5-1-2015

Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory

Anne Chao
National Tsing Hua University

T. C. Hsieh
National Tsing Hua University

Robin L. Chazdon
University of Connecticut

Robert K. Colwell
University of Connecticut

Nicholas J. Gotelli
University of Vermont

See next page for additional authors

Recommended Citation

Chao A, Hsieh TC, Chazdon RL, Colwell RK, Gotelli NJ. Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. Ecology. 2015 May;96(5):1189-201.

Ecology, 96(5), 2015, pp. 1189–1201

Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory

Anne Chao,[1,6] T. C. Hsieh,[1] Robin L. Chazdon,[2,3] Robert K. Colwell,[2,4] and Nicholas J. Gotelli[5]

[1]*Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043 Taiwan*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269 USA*
[3]*Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309 USA*
[4]*University of Colorado Museum of Natural History, Boulder, Colorado 80309 USA*
[5]*Department of Biology, University of Vermont, Burlington, Vermont 05405 USA*

Abstract. Based on a sample of individuals, we focus on inferring the vector of species relative abundance of an entire assemblage and propose a novel estimator of the complete species-rank abundance distribution (RAD). Nearly all previous estimators of the RAD use the conventional "plug-in" estimator $\hat{p}_i$ (sample relative abundance) of the true relative abundance p_i of species i. Because most biodiversity samples are incomplete, the plug-in estimators are applied only to the subset of species that are detected in the sample. Using the concept of sample coverage and its generalization, we propose a new statistical framework to estimate the complete RAD by separately adjusting the sample relative abundances for the set of species detected in the sample and estimating the relative abundances for the set of species undetected in the sample but inferred to be present in the assemblage. We first show that $\hat{p}_i$ is a positively biased estimator of p_i for species detected in the sample, and that the degree of bias increases with increasing relative rarity of each species. We next derive a method to adjust the sample relative abundance to reduce the positive bias inherent in $\hat{p}_i$. The adjustment method provides a nonparametric resolution to the longstanding challenge of characterizing the relationship between the true relative abundance in the entire assemblage and the observed relative abundance in a sample. Finally, we propose a method to estimate the true relative abundances of the undetected species based on a lower bound of the number of undetected species. We then combine the adjusted RAD for the detected species and the estimated RAD for the undetected species to obtain the complete RAD estimator. Simulation results show that the proposed RAD curve can unveil the true RAD and is more accurate than the empirical RAD. We also extend our method to incidence data. Our formulas and estimators are illustrated using empirical data sets from surveys of forest spiders (for abundance data) and soil ciliates (for incidence data). The proposed RAD estimator is also applicable to estimating various diversity measures and should be widely useful to analyses of biodiversity and community structure.

Key words: Good-Turing theory; relative abundance; sample coverage; species abundance distribution (SAD); species-rank abundance distribution (RAD).

Introduction

Most plant and animal assemblages are characterized by a few common species and many uncommon or rare species. A major research aim of ecology is to understand the mechanisms and processes that generate and shape the differences among species abundances (Whittaker 1965, 1970, 1972; see McGill et al. 2007 for a review). A broad array of conceptual and methodological frameworks has been proposed to model and interpret species abundance patterns among assemblages. These previous approaches encompass a wide range of biological and statistical models, from classic analyses of the log series (Fisher et al. 1943), log-normal distribution (Preston 1948), and broken-stick distribution (MacArthur 1957, 1960) to more recent treatments of mechanistic neutral (Caswell 1976, Hubbell 2001) and niche-partitioning (Sugihara 1980, Tokeshi 1990) models; see Magurran (2004) and Magurran and McGill (2011) for overviews.

In this study, we mainly focus on inferring the relative abundance or frequency of every species in an entire focal assemblage, including species undetected by sampling. Based on a sample of n individuals, ecologists often use the conventional "plug-in" estimator ($\hat{p}_i = X_i/n$, sample relative abundance/frequency) to estimate the true relative abundance p_i or probability of species i, where X_i is the number of individuals observed of species i in the sample. These sample relative abundances have routinely been used to compute species diversity and evenness measures

Manuscript received 23 March 2014; revised 15 August 2014; accepted 3 October 14. Corresponding Editor: B. D. Inouye.
[6] E-mail: chao@stat.nthu.edu.tw

(Magurran 2004) and to obtain the empirical plots of the species abundance distribution (SAD) and species-rank abundance distribution (RAD); see McGill et al. (2007). The empirical RAD curve depicts a so-called Whittaker (1965) plot: the sample relative abundance on the y-axis (often with a $\log_{10}$-transformation to accommodate several orders of magnitude), with the species list, ranked from the most abundant species to the least abundant, on the x-axis. Based on a sample of species abundances from an assemblage, we propose a new statistical framework for inferring the SAD/RAD of the entire assemblage. We focus on the RAD estimation because the RAD conveys the same information as the SAD, and the RAD can be used visually to demonstrate the advantages of our approach and to reveal the novelty of our method.

Beginning with seminal work by R. A. Fisher and F. W. Preston in the 1940s, ecologists have fit various statistical models to species or species-rank abundance data; see Magurran (2004) for a review. These distribution-fitting approaches to estimating the complete RAD are entirely dependent on the use of the plug-in estimator for detected species. This approach seems natural and intuitive, because the sample relative abundance is considered to be an unbiased estimator of the true species relative abundance under popular sampling models (Lehmann and Casella 1998). As we explain by simple examples and statistical theory, "unbiasedness" can be achieved only by averaging out all possible species occurrences, including both nonzero occurrences (which are detected in the sample) and zero occurrences (which are not). In nearly all practical applications, however, data consist of the detected species only. The undetected species cannot be included in the data because we do not know whether or not the focal assemblage includes any unobserved species.

This study first addresses the following questions: given the detection of a species in a sample, is its sample relative abundance an unbiased estimator of that species' true relative abundance? If not, can the bias be reduced or eliminated? These questions are related to a longstanding challenge in community ecology of characterizing the relationship between the SAD in the entire assemblage and the observed SAD in a sample. Most previous approaches (e.g., Dewdney 2000, Green and Plotkin 2007) are based on a parametric assumption about the SAD of the entire assemblage. In this study, we provide a simple and transparent nonparametric relationship. For any species detected in the sample, we demonstrate that the plug-in estimator is a positively biased estimator of the true relative abundance of the species when the sample is not complete. We provide a method to reduce this inherent positive bias.

The next question this study addresses is, without assuming a particular statistical distribution for the underlying SAD/RAD, is it feasible to estimate the relative abundances of the undetected species? In Preston's (1948) pioneering work, a log-normal model was used to estimate the portion of the assemblage behind a lower limit of observed abundance that he called the "veil line." The fitted log-normal distribution is used to push back the veil line to estimate the number and relative proportions of the undetected species. But Preston's analysis depends on the restrictive assumption of a known log-normal model. In different contexts, Gotelli et al. (2010) and Chazdon et al. (2011) addressed this problem in a nonparametric way, but it has not previously been applied to the estimation of the complete RAD.

Here, we describe a general method for estimating the RAD for both detected and undetected species to address these questions. Our method is based on the Good-Turing sample coverage theory and a generalization of that theory that is derived for the first time in this study. The basic theory was originally developed by A. Turing and I. J. Good for their famous cryptographic analyses during World War II. Turing never published this theory, but gave permission to Good to publish it (Good 1953, 2000). Good and Turing discovered that the total probabilities (total true relative abundances) for those species detected in a sample (sample coverage) can be very accurately estimated based only on the sample data themselves. This result implies that the complement of sample coverage (the total probabilities for those species undetected in the sample; coverage deficit sensu Chao and Jost [2012]) can also be very accurately estimated. However, as we will show, this information, although essential, is not in itself sufficient to properly adjust for the biases caused by using the plug-in estimator $\hat{p}_i$ of species relative abundance, nor is it sufficient to accurately estimate the relative abundances for undetected species. We generalize the Good-Turing sample coverage theory to show that there are other aspects of undetected species that we can estimate accurately, and that these measures of information are required to construct a complete RAD.

We separately estimate the RAD for species detected and undetected in a sample. Based on the Good-Turing sample coverage theory and its generalization, we show how to adjust the sample relative abundance of each detected species to better estimate its true relative abundance. Using an estimate of the number of undetected species in the sample (the Chao1 estimator; Chao 1984), we assume that the functional form of the relative abundances of undetected species follows a simple geometric series model (although any other models or distributions could be used instead) and derive an estimated RAD for undetected species. We then combine the adjusted relative abundances for detected species and the estimated part for undetected species to obtain an estimator of the complete RAD (or SAD). Using simulations, we compare the empirical RAD based on $\hat{p}_i$ and the proposed, estimated RAD.

TABLE 1. A simple simulation to illustrate the problem with the conventional plug-in estimator (i.e., sample relative abundance) for 10 species (labeled A–J) and their true relative abundances.

Species ID	Sample relative abundances with sample size $n = 100$				Average		True relative abundance
	1	2	3	10 000	Conditional	Unconditional	
A	0.3	0.41	0.29	0.32	0.3009	0.3009	0.3
B	0.08	0.11	0.13	0.08	0.0998	0.0998	0.1
C	0.42	0.34	0.38	0.44	0.3998	0.3998	0.4
D	0.02	0.01	0.05	0.02	0.0315	0.0301	0.03
E	0.1	0.02	0.06	0.04	0.0500	0.0498	0.05
F	0.06	0.07	0.07	0.06	0.0649	0.0648	0.065
G		0.01	0.01	0.01	0.0269	0.0248	0.025
H	0.01	0.01		0.01	0.0194	0.0150	0.015
I		0.02	0.01	0.02	0.0157	0.0099	0.01
J	0.01				0.0127	0.0051	0.005

Notes: A total of 10 000 samples of size 100 were generated from the assemblage. Among the 10 000 samples, the species sample relative abundances for the first three samples and the last sample are shown. A blank cell means that a species was not detected in that sample. For each particular species, the averages of sample relative abundances over 10 000 samples are shown as the unconditional average, i.e., those samples in which that species was not detected (as shown by a blank, for which the species' sample relative abundance is thus simply 0) are also counted in the divisor (the number of samples counted) to compute the average. The conditional average was obtained by averaging only those samples in which the species was detected, i.e., those samples in which that species was not detected are not included in the divisor. For each species, only the divisor differs between the two averages.

Most biological survey data can be classified as abundance data (in which individuals are randomly selected) or incidence data (in which sampling units are randomly selected). For the latter, the sampling unit is often a trap, net, quadrat, plot, or timed survey. For incidence data, the abundance of each species is not recorded; only its detection or non-detection in each sampling unit. Although our study deals primarily with abundance data, we briefly discuss parallel derivations that extend our approach to incidence data.

We illustrate the application of our estimators to an empirical data set of pitfall trap catches of temperate forest spiders for abundance data (Sackett et al. 2011), and a data set of soil ciliates for incidence data based on soil samples (Foissner et al. 2002). The formulas for our estimated RAD are relatively simple to calculate and should improve estimation for a variety of ecological questions in which an estimator of the true RAD is desired. We discuss the potential application of our method to the estimation of various diversity measures derived from the RAD and the assessment of sampling errors of complicated estimators.

Problems with Sample Relative Abundances for Detected Species

Assume that there are S species in the assemblage and that the true species relative abundances or probabilities are $(p_1, p_2, \ldots, p_S)$, $\sum_{i=1}^{S} p_i = 1$. Here, p_i can also be interpreted as the probability that any individual is classified to the ith species. Assume a random sample of n individuals is selected with replacement. Let X_i denote the sample abundance of the ith species in the sample, $i = 1, 2, \ldots, S$. Then $(X_1, X_2, \ldots, X_i, \ldots, X_S)$ is a multinomial distribution with parameters $(p_1, p_2, \ldots, p_i, \ldots, p_S)$, where $\sum_{X_i \geq 1} X_i = n$. Only those species with abundance $X \geq 1$ are detected in sample; those species with abundance $X = 0$ are undetected in sample and are therefore not included in the data.

We use a simple example to explain the problem with the familiar plug-in estimator of relative abundances. Assume that an assemblage consists of 10 species labeled A, B, ..., I, J, as in Table 1, with $p = 0.3$, 0.1, 0.4, 0.03, 0.05, 0.065, 0.025, 0.015, 0.010, and 0.005, respectively (Table 1). Some species are common and some are rare. Assume we take a random sample of 100 individuals, with replacement, from this assemblage. The expected abundances for the 10 species would be 30, 10, 40, 3, 5, 6.5, 2.5, 1.5, 1.0, and 0.5, respectively. However, some of the species with small expected abundances will likely be undetected in any particular sample. We generate 10 000 samples, each with sample size 100. Of the 10 000 samples, we illustrate in Table 1 the sample relative abundances for the first three samples and the last sample. Note that in each sample, some species are not detected. For example, in the first sample, species G and I are not detected (and are thus indicated as blank in Table 1).

For each species, we can calculate two types of averages or expectations for the sample relative abundance: the conditional (on detection) average and the unconditional average. The unconditional average is obtained by averaging over all 10 000 samples, including both detected and undetected species in the calculation: if a species occurs in a particular sample, the sample relative abundance is used in computing the average; if a species does not occur in a particular sample, its estimated relative abundance is 0. The divisor for this

unconditional average is always 10 000 for each species. However, in practice, in a single sample, we can only obtain sample relative abundances conditional on those species that are detected in that sample. This conditional average is obtained by averaging over only those samples in which that species is detected. Therefore, the divisor for the conditional average for rare species may be less than 10 000, because not all samples are included in the divisor.

Table 1 reveals that, for all species, the unconditional averages are very close to the true relative abundances. For abundant species, which are likely to be observed in nearly all samples (such as species A–F), the conditional and unconditional averages are almost identical. For rare species, however, which will be found in few or no samples, the conditional averages are consistently higher than the true relative abundances. For rare species G–J in Table 1, with relative abundances 0.025, 0.015, 0.010, and 0.005, the corresponding conditional averages are 0.0269, 0.0194, 0.0157, and 0.0127. These results imply that the sample relative abundance for any detected species overestimates its true value. The level of overestimation is not uniform, but scales inversely with abundance: estimates for rare detected species are more severely biased than estimates for common detected species. Sample relative abundances do not need to be adjusted for abundant species, but sample relative abundances for rare species have substantial positive relative biases and should be properly adjusted.

Statistical explanation

The level of overestimation of the sample relative abundance for any detected species can be seen by examining the following theoretical conditional average or statistical expectation (Chazdon et al. 2011: Appendix C):

$$E(\hat{p}_i \mid X_i > 0) = E\left(\frac{X_i}{n} \mid X_i > 0\right) = \frac{p_i}{1-(1-p_i)^n} \quad (1)$$

which is the expected proportion of individuals in a sample of size n that represent species i, given that species i has been detected in sample. The denominator $1 - (1 - p_i)^n$ in Eq. 1 is $P(X_i > 0)$, the probability of detection of species i in the sample. Because this denominator is always less than 1, Eq. 1 proves that the sample relative abundance for any detected species consistently overestimates the true probability p_i.

When p_i is relatively large, the denominator $1 - (1 - p_i)^n$ tends to 1, because the species is sufficiently abundant that it would be observed in any sample. Therefore, for relatively common species, the sample relative abundance X_i/n works well as an estimate of p_i, and almost no adjustment is required. In contrast, when p_i is very small, the denominator $1 - (1 - p_i)^n$ is much less than 1, which generates a substantial bias. For example, with a sample size of 100, the probability of detecting the rarest species in Table 1 is $1 - (1 - p_i)^n = 1 - (1 - 0.005)^{100} = 0.394$. The conditional average is $0.005/0.394 = 0.0127$, more than double the correct value (0.005). This theoretical value of 0.0127 is further confirmed by our simulation result (Table 1).

Now we can connect the foregoing discussion to the classic unbiasedness of sample relative abundance in the following sense. For any species i, it will be detected in the sample with probability $P(X_i > 0) = 1 - (1 - p_i)^n$ or it will be undetected with probability $P(X_i = 0) = (1 - p_i)^n$. Then on average, we have

$$\begin{aligned} E\left(\frac{X_i}{n}\right) &= E\left(\frac{X_i}{n} \mid X_i > 0\right) P(X_i > 0) \\ &\quad + E\left(\frac{X_i}{n} \mid X_i = 0\right) P(X_i = 0) \\ &= \frac{p_i}{1-(1-p_i)^n} \times [1 - (1-p_i)^n] \\ &\quad + 0 \times (1-p_i)^n \\ &= p_i. \end{aligned}$$

This (unconditional) expectation, which is valid for all species in the complete assemblage, considers both detection and non-detection and implies unbiasedness.

Simulation Part I: the Empirical RAD

We use a suite of simple simulations to illustrate the undersampling bias with the empirical RAD when sample size is not large enough to detect all species. We simulated data from two theoretical abundance distributions (the Zipf-Mandelbrot model and the log-normal model) and treated four large empirical diversity surveys as the complete assemblages. For the latter cases, the species-rank abundance distribution from each survey was assumed to be the "true" complete distribution; this true RAD was then compared with the empirical RADs obtained from simulated samples of several sample sizes. Here we report in detail only the simulation results for the Zipf-Mandelbrot model for illustration. See Appendix A for simulation results of other scenarios.

In the Zipf-Mandelbrot model, we fix the number of species at 200 and the true relative abundance takes the form $p_i = c/(2 + i)$, $i = 1, 2, \ldots, 200$, where c is a normalized constant such that the sum of the relative abundances is 1. In Fig. 1a, we compare the true complete RAD of the entire assemblage (light blue line) and the empirical RAD based on 200 simulated data sets of sample sizes 200, 400, and 800 (200 superimposed dark blue lines, each line corresponding to an empirical RAD for each generated data set). When the sample sizes are not large enough ($n = 200$ and 400) to detect all species, only about half of the complete RAD can be revealed empirically from the simulated data. Even for a large sample size ($n = 800$), most of the empirical RAD curves still cannot unveil the complete "tail" of the true RAD. Although the observed species (say there are K of them) in a sample may not correspond to the first K species in the true RAD, most of the empirical RAD curves lie above the true

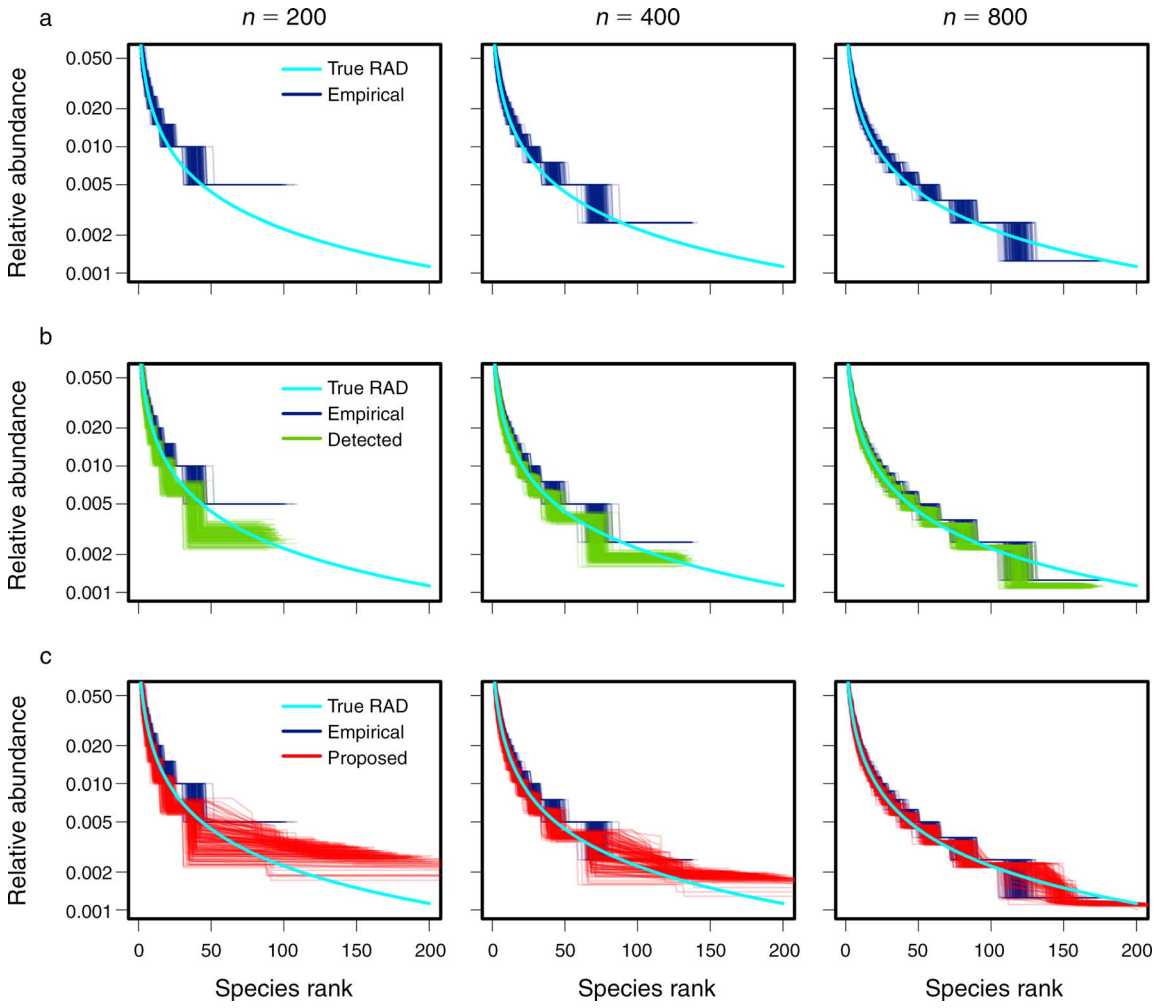

FIG. 1. (a) Comparison of the true species-rank abundance distribution (RAD) of the complete assemblage (light blue line) and the empirical RAD curves (superimposed dark blue lines with 200 replications). (b) Comparison of the true RAD, empirical RAD, and adjusted RAD curves for detected species only (superimposed green lines with 200 replications). (c) Comparison of the true RAD, empirical RAD, and estimated RAD curves for both detected and undetected species (superimposed red lines with 200 replications). For each of the sample sizes 200 (left panels), 400 (middle panels), and 800 (right panels), 200 data sets were generated from the Zipf-Mandelbrot model; thus there are 200 estimated RADs (200 dark blue lines, 200 green lines, and 200 red lines). Note that the x-axis is the species list, ranked from most to least abundant, and the y-axis (relative abundance) is displayed on a $\log_{10}$ scale.

RAD, signifying that the positive bias is associated with the empirical RAD for the detected species, as predicted from our theory (Eq. 1) and shown by an example (Table 1).

GOOD-TURING SAMPLE COVERAGE THEORY AND A GENERALIZATION

Sample coverage and coverage deficit

Let f_k be the number of species represented by exactly k individuals in the sample, $k = 0, 1, \ldots, n$; we refer to f_k as the abundance frequency counts. In particular, f_1 is the number of species represented by exactly one individual (singletons) in the sample, and f_2 is the number of species represented by exactly two individuals (doubletons). The unobservable frequency f_0 denotes the number of species present in the entire assemblage but not detected in the sample. Good and Turing discovered a surprisingly simple estimator for the sample coverage (C; a measure of sample completeness, as defined in *Introduction*) that is a simple function of the number of singletons and the sample size if $f_1 > 0$

$$\hat{C} = 1 - \frac{f_1}{n}. \tag{2a}$$

When f_1, $f_2 > 0$, an improved Turing's coverage estimator (Chao and Jost 2012) is

$${}^1\hat{C} = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right]. \tag{2b}$$

Here, the leading superscript 1 in ${}^1\hat{C}$ refers to the first-

order sample coverage of our generalization (*A generalization*). This improved coverage estimator incorporates information about the doubletons. It is improved in the sense that this coverage estimator generally has smaller mean squared error than Good-Turing's estimator. In the following derivation, we will adopt this more accurate estimator. Subtracting the sample coverage estimator from unity gives the estimator of coverage deficit

$$ {}^1\hat{C}_{\text{def}} = 1 - {}^1\hat{C} = \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right]. \qquad (2c) $$

A tiny percentage of coverage can nevertheless contain a very large number of rare species. The estimated coverage deficit is not an estimate of the number or proportion of undetected species, but rather it is an estimate of the proportion of the total number of individuals in the assemblage that belong to the undetected species. For this reason, extremely rare, undetected species do not make a significant contribution to that proportion, even if there are many such species. This distinction intuitively explains why the estimation of species richness in highly diverse assemblages is so statistically challenging, even though sample coverage for the same data can be accurately estimated.

The coverage estimator and its complement make it possible to adjust the sample relative abundance for detected species and to infer the relative abundance of undetected species. This approach allows models with one parameter, which are useful for assessing sampling variances in some inference problems (Chao et al. 2013: Appendix S2; Chao et al. 2014: Appendix G). However, the one-parameter models are not flexible enough to provide accurate estimators for the complete RAD. For this reason, in this study we extend the Good-Turing concept of coverage for the first time, and develop improved models for estimating species abundances beyond the veil line.

A generalization

Lande et al. (2000) commented, "without regard to the species abundance distribution, the only aspect of unobserved species that can be accurately extrapolated is their total frequency in a community [i.e., coverage deficit], using the number of singletons divided by sample size." In addition to coverage deficit, however, there are other aspects of undetected species that we can measure accurately. To show this, we first generalize the concept of sample coverage to the rth order sample coverage ${}^rC$ as

$$ {}^rC = \frac{\sum_{i\in\text{detected}} p_i^r}{\sum_{i=1}^{S} p_i^r} = \frac{\sum_{i=1}^{S} p_i^r\, I(X_i>0)}{\sum_{i=1}^{S} p_i^r}, \qquad r = 1, 2, \ldots, $$

where indicator function $I(A) = 1$ if event A occurs, and 0 otherwise. The coverage ${}^rC$ is the fraction of the rth power of the true relative abundances of those species detected in sample. For $r = 1$, ${}^1C$ reduces to Good-Turing's sample coverage, and its estimator is given in Eq. 2b. For $r = 2$, ${}^2C$ is the fraction of the squared true relative abundances of the detected species; it quantifies the sample completeness for very abundant or dominant species. When $f_2, f_3 > 0$, ${}^2C$ can be accurately estimated by (see Appendix B for derivation)

$$ {}^2\hat{C} = 1 - \frac{2f_2}{\sum_{X_i\geq 2} X_i(X_i-1)}\left[\frac{(n-2)f_2}{(n-2)f_2 + 3f_3}\right]^2. \qquad (3a) $$

We define the rth order coverage deficit as ${}^rC_{\text{def}} = 1 - {}^rC$. For $r = 1$, ${}^1C_{\text{def}}$ reduces to the coverage deficit defined in Chao and Jost (2012), and its estimator is given in Eq. 2c. For $r = 2$, ${}^2C_{\text{def}}$ can be accurately estimated by

$$ {}^2\hat{C}_{\text{def}} = \frac{2f_2}{\sum_{X_i\geq 2} X_i(X_i-1)}\left[\frac{(n-2)f_2}{(n-2)f_2 + 3f_3}\right]^2. \qquad (3b) $$

As we will see, the estimators for the first- and second-order sample coverages and their deficits make possible two-parameter models for inferring the complete RAD. As proved in Appendix B, if the abundance frequency counts up to f_{r+1} are all nonzero, then ${}^1C, {}^2C, \ldots, {}^rC$ and their deficits can be accurately and efficiently estimated. Thus, in addition to the coverage deficit, we have more information (i.e., higher orders of coverage deficits including ${}^2C_{\text{def}}, {}^3C_{\text{def}}, \ldots, {}^rC_{\text{def}}$) about the undetected species. This information can be used to help estimate the complete RAD.

Unveiling the Complete RAD

Adjusting the sample relative abundances for detected species

Based on Eq. 1, we have

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

If we replace the expected value in this equation with the observed data, then for $X_i > 0$ (i.e., a detected species), we have the following approximation:

$$ p_i \approx \frac{X_i}{n}[1 - (1-p_i)^n] \approx \frac{X_i}{n}[1 - \exp(-np_i)]\,. \qquad (4a) $$

This formula shows that the approximate adjustment factor for the sample relative abundance would be $[1 - (1 - p_i)^n] \approx [1 - \exp(-np_i)]$, which depends mainly on the product of n and p_i. Note that the (unconditional) expected abundance of species i in the sample is np_i, i.e., $E(X_i) = np_i$. However, as we have already argued, the adjustment factor $[1 - (1 - p_i)^n]$ cannot be estimated simply by substituting the sample relative abundance X_i/n for p_i, because the sample relative abundance does not estimate p_i well for rare species. Similarly, replacing np_i

in the adjustment factor $[1 - \exp(-np_i)]$ by the observed abundance X_i for each individual species i does not provide a good estimate. Instead, we introduce two parameters, λ and θ, to the adjustment factor. From Eq. 4a, we assume that parameter $\lambda > 0$ and parameter $0 < \theta \leq 1$, so that for $X_i > 0$, $p_i \approx (X_i/n)(1 - \lambda e^{-\theta X_i})$. Here, parameter θ is restricted to be in [0, 1] because X_i for a detected species overestimates np_i. The special case of $\theta = 1$ reduces to the one-parameter approach discussed in Chao et al. (2013: Appendix S2) and Chao et al. (2014: Appendix G). Here, we adopt a more flexible two-parameter model which performs better for estimating the complete RAD in benchmark simulations. Next, we obtain parameters λ and θ from the estimated first-order (Eq. 2b) and second-order (Eq. 3a) sample coverage by the following equations:

$$\sum_{i \in \text{detected}} p_i \approx \sum_{X_i \geq 1} \frac{X_i}{n}(1 - \lambda e^{-\theta X_i}) = {}^1\hat{C} \qquad (4b)$$

$$\sum_{i \in \text{detected}} p_i^2 \approx \sum_{X_i \geq 1} \left[\frac{X_i}{n}(1 - \lambda e^{-\theta X_i})\right]^2 = {}^2\hat{C} \times \frac{\sum_{X_i \geq 2} X_i(X_i - 1)}{n(n-1)}. \qquad (4c)$$

The rightmost term in Eq. 4c is an unbiased estimator of $\sum_{i=1}^{S} p_i^2$ (i.e., the denominator of ${}^2C$). Let $\hat{\lambda}$ and $\hat{\theta}$ denote the solution of λ and θ, respectively, in this system of nonlinear equations. If the solution $\hat{\theta}$ is out of the range of [0, 1], then we replace it by 1 so that the model reduces to the one-parameter case. The proposed adjusted relative abundance of species i (with $X_i > 0$) is

$$\tilde{p}_i = \frac{X_i}{n}(1 - \hat{\lambda} e^{-\hat{\theta} X_i}). \qquad (4d)$$

This is a unified adjustment formula that is valid for all species abundance distributions. In Appendix A, we show by simulations that the adjusted estimator reduces substantial bias inherent in the plug-in estimator and has a smaller root mean squared error. The proposed adjustment scales inversely with the sample abundance in the following sense: for abundant species, with correspondingly large sample abundance X_i, the adjustment factor $1 - \hat{\lambda}\exp(-\hat{\theta}X_i)$ approaches unity. Thus, virtually no adjustment is needed for abundant species, whereas for rare species, the adjustment factor can be much less than 1. The smaller the abundance X_i, the smaller the adjustment factor and the larger its effect.

Our adjustment formula (Eq. 4d) also provides a simple nonparametric relationship between the plug-in estimator (X_i/n) calculated for a species in a sample and its estimated true relative abundance in the entire assemblage. Note that the formula is a function of sample abundances of all detected species, not merely the sample abundance of species i. This is because, given the sample coverage estimates $({}^1\hat{C}, {}^2\hat{C})$, other species also carry information about species i via $\hat{\lambda}$ and $\hat{\theta}$, which are functions of all sample frequencies (by Eqs. 4b and 4c). Thus, our adjustment formula "borrows strength" from the observed abundances of other species. We will discuss how to assess the sampling error of the adjusted estimator after we obtain an estimator for the complete RAD.

Estimating the relative abundances of undetected species

As discussed, it is difficult to accurately estimate the number of undetected species in an incomplete sample if there are many, almost-undetectable species in a hyper-diverse assemblage. Practically, an accurate lower bound for species richness is preferable to an inaccurate point estimator. A widely used nonparametric lower bound developed by Chao (1984) uses only the information on rare species (numbers of singletons and doubletons) to estimate the number of undetected species in samples, as rare, detected species contain nearly all information about the number of undetected species. This lower bound for the number of undetected is universally valid for any species abundance distribution and has the following form:

$$\hat{f}_0 = \begin{cases} \dfrac{(n-1)}{n}\dfrac{f_1^2}{2f_2} & \text{if } f_2 > 0 \\ \dfrac{(n-1)}{n}\dfrac{f_1(f_1-1)}{2} & \text{if } f_2 = 0. \end{cases} \qquad (5a)$$

See Chao and Chiu (2012) for a recent review. Because the number of species must be an integer in later derivations, we define $\hat{f}_0$ hereafter to be the smallest integer that is greater than or equal to the value computed from Eq. 5a. The empirical RAD ignores the tail, which includes at least $\hat{f}_0$ species. Although this is a lower bound, when sample size is large enough, this lower bound approaches the true number of undetected species. Based on Eq. 5a, we propose a robust method to estimate the species RAD for the undetected species.

We must assume a functional form for the rank abundances of the undetected tail. There are many options for a functional form, and our method is applicable to any functional form. Here, we adopt more flexible, two-parameter models. Because the method is applied to only the undetected tail part of the true RAD, where all relative abundances are low, a simple functional form with estimable parameters is preferable. A natural assumption is that the abundance distribution of the undetected species is a two-parameter geometric series

$$p_i = \alpha\beta^i, i = 1, 2, \ldots, \hat{f}_0 \qquad (5b)$$

where α is a normalized constant (see Eq. 6a) and β is a positive decay factor. If all relative abundances for undetected species are approximately equal, then the parameter β is close to 1.

Based on the coverage deficits of the first- and second-order (Eqs. 2c and 3b), we have the following two

equations in terms of parameters α and β for the undetected species:

$$\sum_{i\in \text{undetected}} p_i \approx \sum_{i=1}^{\hat{f}_0} \alpha\beta^i = {}^1\hat{C}_{\text{def}} \tag{6a}$$

$$\sum_{i\in \text{undetected}} p_i^2 \approx \sum_{i=1}^{\hat{f}_0} (\alpha\beta^i)^2 = {}^2\hat{C}_{\text{def}} \times \frac{\sum_{X_i\geq 2} X_i(X_i-1)}{n(n-1)}. \tag{6b}$$

Let $\hat{\alpha}$ and $\hat{\beta}$ be the solution of this system of nonlinear equations. The proposed estimated relative abundances for the undetected species are

$$\tilde{p}_i = \hat{\alpha}\hat{\beta}^i,\ i = 1, 2, \ldots, \hat{f}_0. \tag{6c}$$

Combining the adjustment method for detected species (Eq. 4d) and the estimated relative abundances for undetected species (Eq. 6c), we can construct a complete RAD based on a sample. See *Discussion* for other possible parametric assumptions about the functional form of the relative abundances of undetected species. To examine the performance of the estimated RAD based on simulations, we first illustrate the estimation procedure step-by-step for an example so that the simulation plots can be better understood.

Example (Abundance Data)

Sackett et al. (2011) collected species abundance data for samples of spiders from four experimental forest-canopy-manipulation treatments at the Harvard Forest (Massachusetts, USA). The treatments were established to study the long-term consequences of loss of the dominant forest tree, eastern hemlock (*Tsuga canadensis*), caused by a nonnative insect, the hemlock woolly adelgid (*Adelges tsugae*; Ellison et al. 2010). To illustrate our method, we use the data of the Hemlock Girdled treatment, in which bark and cambium of hemlock trees were cut and the trees left in place to die, to mimic tree mortality by adelgid infestation. In this experimental treatment, 26 spider species were represented by a total of 168 individuals. The nonzero abundance frequency counts are $f_1 = 12, f_2 = 4, f_4 = 1, f_6 = 2, f_8 = f_9 = 1, f_{15} = 2$, and $f_{17} = f_{22} = f_{46} = 1$. The first- and second-order sample coverage estimates are respectively 92.89% and 99.77%; the corresponding coverage deficits are thus respectively 7.11% and 0.23%. Our estimation procedure includes the following four steps (see Fig. 2): (1) Construct the adjusted RAD for the detected species. For the 26 detected species, first plot the empirical RAD, as shown in Fig. 2a (white plus gray bars). Then use Eqs. 4b and c to obtain $\hat{\lambda} = 0.2980$ and $\hat{\theta} = 0.1267$, and substitute these estimates into Eq. 4d to adjust the sample relative abundances for each detected species downward, as shown in Fig. 2a (white bars). (2) Estimate the RAD for undetected species: based on the Chao1 estimator, which uses the observed numbers of singletons and doubletons, estimate the number of undetected species as $\hat{f}_0 = 18$ species (SE = 13.4). The undetected species are labeled Undetected.1 to Undetected.18 in Fig. 2b. For the 18 undetected species, use Eqs. 6a and 6b to obtain $\hat{\alpha} = 0.0045$ and $\hat{\beta} = 0.9865$, then substitute these two estimates into Eq. 6c to estimate their relative abundances, as shown in Fig. 2b. (3) Combine the adjusted RAD for the detected species in (1) and the estimated RAD for the undetected species in (2) to obtain a complete RAD, as shown in Fig. 2c. A full list of the estimated species relative abundances for the complete RAD is given in Appendix C. (4) In (1) through (3), we use bar plots for clearer illustration. Conventionally, only line plots as those plotted in Fig. 1 are sufficient for comparison. In Fig. 2d, we provide the line plots for the empirical RAD and the proposed RAD estimator.

In (1), notice from Fig. 2 that, for abundant species, virtually no adjustment is needed, whereas the adjustment for rare species is substantial and that scales inversely with the sample abundance. In (2), our estimated number of undetected species is only a lower bound, implying that there may have been additional undetected species, but they cannot be statistically estimated from our inference, so they are treated as having negligible abundances. See *Discussion* for further explanation. In this example (Fig. 2c), the estimated relative abundance for the most abundant of the undetected species is slightly larger than the adjusted species relative abundances of the least-abundant detected species (i.e., singletons). Our analysis shows that the empirical RAD curve differs greatly from the proposed RAD curve in the tail distribution. When there are undetected species, as will be confirmed by simulations in *Simulation II: The complete estimated RAD*, our proposed approach unveils the tail distribution and provides a more complete picture of the true RAD. In Appendix C, as alternatives to the geometric series, we present a Poisson log-normal and a broken-stick model for the relative abundances of the undetected species. See *Discussion* for more details.

Sampling Variances of our Estimators

In the estimated complete RAD, there are $S_{\text{obs}} + \hat{f}_0$ species, where S_{obs} denotes the number of observed species in the sample. This estimated RAD mimics the profile of the complete assemblage. We can thus assess the sampling error of any estimator of a parameter by bootstrapping or resampling the estimated RAD. For example, we can approximate the sampling variance of the adjusted estimator $\tilde{p}_i$ (Eq. 4d) for any particular detected species *i*. For each bootstrap replication, we generate a random sample of *n* individuals from the estimated RAD, with replacement, yielding a new set of species sample abundances (here we retain only those sets in which species *i* is detected, because the estimating target is the relative abundance of a detected species). Based on this new set, we then calculate $({}^1\hat{C},\ {}^2\hat{C})$ to obtain new estimates $(\hat{\lambda}, \hat{\theta})$. All these new statistics are

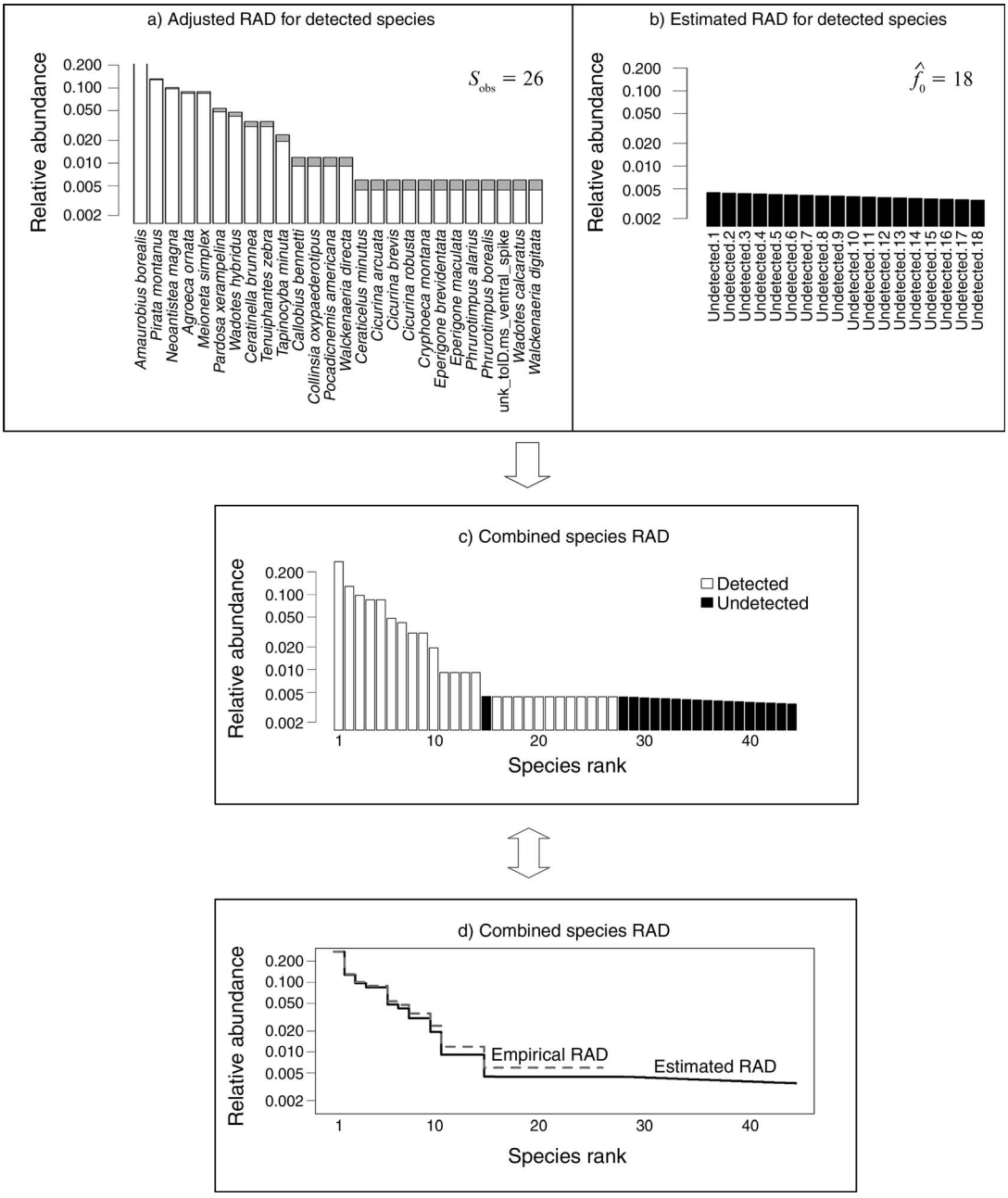

FIG. 2. Combining the adjusted RAD for detected species and the estimated RAD for undetected species based on the abundance data of forest spiders (Sackett et al. 2011). All y-axes are on a $\log_{10}$ scale. (a) The bar plot for the empirical RAD (white plus gray bar) and adjusted RAD (white bars) for detected species (S_{obs}). (b) The bar plot of the estimated RAD for undetected species, where the estimator of undetected species is $\hat{f}_0 = 18$. The undetected species are indexed by Undetected.1 to Undetected.18. (c) Combining the two bar plots in (a) and (b) to construct a complete RAD. (d) The empirical RAD curve is compared with the proposed RAD curve in conventional line plots.

then substituted into Eq. 4d to obtain a bootstrap estimate $\tilde{p}_i^*$, based on the generated sample. The procedure is replicated to obtain B bootstrap estimates $\{\tilde{p}_i^{*1}, \tilde{p}_i^{*2}, \ldots, \tilde{p}_i^{*B}\}$ ($B = 1000$ is suggested in confidence interval construction). The bootstrap variance estimator of our estimator in Eq. 4d is the sample variance of these B estimates. Moreover, the 2.5% and 97.5% percentiles of these B bootstrap estimates can be used to construct a 95% confidence interval. See Appendix C: Table C1 for the bootstrap SE of the adjusted estimator $\tilde{p}_i$ for each detected species in the spider example.

Similar procedures can be used to derive variance estimators for any other estimators (e.g., estimators of sample coverages and their deficits) and to construct the

associated confidence intervals. For our proposed estimator $\tilde{p}_i$ for undetected species (Eq. 6c), however, sampling variance cannot be assessed because the estimated number of undetected species varies with bootstrap samples and the identities of those undetected species are unknown and thus there is no pre-specified target species.

Simulation Part II: the Estimated RAD

The adjusted RAD for detected species (Fig. 1b)

For the scenario considered in Fig. 1a, in addition to the true RAD curve (light blue line) and the empirical RADs (dark blue lines), we now superimpose in Fig. 1b the estimated RADs (green lines) for detected species based on 200 data sets of sample sizes 200, 400, and 800. Thus, there are 200 additional superimposed green lines for each sample size. Most of the green lines are below the empirical RAD, showing the reduction of the positive biases associated with sample relative frequencies for detected species.

The complete estimated RAD (Fig. 1c)

In Fig. 1c, we compare the true RAD curve (light blue line), the empirical RADs (dark blue lines), and the estimated complete RADs including both detected and undetected species (red lines) based on 200 data sets of sample sizes 200, 400, and 800. For sample sizes 200 and 400, the improvement with the estimated RAD is clearly seen: the tail of the true RAD can be revealed, although our estimated tail of RADs for a sample size of 200 unavoidably overestimates the true lines to some extent (i.e., data do not provide sufficient information to accurately infer very small relative abundances; see Appendix A: Fig. A2). When sample size is increased to 400, the proposed RAD curves closely trace the RAD of the complete assemblage; for a sample size of $n = 800$, all the proposed RAD curves match closely with the true RAD curve.

Extension to Incidence Data

Our statistical framework for abundance data can be extended to incidence data by parallel derivations. Here we only outline the extension; all details are provided in Appendix D. Following the notation and terminology used in Colwell et al. (2012) and Chao et al. (2014), we assume that in the focal assemblage there are S species indexed by 1, 2, ..., S. For any sampling unit, assume that the ith species has its own unique incidence (or occurrence) probability π_i that is constant for any randomly selected sampling unit. The incidence probability π_i is the probability that species i is detected in a sampling unit. This incidence probability π_i is analogous to p_i, but $\sum_{i=1}^{S} \pi_i$ may be greater than unity.

As with abundance data, we can similarly define the species incidence distribution (SID) and the corresponding species-rank incidence distribution (RID) for the set $(\pi_1, \pi_2, \ldots, \pi_S)$ of the S species. Our goal here is to estimate the RID based on incidence data of a set of sampling units. Assume that a set of T sampling units are randomly selected from the study area, with replacement. The underlying data consist of a species-by-sampling-unit incidence matrix $\{W_{ij}; i = 1, 2, \ldots, S, j = 1, 2, \ldots, T\}$ with S rows and T columns; here $W_{ij} = 1$ if species i is detected in sampling unit j, and $W_{ij} = 0$ otherwise. Under our assumption that the probability of detecting species i in any sampling unit is a constant π_i, $i = 1, 2, \ldots, S$, the variable W_{ij} for all j follows a Bernoulli distribution with parameter $\pi_i = P(W_{ij} = 1)$, $i = 1, 2, \ldots, S$. Let Y_i be the number of sampling units in which species i is detected, $Y_i = \sum_{j=1}^{T} W_{ij}$; here Y_i is referred to as the sample species incidence frequency and is analogous to X_i in the abundance data. Species present in the assemblage but not detected in any sampling unit yield $Y_i = 0$.

Denote the incidence frequency counts by $(Q_0, Q_1, \ldots, Q_T$, where Q_k is the number of species that are detected in exactly k sampling units in the data, $k = 0, 1, \ldots, T$. Here Q_k is analogous to f_k in the abundance data. The unobservable zero frequency count Q_0 denotes the number of species among the S species present in the assemblage that are not detected in any of the T sampling units. Also, Q_1 represents the number of unique species (those that are detected in only one sampling unit), and Q_2 represents the number of duplicate species (those that are detected in only two sampling units).

Define the sample incidence probability of species i as $\hat{\pi}_i = Y_i/T$ (the plug-in estimator); the empirical RID is based on $\hat{\pi}_i$. Since Y_i, $i = 1, 2, \ldots, S$ follows a binomial distribution with the total number T and the detection probability π_i, a formula parallel to Eq. 1 can be derived

$$E(\hat{\pi}_i \mid Y_i > 0) = E\left(\frac{Y_i}{T} \mid Y_i > 0\right) = \frac{\pi_i}{1 - (1 - \pi_i)^T}. \quad (7)$$

We can similarly define the general rth sample coverage and its deficits for the incidence probabilities $(\pi_1, \pi_2, \ldots, \pi_S)$ based on the sample species incidence frequencies $(Y_1, Y_2, \ldots, Y_S)$. Then, derivation steps parallel to those for abundance data lead to the following adjusted incidence probability for a detected species: $\tilde{\pi}_i = (Y_i/T)(1 - \hat{\lambda}e^{-\hat{\theta} Y_i})$ for $Y_i > 0$, where $\hat{\lambda}$ and $\hat{\theta}$ are solved from two nonlinear equations involving the estimated sample coverage of the first two orders (see Appendix D).

To estimate the RID for undetected species, we first apply the Chao2 estimator (Chao 1987) to obtain an estimated lower bound on the number of undetected species in T sampling units

$$\hat{Q}_0 = \begin{cases} \dfrac{(T-1)}{T}\dfrac{Q_1^2}{2Q_2} & \text{if } Q_2 > 0 \\ \dfrac{(T-1)}{T}\dfrac{Q_1(Q_1-1)}{2} & \text{if } Q_2 = 0. \end{cases} \quad (8)$$

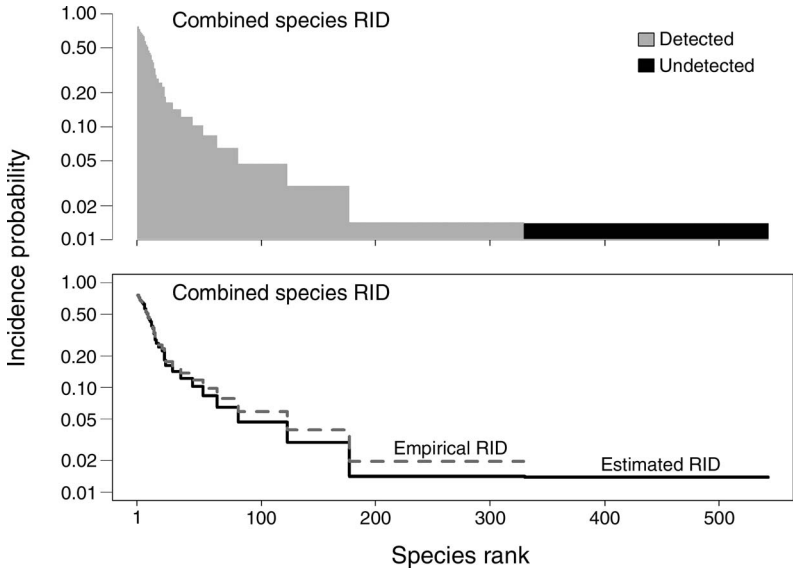

FIG. 3. The empirical and estimated species-rank incidence distribution (RID) for incidence data of soil ciliates (Foissner et al. 2002) in 51 soil samples. The RID is depicted by a simple bar plot (upper panel) and by a line plot (lower panel). All *y*-axes are on a $\log_{10}$ scale. These two plots correspond to Fig. 2c and d, respectively. The plots corresponding to Fig. 2a and b are not shown due to large numbers of detected species (331) and undetected species (209).

Assuming a geometric series for the incidence probabilities for the undetected species, we can obtain the proposed incidence probabilities for the undetected species: $\tilde{\pi}_i = \hat{\alpha}\hat{\beta}^i$, $i = 1, 2, \ldots, \hat{Q}_0$. Here $\hat{\alpha}$ and $\hat{\beta}$ are solved from two nonlinear equations involving the estimated sample coverage deficits of the first two orders (see Appendix D). Combining the adjusted incidence probabilities for detected species and the estimated incidence probabilities for undetected species, we can construct a complete RID based on incidence data.

EXAMPLE (INCIDENCE DATA)

We use soil ciliate data collected by Foissner et al. (2002) to illustrate our approach for incidence data. A total of 51 soil samples were collected in Namibia and the detection or non-detection of soil ciliate species was recorded in each sample. Detailed sampling locations, procedures, and species identifications are described in Foissner et al. (2002). In total, 331 species were detected in 51 soil samples. The first 14 incidence frequency counts $(Q_1$–$Q_{14}) = (150, 53, 42, 18, 12, 9, 10, 7, 6, 1, 0, 2, 3, 2)$, and a full list of the data are given in Appendix D. The first- and second-order sample coverage estimates are 88.45% and 99.38%, respectively; the corresponding coverage deficits are 11.55% and 0.62%, respectively. 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 (Eq. 8) gives an estimate of 209 (SE = 43.5) for the minimum number of undetected species.

All estimation procedures are parallel to the corresponding steps for abundance data in Fig. 2. However, it is not feasible to present the detailed bar plot for the empirical and adjusted RID for all 331 detected species (Fig. 2a), nor the bar plot of the estimated RID for 209 undetected species (Fig. 2b). Therefore, we simply show the empirical RID and the proposed complete RID by bar and line plots in Fig. 3. It is striking that our proposed RID has a very long tail compared with the empirical RID. This is due to a relatively high proportion of undetected species in the estimated RAD. The Chao2 estimator is a universal lower bound, implying that the complete RID may have an even longer tail, but the incidence probabilities are close to zero and invisible in our plot of the estimated RID; see *Discussion*.

DISCUSSION

We have shown that the empirical RAD using species sample abundances works only when all species are detected in a sample. For an undersampled data set with undetected species, the empirical RAD ignores the set of undetected species, and therefore overestimates the true relative abundances of the set of the detected species (Fig. 1a). We have proposed a general framework to estimate the complete RAD from sample data. Our proposed RAD estimator combines the adjusted RAD (Eq. 4d) for the detected species in samples and the estimated RAD (Eq. 6c) for the undetected species. Both parts are based on Good-Turing's sample coverage theory and its generalization. For any detected species, we have proposed a novel method to adjust its sample relative abundance to reduce its positive bias (Fig. 1b and Appendix A). For the undetected species (which are assumed to have very low relative abundances), we estimate their relative abundances using an estimator of

the number of undetected species. See Fig. 2 for an illustrative example to describe our procedures. With our approach, the complete RAD is unveiled if sample size is large enough (Fig. 1c; Appendix A: Fig. A1).

In our inference procedure for undetected species in samples, we use a universal lower bound, i.e., the Chao1 estimator for abundance data and the Chao2 estimator for incidence data; see Chao (1984, 1987). Thus, we essentially assume that there might be additional extremely rare species in the assemblage, but they cannot be statistically estimated, so their relative abundances are estimated to be zero. Our estimator of the number of undetected species could also be replaced by any other reasonable estimator. We also assume that the relative abundances for the set of undetected species follow a simple geometric series model. Nevertheless, our method is not restricted to this distribution. The assumption of a geometric series can be replaced by any other appropriate distribution. There are many other choices, including the commonly used broken-stick model and the Poisson log-normal model, among others. Appendix C provides estimation procedures for these two additional models. For illustration, we also fitted these two models to the data analyzed in *Example (abundance data)*. The Poisson log-normal model and the geometric model yield almost-identical RAD curves. We emphasize that we use these models only for modeling the undetected tail distribution; unless the assemblage is poorly sampled, the relative abundances of those undetected species (i.e., in the tail of the estimated RAD) are typically very small. Thus, the choice of the model for estimating the relative abundances of undetected species is a minor issue in our approach.

Ecologists have recognized that, although an accurate species richness estimator remains beyond our reach, one aspect of undetected species (the coverage deficit) can be accurately estimated; see Eq. 2c. We show that there are other aspects of undetected species (e.g., the deficits of the second- and higher-order sample coverage) that we can also accurately estimate using the information on frequency counts. In this study, we used the first- and second-order sample coverage and their deficits to construct two-parameter models for inferring RAD. In theory, we could have used higher-order (>2) sample coverages and their deficits to build models with more than two parameters. However, the parameter estimates from such models may be too uncertain to be useful, and may be too unstable to estimate properly.

The concept of the SAD/RAD has been also extended in this study to the corresponding SID/RID for incidence data comprising species detection/non-detection records in each sampling unit; a non-detection of a species in a sampling unit may be due to a true absence or an undetected presence, so this model can be applied not only to surveys of sessile plants but also to surveys of mobile animals in which detection probabilities are less than 1.0. If we consider the special case in which a species can always be detected if it is present in a sampling unit, then the detection/non-detection records become presence/absence data and our model can be connected to a special case of occupancy estimation and modeling (e.g., MacKenzie et al. 2006). In this special case, the incidence probability π_i can be interpreted as occupancy rate of species i in the study area. Our proposed formula $\tilde{\pi}_i = (Y_i/T)(1 - \hat{\lambda}e^{-\hat{\theta}Y_i})$ for detected species provides a nonparametric adjustment to the sample occupancy rate (i.e., Y_i/T) and thus can provide a better estimator of the true occupancy rate in the study area.

Our proposed estimator for the RAD/RID is also potentially useful in other inference problems. For example, the proposed RAD can be used for estimating any diversity measure that is a function of species relative abundances ($p_1, p_2, \ldots, p_S$). An enormous number of diversity measures have been proposed, not only in ecology but also in other disciplines, e.g., genetics, economics, information sciences, physics, and social sciences, among others; see Magurran and McGill (2011). Hill numbers (including the Shannon diversity and Simpson diversity), originally proposed by Hill (1973) have been increasingly used to quantify species diversity. We specifically discuss (in Appendix E) the use of our estimated RAD in the estimation of diversity profiles based on Hill numbers. The resulting profiles significantly improve over the empirical diversity measures mainly because the relative abundances of undetected species can be incorporated.

In another important application, when diversity measures are complicated functions of species sample abundances, and their variances are therefore difficult to estimate analytically, our proposed RAD estimator can be bootstrapped to assess their sampling variances and to construct the associated confidence intervals. This approach was applied to obtain the variances of the estimator given in Eq. 4d (see *Sampling variances of our estimators*) and the diversity estimators (see Appendix E). It has many potential applications in the analyses of beta diversity and related similarity (or differentiation) measures based on species relative abundances.

All the estimation procedures and estimators proposed in this study are featured in the freeware application JADE (joint species-rank abundance distribution/estimation; *available online*).[7] The R scripts for JADE are available in the Supplement.

Acknowledgments

The authors thank the subject editor (Brian Inouye), Lou Jost, and an anonymous reviewer for very helpful and thoughtful suggestions and comments. Their comments inspired us to extend our method to deal with incidence data. A. Chao's work on this project is funded by Taiwan Ministry of Science and Technology under Contract 103-2628-M007-007. T. C. Hsieh is supported by a post-doctoral fellowship, Ministry of Science and Technology, Taiwan. R. L. Chazdon was supported by U.S. NSF DEB 0639393, NSF DEB 1050957, and NSF DEB-1147429. R. K. Colwell was supported by CAPES Ciência sem Fronteiras (Brazil). N. J. Gotelli was

[7] http://chao.stat.nthu.edu.tw/blog/software-download/

supported by U.S. NSF DEB 1257625, NSF DEB 1144055, and NSF DEB 1136644.

Literature Cited

Caswell, H. 1976. Community structure: a neutral model analysis. Ecological Monographs 46:327–354.

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

Chao, A. 1987. Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783–791.

Chao, A., and C. H. Chiu. 2012. Estimation of species richness and shared species richness. Pages 76–111 *in* N. Balakrishnan, editor. Methods and applications of statistics in the atmospheric and earth sciences. Wiley, New York, New York, USA.

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

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

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

Chazdon, R. L., A. Chao, R. K. Colwell, S. Y. Lin, N. Norden, S. G. Letcher, D. B. Clark, B. Finegan, and J. P. Arroyo. 2011. A novel statistical method for classifying habitat generalists and specialists. Ecology 92:1332–1343.

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

Dewdney, A. K. 2000. A dynamical model of communities and a new species-abundance distribution. Biological Bulletin 198:152–165.

Ellison, A. M., A. A. Barker-Plotkin, D. R. Foster, and D. A. Orwig. 2010. Experimentally testing the role of foundation species in forests: the Harvard Forest Hemlock Removal Experiment. Methods in Ecology and Evolution 1:168–179.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology 12:42–58.

Foissner, W., S. Agatha, and H. Berger. 2002. Soil ciliates (Protozoa, Ciliophora) from Namibia (Southwest Africa), with emphasis on two contrasting environments, the Etosha region and the Namib Desert. Denisia 5:1–1459.

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

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

Gotelli, N. J., R. M. Dorazio, A. M. Ellison, and G. D. Grossman. 2010. Detecting temporal trends in species assemblages with bootstrapping procedures and hierarchical models. Philosophical Transactions of the Royal Society B 365:3621–3631.

Green, J., and J. B. Plotkin. 2007. A statistical theory for sampling species abundances. Ecology Letters 10:1037–1045.

Hill, M. 1973. Diversity and evenness: a unifying notation and its consequences. Ecology 54:427–432.

Hubbell, S. P. 2001. A unified theory of biodiversity and biogeography. Princeton University Press, Princeton, New Jersey, USA.

Lande, R., P. J. DeVries, and T. R. Walla. 2000. When species accumulation curves intersect: implications for ranking diversity using small samples. Oikos 89:601–605.

Lehmann, E. L., and G. Casella. 1998. Theory of point estimation. Second edition. Springer-Verlag, New York, New York, USA.

MacArthur, R. H. 1957. On the relative abundance of bird species. Proceedings of the National Academy of Sciences USA 43:293–295.

MacArthur, R. H. 1960. On the relative abundance of species. American Naturalist 94:25–36.

MacKenzie, D. I., J. D. Nichols, J. A. Royle, K. H. Pollock, L. L. Bailey, and J. E. Hines. 2006. Occupancy estimation and modeling: inferring patterns and dynamics of species occurrence. Elsevier, Amsterdam, Netherlands.

Magurran, A. E. 2004. Measuring biological diversity. Blackwell, Oxford, UK.

Magurran, A. E., and B. J. McGill, editors. 2011. Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford, UK.

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

Preston, F. 1948. The commonness and rarity of species. Ecology 29:254–283.

Sackett, T. E., S. Record, S. Bewick, B. Baiser, N. J. Sanders, and A. M. Ellison. 2011. Response of macroarthropod assemblages to the loss of hemlock (*Tsuga canadensis*), a foundation species. Ecosphere 2:art74.

Sugihara, G. 1980. Minimal community structure: an explanation of species abundance patterns. American Naturalist 116: 770–787.

Tokeshi, M. 1990. Niche apportionment or random assortment: species abundance patterns revisited. Journal of Animal Ecology 59:1129–1146.

Whittaker, R. H. 1965. Dominance and diversity in land plant communities. Nature 147:250–260.

Whittaker, R. H. 1970. Communities and ecosystems. MacMillan, New York, New York, USA.

Whittaker, R. H. 1972. Evolution and measurement of species diversity. Taxon 12:213–251.

Supplemental Material

Ecological Archives

Appendices A–E and the Supplement are available online: http://dx.doi.org/10.1890/14-0550.1.sm