>34</th>
<th>36</th>
<th>37</th>
<th>43</th>
<th>44</th>
<th>45</th>
<th>49</th>
<th>51</th>
<th>52</th>
<th>53</th>
<th>54</th>
<th>57</th>
<th>58</th>
<th>60</th>
<th>61</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>64</th>
<th>67</th>
<th>73</th>
<th>76</th>
<th>78</th>
<th>84</th>
<th>89</th>
<th>96</th>
<th>99</th>
<th>109</th>
<th>112</th>
<th>120</th>
<th>122</th>
<th>129</th>
<th>135</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>141</th>
<th>148</th>
<th>149</th>
<th>151</th>
<th>154</th>
<th>177</th>
<th>181</th>
<th>187</th>
<th>190</th>
<th>199</th>
<th>211</th>
<th>221</th>
<th>226</th>
<th>235</th>
<th>239</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>244</th>
<th>246</th>
<th>282</th>
<th>305</th>
<th>306</th>
<th>333</th>
<th>464</th>
<th>560</th>
<th>572</th>
<th>589</th>
<th>604</th>
<th>743</th>
<th>823</th>
<th>2349</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>13</th>
<th>15</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>46</td>
<td>30</td>
<td>16</td>
<td>12</td>
<td>6</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>46</td>
<td>30</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>19</th>
<th>20</th>
<th>25</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>46</th>
<th>52</th>
<th>55</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Table S3.6. Abundance frequency counts of the tree assemblage in LEP older (29 years) second growth forest site in northeastern Costa Rica (Norden et al. 2009; Colwell et al. 2012), 104 species, 1263 individuals, and CV = 2.339. The count $f_k$ is the number of species represented by exactly $k$ individuals in the survey. Diversities: $S = ^0D = 104, ^1D = 30.4, ^2D = 16.2$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>33</td>
<td>15</td>
<td>13</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>16</th>
<th>17</th>
<th>20</th>
<th>22</th>
<th>39</th>
<th>57</th>
<th>72</th>
<th>88</th>
<th>132</th>
<th>133</th>
<th>178</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table S3.7. Abundance frequency counts of the set including nouns in Macaulay’s essay on Bacon (Good 1953). 2048 distinct words, 8045 words, and CV = 2.196. The count $f_k$ is the number of species represented by exactly $k$ individuals in the survey. Diversities: $S = ^0D = 2048, ^1D = 929.1, ^2D = 351.8$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>990</td>
<td>367</td>
<td>173</td>
<td>112</td>
<td>72</td>
<td>47</td>
<td>41</td>
<td>31</td>
<td>34</td>
<td>17</td>
<td>24</td>
<td>19</td>
<td>10</td>
<td>10</td>
<td>13</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>16</th>
<th>17</th>
<th>18</th>
<th>19</th>
<th>20</th>
<th>21</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>25</th>
<th>26</th>
<th>27</th>
<th>28</th>
<th>29</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>3</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>4</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>31</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
<th>45</th>
<th>48</th>
<th>57</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>$k$</th>
<th>65</th>
<th>76</th>
<th>81</th>
<th>89</th>
<th>255</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_k$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Table S3.8. Abundance frequency counts of the set including chess opening in games published in the British Chess Magazine, 1951 (Good 1953). 174 different openings, \( n = 385 \) games, and \( CV = 1.721 \). The count \( f_k \) is the number of species represented by exactly \( k \) individuals in the survey. Diversities: \( S = 0 \), \( 1D = 174 \), \( 2D = 93.5 \), \( 3D = 44.1 \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>13</th>
<th>14</th>
<th>16</th>
<th>23</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_k )</td>
<td>126</td>
<td>22</td>
<td>5</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table S3.9. Abundance frequency counts of a tomato cDNA library (Quackenbush et al. 2001, Mao 2004). 1825 genes sequenced from 2586 expressed sequence tags (ESTs), and \( CV = 0.982 \). The count \( f_k \) is the number of species represented by exactly \( k \) individuals in the survey. Diversities: \( S = 0 \), \( 1D = 1825 \), \( 2D = 1466.8 \), \( 3D = 929.3 \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_k )</td>
<td>1434</td>
<td>253</td>
<td>71</td>
<td>33</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( k )</th>
<th>23</th>
<th>27</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_k )</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table S3.10. Abundance frequency counts of a mainz cDNA library including 1447 expressed sequence tags (ESTs) classified into 794 genes (Köllner & Greulich 1994, Mao 2006), and \( CV = 1.412 \). The count \( f_k \) is the number of species represented by exactly \( k \) individuals in the survey. Diversities: \( S = 0 \), \( 1D = 794 \), \( 2D = 559.4 \), \( 3D = 265.5 \).

<table>
<thead>
<tr>
<th>( k )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>10</th>
<th>14</th>
<th>15</th>
<th>16</th>
<th>22</th>
<th>37</th>
<th>46</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_k )</td>
<td>472</td>
<td>233</td>
<td>36</td>
<td>22</td>
<td>9</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

References


Appendix S4. Additional simulation results

SPECIES ABUNDANCE MODELS

In addition to the log-normal and the Zipf model that are used in the main text, we also considered the other two models in our simulation study, as described below.

(a) The random uniform model with 200 species. The species relative abundances are \( p_i = c a_i \), where \( (a_1, a_2, \ldots, a_{200}) \) are a random sample from a uniform \((0, 1)\) distribution. Then the species relative abundances \( (p_1, p_2, \ldots, p_{200}) \) represent the true species relative abundances of the complete assemblage and are kept as fixed in the simulation procedures. CV = 0.541 for the set we used in our simulations.

(b) The broken-stick model (MacArthur 1957) with 200 species. The species relative abundances are \( p_i = c a_i \), where \( (a_1, a_2, \ldots, a_{200}) \) are a random sample from an exponential distribution. Or equivalently, \( (p_1, p_2, \ldots, p_{200}) \) is a Dirichlet distribution with parameter 1. CV = 0.960 for the set we used in our simulations.

All simulation procedures and comparisons are the same as those described in the main text. So the details are omitted here. The results for the random uniform and the broken-stick model are shown respectively in Fig. S4.1 and Fig. S4.2.

ACTUAL SURVEYS ARE USED TO DEFINE THE COMPLETE TRUE ASSEMBLAGE

We used the ten large survey data sets tabulated in Appendix S3 as the “true” assemblages and generated data sets from these assemblages in our simulation study. The surveys by Miller & Wiegert (1989) and Norden et al. (2009) for the LSUR younger second growth forest site are presented in the main text. The simulation results for the other eight surveys are shown in Fig. S4.3–Fig. S4.10 below. See the main text for all simulation procedures.

All plots reveal similar patterns as those presented in the main text and lead to generally consistent findings and conclusions; we thus omit the comparison details.
The random uniform model with 200 species, CV = 0.541

Fig. S4.1. Simulation comparison of the true diversity profile, empirical diversity profiles and the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on 200 replications. Data sets were generated from the random uniform model with 200 species. \( n = 200 \) (upper panels) and \( n = 400 \) (lower panels).
The broken-stick model with 200 species, CV = 0.960

Fig. S4.2. Simulation comparison of the true diversity profile, empirical diversity profiles and the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on 200 replications. Data sets were generated from the broken-stick model with 200 species. \( n = 200 \) (upper panels) and \( n = 400 \) (lower panels).
Data from Fisher, Corbet & Williams (1943), CV=1.435

True diversities (black line): $S = ^0D = 620, ^1D = 309.7, ^2D = 202.9$;
Mean empirical diversities (light blue line): $^0D_{obs} = 413.5, ^1D_{obs} = 264.4, ^2D_{obs} = 183.6$;
Mean proposed diversities (red line): $^0\hat{D} = 519.2, ^1\hat{D} = 305.2, ^2\hat{D} = 202.0$.

Fig. S4.3. Simulation comparison of the true diversity profile, empirical diversity profiles and the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on 200 replications. Data sets were generated from the Malayan butterfly survey (Fisher, Corbet & Williams 1943), 620 species, 9031 individuals. \(n = 2000\) (upper panels) and \(n = 4000\) (lower panels).
Data from Fisher, Corbet & Williams (1943), CV = 2.915

True diversities (black line): \(S = {^0}D = 240, {^1}D = 59.2, {^2}D = 25.4\);

Mean empirical diversities (light blue line): \(0 D_{obs} = 175.8, 1 D_{obs} = 56.7, 2 D_{obs} = 25.0\);

Mean proposed diversities (red line): \(0 \hat{D} = 209.4, 1 \hat{D} = 58.9, 2 \hat{D} = 25.2\).

Fig. S4.4. Simulation comparison of the true diversity profile, empirical diversity profiles and the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on 200 replications. Data sets generated from the Lepidoptera trapping at Rothamsted Station (Fisher, Corbet & Williams 1943), 240 species, 15609 individuals. \(n = 3000\) (upper panels) and \(n = 4000\) (lower panels).
Data from Norden et al. (2009), Colwell et al. (2012), CV=1.545

True diversities (black line): $S = {}^0D = 152$, $^1D = 74.6$, $^2D = 45.1$;
Mean empirical diversities (light blue line): $^0D_{obs} = 65.0$, $^1D_{obs} = 47.7$, $^2D_{obs} = 35.3$;
Mean proposed diversities (red line): $^0\tilde{D} = 114.3$, $^1\tilde{D} = 71.8$, $^2\tilde{D} = 46.1$.

Fig. S4.5. Simulation comparison of the true diversity profile, empirical diversity profiles and the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on 200 replications. Data sets were generated from the tree assemblage in LEP old growth forest site in northeastern Costa Rica (Norden et al. 2009; Colwell et al. 2012), 152 species, 943 individuals. $n = 150$ (upper panels) and $n = 300$ (lower panels).
Data from Norden et al. (2009), Colwell et al. (2012), CV = 2.339

True diversities (black line): $S = ^0D = 104, ^1D = 30.4, ^2D = 16.2$; 
Mean empirical diversities (light blue line): $^0D_{obs} = 33.6, ^1D_{obs} = 21.0, ^2D_{obs} = 14.5$; 
Mean proposed diversities (red line): $^0\hat{D} = 72.4, ^1\hat{D} = 29.9, ^2\hat{D} = 16.9$.

Fig. S4.6. Simulation comparison of the true diversity profile, empirical diversity profiles and the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on 200 replications. Data sets were generated from the tree assemblage in LEP older (29 years) second growth forest site in northeastern Costa Rica (Norden et al. 2009; Colwell et al. 2012), 104 species, 1263 individuals. $n = 100$ (upper panels) and $n = 200$ (lower panels).
**Data from Good (1953), CV = 2.196**

True diversities (black line): \( S = 0D = 2048, 1D = 929.1, 2D = 351.8 \);

Mean empirical diversities (light blue line): \( 0D_{obs} = 1369.3, 1D_{obs} = 743.4, 2D_{obs} = 329.7 \);

Mean proposed diversities (red line): \( 0D = 1899.3, 1D = 920.7, 2D = 353.0 \).

---

**Fig. S4.7.** Simulation comparison of the true diversity profile, empirical diversity profiles and the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on 200 replications. Data sets were generated from the set including nouns in Macaulay’s essay on Bacon (Good 1953), 2048 distinct words, 8045 words. \( n = 5000 \) (upper panels) and \( n = 7000 \) (lower panels).
Data from Good (1953), CV = 1.721

True diversities (black line): $S = 0D = 174, 1D = 93.5, 2D = 44.1$;
Mean empirical diversities (light blue line): $0D_{obs} = 89.3, 1D_{obs} = 59.5, 2D_{obs} = 36.7$;
Mean proposed diversities (red line): $0\hat{D} = 164.9, 1\hat{D} = 92.6, 2\hat{D} = 45.0$.

$n = 200$

True diversities (black line): $S = 0D = 174, 1D = 93.5, 2D = 44.1$;
Mean empirical diversities (light blue line): $0D_{obs} = 110.9, 1D_{obs} = 68.0, 2D_{obs} = 39.0$;
Mean proposed diversities (red line): $0\hat{D} = 170.9, 1\hat{D} = 93.4, 2\hat{D} = 44.9$.

$n = 300$

Fig. S4.8. Simulation comparison of the true diversity profile, empirical diversity profiles and the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on 200 replications. Data sets were generated from the set including chess opening in games published in the British Chess Magazine, 1951 (Good 1953), 174 different openings, 385 games. $n = 200$ (upper panels) and $n = 300$ (lower panels).
Data from Quackenbush et al. (2001) and Mao (2004), CV=0.982

True diversities (black line): $S = 0D = 1825$, $1D = 1466.8$, $2D = 929.3$;
Mean empirical diversities (light blue line): $D_{obs} = 702.6$, $D_{obs} = 606.2$, $D_{obs} = 481.7$;
Mean proposed diversities (red line): $\hat{D} = 1652.2$, $\hat{D} = 1417.1$, $\hat{D} = 933.4$.

Fig. S4.9. Simulation comparison of the true diversity profile, empirical diversity profiles and
the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on
200 replications. Data sets generated from a tomato cDNA library (Quackenbush et al. 2001,
Mao 2004), 1825 genes sequenced from 2586 expressed sequence tags (ESTs). $n = 1000$ (upper
panels) and $n = 1500$ (lower panels).
Data from Köllner & Greulich (1994) and Mao (2006), CV=1.412

True diversities (black line): $S = 0D = 794, 1D = 559.4, 2D = 265.5$;
Mean empirical diversities (light blue line): $0D_{obs} = 274.5, 1D_{obs} = 226.5, 2D_{obs} = 161.3$;
Mean proposed diversities (red line): $0\hat{D} = 678.5, 1\hat{D} = 532.6, 2\hat{D} = 273.1$.

Fig. S4.10. Simulation comparison of the true diversity profile, empirical diversity profiles and the proposed diversity profiles (left panels), and comparison of RMSEs (right panels) based on 200 replications. Data sets generated from a mainz cDNA library including 1447 expressed sequence tags (ESTs) classified into 794 genes (Köllner & Greulich 1994; Mao 2006). $n = 400$ (upper panels) and $n = 800$ (lower panels).
References


**Appendix S5. More comparison with the empirical profiles**

As discussed in the main text, our diversity and entropy profile estimators are mainly designed to take into account the effect of the undetected species in samples. Thus, our real data analysis and all simulation comparisons presented in the main text and in Appendix S4 are 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 $q > 3$. Here, we provide some theoretical investigations and use the application example discussed in the main text for numerical comparisons.

To compare the two diversity and entropy profiles for $q > 3$, without losing of generality, we consider the case that $q$ is an integer. Our bias-corrected approach starts with the unbiased estimation of the $q$-th order basic sum, $\sum_{i=1}^{S} r_{i}^{q}$. Under a multinomial model for species frequencies, statistical theory implies that the unique minimum variance unbiased estimator for the $q$-th order basic sum is

$$\sum_{X_{i} \geq q} \frac{X_{i}(X_{i}-1)...(X_{i}-q+1)}{n(n-1)...(n-q+1)}.$$

For given data, this estimator is positive if $q \leq \max X_{i}$, where $\max X_{i}$ denotes the maximum species frequency; this estimator is zero if $\max X_{i} \leq q \leq n$, where $n$ denotes the sample size. Statistical theory also implies that no unbiased estimator exists if $q > n$.

Our proposed bias-corrected diversity estimator of an integer order $q \geq 2$ takes a negative exponent of the unbiased estimator given in (S5.1):

$$q \hat{D} = \left( \sum_{X_{i} \geq q} \frac{X_{i}(X_{i}-1)...(X_{i}-q+1)}{n(n-1)...(n-q+1)} \right)^{\frac{1}{1-q}}.$$

See eqn 7c of the main text. If $q > \max X_{i}$, then the unbiased estimator for the basic sum is either zero or does not exist, implying in such cases there exists no positive bias-corrected
estimator. Thus, the maximum order of $q$ that we have a bias-corrected estimator for diversity is the maximum species frequency of the given data. Unlike the empirical diversity profile that we can consider an order of infinity, our bias-corrected approach does not exist for the order of infinity.

The empirical diversity of order $q$ is the observed diversity which can be expressed as

$$qD_{\text{obs}} = \left( \sum_{X \geq 1} \left( \frac{X}{n} \right)^q \right)^{1/(1-q)}. \tag{S5.3}$$

The proposed and the empirical estimators of diversity of order $q$ should yield close results for any species frequency $X$ provided

$$\log \frac{X(X-1)\ldots(X-q+1)}{n(n-1)\ldots(n-q+1)} \approx \log \left( \frac{X}{n} \right)^q. \tag{S5.4}$$

Notice that

$$\frac{X(X-1)\ldots(X-q+1)}{n(n-1)\ldots(n-q+1)} = \left( \frac{X}{n} \right)^q \prod_{k=1}^{q-1} \left( 1 - \frac{k}{X} \right) \prod_{k=1}^{q-1} \left( 1 - \frac{k}{n} \right).$$

Taking the logarithm of the both sides and expanding the logarithm function, we obtain

$$\log \frac{X(X-1)\ldots(X-q+1)}{n(n-1)\ldots(n-q+1)} \approx \log \left( \frac{X}{n} \right)^q - \frac{q(q-1)}{2X} + \frac{q(q-1)}{2n}.$$

If $q(q-1)$ is less than $X$ and $n$ is large, then the discrepancy between the two terms in (S5.4) is less than 0.5 (and thus can be rounded to 0) so that the approximation formula (S5.4) can be heuristically considered to be satisfactory for most applications. Since the diversity of a large order $q > 3$ is dominated by the largest relative abundance and those of the same magnitude, implying that our 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). For example, the maximum frequency for Osa old-growth beetle species data is 42, giving $R = 6.5$; the corresponding maximum frequency for Osa second-growth forest is 77, giving $R = 8.8$. The numerical results are shown in Table S5.1 below for $q$ between 0 and 10. We have also tested other real surveys listed in Appendix S3 and all surveys yield generally consistent results.
Our investigations suggest that, 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 $R$, the 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.

Table S5.1. (This is an expanded version of Table 2 in the main text.) Comparison of the empirical and proposed diversity estimates for $q = 0$ to 3 in increments of 0.5, and also from $q = 3$ to 10 in increments of 1 (all with s.e. based on a bootstrap method of 1000 replications) for beetle species data in the second-growth and old-growth sites (Janzen 1973). In Table 2 and Fig. 3 of the main text, the comparisons are made only for the order $q$ between 0 and 3; here we give the results for $q$ between 3 and 10 to show that the empirical and proposed estimates differ little beyond $q = 2$.

<table>
<thead>
<tr>
<th>Diversity order</th>
<th>Old-growth site</th>
<th>Second-growth site</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Empirical diversity (s.e.)</td>
<td>Proposed diversity (s.e.)</td>
</tr>
<tr>
<td>$q = 0$</td>
<td>112.0 (6.4)</td>
<td>463.3 (106.1)</td>
</tr>
<tr>
<td>$q = 0.5$</td>
<td>86.3 (6.6)</td>
<td>290.5 (56.7)</td>
</tr>
<tr>
<td>$q = 1$</td>
<td>58.3 (6.1)</td>
<td>123.5 (18.9)</td>
</tr>
<tr>
<td>$q = 1.5$</td>
<td>35.6 (5.1)</td>
<td>47.2 (8.2)</td>
</tr>
<tr>
<td>$q = 2$</td>
<td>22.5 (4.2)</td>
<td>24.8 (5.2)</td>
</tr>
<tr>
<td>$q = 2.5$</td>
<td>16.2 (3.2)</td>
<td>17.0 (3.6)</td>
</tr>
<tr>
<td>$q = 3$</td>
<td>13.0 (2.5)</td>
<td>13.5 (2.7)</td>
</tr>
<tr>
<td>$q = 4$</td>
<td>10.0 (1.8)</td>
<td>10.4 (2.0)</td>
</tr>
<tr>
<td>$q = 5$</td>
<td>8.7 (1.5)</td>
<td>9.1 (1.7)</td>
</tr>
<tr>
<td>$q = 6$</td>
<td>8.0 (1.4)</td>
<td>8.5 (1.6)</td>
</tr>
<tr>
<td></td>
<td>( q = 7 )</td>
<td>( q = 8 )</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>( a )</td>
<td>7.5 (1.3)</td>
<td>7.2 (1.2)</td>
</tr>
<tr>
<td>( b )</td>
<td>8.1 (1.5)</td>
<td>7.9 (1.4)</td>
</tr>
<tr>
<td>( c )</td>
<td>16.9 (1.0)</td>
<td>16.4 (1.0)</td>
</tr>
<tr>
<td>( d )</td>
<td>17.7 (1.1)</td>
<td>17.3 (1.2)</td>
</tr>
</tbody>
</table>

**Reference**


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

In our proposed diversity and entropy profile estimators for individual-based abundance data, we assume that individuals are randomly selected with replacement, in which individuals can be repeatedly observed. In this appendix, we consider the behavior of our diversity profile estimator under another type of sampling scheme, sampling without replacement. This sampling scheme is widely used in trap and net surveys when individuals such as insects are killed when sampled, so that no sampled individual can be repeatedly observed.

Under sampling with replacement, species sample frequencies are typically modeled as a multinomial distribution in terms of species relative abundances, as we discussed in the main text. When sampling is done without replacement, and the goal of the investigation is to calculate the diversity of a finite set of objects, dependence among individuals is induced. In this case, any adequate model for species sample frequencies should be formulated in terms of species absolute abundances. A natural choice of such a model is the generalized hypergeometric model (Chao & Lin (2012). For species richness, Chao and Lin (2012) derived an estimator under this model; they also concluded that if sampling is done without replacement, the traditional estimators derived under sampling with replacement tend to overestimate richness when sampling fractions (ratio of the sample size to the population size) are high, and do not converge to the true species richness when the sampling fraction approaches one. In this appendix, we report simulation results to show that a similar conclusion is also valid for diversity profiles for $q > 0$.

The two large surveys presented in the main text were used to define the true assemblages: (a) Miller & Wiegert (1989) plant species survey with 188 species, 1008 individuals, and (b) Norden et al. (2009) tree species survey with 76 species, 1020 individuals. See the main text for details. The species abundance frequency counts for these two surveys are respectively
given in Table S4.1 and Table S4.2 (Appendix S4). These counts are used as the species absolute abundances of the complete assemblage in the simulations.

We examine the behavior of our diversity profile estimator under both sampling schemes. For sampling without replacement, we report five cases (10%, 30%, 50%, 70%, 90%) in Fig. S6.1. For example, when the sampling fraction = 10% from the assemblage defined by the Miller & Wiegert (1989) survey data, we randomly selected of a sample of size = 101 individuals (1008×10% ~ 101) from the assemblage without replacement. The same size sample of 101 individuals was also selected for sampling with replacement. For each sampling fraction, we generated 200 sub-samples under both sampling schemes. In the left panels of Fig. S6.1 we compare, for $q$ between 0 and 3, the average values (over 200 simulation trials) of our diversity profile estimators under the two sampling schemes with the known diversity profiles. The sample root mean squared errors (RMSE) based on 200 estimates are provided in the right panels. Our general findings are summarized below.

(1) First notice from all the left panels in Fig. S6.1 that the estimated diversity profile under sampling without replacement is always above the corresponding profile under sampling with replacement. For species richness ($q = 0$), when sampled individuals cannot be repeatedly observed, the probability of observing rare species becomes relatively higher in subsequent selections. Thus, more undetected species would be inferred in our statistical estimation. A similar explanation applies to the effective number of species when $q > 0$.

(2) As expected by statistical theory, the performances of our estimator under both sampling schemes are generally similar when sampling fraction is low (10%); the two diversity profiles are close and the two RMSE curves differ little.

(3) Although our proposed diversity profiles derived under sampling with replacement underestimate the true profiles when sample sizes are relatively low, the profiles converge correctly to the true curve when sample size is increased. The magnitude of bias and RMSE all show the expected decreasing pattern when the sample size becomes large.
(4) Under sampling without replacement, the proposed profiles are monotonically increasing with the sampling fraction, but they overestimate the true species richness when the sampling fraction is > 30%. They do not converge to the true profile when sampling fraction approaches one. This is consistent with the conclusion for species richness (Chao & Lin 2012). Besides, the magnitude of bias and RMSE may exhibit counter-intuitive pattern, i.e., they may increase with sampling fractions.

Our simulations suggest that when sampling fraction is relatively high (say, > 30%), research is needed to develop new estimators for sampling data without replacement. Chao & Lin (2012) derived a species richness estimator under sampling without replacement and their estimator is in terms of sampling fraction; it is worthwhile to derive for $q > 0$ the diversity profile estimators that incorporate sampling fraction.
(a) Miller & Wiegert (1989) survey, 188 species, 1008 individuals, CV = 1.563

Sample size = 101, sampling fraction = 10%

Sample size = 302, sampling fraction = 30%
(b) Norden et al. (2009) survey, 76 species, 1020 individuals, CV = 2.305

Sample size = 102, sampling fraction = 10%

Sample size = 306, sampling fraction = 30%
Sample size = 510, sampling fraction = 50%

Sample size = 714, sampling fraction = 70%

Sample size = 918, sampling fraction = 90%
**Fig. S6.1.** Simulation comparison of the proposed diversity profiles under two sampling schemes 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 under both sampling with and without replacement. The left panels compare the true diversity profile (black line), and the mean proposed diversity profiles under sampling with replacement (solid red line) and without replacement (green dashed line). The right panels compare the root mean squared errors (RMSEs) of the proposed diversity profiles based on 200 replications.

**References**
Appendix S7. Diversity profile estimation for multiple incidence data

INICIDENCE DATA AND MODEL

The model formulation and sampling models in this Appendix generally follow Colwell et al. (2012) and Chao et al. (2014). Assume that there are $S$ species in the assemblage and there are $T$ sampling units. The detection or non-detection of each species within each sampling unit is recorded to form a species-by-sampling-unit incidence matrix ($W_{ij}$) with $S$ rows and $T$ columns. The value of the element $W_{ij}$ of this matrix is 1 if species $i$ is detected in the $j$-th sampling unit, and 0 if it is undetected. The row sum of the incidence matrix $Y_i = \sum_{j=1}^{T} W_{ij}$ denotes the incidence-based sample frequency of species $i$, $i = 1, 2, ..., S$. Here, $Y_i$ is analogous to $X_i$ in the individual-based abundance data discussed in the main text.

We assume that the $i$-th species has its own unique incidence probability $\pi_i$ that is constant for any randomly selected sampling unit. Each element $W_{ij}$ in the incidence matrix is a Bernoulli random variable, so the probability distribution for the incidence matrix is

$$ P(W_{ij} = w_{ij}; i = 1, 2, ..., S, j = 1, 2, ..., T) = \prod_{i=1}^{S} \prod_{j=1}^{T} \pi_i^{w_{ij}} (1 - \pi_i)^{1-w_{ij}} = \prod_{i=1}^{S} \pi_i^{Y_i} (1 - \pi_i)^{T-Y_i}. $$

The model is equivalent to a binomial product model for the row sums $(Y_1, Y_2, ..., Y_S)$:

$$ P(Y_i = y_i; i = 1, 2, ..., S) = \prod_{i=1}^{S} \binom{T}{Y_i} \pi_i^{Y_i} (1 - \pi_i)^{T-Y_i}. \quad (S7.1) $$

Let $S_{obs}$ denote the number of observed species in the sample, and $Q_k$ denote the incidence frequency counts, the number of species that are detected in exactly $k$ sampling units, $k = 0, 1, ..., T$. In other words, the count $Q_k$ is the number of species each represented exactly $k$ times in the incidence matrix. Here $Q_k$ is analogous to $f_k$ in the abundance data: $Q_0$ represents the number of species present in the assemblage but are not undetected species in the sample, $Q_1$...
represents the number of unique species (those that are each detected in only one sampling unit) and \( Q_2 \) represents the number of duplicate species (those that are each detected in exactly two sampling units). The total number of incidences \( U \) recorded in the \( T \) sampling units is analogous to the sample size \( n \) in the abundance data. However, \( n \) is fixed for abundance data whereas \( U \) here is a random variable and can be expressed as \( U = \sum_{i=1}^{T} kQ_i = \sum_{i=1}^{S} Y_i \).

**HILL NUMBERS FOR INCIDENCE DATA**

Based on the model in eqn (S7.1), \( \sum_{i=1}^{S} \pi_i \) may be greater than 1. Chao et al. (2014) first normalized each parameter \( \pi_i \) (i.e. divide each \( \pi_i \) by the sum \( \sum_{j=1}^{S} \pi_j \)) to yield \( \lambda_i = \pi_i / \sum_{j=1}^{S} \pi_j \), the relative incidence of the \( i \)-th species in the assemblage. This relative incidence is assumed to be the same for any randomly selected sampling unit. Chao et al. (2014) defined the Hill numbers of order \( q \) for incidence data as:

\[
\hat{D}_{q} = \left( \frac{\sum_{i=1}^{S} \pi_i}{\sum_{j=1}^{S} \pi_j} \right)^{q} = \left( \sum_{i=1}^{S} \lambda_i \right)^{\frac{1}{1-q}}, \quad q \geq 0, \ q \neq 1. \tag{S7.2}
\]

The parameter \( q \) in eqn (S7.2) determines the sensitivity of \( \hat{D}_{q} \) to the relative incidences. If all the incidence probabilities \( (\pi_1, \pi_2, \ldots, \pi_S) \) are identical, then Hill numbers of all orders become species richness. The Hill number \( \hat{D}_{q} \) for incidence data is interpreted as the effective number of *equally frequent* species in the assemblage from which the sampling units are drawn; see Chao et al. (2014). That is, if \( \hat{D}_{q} = y \), then the diversity of the assemblage is the same as that of an idealized assemblage with \( y \) species all of equal probability of incidence.

Eqn (S7.2) yields species richness for incidence data when \( q = 0 \). The limit of \( \hat{D}_{q} \) as \( q \) tends to 1 exists and gives

\[
\hat{D} = \lim_{q \to 1} \hat{D}_{q} = \exp \left( -\sum_{i=1}^{S} \frac{\pi_i}{\sum_{j=1}^{S} \pi_j} \log \frac{\pi_i}{\sum_{j=1}^{S} \pi_j} \right), \tag{S7.3}
\]

which is equal to the Shannon diversity for incidence data, i.e., the exponential of Shannon entropy based on the relative incidences in the assemblage. When \( q = 2 \), eqn (S7.2) becomes
\[ 2D = 1 \sum_{i=1}^{S} \left( \frac{\pi_i}{\sum_{j=1}^{S} \pi_j} \right)^2, \]  
(S7.4)

which is the Simpson diversity for incidence data, i.e., the inverse Simpson concentration based on relative incidences. By analogy to the case for abundance data, a plot of \( D_q \) versus \( q \geq 0 \) completely characterizes the species diversity of an assemblage with \( S \) species and incidence probabilities \( (\pi_1, \pi_2, \ldots, \pi_S) \).

**DIVERSITY ACCUMULATION CURVE AND SLOPES**

As with the abundance data, let \( S(t) \) be the expected number of species when \( t \) sampling units are taken (with replacement) from the assemblage. Given species incidence probabilities, we obtain

\[ S(t) = \sum_{i=1}^{S} [1 - (1 - \pi_i)^t] = S - \sum_{i=1}^{S} (1 - \pi_i)^t, \quad t = 0, 1, 2, \ldots, \]  
(S7.5)

with \( S(0) = 0 \) and \( S(1) = \sum_{j=1}^{S} \pi_j \). The incidence-based species accumulation curve (SAC) plots the expected species richness \( S(t) \) versus the number of sample units \( t \). The horizontal asymptote of this curve as \( t \) tends to infinity is the true species richness. The slope of the line connecting two adjacent points \((t, S(t))\) and \((t+1, S(t+1))\) is

\[ \Delta(t) = \frac{S(t+1) - S(t)}{(t+1) - t} = \sum_{i=1}^{S} \pi_i (1 - \pi_i)^t, \]  
(S7.6)

with the initial value \( \Delta(0) = \sum_{j=1}^{S} \pi_j \).

Based on data of \( T \) sampling units with non-zero species incidence-based frequencies \((Y_1, Y_2, \ldots, Y_S)\) and incidence frequency counts \((Q_1, Q_2, \ldots, Q_T)\), the minimum variance unbiased estimator for the expected slope \( \Delta(t) \) exists for any size \( t \) less than \( T \):

\[ \hat{\Delta}(t) = \sum_{l=1}^{T} \frac{Y_l}{T} \left[ \begin{array}{c} T - Y_l \\ t \\ \end{array} \right] = \sum_{l=1}^{T} \frac{Y_l - 1}{T} \left[ \begin{array}{c} T - t - 1 \\ T \\ \end{array} \right], \quad t < T. \]  
(S7.7)
Define

\[
A^* = \begin{cases} 
  2Q_2/[(T-1)Q_1 + 2Q_2] & \text{if } Q_2 > 0 \\
  2/[(T-1)(Q_1 - 1) + 2] & \text{if } Q_2 = 0, Q_1 \neq 0 \\
  1 & \text{if } Q_2 = Q_1 = 0
\end{cases}
\]

Following a similar approach as that in Chao & Jost (2012), we have the following estimator for \( \Delta(T + m) \):

\[
\hat{\Delta}(T + m) = \frac{Q}{T} (1 - A^*)^{m+1}, \quad m \geq 0.
\]  

**ESTIMATION OF DIVERSITY PROFILE**

The empirical profiles for incidence data are obtained by substituting species sample relative incidence proportions, \( \hat{\pi}_i = \hat{\pi}_i / \sum_{j=1}^{S} \hat{\pi}_j = (Y_i / T) / \sum_{j=1}^{S} (Y_j / T) = \sum_{j=1}^{S} Y_j \equiv Y_i / U \), into Hill numbers for \( \lambda_i = \pi_i / \sum_{j=1}^{S} \pi_j \). As with the case of abundance data, our simulations have also revealed that the empirical profiles typically underestimate the true population profile for low values of \( q \) because samples usually miss of the assemblage’s species due under-sampling. Our derivation in the main text for the abundance data can be extended by parallel steps to obtain an accurate diversity profile estimator for incidence data. As in eqn 3a of the main text, we first express the \( q \)-th order basic sum of the species relative incidence probabilities as:

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

\[
= \sum_{i=1}^{S} \sum_{t=0}^{\infty} \pi_i \left( \frac{q-1}{t} \right) (-1)^t (1 - \pi_i)^t = \sum_{i=0}^{\infty} \left( \frac{q-1}{t} \right) (-1)^t \Delta(t).
\]  

Then the diversity of order \( q \) can be expressed as a non-linear function of the slopes of the SAC:

\[
q^D = \left( \frac{\sum_{i=1}^{S} \pi_i}{\sum_{j=1}^{S} \pi_j} \right)^{1/(1-q)} = \left( \frac{\sum_{j=1}^{S} \pi_j}{\sum_{i=1}^{S} \pi_j} \right)^{q/(q-1)} \left( \sum_{j=0}^{q-1} \left( \frac{q-1}{t} \right) (-1)^t \Delta(t) \right)^{1/(1-q)}, \quad q \geq 0.
\]
Because \( E(U) = T \sum_{j=1}^{S} \pi_j \), an unbiased estimator of the total incidence probabilities, \( \sum_{j=1}^{S} \pi_j \), is \( U/T \). Substituting \( U/T \) into eqn (S7.10) for \( \sum_{j=1}^{S} \pi_j \) and also replacing the slopes there by estimators in eqn (S7.7) and eqn (S7.8), we obtain the proposed diversity estimator:

\[
\hat{D}^q = \left( \frac{U}{T} \right)^{q/(q-1)} \left( \sum_{j=1}^{S} \left( \frac{T - 1}{T} \right)^j \right) \left( \sum_{j=1}^{S} \left( \frac{T - 1}{T} \right)^j \right)^{1/(1-q)}
\]

, \( q \geq 0 \).

After some algebra as we did for the abundance data, the above is simplified to a finite sum:

\[
\hat{D}^q = \left( \frac{U}{T} \right)^{q/(q-1)} \left( \sum_{j=1}^{S} \left( \frac{T - 1}{T} \right)^j \right) \left( \sum_{j=1}^{S} \left( \frac{T - 1}{T} \right)^j \right)^{1/(1-q)}
\]

We consider three special cases:

(1) For \( q = 0 \), eqn (S7.11) gives a species richness estimator \( \hat{D}^0 = \hat{S} = \sum_{t=0}^{\infty} \hat{A}(t) = T \hat{A}(t) \), and we can prove that \( \sum_{t=0}^{\infty} \hat{A}(t) = S_{\text{obs}} \) (the observed species richness in the sample) and \( \sum_{t=0}^{\infty} \hat{A}(t) = \hat{Q}_0 \), where

\[
\hat{Q}_0 = \begin{cases} 
\frac{(T-1) \frac{Q^2}{T}}{(2Q_2)}, & \text{if } Q_2 > 0; \\
\frac{(T-1) \frac{Q(Q-1)}{2}}{T}, & \text{if } Q_2 = 0.
\end{cases}
\]

Thus the species richness estimator becomes the Chao2 estimator (Chao 1987).

(2) For \( q = 1 \), eqn (S7.11) reduces to the exponential of the entropy estimator derived in Chao, Wang & Jost (2013, their Appendix S6). That is,

\[
\hat{D}^1 = \exp(\hat{H}) = \exp\left( \frac{T}{U} \hat{H}_0 + \log \frac{U}{T} \right),
\]

where
\[
\hat{H}_0 = \sum_{k=1}^{T-1} \frac{1}{k} \sum_{1 \leq i \leq T-k} \left( \frac{T - Y_i}{k} \right) + \frac{Q_i}{T} \sum_{m=0}^{\infty} \frac{1}{T + m} (1 - A^*)^{m+1}
\]

\[
= \sum_{1 \leq i \leq T} \frac{Y_i}{T} \left( \sum_{k=1}^{T-1} \frac{1}{k} \right) + \frac{Q_i}{T} (1 - A^*)^{-T+1} \left( - \log A^* - \sum_{r=1}^{T-1} (1 - A^*)^r \right).
\]

(S7.13)

(3) If \( q = 2, 3, \ldots, n \), we have

\[
Q^1_q = \left( \frac{T}{U} \right)^q \sum_{k=0}^{q-1} \left( \frac{q-1}{k} \right) (-1)^k \hat{A}(k) = \left( \frac{T}{U} \right)^q \sum_{i=q}^{T} \frac{Y_i(Y_i-1)...(Y_i-q+1)}{T(T-1)...(T-q+1)}
\]

The right hand side of the above formula is positive only when \( q \leq \text{max } Y_i \) under which our estimator becomes the nearly unbiased estimator:

\[
\hat{D}^1_q = \left( \frac{U}{T} \right)^q \left( \sum_{k=0}^{q-1} \left( \frac{q-1}{k} \right) (-1)^k \hat{A}(k) \right)^{1/(1-q)} = \left( \frac{U}{T} \right)^q \left( \sum_{i=q}^{T} \frac{Y_i(Y_i-1)...(Y_i-q+1)}{T(T-1)...(T-q+1)} \right)^{1/(1-q)}.
\]

(S7.14)

In the special case of \( q = 2 \), we have the estimator

\[
2 \hat{D} = \left( \frac{U}{T} \right)^2 \left( \sum_{i=1}^{2} \frac{Y_i(Y_i-1)}{T(T-1)} \right)^{-1} = \frac{(1-1/T)U^2}{\sum_{i=1}^{2} Y_i(Y_i-1)}.
\]

As with the abundance data discussed in the main text, we suggest that our proposed diversity profiles be plotted for the low orders of \( q \) (\( 0 \leq q \leq 3 \)). A bootstrap method similar to that used for abundance data can be applied to obtain a variance estimator for our proposed diversity estimator; see Chao et al. (2014, Appendix G) for details.

AN EXAMPLE

Foissner and colleagues collected a total of 51 soil samples in Namibia, southwest Africa. The presence or absence of soil ciliate species was recorded in each soil sample. Here, the “sampling unit” corresponds to a standardized soil sample. Detailed sampling locations, procedures, and species identifications are described in Foissner et al. (2002). Among the 51 soil samples, there were 331 detected species. The incidence frequency counts for the ciliate
data are summarized in the following table (Foissner et al. 2002) where $Q_k$ denotes number of species that are detected in exactly $k$ soil samples:

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_k$</td>
<td>150</td>
<td>53</td>
<td>42</td>
<td>18</td>
<td>12</td>
<td>9</td>
<td>10</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>


<table>
<thead>
<tr>
<th>$k$</th>
<th>17</th>
<th>19</th>
<th>20</th>
<th>22</th>
<th>23</th>
<th>24</th>
<th>26</th>
<th>27</th>
<th>29</th>
<th>32</th>
<th>33</th>
<th>34</th>
<th>35</th>
<th>37</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Q_k$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

As with the analysis for abundance data in the main text, we give in Table S7.1 the empirical and proposed diversity estimates and their standard errors (s.e.) for $q = 0$ to 3 in increments of 0.5. The corresponding continuous profile plots along with 95% confidence intervals are shown in Fig. S7.1. All s.e. estimates in Table S7.1 and the confidence intervals in Fig. S7.1 were calculated based on a bootstrap method of 1000 replications using the R script provided in Appendix S8.

Foissner et al. (2002) conjectured that there were still many species present in the study area that were not detected in the 51 soil samples. The Chao2 estimator of the number of undetected species based on eqn (S7.12) gives an estimate of 208.1 (s.e. 43.5) for the minimum number of undetected species. Thus, in our diversity profile for $q = 0$, the empirical species richness is 331 whereas our estimated species richness is 331 + 208.1 = 539.1 (s.e. 36.6). For each order of $q$ between 0 and 1.3, our estimated diversity estimates are significantly higher than the corresponding empirical diversity because the two confidence intervals do not overlap (Fig. S7.1). For $q > 1.3$, the estimated profile is still above the empirical curve, but the two intervals overlap.

Our bias correction method for $q = 2$ suggests that nearly all of the very frequent species have been detected in the sample. For common species ($q = 1$), there are still undetected species. Our method for $q = 0$ reveals that a relatively high proportion of rare species is not detected species in the sample. The Chao2 estimator is a universal lower bound, implying that the true number of species richness is likely to be higher than 539.1. Our method greatly
removes the negative bias associated with the empirical diversity and draws conclusions that conform to biologists’ expectation. Thus, for incomplete incidence data, we recommend the use of our proposed diversity estimation method for practical applications.

**Table S7.1.** The empirical and proposed diversity estimates for $q = 0$ to $3$ in increments of $0.5$ (with s.e. based on a bootstrap method of 1000 replications) for soil ciliate data in 51 samples collected from Namibia, southwest Africa (Foissner et al. 2002).

<table>
<thead>
<tr>
<th>Diversity order</th>
<th>Soil ciliate data</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Empirical diversity (s.e.)</td>
<td>Proposed diversity (s.e.)</td>
<td></td>
</tr>
<tr>
<td>$q = 0$</td>
<td>331 (9.4)</td>
<td>539.1 (35.2)</td>
<td></td>
</tr>
<tr>
<td>$q = 0.5$</td>
<td>242.7 (7.2)</td>
<td>351.2 (15.8)</td>
<td></td>
</tr>
<tr>
<td>$q = 1$</td>
<td>172.0 (5.6)</td>
<td>214.0 (7.6)</td>
<td></td>
</tr>
<tr>
<td>$q = 1.5$</td>
<td>125.1 (4.4)</td>
<td>139.6 (5.2)</td>
<td></td>
</tr>
<tr>
<td>$q = 2$</td>
<td>97.3 (3.7)</td>
<td>103.3 (4.1)</td>
<td></td>
</tr>
<tr>
<td>$q = 2.5$</td>
<td>81.1 (3.1)</td>
<td>84.3 (3.4)</td>
<td></td>
</tr>
<tr>
<td>$q = 3$</td>
<td>71.0 (2.8)</td>
<td>73.3 (3.0)</td>
<td></td>
</tr>
</tbody>
</table>
**Fig. S7.1.** The empirical (dotted lines) and proposed (solid lines) diversity profiles for soil ciliate data 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.

**References**


