Ecology, 93(12), 2012, pp. 2533–2547

Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size

Anne Chao[1,3] and Lou Jost[2]

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

Abstract. We propose an integrated sampling, rarefaction, and extrapolation methodology to compare species richness of a set of communities based on samples of equal completeness (as measured by sample coverage) instead of equal size. Traditional rarefaction or extrapolation to equal-sized samples can misrepresent the relationships between the richnesses of the communities being compared because a sample of a given size may be sufficient to fully characterize the lower diversity community, but insufficient to characterize the richer community. Thus, the traditional method systematically biases the degree of differences between community richnesses. We derived a new analytic method for seamless coverage-based rarefaction and extrapolation. We show that this method yields less biased comparisons of richness between communities, and manages this with less total sampling effort. When this approach is integrated with an adaptive coverage-based stopping rule during sampling, samples may be compared directly without rarefaction, so no extra data is taken and none is thrown away. Even if this stopping rule is not used during data collection, coverage-based rarefaction throws away less data than traditional size-based rarefaction, and more efficiently finds the correct ranking of communities according to their true richnesses. Several hypothetical and real examples demonstrate these advantages.

Key words: *diversity; extrapolation; interpolation; prediction; rarefaction; replication principle; sample coverage; species accumulation curve; species richness; standardization.*

Introduction

Species richness is the oldest, simplest, and most popular diversity measure, and its mathematical properties are especially intuitive (Magurran 2004, Magurran and McGill 2011, Gotelli and Colwell 2001, 2011, Colwell et al. 2012). Unfortunately, the statistical properties of species richness are much worse than those of any other common diversity measure (Colwell and Coddington 1994, Chao 2005, Gotelli and Chao 2013), since it is as sensitive to rare and hard-to-detect species as to common ones. In communities with many rare species, like tropical arthropod, orchid (see Plate 1), or soil microbe communities, sample richness might never stabilize as sample size increases, under any realistic sampling scheme (Coddington et al. 2009). Sample richnesses of such communities depend strongly on sample size.

To control for this dependence when comparing the richnesses of different communities, ecologists use rarefaction to down-sample the larger samples until they are the same size as the smallest sample (Sanders 1968, Hurlbert 1971, Simberloff 1972, Gotelli and Colwell 2001, 2011). Ecologists then compare the richnesses of these equally large samples. However, this is usually not a "fair" comparison. Samples standardized by size will usually have different degrees of completeness, depending on the species–abundance distributions of the communities being compared. A temperate-zone tree community with 10 species might be completely characterized by a sample of 100 individuals, but the same size sample would greatly underestimate the richness of a tropical rain forest with 300 tree species. If we compared the richnesses of these two communities using samples standardized to 100 individuals, the richnesses we obtain would not give a meaningful idea of the real degree of difference between the communities' richnesses. In fact, the degree of difference observed would depend almost entirely on the sample size used, rather than on some real characteristics of the communities being compared. As Peet (1974) pointed out, richness estimates based on fixed sample sizes necessarily compress the ratio of their richnesses, and our examples in the *Replication principle for samples standardized by coverage* section below show that this compression can be severe.

The solution is to compare samples of equal completeness, not equal size. Recently, Alroy (2009, 2010*a*, *b*), Jost (2010), and Chao (*personal correspondence* in Jost 2010) noticed that when samples are standardized by their coverage (a measure of sample completeness; see next section) instead of by their size, the estimated richnesses approximately satisfy a replication principle, which is an essential property for characterizing diversity. Species richness itself has this

Manuscript received 29 October 2011; revised 7 May 2012; accepted 12 June 2012. Corresponding Editor: F. He.
[3] E-mail: chao@stat.nthu.edu.tw

property, but estimated richnesses based on samples of fixed size do not. Because coverage-based estimated richness preserve this important property of species richness, the ratio (or any measure of the degree of difference) between coverage-based richness estimates of any two communities will be more representative of the true relationship between the diversities of the communities, compared with the ratio based on traditional richness estimates at fixed sample size. The ratio will show much less compression than the ratio based on estimates at fixed sample size, and in special cases there will be no compression at all, even for small samples. This cannot be achieved by the traditional approach. The estimated richnesses at fixed coverage can meaningfully be compared in other ways (such as taking their absolute difference or relative difference) as well, since they estimate properties of a fixed fraction of each community, as explained in the next section.

Alroy (2009, 2010*a*, *b*) called his coverage-based rarefaction approach "shareholder quorum subsampling," and implemented it by randomly drawing specimens until the desired coverage level was reached, with the coverage obtained at any one sampling level being estimated from a combination of the observed species frequencies and the overall sample's coverage. Jost (2010) also proposed an algorithmic method for estimating richness at a given coverage. Here we derive for the first time an unbiased analytical rarefaction formula for estimating richness at a given coverage, and also present a new, unbiased algorithmic technique for rarefying samples to a fixed sample coverage instead of a fixed size. We also show how our method leads to a well-known adaptive stopping rule for sampling, so that ecologists need not waste any sampling effort (and, in fact, do not need to rarefy their samples at all).

If our recommended stopping rule is not used, coverages will generally differ among samples. Then rarefaction of these samples to a standard coverage will end up throwing away data, just like traditional rarefaction. This frustrating limitation of rarefaction (whether size- or coverage-based) can be overcome by extrapolating the species-accumulation curves of the less complete samples instead of rarefying the more complete samples, or by a mixture of rarefaction of the largest samples and extrapolation of the smaller ones, always with the aim of comparing samples of equal coverage. Ecologists have often used curve-fitting methods to obtain extrapolated richness estimates (estimates of the richness that would be expected if sample size were increased by a given amount); see Colwell and Coddington (1994) for an overview. Good and Toulmin (1956) provided the foundation for sampling theory-based extrapolation methods. Complicated mixture extrapolation models were developed by Colwell et al. (2004) and Mao et al. (2005). Recently, Colwell et al. (2012) unified the approach based on standardized sample size to link rarefaction and extrapolation curves, which helps solve this issue of throwing away data. We here propose for the first time an analytic formula to extrapolate a sample to a higher coverage instead of larger sample size. Thus, a unified coverage-based sampling curve that integrates both rarefaction and extrapolation can be constructed. This unified curve uses more data and provides more meaningful information than traditional approaches. We illustrate our approach with examples demonstrating its advantages.

Sample Coverage and the Species Accumulation Curve

The concept of "sample coverage" (or simply "coverage") was originally developed for cryptographic analyses during World War II by the founder of modern computer science, Alan Turing, and by his colleague I. J. Good (Good 1953, 2000). It is a measure of sample completeness, giving (in our context) the proportion of the total number of individuals in a community that belong to the species represented in the sample. Subtracting the sample coverage from unity gives the proportion of the community belonging to unsampled species; we call this the "coverage deficit." The coverage deficit of the sample is also the probability that a new, previously unsampled species would be found if the sample were enlarged by one individual.

An idealized example will help illustrate the coverage concept. Suppose Community A hosts a terrestrial arthropod community with 50 species. Species 1 has relative abundance 0.3, species 2 has relative abundance 0.1, species 3 through 5 each has relative abundances of 0.05, and species 6 through 50 each have relative abundances of 0.01. We can write the relative abundances of the community's species as $\{0.3, 0.1, 0.05 \times 3, 0.01 \times 45\}$. Suppose a random sample of 20 individuals is taken with replacement from this community. Assume we observe the most abundant 12 species, but not the other 38 species (this is just the most likely combination, but we could observe many other combinations). Then the sample coverage of this specific sample is $0.3 + 0.1 + 0.05 \times 3 + 0.01 \times 7 = 62\%$ because these 12 species in the sample, taken together, constitute 62% of the total number of individuals in the community. The coverage deficit is $100\% - 62\% = 38\%$, meaning that 38% of the individuals in the community belongs to species that were not detected by the sample. These numbers are objective indicators of the completeness of the sample.

Ecologists already make heavy use of this coverage concept as a measure of sample completeness, perhaps without realizing it. We often judge the completeness of the sample by looking at the final slope of the rarefaction curve calculated from the species frequency data (e.g., Schloss and Handelsman 2004, Tringe et al. 2005). The sample is considered to be nearly complete if and only if this slope is small. This slope is the expected rise of the curve if one individual is added to the sample; in other words, it is the probability that the next sampled individual is a new species not previously sampled (Olszewski 2004). This is just the coverage deficit of the

sample. Coverage thus already forms the main criterion by which ecologists judge the completeness of a sample whose true richness is unknown.

Theoretical relationship

Let S_m be the number of species in a given sample of size m. Good (1953) showed that

$$E(S_m) = \sum_{i=1}^{S}[1 - (1 - p_i)^m] = S - \sum_{i=1}^{S}(1 - p_i)^m \qquad (1)$$

where S is the total number of species, and p_i is the relative abundance of the ith species. (Our methods in this paper are also valid in a more general framework, in which p_i is interpreted as the detection rate of the ith species, i.e., a normalized product of relative abundance and individuals' detectability. Both relative abundance and individuals' detectability are allowed to vary with species. For simplicity, we present all our derivations assuming p_i is simply the relative abundance of the ith species.) A traditional size-based expected species accumulation curve (SAC) plots $E(S_m)$ with respect to sample size m. For our Community A with relative abundances $\{0.3, 0.1, 0.05 \times 3, 0.01 \times 45\}$, Eq. 1 tells us we would expect to find 12 species in our sample of 20 individuals (Table 1). If we consider all possible combinations of sampled individuals, the expected value of the coverage of a sample of size m is given by (Good 1953, Robbins 1968) as follows:

$$E(C_m) = \sum_{i=1}^{S} p_i[1 - (1 - p_i)^m] = 1 - \sum_{i=1}^{S} p_i(1 - p_i)^m \qquad m > 0. \qquad (2)$$

For our Community A, the expected coverage for $m = 20$ is 57%, so the expected coverage deficit is 43%. The connection between the coverage deficit and the slope of the expected SAC for any sample size m can be found by combining Eqs. 1 and 2 to obtain the following "slope-coverage relationship":

$$1 - E(C_m) = E(S_{m+1}) - E(S_m) \qquad \text{with } m > 0. \qquad (3)$$

The right-hand side is the slope of the species accumulation curve (the expected change along the y-axis, $E(S_{m+1}) - E(S_m)$, divided by the corresponding change in the x-axis, which is a sample increment of one individual). This proves that the slope of the expected SAC for a sample of size m is equal to the expected coverage deficit of the sample.

Replication principle for samples standardized by coverage

Estimates of species richness standardized by coverage preserve an important property of species richness, a kind of replication principle, as first noted by Alroy (2010*a*) and Jost (2010). This property is critical for properly judging the relative diversities of multiple communities. We explain this property by means of the following example.

TABLE 1. The expected number of species observed in any sample of size m, $E(S_m)$, in Community A and Community A + B and their ratio for $m = 5, 10, 20, \ldots, 2000$.

Sample size, m	$E(S_m)$		Ratio
	Community A	Community A + B	
5	4.13	4.51	1.09
10	7.13	8.15	1.14
20	**12.00**	**14.17**	**1.18**
30	16.03	19.31	1.21
40	19.50	23.91	1.23
50	**22.54**	**28.11**	**1.25**
60	25.24	31.97	1.27
70	27.65	35.56	1.29
80	29.81	38.91	1.31
90	31.76	42.04	1.32
100	**33.51**	**44.99**	**1.34**
200	43.97	66.94	1.52
300	47.79	79.99	1.67
400	49.19	87.88	1.79
500	49.70	92.66	1.86
600	49.89	95.55	1.92
700	49.96	97.31	1.95
800	49.99	98.37	1.97
900	49.99	99.01	1.98
1000	49.99	99.40	1.99
2000	49.99	99.99	2.00

Note: The three species ratios shown in Fig. 1a are in boldface.

Tropical forests typically have distinct canopy and understory butterfly assemblages (DeVries and Walla 2001). Suppose Community A from our coverage example is the canopy butterfly assemblage, with species abundance distribution $\{0.3, 0.1, 0.05 \times 3, 0.01 \times 45\}$. Suppose the understory assemblage, Community B, has the same species abundance distribution, $\{0.3, 0.1, 0.05 \times 3, 0.01 \times 45\}$, and no species shared with the canopy community. Suppose both communities have the same density of individuals.

We compare two forests, one which has both a canopy and understory community A + B and another that, because of human alternations to the understory, has only a canopy community A. Community A + B is twice as diverse as Community A, not just in terms of species richness, but also in terms of Shannon entropy, Simpson indices, or any other standard diversity measure, when these measures are converted to diversities (effective number of species; Jost 2007). An ideal comparison method should therefore yield a diversity ratio of 2.0 for Community A + B relative to Community A.

As discussed in the *Theoretical relationship* section, a sample of 20 individuals from Community A is expected to contain 12 species (from Eq. 1); see Table 1. If we use this same sample size for Community A + B, or if we rarefy a larger sample of Community A + B down to 20 individuals as ecologists would usually do, we would not expect to find twice as many species (24 species), but only 14 species (Eq. 1, Table 1). From these samples, we would estimate that the ratio of the diversities of the two

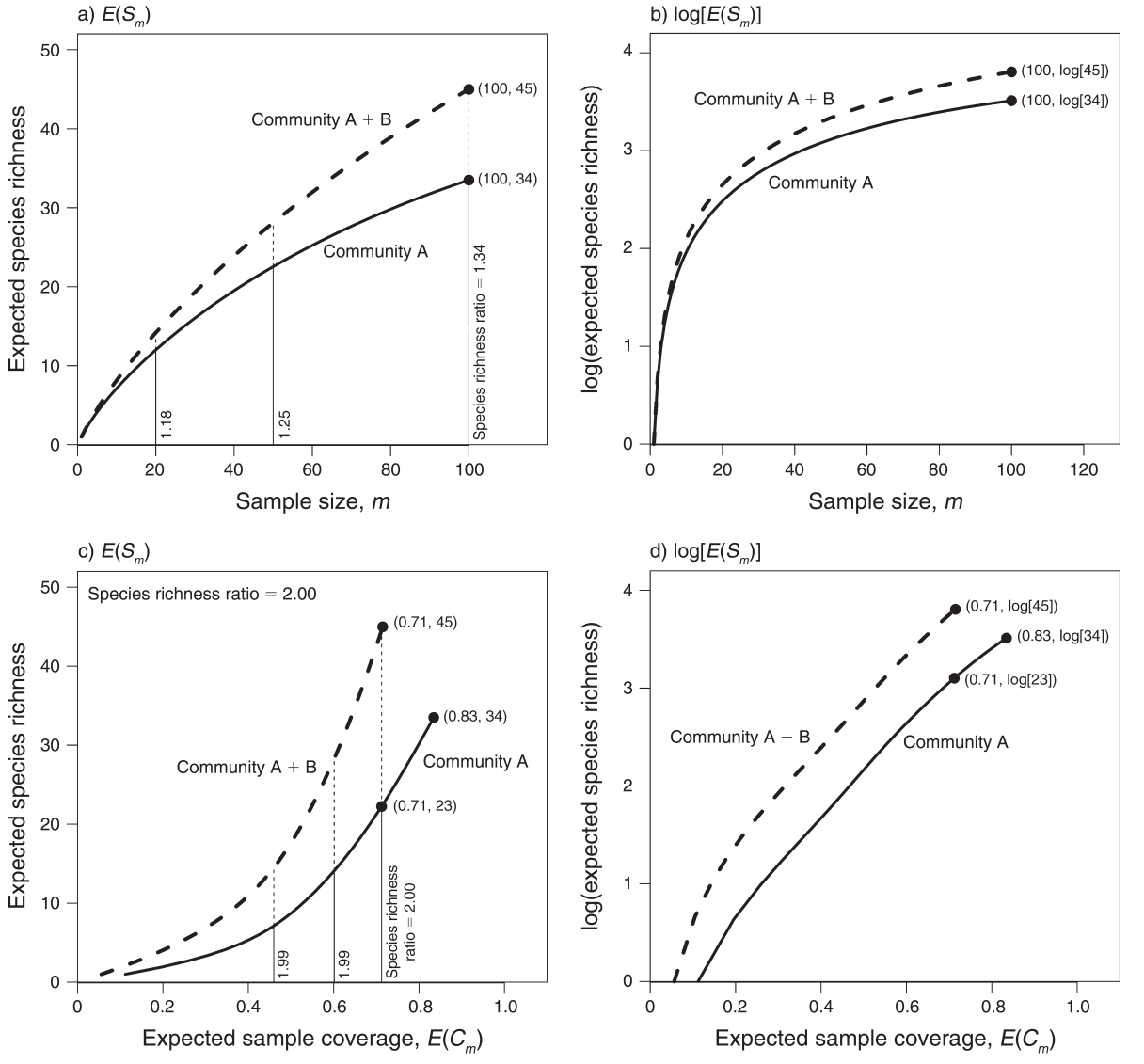

FIG. 1. (a) The plot of the expected species richness, $E(S_m)$, with respect to sample size m, and (c) with respect to the expected sample *coverage*, $E(C_m)$, for Community A (solid line) and Community A + B (dashed line). (b) The plot of log[$E(S_m)$] with respect to sample size m, and (d) with respect to the expected sample coverage, $E(C_m)$, for Community A (solid line) and Community A + B (dashed line). The species richness ratio is calculated for sample sizes $m = 20$, 50, and 100 in panel (a). These three sizes correspond to sample coverage 0.46, 0.60, and 0.71, respectively, in Community A + B (the curve with the less coverage; see Table 2), and the three corresponding ratios for these three coverages are shown in panel (c). Note that in panel (c), for any given sample coverage, the value in the dashed line is approximately double the value in the solid line. The plot of the logarithm of species richness with respect to sample coverage in panel (d) clearly shows two approximately parallel lines. The numbers in parentheses show the x- and y-axis coordinates for each point.

communities is only $14/12 = 1.18$, not even close to the true ratio of 2.00. Severe underestimation of the degree of difference in richnesses of these two communities continues even for quite large sample sizes (Table 1, Fig. 1a, b). Only when sample size exceeds 400 does the estimated diversity ratio between the communities come within 10% of the correct value.

However, if our sample of Community A + B is standardized to the same *coverage* as our sample from Community A (rather than the same sample size), a remarkable thing happens. The sample from Community A + B will be approximately twice as rich as the sample from Community A, on average, even for very small sample sizes (Table 2, Fig. 1c, d). This can be seen by noting that, to achieve a coverage of 0.57 (the coverage of the 20-individual sample from Community A), a sample of Community A + B would have to be twice as big as the sample from Community A, and would most likely include its 24 most abundant species. The average number of species in a sample of size 40

TABLE 2. The expected sample coverage, $E(C_m)$, and expected species richness, $E(S_m)$, by matching coverage in Community A and Community A + B.

$E(C_m)$		Sample size m		$E(S_m)$		
Community A	Community A + B	Community A	Community A + B	Community A	Community A + B	Ratio
0.35	0.34	5	10	4.13	8.15	1.98
0.46	**0.46**	**10**	**20**	**7.13**	**14.17**	**1.99**
0.57	0.56	20	40	12.00	23.91	1.99
0.60	**0.60**	**25**	**50**	**14.09**	**28.11**	**1.99**
0.63	0.63	30	60	16.03	31.97	1.99
0.68	0.68	40	80	19.50	38.91	2.00
0.72	**0.71**	**50**	**100**	**22.54**	**44.99**	**2.00**
0.75	0.75	60	120	25.24	50.39	2.00
0.77	0.77	70	140	27.65	55.21	2.00
0.80	0.80	80	160	29.81	59.54	2.00
0.82	0.82	90	180	31.76	63.43	2.00
0.83	0.83	100	200	33.51	66.94	2.00
0.90	0.90	150	300	40.03	79.99	2.00

Note: The sample size in Community A + B is double the sample size in Community A. "Ratio" is calculated as the ratio of $E(S_m)$ of Community A + B to $E(S_m)$ of Community A. The three species ratios shown in Fig. 1c are in boldface.

from Community A + B is 23.91, from Eq. 1 and Table 2. The expected richness of the sample (for $m = 40$) from Community A + B will thus be twice that of the sample (for $m = 20$) from Community A. In general, richness estimates based on coverage obey the following *replication principle* (Alroy 2010*a*, Jost 2010), which is also obeyed by species richness itself: If Community 2 consists of K replicates of Community 1 (each replicate with the same species abundance distribution and density as Community 1, but with no species in common between replicates), then on the average, a sample from Community 2 will have an expected richness approximately K times that of a sample from Community 1 when both samples have the same coverage. The proof of this property and its generalization is given in Appendix A. This property is a necessary condition for diversity estimators to behave intuitively in ratio comparisons and other comparison criteria.

Even in the general case, when one community is not an exact multiple of the other (in the sense of the replication principle just described), standardizing on coverage produces more useful and meaningful comparisons of species richness and other diversity measures. When we compare samples with the same coverage, we are making sure that samples are equally complete and that the unsampled species constitute the same proportion of the total individuals in each community. Although we are usually unable to compare true species richness due to incomplete sampling, we can now compare species richness for the same proportion of each community's individuals. Therefore, our comparison is based on a community's characteristic, rather than surveyor's sampling efforts.

Estimating coverage from data

In practical applications, sample coverage must be estimated from data. This might seem to require advance knowledge of the true relative abundances of all the species in the community. However, contrary to most people's intuition, sample coverage can be very accurately and efficiently estimated using only information contained in the sample itself, as long as the sample is reasonably large (Good 1953, Robbins 1968, Esty 1983, 1986). For a given original sample of size n (which will be referred to as "reference sample"), a commonly used estimator of sample coverage is simply $1 - f_1/n$, where f_1 is the number of singletons (species each represented by exactly one individual in the reference sample). This estimator was originally discovered by Alan Turing (Good 1953). Robbins (1968) showed that the mean squared error of Turing's estimator is less than $1/n$, indicating it is quite accurate if n is large. Esty (1983) and Zhang and Zhang (2009) proved asymptotic normality of Turing's estimator, and provided a confidence interval for sample coverage. Chao et al. (1988) derived a less biased, but more complicated estimator, which was rediscovered by Zhang and Huang (2007). A Bayesian nonparametric approach was proposed in Lijoi et al. (2007). Using information on both f_1 and f_2 (the number of doubletons in the sample), we adopt here the following improved coverage estimator for a reference sample of size n (Chao and Shen 2010):

$$\hat{C}_n = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2}\right]. \quad (4a)$$

This coverage estimator generally has smaller mean squared error than Turing's estimator. An asymptotic approach yields an approximate variance estimator and the associated confidence interval to reflect sampling uncertainty (Chao and Shen 2010).

Coverage-based stopping rule

We have shown that samples of equal completeness, not equal size, should be compared across communities if we want to make inferences about the relative differences in richness between communities rather than samples. This suggests that ecologists should sample each community to the same degree of completeness, as

measured by the coverage estimate $1 - f_1/n$ or the more accurate Eq. 4a. When a sample reaches a predetermined value of this coverage estimate, which can be easily calculated in the field, sampling stops. This stopping rule was theoretically justified by an optimal stopping theory (Rasmussen and Starr 1979). Such equally complete samples from different communities can be compared directly, without any need for rarefaction (see Example 2 below).

Coverage-based Rarefaction (Interpolation)

Let C_m be the coverage for subsamples of size m ($m < n$) from the reference sample of size n. Under the model that the species sample frequencies $(X_1, X_2, \ldots, X_S)$ follow a multinomial distribution with a cell total n and probabilities $\{p_1, p_2, \ldots, p_S\}$, the unique minimum-variance unbiased estimator of the expected coverage $E(C_m)$ is (see proof in Appendix B):

$$\hat{C}_m = 1 - \sum_{X_i \geq 1} \frac{X_i}{n} \frac{\binom{n - X_i}{m}}{\binom{n-1}{m}} \qquad m < n \qquad (4b)$$

where X_i is the number of individuals observed for species i in the reference sample so that $\sum_{i=1}^{S} X_i = n$. This new equation yields exact values that previously could only be estimated using the algorithmic approaches suggested by Alroy (2010*a*) and Jost (2010). This equation is analogous to traditional rarefaction equation giving the estimated species richness $\hat{S}_m$ for a subsample size m as follows:

$$\hat{S}_m = S_{\text{obs}} - \sum_{X_i \geq 1} \frac{\binom{n - X_i}{m}}{\binom{n}{m}} \qquad m < n \qquad (5)$$

where S_{obs} is the number of species observed in the reference sample of size n. The estimator $\hat{S}_m$ is the unique minimum variance unbiased estimator for $E(S_m)$, the expected number of species that would be observed in a subsample size of m, $m < n$ (Hurlbert 1971, Smith and Grassle 1977); see Eq. 1. Since $\hat{S}_m$ is an unbiased estimator, its expected value is simply $E(S_m)$, which satisfies the replication principle when the expected coverage is standardized, as we prove in Appendix A. For the estimator $\hat{S}_m$, Colwell et al. (2012) derived an analytic variance formula in terms of an estimated number of undetected species. Their variance estimator works well for a large rarefied size m, but tends to overestimate and thus produces a conservative confidence interval when the size m is not large relative to n. To remedy this problem, we suggest a bootstrap method to construct an estimated variance and the associated confidence interval for any given sample coverage; see Appendix C for a description. For both analytic and bootstrap approaches, the reference sample size n should be large enough so that the undetected species richness and sample coverage can be adequately estimated. A rough guideline is that the estimated coverage value (Eq. 4a) should be at least 50% (Chao and Lee 1992).

Based on Eqs. 4b and 5, we show that the traditional rarefaction estimator $\hat{S}_m$ is related to the coverage deficit estimator by the following relation (see Appendix B):

$$1 - \hat{C}_m = \hat{S}_{m+1} - \hat{S}_m \qquad m < n. \qquad (6)$$

This shows that the estimated coverage deficit after m individuals have been sampled is exactly equal to the slope of the line connecting the two points $(m, \hat{S}_m)$ and $(m + 1, \hat{S}_{m+1})$ in a traditional rarefaction curve. This relationship between the data-based estimators of coverage and richness is identical to the relationship between the theoretical expected coverage and richness given by Eq. 3. This means that, based on data, two samples rarefied down to sizes m_1 and m_2, respectively, are equally complete if and only if the slopes of their rarefaction curves (at x-axis of m_1 and m_2, respectively) are the same.

Analytical approach

Our Eq. 4b gives the estimated coverage $\hat{C}_m$ as a function of sample size m, and Eq. 5 gives the estimated richness $\hat{S}_m$ of a sample of size $m < n$. The coverage-based analytic rarefaction curve is obtained by simply plotting the species estimate $\hat{S}_m$ with respect to the sample coverage estimate $\hat{C}_m$ for $m < n$. For $m = n$, we plot S_{obs} with respect to $\hat{C}_n$ which is given in Eq. 4a. See examples in *Applications*.

This method is valid as long as the sampling method actually used in the field does not significantly alter the population's species abundance distribution. Thus, it is always valid if sampling is done with replacement, and it is also valid for sampling without replacement if the communities are much larger than the samples (as is usually the case).

Unbiased algorithm

The mathematical theory for sampling also leads to an algorithmic method to obtain the coverage-based rarefaction curve (See Appendix D). For a sample size $m < n$, we take $m + 1$ individuals *without* replacement from the original sample, and record the number of singletons in this subsample. Appendix D provides theoretical justification that this algorithm is unbiased in the sense that, after a large number of replications, one minus the average proportion of singletons in a subsample of size $m + 1$ tends to $\hat{C}_m$ in Eq. 4b. We have tested this algorithm with many sets of empirical data. In all data sets, the analytic $\hat{C}_m$ and the simulated value perfectly match each other when the number of replications is sufficiently large.

Although the analytic method in Eq. 4b is sufficient to construct coverage-based rarefaction curves, the unbiased algorithmic method can be useful for estimating other measures; see Discussion. Alroy (2009, 2010*a*, *b*)

proposed a different algorithmic technique that allows users to pre-specify a desired value of C_m (called a subsampling quorum) and obtain a corresponding richness estimate. In Appendix D, our algorithm is theoretically proved to be unbiased under a commonly used sampling model, and statistical estimation theory implies that our approach is the unique minimum variance unbiased estimator.

Coverage-Based Extrapolation (Prediction)

Analytic approach

For extrapolation, only an analytic approach is feasible. In a traditional size-based sampling curve, given the data for a reference sample of size n, the extrapolation problem is to predict the expected number of species $E(S_{n+m^*})$ in an augmented sample of $n + m^*$ individuals from the community ($m^* > 0$). Good and Toulmin (1956) derived a prediction formula, but their estimator lacks some theoretical properties of the prediction function (Boneh et al. 1998) and may take negative values or become extremely large if $m^* > n$; see Chao and Shen (2004) for examples. The predictors proposed in Colwell et al. (2004) and Mao et al. (2005), although theoretically useful and flexible, are based on rather complicated mixture models. Recently, Colwell et al. (2012) linked the interpolation (rarefaction) curve given in Eq. 5 and the extrapolation (prediction) curve proposed in Shen et al. (2003) to yield a single smooth curve meeting at the reference sample. Here, we slightly modify the approach of Shen et al. (2003) and consider the following more accurate predictor for the species richness in an augmented sample with size $n + m^*$:

$$\hat{S}_{n+m^*} = S_{\text{obs}} + \hat{f}_0 \left[1 - \left(1 - \frac{f_1}{n\hat{f}_0 + f_1} \right)^{m^*} \right] \tag{7}$$

where $\hat{f}_0$ is an estimator for f_0 (the number of undetected species). Colwell et al. (2012) suggested that $\hat{f}_0$ can be obtained by using the Chao1 estimator (Chao 1984), as follows:

$$\hat{f}_0 = \frac{(n-1)}{n} \frac{f_1^2}{(2f_2)} \quad \text{for } f_2 > 0$$

or

$$\hat{f}_0 = \frac{(n-1)}{n} \frac{f_1(f_1-1)}{2(f_2+1)} \quad \text{for } f_2 = 0. \tag{8}$$

A similar bootstrap method as we used for rarefaction (Appendix C) can be applied to obtain a variance estimator of $\hat{S}_{n+m^*}$ and confidence interval of $E(S_{n+m^*})$. When m^* tends to infinity, the extrapolated estimator approaches $S_{\text{obs}} + \hat{f}_0$, implying the Chao1 species richness estimator is the asymptotic value of our extrapolation formula. The first-order approximation of $E(\hat{S}_{n+m^*})$ (i.e., replace all observed data in Eq. 7 by their expected values) still obeys the replication principle in the following sense (as proved in Appendix E): Assume Community 2 consists of K replicates of Community 1. If the sample size in Community 2 is K times that in Community 1 (so that the expected sample coverage is standardized), then $E(\hat{S}_{K(n+m^*)})$ in Community 2 for an augmented sample of size $K(n + m^*)$ is approximately K times of $E(\hat{S}_{n+m^*})$ in Community 1 for an augmented sample of size $n + m^*$.

To plot the extrapolation part of the coverage-based sampling curve, we additionally need to derive an estimator for the expected coverage $E(C_{n+m^*})$ in an augmented sample of $n + m^*$ individuals from the community ($m^* > 0$). Using similar arguments as in Chao et al. (2009), we can obtain an estimator of $E(C_{n+m^*})$ when $f_1, f_2 > 0$ as follows (see Appendix E for derivation details):

$$\hat{C}_{n+m^*} = 1 - \frac{f_1}{n} \left[\frac{(n-1)f_1}{(n-1)f_1 + 2f_2} \right]^{m^*+1}. \tag{9a}$$

When m^* tends to infinity, the extrapolated coverage estimator approaches unity, indicating a complete coverage. When $m^* = 0$, it reduces to the sample coverage estimator for the reference sample as given in Eq. 4a. When n is large, it follows from Eq. 4a that the coverage deficit for an augmented size of $n + m^*$ is reduced by a factor of approximately $\exp[-2m^* f_2/(nf_1)]$:

$$1 - \hat{C}_{n+m^*} \approx (1 - \hat{C}_n)\exp[-2m^* f_2/(nf_1)]. \tag{9b}$$

With the modified estimator in Eq. 7 and the predicted coverage formula in Eq. 9a, we can prove (Appendix E)

$$1 - \hat{C}_{n+m^*} = \hat{S}_{n+m^*+1} - \hat{S}_{n+m^*}. \tag{10}$$

Thus, the theoretical slope coverage relationship given by Eq. 3 is valid not only for the rarefaction part (Eq. 6), but also for the extrapolation part (Eq. 10).

An integrated coverage-based sampling curve includes the rarefaction part (which plots $\hat{S}_m$ with respect to $\hat{C}_m$, $m < n$; see Eqs. 4b and 5) and the extrapolation part (which plots $\hat{S}_{n+m^*}$ with respect to $\hat{C}_{n+m^*}$ for $m^* > 0$; see Eqs. 7 and 9a); both parts join smoothly at the reference point ($\hat{C}_n$, S_{obs}); see Eq. 4a. As will be shown in our examples, the bootstrap confidence intervals in the two parts also join smoothly.

Like most statistical theory-based predictors, the performance of our extrapolated estimator $\hat{S}_{n+m^*}$ depends on the range of the extrapolation. For a short-range prediction, the prediction bias with respect to the parameter $E(S_{n+m^*})$ is negligible, but the magnitude of the bias increases when the prediction range becomes large. The variance of the extrapolation generally depends on the amount of data in the reference sample. If our goal is to rank the species richnesses among multiple communities, the extrapolated size in each sample could be extended to several times of the reference sample. However, for the goal of estimating how much richer one community is compared to another, we suggest that the extrapolation formula be applied at most only up to a doubling of reference

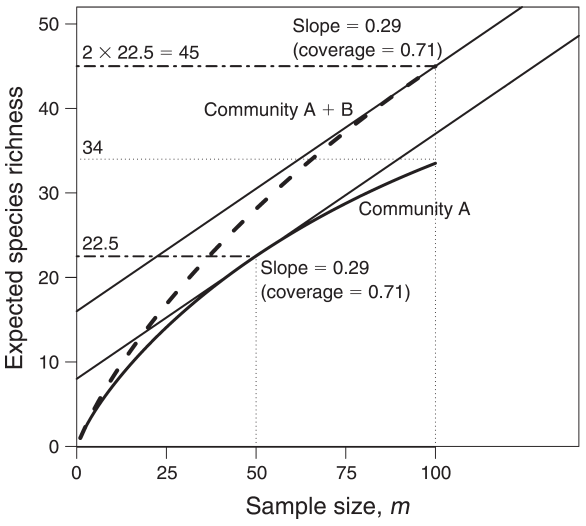

FIG. 2. Graphic comparison of the two expected species accumulation curves (SACs) in Fig. 1a. Assume that the sample size for each community is 100. Community A + B (dashed line, as in Fig. 1a) is the curve with the higher terminal slope, 0.29 (or the lower coverage = 0.71). The slope at $m = 100$ for Community A + B is approximately the same as the slope at $m = 50$ for Community A (as shown by the two parallel lines). Therefore, these two samples ($m = 50$ for Community A and $m = 100$ for Community A + B) are equally complete. In this special case, the species richness (45) at $m = 100$ for Community A + B is approximately twice the species richness (22.5) at $m = 50$ for Community A, as it should be.

sample size. That is, we recommend m^* should not exceed n in Eqs. 7 and 9a. This means from Eq. 9b that the coverage deficit is reduced by an approximate factor no less than $\exp(-2f_2/f_1)$.

Comparison of Multiple Samples

If comparison is only based on rarefaction, we first construct the coverage-based rarefaction curve for each sample. Then we identify the curve with the lowest final sample coverage; all other samples have higher coverage and will have to be rarefied down to the coverage of this sample. On each of the other curves, we locate the point with the same coverage, and we find the species richness corresponding to that point. The set of species richnesses obtained in this way are based on equal-coverage samples, and can be legitimately compared with each other. With this curve in hand for each sample, we can choose any sample coverage desired (as long as it is less than the lowest coverage) and read off the corresponding species richnesses at that coverage for each of the samples being compared.

Alternatively, we can use the traditional size-based rarefaction curves to rarefy to a fixed sample coverage. We find the rarefaction curve with the steepest terminal slope. All other rarefaction curves will have to be rarefied down until they match this steepest slope. Once the points of equal slope have been identified, the richnesses at those points can be compared, since they have equal coverages. See Fig. 2 for a numerical example.

However, the above rarefaction-based comparison discards data and information from larger samples, and comparison can be made only up to the lowest coverage. By combining rarefaction and extrapolation, we can obtain a more informative comparison among multiple samples. We first construct for each sample the integrated coverage-based sampling curve up to a maximum coverage value, approximately $1 - (f_1/n)\{(n - 1)f_1/[(n - 1)f_1 + 2f_2]\}^{n+1} \approx 1 - (1 - \hat{C}_n)\exp(-2f_2/f_1)$, corresponding to a doubling of the reference sample size (i.e., $m^* \leq n$ in the predictor shown in Eqs. 7 and 9a). We select the lowest among these maximum coverages as our "base coverage." We then obtain the estimated species richnesses of all communities at that base coverage; these estimates can be legitimately compared across communities. Similar comparison can be made for any coverage less than the base coverage. Finally, the bootstrap method (Appendix C) is used to construct 95% confidence intervals for the expected interpolated and extrapolated species richnesses, for any given sample coverage less than or equal to the base coverage. Currently, we do not consider the variation of our sample coverage estimator in constructing these confidence intervals, because our sample coverage estimator in Eqs. 4a, 4b, or 9a (for $m^* < n$) is generally accurate for large reference sample sizes. The effect of its variation on the confidence intervals and related inferences is thus limited.

Based on 95% confidence intervals, rigorous statistical comparison can be performed not only for rarefaction but also for extrapolated richness values. For any fixed sample coverage less than or equal to the base coverage, if the confidence intervals do not overlap, then significant differences at a level of 5% among the expected species richnesses (whether interpolated or extrapolated) are guaranteed. Examples are provided in the next section. However, overlapped intervals do not imply nonsignificance (Schenker and Gentleman 2001). For any fixed coverage, typical multiple comparisons can be performed to test whether expected species richnesses are significantly different, but simultaneously comparing entire sampling curves needs further research. See Colwell et al. (2012) for details.

Nevertheless, it is important to realize that the null hypothesis of no difference in expected richnesses will virtually always be false in real communities, so statistically significant differences can always be found if sample size is large enough. A more meaningful approach is to estimate an easily interpretable measure of the degree of difference between the richnesses of the communities at a given coverage, and provide confidence intervals for it. The richness ratio $\hat{S}_1/\hat{S}_2$ *at fixed coverage* is a good measure for this. Then the variance estimator for each estimated richness can be used to produce confidence intervals for their expected ratio. If the variance estimator for the estimated richness $\hat{S}_i$ is

$\text{var}(\hat{S}_i)$, then the 95% confidence interval of the expected species ratio *at the same coverage* is $(\hat{S}_1/\hat{S}_2) \pm 1.96\ \text{SE}(\hat{S}_1/\hat{S}_2)$, where $\text{SE}(\hat{S}_1/\hat{S}_2)$ denotes an approximate standard error of $\hat{S}_1/\hat{S}_2$ and $\text{SE}(\hat{S}_1/\hat{S}_2) \approx (\hat{S}_1/\hat{S}_2)[\text{var}(\hat{S}_1)/\hat{S}_1^2 + \text{var}(\hat{S}_2)/\hat{S}_2^2]^{1/2}$.

Applications

We used two real ecological data sets to show how to construct coverage-based rarefaction and extrapolation curves using our analytic method. We have developed an R program (R Development Core Team 2008) to compute the proposed rarefaction/extrapolation curves and the associated 95% confidence intervals. The program will be implemented in iNEXT (interpolation/extrapolation; available from the author).

Example 1

Janzen (1973*a*, *b*) tabulated many data sets of tropical foliage insects from sweep samples in Costa Rica. We selected two beetle data sets to compare beetle species richness between an old-growth forest site and a second-growth site. The data are given in Appendix F: Table F1. 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 $f_2 = 17$. Based on Eq. 4a, the sample coverage estimate for the reference sample is 93% (SE = 0.74%). In the Osa old-growth site, there were 112 species, 237 individuals, and $f_1 = 84$, $f_2 = 10$, yielding a coverage estimate of 65% (SE = 3.5%).

In the raw data, fewer species (112 vs. 140) were found in the old-growth site than in the second-growth site in the reference samples; see Fig. 3a. Traditional size-based rarefaction analysis would draw the second-growth sample size down to 237 (Fig. 3b for a size of 237) and conclude that, for a standardized size of 237, the old-growth site has more species (112 vs. 70), but that the ratio of old-growth richness to second-growth richness is only 1.60.

In Fig. 3a, we show the size-based rarefaction and extrapolation curves, which connect smoothly at the reference points. For the Osa second-growth site, the extrapolation is extended to a sample size of 2000 (about double of the reference sample size) based on our criterion for reliable extrapolation. For the Osa old-growth site, the extrapolation is extended to 500 (about double of the reference sample size) based on the same criterion. (Extrapolation beyond double the sizes of the reference samples theoretically could be computed, but they may be subject to some biases and should be used with caution in estimating species richness ratios or other measures. Colwell et al. [2012] extrapolated for this case to a much larger size of 1200 because their goal was mainly to rank species richnesses.) Since data are relatively sparse in the old-growth site, this extrapolation inevitably yields wider confidence intervals than those for the second-growth site for any fixed size. We can compute the beetle species richness ratio of the two sites for any sample size less than or equal to 500. We compare the ratio for three sample sizes (100, 237, and 500) in Fig. 3b. Except for the initial small sizes, the old-growth and second-growth confidence intervals do not overlap for any sample size considered. This implies that beetle species richness for any sample size is significantly greater in the old-growth site than that in the second-growth site for sample size up to 500 individuals (Fig. 3b). However, more important than the statistical significance is the estimate of the ratio of the old growth richness to the second growth richness. The richness ratios for these three sizes (100, 237, and 500 individuals) are 1.32 (SE = 0.07), 1.60 (SE = 0.10), and 1.92 (SE = 0.16). All computational details are provided in Appendix F: Table F2.

The sample coverage estimates show that these two samples have very different degrees of completeness. In the old-growth forest site, the sample only covers 65% of the population, whereas the sample in the second-growth site covers 93%. For the old-growth site, when the sample size is extended from 237 to 500 (as in Fig. 3a), the sample coverage is extended from 65% to 73% (as in Fig. 3c) using the formula in Eq. 9a. Since the coverage for the reference sample in the second-growth site is 93%, our "base coverage" is 73%. In other words, if our goal is to compare these two sites, only rarefaction is needed in the secondary-growth site. However, if the goal is to predict the richness of the second-growth site, we can extend the size from 976 to 2000 as we did in the size-based approach. The coverage is extended from 93% to 96%, only an increment of 3%. In Fig. 3c, we show the coverage-based rarefaction and extrapolation. We can compare the two sites for any community coverage less than or equal to 73%; for coverage higher than 73%, our data do not provide sufficient information to supply reliable information about the ratio of their richnesses.

The three sample sizes considered in size-based comparisons (Fig. 3b) correspond to sample coverages of 0.55, 0.65, and 0.73 in the old-growth forest (the sample with the lower coverage), and the corresponding species richness ratios for these coverages are 2.56 (SE = 0.13), 3.58 (SE = 0.23), and 4.68 (SE = 0.37), respectively (Fig. 3d). These ratios are much higher than those obtained in size-based sampling curves. For two equally complete samples of 55% coverage, the species ratio is 2.56, while for two equally complete samples of 73% coverage, the species ratio increases to 4.68. Unlike the sized-based standardization in which size is determined by samplers, here the coverage-based standardization compares equal population fractions of each community. The population fraction is a community-level characteristic that can be reliably estimated from data. All computational details for coverage-based curves are provided in Appendix F: Table F3. Except for very low coverage values, the old-growth and second-growth confidence intervals do not overlap for any fixed

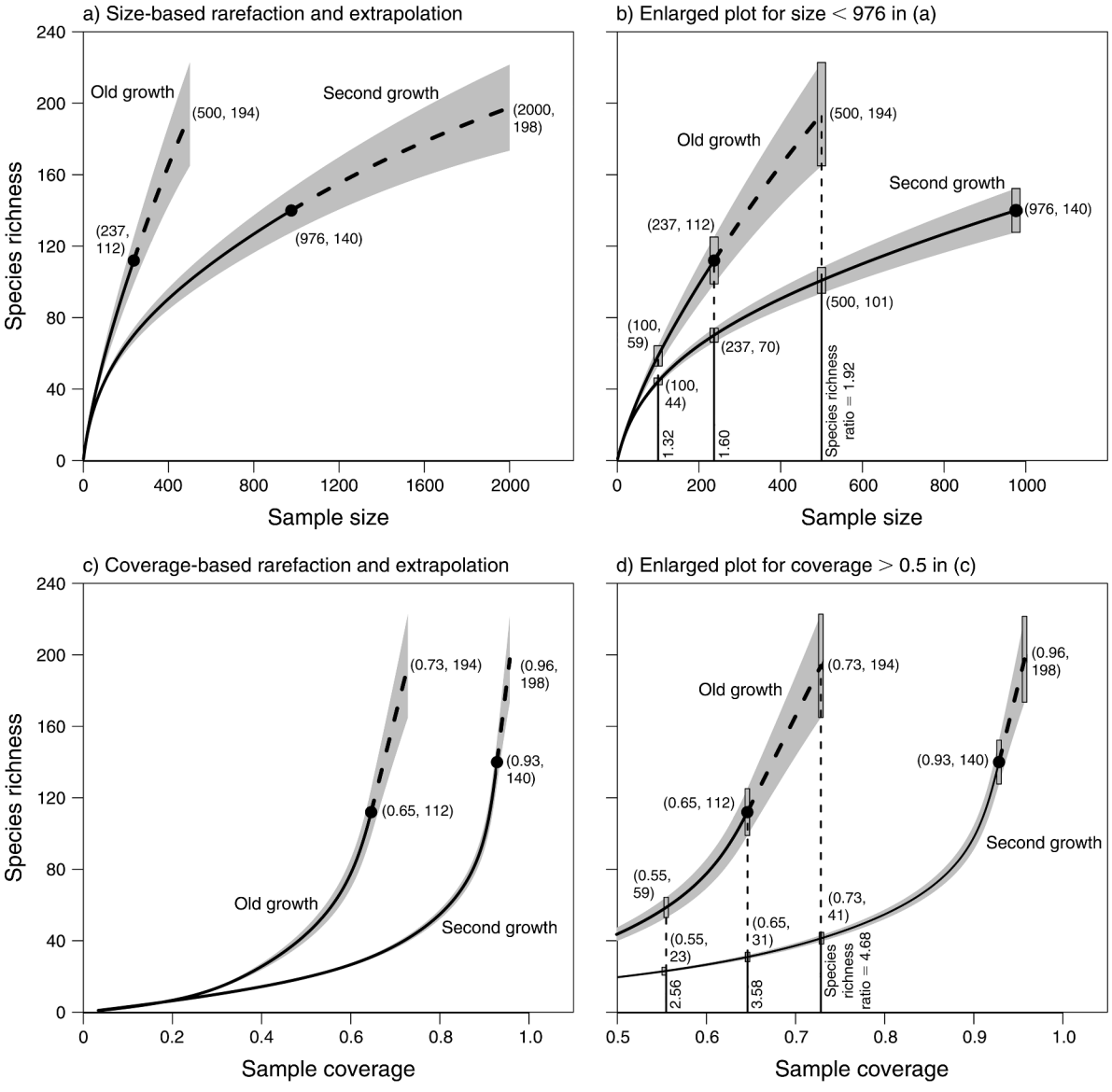

FIG. 3. (a, b) Size-based rarefaction (solid curves) and extrapolation (dashed curves) and (c, d) the proposed coverage-based rarefaction/extrapolation curve with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) comparing beetle species richness in an old-growth forest site and a second-growth vegetation site (Janzen 1973*a*, *b*). Reference samples are indicated by solid black dots. For species richness, numbers are rounded to the nearest integer for display simplicity, but all significant figures were retained in all calculations. The species richness ratio is calculated for sample size $m = 100$, 237, and 500 in panel (b). These three sample sizes correspond to sample coverages of 0.55, 0.65, and 0.73, respectively, in the old-growth forest (the sample with the lower coverage), and the three corresponding species richness ratios for these three coverages are shown in panel (d). See Appendix F (Tables F2 and F3) for computational details. The numbers in parentheses show the *x*- and *y*-axis coordinates for each point. The small rectangles represent the 95% confidence interval of species richness specifically for three sample sizes in panel (b), and three sample coverage values in panel (d).

coverage less than or equal to 73%, again implying significant differences in beetle species richnesses.

Example 2

Baker et al. (2002) surveyed bird richness and abundance in three types of habitats (woodland, heath, and ecotone) in South Australia. The reader is referred to Baker et al. (2002) for data details. We analyzed their data from the pure woodland and pure heath plots. In the woodland plot, there were 69 species, 2482 individuals, $f_1 = 12$, $f_2 = 3$, and an estimated sample coverage of 99.5% (SE = 0.13%). In the heath plot, there were 40 species, 1295 individuals, $f_1 = 6$, $f_2 = 4$, and an estimated sample coverage of 99.5% (SE = 0.18%). The two samples have nearly identical sample coverages. In Fig. 4a, we show the rarefaction curve for both plots, as

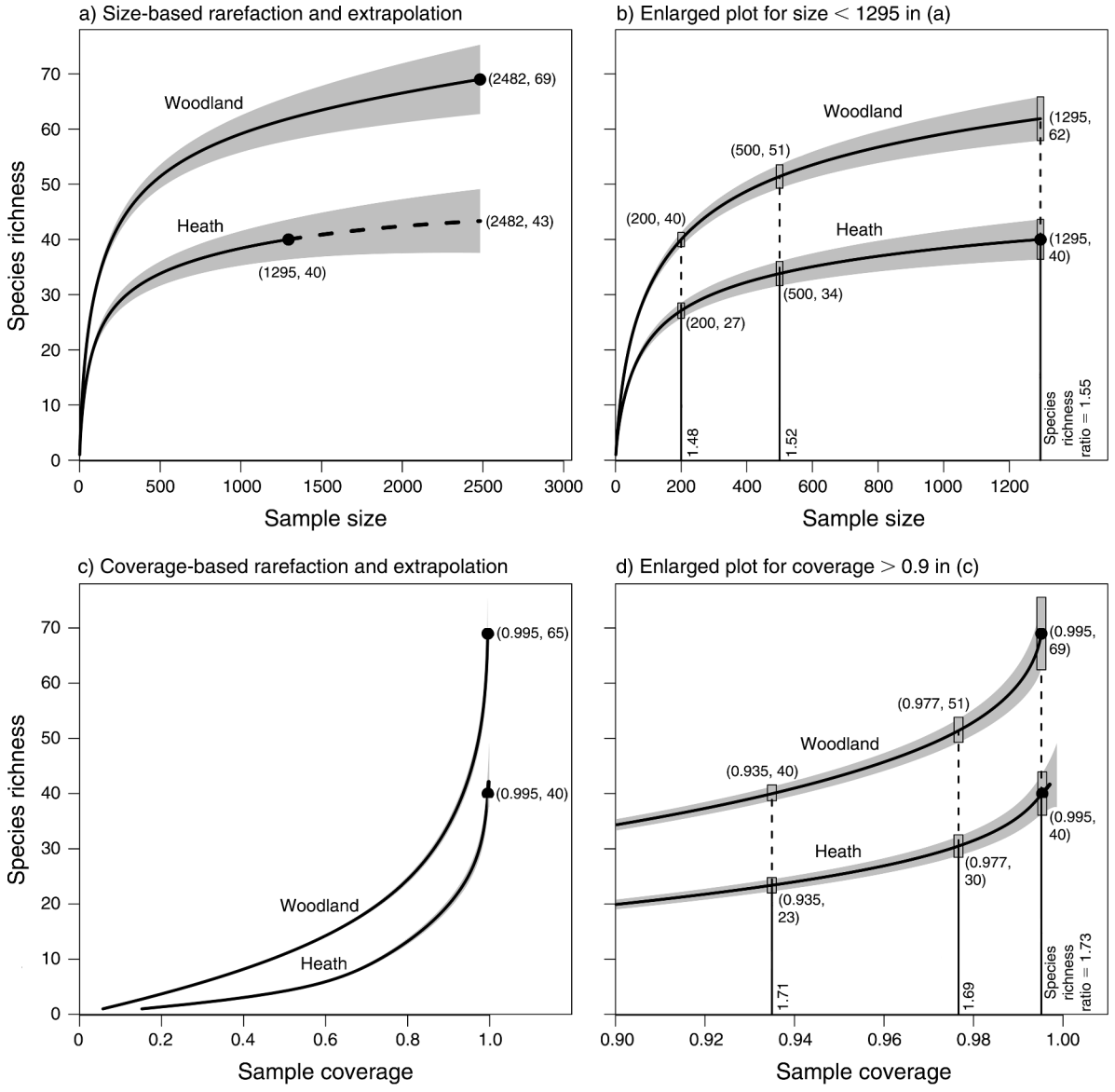

FIG. 4. (a, b) Size-based rarefaction (solid curves) and extrapolation (dashed curves) and (c, d) the proposed coverage-based rarefaction/extrapolation curve with 95% confidence intervals (shaded areas, based on a bootstrap method with 200 replications) comparing bird species richness in woodland and heath habitats (Baker et al. 2002). Reference samples are indicated by solid black dots. For species richness, numbers are rounded to the nearest integer for display simplicity, but all significant figures were retained in all calculations. Note that the extrapolation in the heath plot increases the sample coverage only 0.4% when the sample size is increased from 1295 to 2482. So the coverage-based extrapolation part in panels (c) and (d) is almost invisible. The species richness ratio is calculated for sample sizes $m = 200$, 500, and 1295 in panel (b). These three sizes correspond to sample coverage 0.935, 0.977, and 0.995, respectively, in the woodland (the sample with the lower coverage), and the three corresponding species richness ratios for these three coverages are shown in panel (d). See Appendix F (Tables F4 and F5) for computational details. The numbers in parentheses show the x- and y-axis coordinates for each point. The small rectangles represent the 95% confidence interval of species richness specifically for three sample sizes in panel (b), and three sample coverage values in panel (d).

well as the extrapolation curve in the heath plot from size 1295 to 2482 (the reference sample size of the woodland plot). However, this extrapolation in the heath plot increases the sample coverage from 99.5% to 99.9%, only a 0.4% increase. So the coverage-based extrapolation part in Fig. 4c and 4d is almost invisible. In the woodland plot, if the sample size is doubled, the coverage increases from 99.5% to 99.7%, only a 0.2% increase. Thus, we didn't extrapolate in the woodland plot, and our comparison will be focused on rarefaction of these samples.

Traditional rarefaction analysis draws the sample size of the woodland plot down to 1295 (Fig. 4a, b) and concludes that, for a standardized count of 1295, the

woodland's estimated richness is 1.55 times that of the heath's (62 vs. 40 species). However, the two samples have almost identical sample coverage values (99.5%), implying they are equally complete. Therefore, the raw data by themselves can be directly compared. The estimated woodland richness is 1.73 times that of the heath (69 vs. 40 species). In the size-based rarefaction curve (Fig. 4b), the species richness ratios for sample sizes (200, 500, 1295) are, respectively, 1.48 (SE = 0.05), 1.52 (SE = 0.06), and 1.55 (SE = 0.08). These sample sizes (200, 500, 1295) correspond to sample coverages of 0.935, 0.977, 0.995 in the woodland site (the sample with the lower coverage), and the corresponding species richness ratios for these coverages are 1.71 (SE = 0.05), 1.69 (SE = 0.06), and 1.73 (SE = 0.10), respectively. All computational procedures are parallel to those for Example 1 and are provided in Appendix F: Tables F4 and F5.

Discussion

The purpose of rarefaction and extrapolation is to make fair comparisons between incomplete samples. Although traditional size-based rarefaction and extrapolation, in which the samples are all standardized to equal size, provides useful sampling information, we have argued that it is often more informative to standardize them to equal coverage. This ensures that we are comparing samples of equal quality and equal completeness, and allows us to make more robust and detailed inferences about the sampled communities.

This coverage-based standardization dovetails very well with the way that ecologists actually choose the sample size for their studies. Ecologists normally take larger samples from more diverse communities, and smaller samples from communities with few rare species. Often sampling is continued until the slope of the rarefaction curve decreases to some predetermined value (as in *Example 2* above; Schloss and Handelsman 2004, Tringe et al. 2005). Because the slope of the rarefaction curve equals the coverage deficit, the samples thus obtained from different communities will have equal coverages, so coverage-based standardization will not throw any of that data away. In contrast, traditional rarefaction based on sample size has no connection with this common and sensible sampling strategy, and will often require ecologists to throw away much of their hard-earned data, in reducing samples to the lowest common size.

This method of comparing coverage-based samples from multiple communities has many other important advantages over traditional size-based method. Most important is that it gives more meaningful information about the degree of the differences in diversity among the communities being compared. In traditional size-based comparison, the ratio of the richnesses of the rarefied samples from any two communities is necessarily close to unity when one of the samples is small, and it will depend strongly on sample size if the communities differ greatly in diversity. That means we can't make an objective judgment about how much more diverse one community is compared to the other. Perhaps for this reason, ecologists have often limited themselves to merely ranking communities according to their rarefied richnesses, instead of estimating the magnitude of the differences in the diversities of the communities. Using coverage-based standardization, however, there is no intrinsic bias with respect to sample size in this diversity ratio, unlike the ratio produced by traditional size-based method. The estimated richnesses of the communities at fixed interpolated or extrapolated coverage obey a replication principle; when one community is unambiguously K times as diverse as another (Hill numbers of all orders q for the first community are K times the corresponding Hill numbers for the second community), the ratio of their estimated richnesses is approximately K, even for very small sample sizes, and is approximately independent of sample size. This means ecologists can now go beyond ranking and discuss the more interesting question of the actual magnitudes of the differences in diversity between communities. Although in this paper we primarily focus on species richness ratio to characterize the degree of difference, all methods and discussion can be extended to other comparison criterion such as absolute or relative difference based on a fixed fraction of community individuals.

Yet even for the limited question of ranking communities according to their true species richness, coverage-based standardization is superior to traditional standardization. If size-based sampling curves (rarefaction plus extrapolation) for two communities do not cross for any finite sample size greater than 1, then the two coverage-based curves also do not cross at any finite coverage less than 1 beyond the base point. The reverse is also true. Thus, the two types of curves always give the same qualitative ordering of species richness. If crossing occurs, then size-based and coverage-based curves have exactly the same number of crossing points, but we have proven that the coverage-based method is always more efficient (needing smaller sample sizes in each community) than the traditional method for detecting any specific crossing point. For example, suppose we have two communities: Community 1 includes 200 equally abundant species. Community 2 includes 400 species, but the community is dominated by only three species, and the relative abundance distribution is $\{0.2, 0.15, 0.15, 0.00126 \times 397\}$. For size-based SACs, both sample sizes must be at least 476 (952 individuals in total; 181 species in both communities will be observed at this common size) to detect the correct ranking of the communities in terms of species richness. For coverage-based SACs, only an expected sample coverage of 0.66 (sample size 216 from Community 1, and sample size 314 from Community 2, 530 individuals in total; 132 species) is needed to detect the correct ranking. Our methodology reduces the investigator's sampling effort by half. See Appendix G (Figs. G3 and G4) for the

PLATE 1. Unexpected evolutionary radiation of the orchid genus *Teagueia* Luer in the upper Rio Pastaza watershed near Baños, Ecuador. All are new species discovered in the last 12 years by Jost and his students. The most important sites for these and other local endemic species are now protected by a chain of private reserves established by the EcoMinga Foundation. Though this area has been sampled by botanists and zoologists for 150 years, new species (especially orchids and frogs) continue to be discovered at a high rate, showing that the sample coverage is only moderate for some taxonomic groups. Since there are many unseen species, the methods described in our paper would be needed to make fair comparisons between site richnesses. Photo credit: L. Jost.

size- and coverage-based SACs for this example; the proofs for the conclusions in this paragraph, and more comparisons of the two types of standardizations, are also provided in that appendix.

Nevertheless, it is important to reiterate that the main purpose of rarefaction and extrapolation (of either kind) is to make meaningful and statistically reliable comparisons of well-defined subsets of each community, not to rank communities according to their true species richnesses (Lande et al. 2000). There is no way to be sure that species accumulation curves do not cross at sample sizes greatly exceeding the actual sample sizes or the extrapolated sizes. Thus, even though coverage-based standardization will find the correct ranking using smaller sample sizes than traditional standardization, we cannot know whether any given sample sizes are large enough to be past the crossing point(s). Even with extrapolation, there is never a guarantee that the ranking obtained will not reverse itself for very large sample sizes. Ranking according to true species richness is not statistically possible if all we have are limited samples; the best we can do is to establish lower bounds for the richnesses of each community (Bunge and Fitzpatrick 1993, Chao 2005). There can always be some vanishingly rare species hiding somewhere.

We should not lose sleep over this. We can always estimate the coverage deficit of our samples, which tells us what proportion of the community's individuals are made up of those hidden species. The coverage (Good 1953, Engen 1978, Bunge and Fitzpatrick 1993, Lande et al. 2000) and its deficit are the only aspects of unobserved or hidden species that can be accurately estimated by sample data. If the coverage deficit shows that the undetected species make up a vanishingly small proportion of the community's individuals, why worry about them? The richnesses returned by coverage-based methods (like those of size-based methods) are actually more like robust frequency-dependent diversity measures, in that they are sensitive to the degree of dominance in the communities. This sensitivity decreases as sample size or coverage increases. Coverage-based curves therefore provide information about community

structure, somewhat like diversity profiles (Tóthmérész 1995). See Appendix H for an example of inferences about community structure and some basic theoretical properties for coverage-based curves.

Our presentation here is focused on the sampling curve produced by standardizing sample coverage when a sample of individuals is taken from a community. Gotelli and Colwell (2001) distinguished two types of rarefaction curves: individual-based (the sampling unit is an individual) and sample-based (the sampling unit is a sample or quadrat). The analytic rarefaction and extrapolation formulas proposed in this paper for standardizing sample coverage can be extended to handle sample-based data. This extension, with applications, will be reported elsewhere.

For coverage-based rarefaction, we have presented a graphical method (see Fig. 2), an analytic method, and a new algorithmic method. Algorithmic methods are flexible enough to estimate other quantities besides species richness for a fixed coverage level. This would permit calculation of many evenness and inequality measures (Ricotta 2003, Jost 2010), as well as other diversity measures for the fixed-coverage sample. Evenness and inequality have always been difficult to measure because they are so sensitive to species richness, which is, in turn, very sensitive to sample size. However, their values on a fixed-coverage sample are well defined and can be estimated to any desired precision. We are currently working on the details of this approach.

Acknowledgments

The authors thank John Alroy, Robert Colwell, Nicholas Gotelli, Glenda Mendieta-Leiva, and two anonymous reviewers for carefully reading and editing an earlier version and providing very helpful and insightful comments, which substantially improved this paper. Especially, Alroy and Gotelli suggested plotting of a semi-log plot in Fig. 1, Gotelli and Colwell's comments led to the material in Appendices G and H; one reviewer suggested including extrapolation, and the other reviewer provided some relevant references. We also thank T. C. Hsieh and K. H. Ma for discussion and computational help in graphics. A. Chao is supported by Taiwan National Science Council under Contract 100-2118-M007-006-MY3. L. Jost acknowledges support from John V. Moore through a grant to the Population Biology Foundation.

Literature Cited

Alroy, J. 2009. A deconstruction of Sepkoski's Phanerozoic marine evolutionary faunas based on new diversity estimates. Page 507 *in* Geological Society of America abstracts with programs. GSA annual meeting, Portland, Oregon, USA. Geological Society of America, Boulder, Colorado, USA.

Alroy, J. 2010*a*. Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. Palaeontology 53:1211–1235.

Alroy, J. 2010*b*. The shifting balance of diversity among major marine animal groups. Science 329:1191–1194.

Baker, J., K. French, and R. J. Whelan. 2002. The edge effect and ecotonal species: bird communities across a natural edge in southeastern Australia. Ecology 83:3048–3059.

Boneh, S., A. Boneh, and R. J. Caron. 1998. Estimating the prediction function and the number of unseen species in sampling with replacement. Journal of the American Statistical Association 93:372–379.

Bunge, J., and M. Fitzpatrick. 1993. Estimating the number of species: A review. Journal of the American Statistical Association 88:364–373.

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

Chao, A. 2005. Species estimation and applications. Pages 7907–7916 *in* S. Kotz, N. Balakrishnan, C. B. Read, and B. Vidakovic, editors. Encyclopedia of statistical sciences. Wiley, New York, New York, USA.

Chao, A., R. K. Colwell, C. W. Lin, and N. J. Gotelli. 2009. Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90:1125–1133.

Chao, A., and S. M. Lee. 1992. Estimating the number of classes via sample coverage. Journal of the American Statistical Association 87:210–217.

Chao, A., S. M. Lee, and T. C. Chen. 1988. A generalized Good's nonparametric coverage estimator. Chinese Journal of Mathematics 16:189–199.

Chao, A., and T. J. Shen. 2004. Non-parametric prediction in species sampling. Journal of Agricultural, Biological and Environmental Statistics 9:253–269.

Chao, A., and T. J. Shen. 2010. Program SPADE: species prediction and diversity estimation. Program and user's guide. CARE, Hsin-Chu, Taiwan. http://chao.stat.nthu.edu.tw/softwareCE.html

Coddington, J. A., I. Agnarsson, J. A. Miller, M. Kuntner, and G. Hormiga. 2009. Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. Journal of Animal Ecology 78:573–584.

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.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society of London B 345:101–118.

Colwell, R. K., C. X. Mao, and J. Chang. 2004. Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology 85:2717–2727.

DeVries, P. J., and T. R. Walla. 2001. Species diversity and community structure in neotropical fruit-feeding butterflies. Biological Journal of the Linnean Society 74:1–15.

Engen, S. 1978. Stochastic abundance models. Chapman and Hall, London, UK.

Esty, W. W. 1983. A normal limit law for a nonparametric estimator the coverage of a random sample. Annals of Statistics 11:905–912.

Esty, W. W. 1986. The efficiency of Good's nonparametric coverage estimator. Annals of Statistics 14:1257–1260.

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.

Good, I. J., and G. H. Toulmin. 1956. The number of new species, and the increase in population coverage, when a sample is increased. Biometrika 43:45–63.

Gotelli, N. J., and A. Chao. 2013. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. *In* S. A. Levin, editor. The encyclopedia of biodiversity. Second edition. Elsevier, New York, New York, USA, *in press*.

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

Gotelli, N. J., and R. K. Colwell. 2011. Estimating species richness. Pages 39–54 *in* A. Magurran and B. McGill, editors.

Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford, UK.
Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology 52:577–586.
Janzen, D. H. 1973*a*. Sweep samples of tropical foliage insects: description of study sites, with data on species abundances and size distributions. Ecology 54:659–686.
Janzen, D. H. 1973*b*. Sweep samples of tropical foliage insects: effects of seasons, vegetation types, elevation, time of day, and insularity. Ecology 54:687–708.
Jost, L. 2007. Partitioning diversity into independent alpha and beta components. Ecology 88:2427–2439.
Jost, L. 2010. The relation between evenness and diversity. Diversity 2:207–232.
Lande, R., P. J. DeVries, and T. R. Walla. 2000. When species accumulation curves intersect: implications for ranking diversity using small samples. Oikos 89:601–605.
Lijoi, A., R. H. Mena, and I. Prünster. 2007. Bayesian nonparametric estimation of the probability of discovering a new species. Biometrika 94:769–786.
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.
Mao, C. X., R. K. Colwell, and J. Chang. 2005. Estimating species accumulation curves using mixtures. Biometrics 61:433–441.
Olszewski, T. D. 2004. A unified mathematical framework for the measurement of richness and evenness within and among multiple communities. Oikos 104:377–387.
Peet, R. K. 1974. The measurement of species diversity. Annual Review of Ecology and Systematics 5:285–307.
R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.
Rasmussen, S. L., and N. Starr. 1979. Optimal and adaptive stopping in the search for new species. Journal of the American Statistical Association 74:661–667.
Ricotta, C. 2003. On parametric evenness measures. Journal of Theoretical Biology 222:189–197.
Robbins, H. 1968. Estimating the total probability of the unobserved outcomes of an experiment. Annals of Mathematical Statistics 39:256–257.
Sanders, H. L. 1968. Marine benthic diversity: a comparative study. American Naturalist 102:243–282.
Schenker, N., and J. F. Gentleman. 2001. On judging the significance of differences by examining overlap between confidence intervals. American Statistician 55:182–186.
Schloss, P. D., and J. Handelsman. 2004. Status of the microbial census. Microbiology and Molecular Biology Reviews 68:686–691.
Shen, T. J., A. Chao, and C. F. Lin. 2003. Predicting the number of new species in further taxonomic sampling. Ecology 84:798–804.
Simberloff, D. 1972. Properties of the rarefaction diversity measurement. American Naturalist 106:414–418.
Smith, W., and J. F. Grassle. 1977. Sampling properties of a family of diversity measures. Biometrics 33:283–292.
Tóthmérész, B. 1995. Comparison of different methods for diversity ordering. Journal of Vegetation Science 6:283–290.
Tringe, S. G., C. Von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, and J. C. Detter. 2005. Comparative metagenomics of microbial communities. Science 308:554.
Zhang, C.-H., and Z. Zhang. 2009. Asymptotic normality of a nonparametric estimator of sample coverage. Annals of Statistics 37:2582–2595.
Zhang, Z., and H. Huang. 2007. Turing's formula revisited. Journal of Quantitative Linguistics 4:222–241.

Supplemental Material

Appendix A

Replication principle for the expected species richness when expected sample coverage is standardized (*Ecological Archives* E093-238-A1).

Appendix B

An unbiased estimator for the expected sample coverage in coverage-based rarefaction (*Ecological Archives* E093-238-A2).

Appendix C

A bootstrap method to construct the confidence interval of the expected species richness at a given coverage (*Ecological Archives* E093-238-A3).

Appendix D

Theoretical proof of the proposed unbiased algorithm in coverage-based rarefaction (*Ecological Archives* E093-238-A4).

Appendix E

Coverage-based extrapolation (*Ecological Archives* E093-238-A5).

Appendix F

Beetle species frequency data (Janzen 1973*a*, *b*) and computational details for two examples (*Ecological Archives* E093-238-A6).

Appendix G

Comparisons of size- and coverage-based species accumulation curves (*Ecological Archives* E093-238-A7).

Appendix H

Some basic theoretical properties for coverage-based species accumulation curves (*Ecological Archives* E093-238-A8).