Ecology, 85(10), 2004, pp. 2717–2727

INTERPOLATING, EXTRAPOLATING, AND COMPARING INCIDENCE-BASED SPECIES ACCUMULATION CURVES

ROBERT K. COLWELL,[1,4] CHANG XUAN MAO,[2] AND JING CHANG[3]

[1]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269-3043 USA*
[2]*Department of Statistics, University of California, Riverside, California 92521 USA*
[3]*Department of Preventive Medicine, University of Southern California, Los Angeles, California 90089 USA*

Abstract. A general binomial mixture model is proposed for the species accumulation function based on presence–absence (incidence) of species in a sample of quadrats or other sampling units. The model covers interpolation between zero and the observed number of samples, as well as extrapolation beyond the observed sample set. For interpolation (sample-based rarefaction), easily calculated, closed-form expressions for both expected richness and its confidence limits are developed (using the method of moments) that completely eliminate the need for resampling methods and permit direct statistical comparison of richness between sample sets. An incidence-based form of the Coleman (random-placement) model is developed and compared with the moment-based interpolation method. For extrapolation beyond the empirical sample set (and simultaneously, as an alternative method of interpolation), a likelihood-based estimator with a bootstrap confidence interval is described that relies on a sequential, AIC-guided algorithm to fit the mixture model parameters. Both the moment-based and likelihood-based estimators are illustrated with data sets for temperate birds and tropical seeds, ants, and trees. The moment-based estimator is confidently recommended for interpolation (sample-based rarefaction). For extrapolation, the likelihood-based estimator performs well for doubling or tripling the number of empirical samples, but it is not reliable for estimating the richness asymptote. The sensitivity of individual-based and sample-based rarefaction to spatial (or temporal) patchiness is discussed.

Key words: binomial mixture model; Coleman curve; EstimateS; random placement; rarefaction; richness estimation; richness extrapolation; species accumulation curve; species richness.

INTRODUCTION

Ecologists and conservation biologists often need to know the number of species (species richness) found in a designated area, or they need to compare the number of species in different areas. In many cases, however, it is impractical or even impossible to enumerate species directly. Sampling is therefore necessary. Unfortunately, observed species richness within habitats (alpha diversity) is notoriously dependent on sample size, due to sampling effects. In addition, observed richness depends intrinsically on sample size when data from different habitats are successively pooled, due to species turnover (change in species composition or beta diversity). The study of empirical species–area relationships (e.g., Rosenzweig 1995, Scheiner 2003) generally focuses on, or at least admits, the latter source of sample-size dependence. In this paper, we focus instead on the measurement of species richness on local scales, where sampling issues are substantially more important than turnover. In statistical terms, we are concerned with sample sets in which each sample can reasonably be considered a random sample from the same universe. In practical terms, this means that the order of samples in time or their arrangement in space within a sample set is of no importance; in fact, unimportance of sample order is diagnostic of the kinds of sample sets appropriately used by ecologists to assess local (alpha) diversity.

A *species accumulation curve* is the graph of the number of observed species as a function of some measure of the sampling effort required to observe them. (In the broad sense, classical species–area curves, which focus on beta diversity, are thus species accumulation curves, but the curves that we treat in this paper may or may not use area as the measure of effort, and explicitly depict alpha diversity, as just explained.) The sequential accumulation of individuals in a single sample, or the successive pooling of samples from a single sample set, produces a species accumulation curve, but it will not be a smooth curve because of spatial (or temporal) patchiness and simple stochastic effects. For the individuals in a single sample, classical *individual-based* rarefaction may be used to produce a smooth curve that estimates the number of species that would be observed for any smaller number of individuals, under the assumption of random mixing of individuals (Hurlbert 1971, Simberloff 1972, Heck et al. 1975). For replicated sets of samples (sample sets), the expected number of species that would be observed for any smaller number of samples can be estimated by

Manuscript received 22 August 2003; revised 3 April 2004; accepted 4 April 2004. Corresponding Editor: F. He.

[4] E-mail: colwell@uconn.edu

TABLE 1. Empirical data sets used to illustrate methods.

Taxon and method	Species	Samples	Locality	References
Temperate birds, survey data	67	50	N. American Breeding Bird Survey, 1998	Dorazio and Royle (2003)
Rain forest ants, Winkler traps	197	41	La Selva Biological Station, Costa Rica	Longino et al. (2002)
Rain forest seedbank, quadrats	34	121	La Selva Biological Station, Costa Rica	Butler and Chazdon (1998)
Sapling trees (2.5–5.0 cm dbh): tropical old-growth forest, quadrats	60	100	La Selva Biological Station, Costa Rica	R. L. Chazdon (*unpublished data*)
Sapling trees (2.5–5.0 cm dbh): tropical second-growth forest, quadrats	50	100	La Selva Biological Station, Costa Rica	R. L. Chazdon (*unpublished data*)

sample-based rarefaction, under the assumption of random sample order (Gotelli and Colwell 2001). (Although sample-based rarefaction is the more accurate term, ''smoothing the species accumulation curve'' by random resampling is an accurate description for the same concept.)

A sample-based species accumulation curve can be constructed from any empirical species-by-sample matrix. The cells of the empirical matrix may contain species abundances (an *abundance matrix*) or simply presence/absence data (an *incidence matrix*). Of course, any abundance matrix can be transformed to the corresponding incidence matrix by replacing each nonzero cell value by a 1 to indicate presence. Sample-based species accumulation curves, by their nature, depend only on incidence data, even if abundance data are available (which they will not be for some kinds of sample sets).

Until recently, sample-based rarefaction curves had to be constructed by computationally intensive resampling algorithms, such as those used by the freeware application EstimateS (Colwell 1994–2004). The practical need for such tools is indicated by the fact that, as of 2003, >10 000 copies of EstimateS had been downloaded by users in nearly 100 countries and used in scores of published papers. (As another indication, ⟨www.google.com⟩ currently indexes >1000 pages that cite EstimateS). As with individual-based rarefaction, sample-based rarefaction permits comparison of different assemblages at comparable levels of sampling effort. Unfortunately, no adequate method has been available previously for computing confidence intervals for sample-based rarefaction curves, severely limiting their utility for comparing the richness of sample sets.

The construction of a sample-based rarefaction curve can be viewed as a process of *interpolation* from the pooled species richness of the full set of samples to the expected richness of a subset of those samples. The dream of every biologist involved in biotic inventory is the rigorous *extrapolation* of empirical sample-based rarefaction curves to estimate, with confidence intervals, how many species would be found in a larger set of samples from the same assemblage; ideally, extrapolation would yield the asymptotic, ''true'' richness of the assemblage.

In this paper, we outline a unified, statistically rigorous binomial mixture model for incidence patterns in multispecies assemblages. (The full statistical development of the model, with supporting theorems and proofs, appears elsewhere [Mao et al. 2004]). Based on the model, we present simple, analytical formulas for sample-based rarefaction curves and their confidence intervals (interpolation). These formulas completely replace resampling methods for producing sample-based rarefaction curves. In addition, for the first time, we explore a non-curve-fitting extrapolation method, with bootstrap confidence intervals. We illustrate both interpolation and extrapolation using data sets for tropical forest ants and trees, a tropical seed bank, and temperate birds (Table 1).

THE MODEL

Consider a species assemblage with an unknown true richness S sampled by quadrats, traps, lures, seines, dredge hauls, mist nets, or other replicated sampling units. (For model development, we will call these sampling units *quadrats*, to avoid having to use the word *sample* as both noun and verb.) The data for h quadrats are expressed as an S-by-h species–quadrat incidence matrix consisting of the presence indicators Z_{ij}:

$$Z_{ij} = \begin{cases} 1 & \text{if the } i\text{th species is present in the } j\text{th quadrat} \\ 0 & \text{if the } i\text{th species is absent in the } j\text{th quadrat} \end{cases}$$

To develop a theoretical model, we make two statistical assumptions: (1) the ith species has the same probability ϕ_i of being present in each quadrat, and (2) the Z_{ij} are independent, given ϕ_i, over all i and j. The species accumulation function, which gives the expected number of species observed in h quadrats, is the sum of the probabilities, across species, that each species is not absent from all h quadrats:

$$\tau(h) = \sum_{i=1}^{S} [1 - (1 - \phi_i)^h]. \qquad (1)$$

Species with identical presence probabilities ϕ can be considered together as a group. Suppose that there are G such homogeneous *incidence groups* (some or all of which may contain only a single species). For the kth incidence group, let ψ_k be the common *presence probability* (a measure of rarity or commonness) and π_k be the *relative group size,* that is, the number of species in the kth group divided by the total number of species, S. The species accumulation function $\tau(h)$ then becomes

$$\tau(h) = S \sum_{k=1}^{G} \pi_k[1 - (1 - \psi_k)^h]. \qquad (2)$$

The asymptote $\tau(\infty)$, the limit of $\tau(h)$ as the number of quadrats h goes to infinity, is identical to the true richness S. Note that it is possible to rewrite the species accumulation function $\tau(h)$ as

$$\tau(h) = S \sum_{k=1}^{G} \pi_k(1 - e^{-C_k h}) \qquad (3)$$

where $C_k = -\log(1 - \psi_k)$. This reformulation allows us to show that our model is a nonparametric generalization of the classical negative exponential model of Holdridge et al. (1971) and Soberón and Llorente (1993). The negative exponential model assumes that all species share the same presence probability ψ_1 and thus form a single incidence group. Thus we may set $G = 1$, $\pi_1 = 1$, and $C_1 = C$, yielding the classical exponential model:

$$\tau(h) = S(1 - e^{-Ch}). \qquad (4)$$

In the model expressed by Eq. 2, the number of incidence groups G may take any value, group proportions π_k may vary freely (with the simple constraint that $\Sigma_{k=1}^{G}\pi_k = 1$), and the pattern of presence probabilities ψ_k is unconstrained. Therefore, this rigorous sampling-theoretic model is expected to be applicable to a wide variety of taxa with varying relative abundances and patterns of incidence.

Suppose that we take a random sample of H quadrats, called the *empirical sample set*. If the $Z_{ij} = 0$ for all j (all quadrats), then the ith species is not observed in the empirical sample set. The observed data matrix consists of all rows (species) that have at least one $Z_{ij} > 0$ in the S-by-H species–quadrat matrix. Let s_j stand for the number of species found in exactly j quadrats of the empirical sample set. The s_j are called *counts* (frequencies of occurrence categories). Thus s_0 is the number of species present in the target assemblage but not observed in the empirical sample set, s_1 is the number of species found in precisely one quadrat, s_2 is the number of species found in precisely two quadrats, and so on. The observed richness in the empirical sample set is therefore $S_{\mathrm{obs}} = \Sigma_{j=1}^{H}s_j$, and the total number of species, observed and unobserved, is $S = S_{\mathrm{obs}} + s_0$. The observed counts $s_1, s_2 \ldots s_H$, are *sufficient statistics,* because they contain all of the information necessary for estimating richness as a function of sampling effort, $\tau(h)$, as we demonstrate rigorously elsewhere (Mao et al. 2004) and show by example in the next section.

Interpolation (rarefaction)

An intuitive approach to the estimation of $\tau(h)$ at $h < H$, a process here called *interpolation,* is to systematically enumerate all distinct subsets of h quadrats from the H quadrats of the empirical sample set, find the observed richness in each subset of quadrats, and calculate their mean as an estimator for $\tau(h)$. Table 2 shows a simple example of this procedure for two contrasting, hypothetical sample sets. This systematic enumeration procedure is computationally expensive for large h. The randomization procedure used by EstimateS (Colwell 1994–2004) is an approximate alternative to the explicit enumeration procedure. However, as we now demonstrate, neither the enumeration procedure nor the randomization procedure is necessary, because easily calculated, closed-form estimators are available for $\tau(h)$ at $h < H$, together with asymptotic confidence intervals.

For interpolation, there is an unbiased estimator $\tilde{\tau}(h)$ for $\tau(h)$ that is based on the counts s_j, appropriately weighted by combinatorial coefficients. Recalling that $S_{\mathrm{obs}} = \Sigma_{j=1}^{H}s_j$, then

$$\begin{aligned}\tilde{\tau}(h) &= \sum_{j=1}^{H} (1 - \alpha_{jh})s_j \\ &= S_{\mathrm{obs}} - \sum_{j=1}^{H} \alpha_{jh}s_j \qquad h = 1, 2, \ldots, H \qquad (5)\end{aligned}$$

where the combinatorial coefficients α_{jh} are defined by

$$\alpha_{jh} = \begin{cases} \dfrac{(H-h)!(H-j)!}{(H-h-j)!H!} & \text{for } (j + h \le H) \\ 0 & \text{for } (j + h > H). \end{cases}$$

Note that $\alpha_{jh} = \alpha_{hj}$. Because the coefficient α_{jh} in Eq. 5 is 0 for $h = H$, estimated richness for the full empirical sample set $\tilde{\tau}(H) = S_{\mathrm{obs}}$. We consider the observed richness S_{obs} to be measured with error. This approach is critical to the derivation of an unconditioned variance estimator for $\tau(h)$ at $h < H$.

Because $\tilde{\tau}(h)$ is derived by estimating moments (Mao et al. 2004), we refer to it as the *moment-based* estimator of species richness $\tau(h)$. It is the best estimator in the sense that $\tilde{\tau}(h)/S$ achieves the minimum variance among all unbiased estimators for $\tau(h)/S$. The moment-based estimator $\tilde{\tau}(h)$ can be approximated by a normal random variable with mean $\tau(h)$ and variance $\sigma^2(h)$ (Mao et al. 2004). Therefore one can construct approximate 95% confidence intervals $\tilde{\tau}(h) \pm 1.96\,\tilde{\sigma}(h)$ for $\tau(h)$ with

$$\tilde{\sigma}^2(h) = \sum_{j=1}^{H} (1 - \alpha_{jh})^2 s_j - \tilde{\tau}^2(h)/\tilde{S} \qquad (6)$$

where $\tilde{S}$ is an estimator for the unknown total species

TABLE 2. Sample-based rarefaction. The table demonstrates that the observed counts s_1, s_2 . . . s_H are sufficient statistics for sample-based rarefaction.

Example a. Dissociation between species, low richness variation among samples.

Abundance

Quadrats

	A	B	C	D	Σ
Sp 1	6	0	3	0	9
Sp 2	0	1	0	5	6
Sp 3	1	0	0	0	1
Sp 4	0	4	0	0	4
Sp 5	0	0	2	0	2
Sp 6	0	0	0	3	3
Σ	□	□	□	□	

Incidence

Quadrats

	A	B	C	D	Σ
Sp 1	1	0	1	0	2
Sp 2	0	1	0	1	2
Sp 3	1	0	0	0	1
Sp 4	0	1	0	0	1
Sp 5	0	0	1	0	1
Sp 6	0	0	0	1	1
Σ	2	2	2	2	

Sample-based rarefaction

***h* (number quadrats pooled)**

	1		2		3		4	
	A	2	**A+B**	4	**A+B+C**	5	**A+B+C+D**	6
	B	2	**A+C**	3	**A+B+D**	5		
	C	2	**A+D**	4	**A+C+D**	5		
	D	2	**B+C**	4	**B+C+D**	5		
			B+D	3				
			C+D	4				
Mean richness:		2.00		3.75		5.00		6.00

Example b. Association between species, high richness variation among samples.

Abundance

Quadrats

	A	B	C	D	Σ
Sp 1	4	0	5	0	9
Sp 2	3	0	3	0	6
Sp 3	1	0	0	0	1
Sp 4	4	0	0	0	4
Sp 5	0	0	2	0	2
Sp 6	0	0	3	0	3
Σ	12	0	13	0	

Incidence

Quadrats

	A	B	C	D	Σ
Sp 1	1	0	1	0	2
Sp 2	1	0	1	0	2
Sp 3	1	0	0	0	1
Sp 4	1	0	0	0	1
Sp 5	0	0	1	0	1
Sp 6	0	0	1	0	1
Σ	4	0	4	0	

Sample-based Rarefaction

***h* (number quadrats pooled)**

	1		2		3		4	
	A	4	**A+B**	4	**A+B+C**	6	**A+B+C+D**	6
	B	0	**A+C**	6	**A+B+D**	4		
	C	4	**A+D**	4	**A+C+D**	6		
	D	0	**B+C**	4	**B+C+D**	4		
			B+D	0				
			C+D	4				
Mean richness:		2.00		3.75		5.00		6.00

Notes: Because the two contrasting examples share the same counts ($s_1 = 4$ and $s_2 = 2$) for both examples, they yield the same sample-based rarefaction curve (Fig. 5). Any pattern of incidence that produces the same counts will produce the same pattern of mean richness because of combinatorial averaging. Similarly, the individual-based rarefaction curves for the two examples are identical to each other, despite differences in abundance and incidence patterns, because they share the same relative abundance vector (9, 6, 1, 4, 2). Fig. 5 plots the individual-based and sample-based rarefaction curves for the examples based (respectively) on the Abundance and Incidence matrices.

richness S. Bunge and Fitzpatrick (1993) and Colwell and Coddington (1994) review (and EstimateS [Colwell 1994–2004] computes) various richness estimators. A form of the ''Chao2'' richness estimator (Chao 1989, Colwell 1994–2004, Colwell and Coddington 1994, Mao and Lindsay 2003) is a simple option:

$$\tilde{S} = S_{\text{obs}} + \frac{(H - 1)s_1^2}{2Hs_2} \qquad (7)$$

where s_1 is the number of species that occur in a single quadrat and s_2 is the number of species that occur in exactly two quadrats. A highly conservative approach to estimating $\sigma^2(h)$ is to set $\tilde{S} = \infty$, so that the second term in Eq. 6 becomes negligible.

Ugland et al. (2003) independently arrived at a combinatorial interpolation estimator that is mathematically equivalent to Eq. 5, but they derived this result as the expectation of $\tilde{\tau}(h)$ conditioned on the empirical sample set. They also present a variance estimator for $\tilde{\tau}(h)$, using an entirely different approach than Eq. 6,

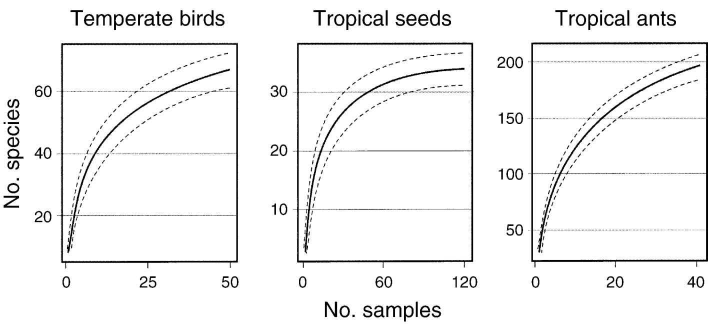

FIG. 1. Sample-based rarefaction (interpolated species accumulation) curves for three empirical data sets from Table 1. Expected species richness values (solid lines) were computed using the moment-based estimator of Eq. 5 with 95% confidence intervals (dashed lines) based on Eqs. 6 and 7.

but because their estimator is the conditional variance, it cannot properly be used to construct confidence intervals.

To illustrate interpolation (sample-based rarefaction) using Eqs. 5 and 6, Fig. 1 plots estimated richness with approximate 95% confidence bands for the bird, seed bank, and ant data sets of Table 1.

Random-placement theory (Coleman 1981, Brewer and Williamson 1994, Colwell and Coddington 1994) might seem to be an alternative approach to estimation of interpolated richness $\tau(h)$, although, to our knowledge, random-placement theory has not been applied previously to incidence data. A Coleman-type, random-placement estimator for incidence data is

$$\tilde{\tau}_*(h) = S_{\text{obs}} - \sum_{j=1}^{H} s_j(1 - h/H)^j. \quad (8)$$

However, the estimator $\tilde{\tau}_*(h)$ is biased. The difference between $\tilde{\tau}(h)$ and $\tilde{\tau}_*(h)$ becomes

$$\tilde{\tau}(h) - \tilde{\tau}_*(h) = \sum_{j=1}^{H} s_j[(1 - h/H)^j - \alpha_{jh}]. \quad (9)$$

It can be shown that $\alpha_{jh} < (1 - h/H)^j$ so that $\tilde{\tau}_*(h) < \tilde{\tau}(h)$, although the difference $\tilde{\tau}(h) - \tilde{\tau}_*(h)$ may often be small. In addition, the variance estimators of Coleman (1981) are conditional, in the sense that uncertainty of sampling is not taken into account (see also Smith and Grassle 1977). Fig. 2 presents the random-placement estimates $\tilde{\tau}_*(h)$ and the differences $\tilde{\tau}(h) - \tilde{\tau}_*(h)$ calculated from the ant data set. The differences are substantial for small h, and can be quite large for particular data sets with high outliers in the incidence counts (Mao et al. 2004).

At the beginning of the previous section (*The Model*), we made two simplifying statistical assumptions, which it is now time to reexamine: (1) the ith species has the same probability ϕ_i of being present in each quadrat, and (2) the Z_{ij} are independent over all i and j. By means of simple but definitive examples, we now show that sample-based rarefaction by Eq. 5 is robust to these assumptions. Table 2 shows two hypothetical examples of empirical sample sets. In each example, six species are distributed among four quadrats. The two examples share the same distribution of counts: in both cases, $s_1 = 4$ and $s_2 = 2$ (four species occur in only one quadrat each, and two species occurs in precisely two quadrats), whereas $s_j = 0$ for all other $j > 0$. Thus the two examples must yield identical sample-based rarefaction curves by Eq. 5, which depends only on the counts s_j. (Ignore the ''Abundance'' matrices in Table 2 for now. They become pertinent in the *Discussion*.)

Now examine the incidence patterns in the two examples, which together bracket the range of possibilities. In Example *a*, the six species are nonrandomly dissociated (they co-occur in the minimum number of quadrats), and there is no variation in total incidence (column totals) among quadrats. In contrast, in Example *b* the six species are maximally associated (they always occur together) with a patchy distribution of overall occurrence (high variation in total incidence among quadrats). Despite these extreme patterns, the mean number of species is identical among all possible combinations of quadrats for $h = 1 \ldots 4$, as shown in the computations on the right side of Table 1. (The corresponding sample-based rarefaction curve is later plotted in Fig. 5.) Clearly, sample-based rarefaction curves, in themselves, are blind to nonrandom occurrence probabilities among quadrats and to non-independence of occurrence among species because of the combinatorial averaging in Eq. 5, which is laid out explicitly on the right side of Table 2. From the statistical point of view, it is necessary to require that the quadrats in the empirical sample set are true ''representatives'' of the set of all possible quadrats. Therefore, the presence probability ϕ_i is understood to be the average presence probability across different quadrats for the ith species.

COMPARISON OF SAMPLE-BASED RAREFACTION CURVES

Now that we are able to estimate rigorous confidence intervals for sample-based rarefaction curves (Fig. 1), the comparison of two or more such curves for different sample sets at comparable sampling effort becomes straightforward. For example, Chazdon and colleagues (R. L. Chazdon, A. Redondo-Brenes, and B. Vilchez-Alvarado, *unpublished data*) sampled old-growth and second-growth rain forest in Costa Rica (Table 1), identifying all stems >1 cm dbh in 100 quadrats (each 10

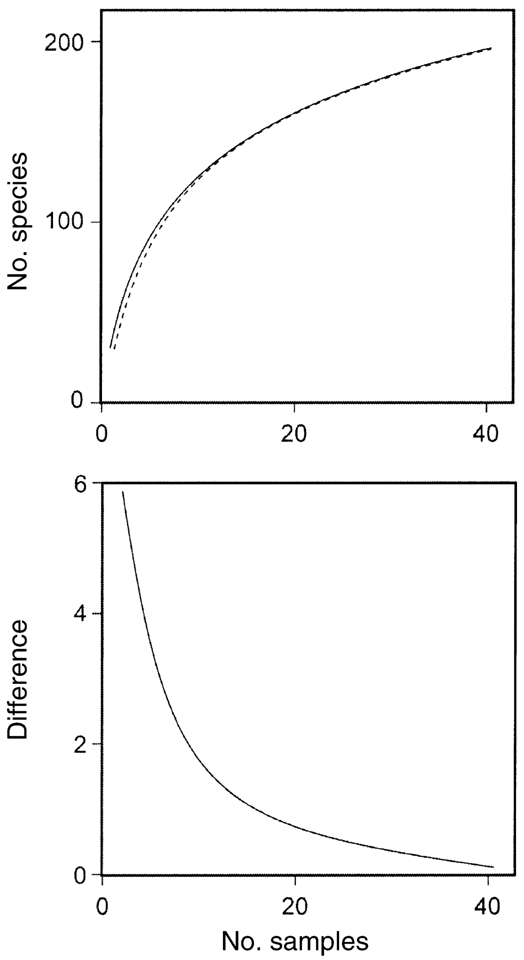

FIG. 2. Richness estimates for the ant data set (Table 1), comparing the moment-based estimator (Eq. 5) with the Coleman-type incidence estimator (Eq. 8). The top plot shows the moment-based estimates $\tilde{\tau}(h)$ (solid line) and Coleman estimates $\tilde{\tau}_*(h)$ (dashed line) as a function of the number of quadrats h. The bottom plot shows the difference $\tilde{\tau}(h) - \tilde{\tau}_*(h)$ as a function of h.

m on a side, arranged on a 50 × 250 m grid) for each forest type. Both graphs in Fig. 3 show the sample-based rarefaction curves with 95% confidence intervals for larger saplings, 2.5–5.0 cm dbh. In upper graph, the x-axis is scaled by accumulated quadrats, whereas the curves in the lower graph are scaled by the number of individual stems accumulated as quadrats are pooled. The two graphs differ because the average density of saplings is considerably greater in second growth (5.1 stems/quadrat) than in old growth (1.8 stems/quadrat), where larger trees predominate.

The upper graph compares species density between the two stands, whereas the lower graph compares species richness (Gotelli and Colwell 2001). In the upper graph, although species density estimates for the old-growth stand are higher than for the second-growth stand at all quadrat accumulation levels (all h), the differences are clearly not significant at $P < 0.05$ because the confidence intervals overlap. When the curves are rescaled by individuals in the lower graph, the difference (in species richness) becomes strongly significant. As discussed in detail by Gotelli and Colwell (2001, and references therein), estimation of species richness (as opposed to species density) often requires rescaling sample-based rarefaction curves by individuals, to adjust for differing densities of individuals.

EXTRAPOLATION

It is often desirable to estimate the number of species that would be found (or would have been found) by taking further samples from an assemblage. Many uses of extrapolation are possible, including informed future allocation of limited time and resources, analysis of historical data for which no further sampling is possible, or the need to statistically ''enlarge'' smaller sample sets for comparison with larger ones at a com-

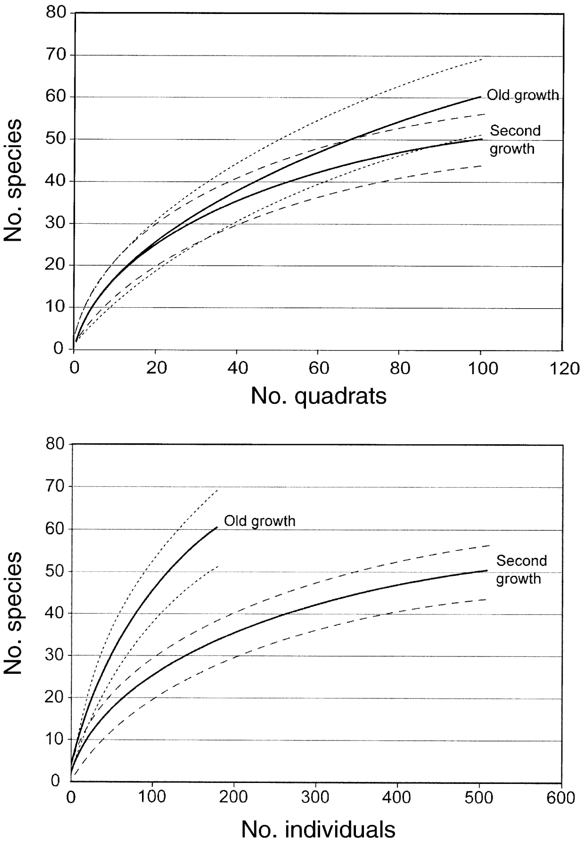

FIG. 3. Comparing species richness between two data sets. The two graphs show the same data for tropical rain forest saplings (Table 1) in old growth (upper solid line with dotted-line 95% confidence interval) vs. second growth (lower solid line with dashed-line 95% confidence interval). The upper graph compares species density (because the x-axis is scaled by quadrats), which does not differ significantly between the two forest stands. The lower graph compares species richness (because the x-axis is scaled by individuals), which differs significantly between stands.

parable level of sampling effort. In terms of our general model, extrapolation involves estimating $\tau(h)$ for $h > H$, where H is the number of quadrats (or other samples) in the empirical data set. The objective then becomes the estimation of the number of additional species, $\tau(h) - \tau(H)$, that would be expected to be found in the additional $h - H$ quadrats. The presentation here is simplified from Mao et al. (2004), where the full mathematical development of extrapolation techniques appears.

Perhaps not surprisingly, the statistical properties of $\tau(h)$ at $h > H$ and $\tau(\infty) = S$ are different from those of $\tau(h)$ at $h \leq H$. Whereas the observed incidence (presence–absence) data provide sufficient information for us to estimate $\tau(h)$ at $h \leq H$ using a simple, moment-based estimator (Eq. 5) with no restrictions on the number of homogeneous incidence groups G, no such simple estimator is available for $h > H$. A likelihood-based method can be developed, however, under the additional constraint that $G \leq H/2$ (Mao et al. 2004). (In practice, this constraint is unlikely to pose a problem, as the empirical examples demonstrate later in this section.)

Our strategy is to develop a function $\theta(h)$ that expresses the expected proportional difference in richness between H quadrats and h quadrats, such that

$$\tau(h) = \tau(H)\theta(h) \tag{10}$$

and, asymptotically,

$$\tau(\infty) = \tau(H)\theta(\infty). \tag{11}$$

The model that we develop applies not only to extrapolation ($h > H$; $\theta(h) > 1$), but also to interpolation ($h < H$; $\theta(h) < 1$), for which the likelihood-based method provides an alternative to the moment-based method described in the *Interpolation* section.

Recall that π_k is the relative group size of the kth incidence group and ψ_k is the common *presence probability* of the species within the kth group. Working from Eq. 2, we define the *mixing weight* for the kth group, ω_k, as

$$\omega_k = \frac{\pi_k[1 - (1 - \psi_k)^H]}{\sum_{m=1}^{G} \pi_m[1 - (1 - \psi_m)^H]} \qquad k = 1, 2, \ldots, G. \tag{12}$$

Now we can specify the desired function $\theta(h)$ as the weighted sum of $k = 1, 2, \ldots, G$ terms as follows:

$$\theta(h) = 1 + \sum_{k=1}^{G} \omega_k \frac{(1 - \psi_k)^H - (1 - \psi_k)^h}{1 - (1 - \psi_k)^H}. \tag{13}$$

As h gets very large, the expression $(1 - \psi_k)^h$ approaches zero, so that

$$\theta(\infty) = 1 + \sum_{k=1}^{G} \omega_k \frac{(1 - \psi_k)^H}{1 - (1 - \psi_k)^H}. \tag{14}$$

Because the true richness for H quadrats, $\tau(H)$, can be estimated by S_{obs}, we need only to estimate $\theta(h)$ and $\theta(\infty)$ or to estimate the parameters ψ_k and ω_k (presence probabilities and mixing weights) used to define $\theta(h)$ and $\theta(\infty)$. To do this, we seek to maximize the log conditional likelihood

$$L = l(\{\omega_k, \psi_k\}_{k=1}^{G}) \tag{15}$$

of the empirical counts $s_1, s_2 \ldots s_H$, given the observed richness S_{obs}. The methods that we recommend to achieve this objective are beyond the scope of this paper, but appear in full in Mao et al. (2004). Here we briefly outline the approach; we then apply it to empirical data sets from Table 1.

Given a particular number of incidence groups G, an expectation maximization (EM) algorithm can be used to maximize the log likelihood L (Eq. 15), yielding a set of estimators for ψ_k and ω_k ($k = 1, 2, \ldots, G$) that are fitted specifically for G groups (Dempster et al. 1977). We begin with $G = 1$, then continue to evaluate the goodness of fit for successive trials with $G = 1,2 \ldots$, using the gradient plot method of Lindsay and Roeder (1992) at each step to assess whether incrementing G will result in any improvement in fit, and the EM algorithm to produce new sets of estimates at each step for the ψ_k and ω_k. A larger number of groups G may increase the log likelihood, but a larger G means that more parameters are used to achieve the improved fit, because the number of independent parameters for the likelihood in Eq. 15 is $2G - 1$. To strike a compromise between goodness of fit and estimation of fewer parameters, we choose the number of groups G that minimizes the AIC (Akaike Information Criterion):

$$\text{AIC}(L_G) = 2G - 1 - 2l(\{\omega_{k,G}, \psi_{k,G}\}_{k=1}^{G}). \tag{16}$$

Once the best estimates for ψ_k and ω_k have been found by this iterative procedure, they are used with Eq. 13 to compute the estimator $\hat{\theta}(h)$, which is then applied to estimate richness for h quadrats, $\hat{\tau}(h)$, by Eq. 10. The same estimates for ψ_k and ω_k can be inserted in Eq. 14 to estimate $\hat{\theta}(\infty)$, producing an estimate of the asymptotic richness $\hat{\tau}(\infty)$ by Eq. 11. Confidence intervals for $\hat{\tau}(h)$ are estimated by taking B bootstrap resamples from the likelihood of the counts $s_1, s_2 \ldots s_H$, given a random $\tilde{S}_{\text{obs}}$ produced as a binomial random variable with size $\hat{\tau}(\infty)$ and probability $1/(1 + \hat{\theta}(\infty))$. For each resample, richness $\hat{\tau}(h)$ (Eq. 10) is computed, the $\hat{\tau}(h)$ estimates are ranked, and the values ranked $0.025B$ and $0.975B$ are recorded as the 95% confidence intervals (Mao et al. 2004). Note that this approach differs fundamentally from the methods of Norris and Pollock (1996), both in theory and in computation. Norris and Pollock's approach incurs such a massive computational load that it is not a practical method for constructing confidence intervals.

Fig. 4 shows results for richness extrapolation to three times the empirical number of quadrats ($3H$) for empirical data sets of Table 1. The richness estimates and confidence intervals in Fig. 3 for both interpolation

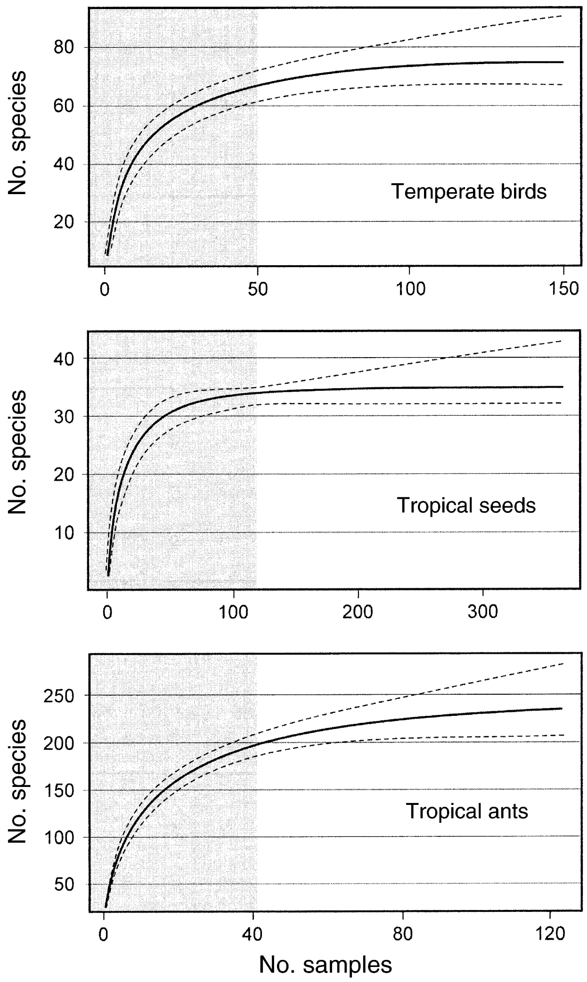

FIG. 4. Extrapolation of species accumulation curves for three empirical data sets (see Table 1) to three times the empirical sample size. Expected species richness values (solid lines) were computed using the likelihood-based estimator of Eq. 10 with 95% bootstrap confidence intervals (for interpolation as well as extrapolation; dashed lines). The shaded area indicates the number of samples in the empirical data set (H).

($h < H$) and extrapolation ($h > H$) were produced using the likelihood-based procedure described previously, with $B = 1000$ for the bootstrap confidence intervals. Table 3 shows the fitted parameters ψ_k and ω_k (presence probabilities and mixing weights) and optimal number of incidence groups G (guided by the AIC). Notice that, for all three empirical examples, G is tiny compared with the number of quadrats H; the restriction that G must be less than $H/2$ is not likely to be a problem for any reasonable level of sampling intensity. Because species are added more and more slowly as H gets larger, it is expected that the optimal value of G will increase much more slowly than H, as well.

We also attempted to apply the method to estimating asymptotic richness S by Eqs. 14 and 11. Unfortunately, but not surprisingly, extrapolation becomes more and more difficult as h gets larger and larger. Note that the confidence intervals become wider and wider as h increases beyond H. Because the likelihood-based estimate for $\tau(\infty)$ often is not reliable (Mao et al. 2004), we recommend limiting the use of extrapolation for extending (say, doubling or tripling) the number of samples in empirical data sets, as demonstrated in Fig. 3.

DISCUSSION

The model introduced in this paper provides a unified theoretical framework for conceptualizing and analyzing species richness in the context of repeated-incidence (presence–absence or occurrence) samples from biological communities. For a given sampling scheme, incidence patterns in samples from natural communities are affected by at least three major sources of heterogeneity. The first and most obvious source is variation among species in overall commonness and rarity (relative abundance), which translates, in general, into variation among species in frequency of occurrence. The second source of heterogeneity is variation among samples in the total abundance of individuals (spatial or temporal aggregation that is concordant among species), which translates into variation among samples in the total number of species occurrences. The third source of heterogeneity is association or dissociation between species, among samples, which translates into nonrandom patterns of co-occurrence of species. The second and third kinds of heterogeneity are often difficult to separate; taken together, they represent what is generally characterized as ''patchiness'' among samples in space or time.

TABLE 3. Incidence groups, presence probabilities, and mixing weights for likelihood-based extrapolation of empirical data sets.

Group (k)	Presence probability ψ_k	Mixing weight ω_k
Temperate birds, $G = 4$		
1	0.0300	0.4589
2	0.1328	0.2787
3	0.2991	0.1401
4	0.5038	0.1224
Tropical seedbank, $G = 4$		
1	0.0195	0.2847
2	0.0633	0.4792
3	0.1773	0.0890
4	0.4066	0.1471
Tropical ants, $G = 5$		
1	0.0252	0.3904
2	0.0908	0.2874
3	0.2893	0.1849
4	0.5465	0.1218
5	0.8584	0.0155

Notes: The tabled values were computed for the extrapolation of the species accumulation curve to three times the empirical sample size (3H), fitted by the likelihood-based method outlined in the text (*Extrapolation*) with the number of incidence groups (G) optimized by AIC. The extrapolated species accumulation curves appear in Fig. 4. The data sets are described in Table 1.

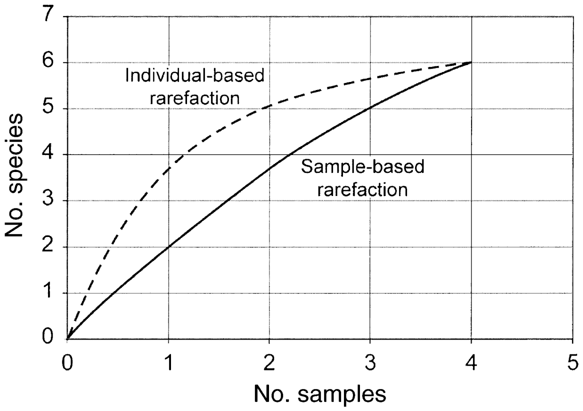

FIG. 5. Individual-based (dashed-line) vs. sample-based (solid-line) rarefaction curves for the hypothetical data of Table 2. The sample-based curve was computed using the moment-based estimator (Eq. 5, or the identical values listed in Table 2). The individual-based curve was computed using the classical rarefaction estimator of Hurlbert (1972).

Individual-based rarefaction models account, explicitly, for the relative abundance of species (Hurlbert 1972). Our incidence-based model allows arbitrary levels of heterogeneity among species in overall occurrence (relative abundance) by treating species occurrences as arising from a mixture of binomial distributions. In this mixture model, each species is assumed to have its own, species-specific presence probability and thus is assumed to follow its own binomial distribution in terms of presences and absences among samples. (This is the least demanding assumption that one might possibly make for incidence data.) In effect, the species-specific binomial distributions are then ''classified'' (by the mixture-fitting algorithm) into groups of approximately homogeneous presence probabilities. The full model is then a mixture of binomial distributions, each weighted by the number of species in its corresponding group (Mao et al. 2004).

Individual-based and sample-based rarefaction make crucially different assumptions about patchiness, which can best be understood by comparing the two approaches for the same data set. Given an empirical abundance matrix, such as the hypothetical examples in Table 2 (left side of each example), the vector of row (species) totals (9, 6, 1, 4, 2, 3 in Table 2) may be used to produce an individual-based rarefaction curve for the sample set. The corresponding incidence matrix (middle of each example in Table 2) may be used to produce a sample-based rarefaction curve (computed on the right side of each example) for the same sample set.

Fig. 5 shows both kinds of rarefaction curves, based on the hypothetical examples in Table 2. When both kinds of rarefaction curves are scaled by the number of pooled samples, the two curves will be identical only if *individuals* of all species occur randomly and independently among the samples in the same sample set. If individuals tend to be (nonrandomly) aggregated among samples (within species), the sample-based rarefaction curve must lie below the abundance-based rarefaction curve (as in Fig. 5) (Coleman 1981, Colwell and Coddington 1994, Gotelli and Colwell 2001). This happens because aggregation of individuals produces an incidence matrix with fewer presences and more absences than a random distribution of the same number of individuals among samples, so that more samples must be pooled in the sample-based curve than in the individual-based curve to reach a given level of richness. To see this, imagine throwing 30 balls (individuals of a single species) at random into 10 boxes, numbered 1–10. Some boxes may be empty, but probably not half of them. Now take all the balls from even-numbered boxes and spread them out among the odd-numbered boxes. The balls are now statistically aggregated, and there are fewer presences and more absences (empty boxes) than before.

Because individual-based rarefaction ignores such patchiness, it thus generally overestimates expected richness for rarefied samples (Fig. 5). Notice that the pattern of species abundance (row totals in the Abundance matrices) in Table 2 is contrived to be identical for the two examples, whereas the actual distribution of individuals among quadrats differs substantially. The individual-based rarefaction curves (Fig. 5) for the two examples are nonetheless identical.

In contrast, sample-based rarefaction curves implicitly reflect empirical levels of within-species aggregation of individuals by considering only incidence, thus providing a realistic estimate of the number of species to be found in sets of real-world samples (Colwell and Coddington 1994, Chazdon et al. 1998, Gotelli and Colwell 2001, Ugland et al. 2003). Suppose that a study area is divided into 1000 quadrats. We sample 50 quadrats at random and count the number of species found in each. A particular species may be in some sampled quadrats and not in others, and its individuals may be nonrandomly aggregated among quadrats. All we require is that the 50 units are truly randomly chosen from the 1000 units, so that the empirically estimated presence probability of a species in these 50 units is close to the true presence probability for that species over the whole 1000 units, for the specified quadrat size and empirical level of individual aggregation. Because aggregated spatial (and temporal) distributions are extremely common, this property of sample-based rarefaction is quite a general one.

The fundamental statistics for estimators based on the model are the counts, or occurrence frequencies, of species over a set of samples. Because the sample-based rarefaction curve depends only on expected (mean) incidence patterns, it can be precisely modeled with these counts for empirical data sets with any amount of patchiness, as demonstrated by the examples in Table 2 and Fig. 5. Some samples may have a high number of occurrences and others may have low numbers (the second source of heterogeneity), but because of combinatorial averaging, there is no effect on the mean rarefaction curve. For the same reason, association or dissociation of species occurrences (the third source of heterogeneity) does not affect the mean rarefaction curve, nor is it reflected in the counts.

Using the general model as a framework, we developed estimators for both interpolation (or sample-based rarefaction) between zero and the richness of the full set of samples in an empirical data set, and extrapolation, or projection of richness beyond the data set to predict the expected number of species in a larger number of samples from the same assemblage.

Interpolation has routinely been carried out in the past, for incidence-based data, by randomly subsampling the data set, retaining sample integrity (rather than pooling all samples, and then drawing out occurrences at random) (Colwell 1994–2004, Colwell and Coddington 1994). Until recently (Ugland et al. 2003, Mao et al. 2004), no analytical method existed for precisely estimating the number of species in a subset of samples from an incidence-based data set. In contrast, the corresponding problem for abundance-based samples (classic rarefaction) was largely solved three decades ago (Hurlbert 1971, Heck et al. 1975). Worse, the problem of setting confidence intervals around sample-based accumulation curves until now has remained entirely unresolved, severely limiting the comparison of curves for different communities or treatments. Our moment-based richness estimator (Eq. 5) for the interpolation problem, with its variance estimator (Eq. 6), solves both of these challenges rigorously, providing precisely the expectation that randomizing sample order produces, with legitimate confidence intervals at all points along the curve. As shown in Fig. 3, these confidence intervals make rigorous comparison between sample-based rarefaction curves possible at last. The moment-based estimator for sample-based rarefaction, with confidence intervals, is included in EstimateS Version 7 (Colwell 1994–2004).

Using the same theoretical framework, we derived the random-placement (Coleman) function for incidence data (Eq. 8), which seems not to have been previously examined. The Coleman curve (Coleman 1981) until now has been applied only to the case of abundance-based (quantitative) samples, for which it approximates the expected curve for individual-based rarefaction (Brewer and Williamson 1994, Colwell and Coddington 1994). Likewise, we found that the incidence-based Coleman curve approximates the true species accumulation (sample-based rarefaction) curve for incidence samples. However, the logic underlying the incidence-based Coleman curve makes sense only for interpolation, the estimator is biased (substantially so if there happen to be any highly dominant species, as indicated by unusually high j for the incidence counts s_j), and available variance estimators are not appropriate for constructing confidence intervals. For these reasons, we prefer the moment-based estimator to the incidence-based Coleman curve for interpolation.

Extrapolation of species accumulation curves previously has been attempted only by fitting functions, such as the asymptotic Michaelis-Menten or various non-asymptotic functions (Soberón and Llorente 1993, Colwell and Coddington 1994). In this paper, for the extrapolation problem, we developed a likelihood-based method that relies on fitting the distribution of the observed counts for the binomial mixture model. The number of incidence groups required for the model is optimized using the Akaike Information Criterion, as a balance between the goodness of fit and the complexity of the model (the number of parameters). Bootstrap confidence intervals can also be computed, as outlined in the *Extrapolation* section. The computations and algorithms for both the expected richness and its bootstrap confidence interval require sophisticated and complex computations, but software for the purpose is available from C. X. Mao.

Using the same likelihood-based method of extrapolation, one can in theory estimate the number of additional species that an infinitely large set of samples from the same assemblage would be expected to yield: the asymptote of the species accumulation curve (Eqs. 11 and 14). In practice, we conclude that our likelihood-based extrapolation method is quite useful for estimation problems that assume doubling or tripling the empirical number of samples. Unfortunately, in its present form, the method does not appear to be a reliable way to estimate asymptotic richness (Mao et al. 2004), but we hope that our efforts may inspire further work on this important but daunting problem.

The likelihood-based method simultaneously models the interpolation problem, and the same bootstrap technique can be used for interpolated richness estimates (sample-based rarefaction), as illustrated in Fig. 4. However, the much simpler and more intuitive moment-based estimator (Eq. 5) and associated confidence intervals (based on Eq. 7) perform just as well (Fig. 1), and are therefore preferred.

Acknowledgments

We are grateful to F. He and two anonymous reviewers for their comments, and to A. M. Ellison for his encouragement. We thank R. L. Chazdon for sharing her unpublished data. This work was supported by US-NSF grant DEB-0072702 to R. K. Colwell.

Literature Cited

Brewer, A., and M. Williamson. 1994. A new relationship for rarefaction. Biodiversity and Conservation **3**:373–379.

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

Butler, B. J., and R. L. Chazdon. 1998. Species richness, spatial variation, and abundance of the soil seed bank of a secondary tropical rain forest. Biotropica **30**:214–222.

Chao, A. 1989. Estimating population size for sparse data in capture–recapture experiments. Biometrics **45**:427–438.

Chazdon, R. L., R. K. Colwell, J. S. Denslow, and M. R. Guariguata. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. Pages 285–309 *in* F. Dallmeier and J. A. Comiskey, editors. Forest biodiversity research, monitoring and modeling: conceptual background and Old World case studies. Parthenon Publishing, Paris, France.

Coleman, B. D. 1981. On random placement and species–area relations. Mathematical Biosciences **54**:191–215.

Colwell, R. K. 1994–2004. EstimateS: statistical estimation of species richness and shared species from samples. ⟨http://viceroy.eeb.uconn.edu/estimates⟩. [Persistent URL: ⟨http://purl.oclc.org/estimates⟩.]

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

Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B **39**: 1–22.

Dorazio, R. M., and J. A. Royle. 2003. Mixture models for estimating the size of a closed population when capture rates vary among individuals. Biometrics **59**:351–364.

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

Heck, K. L., Jr., G. van Belle, and D. Simberloff. 1975. Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. Ecology **56**:1459–1461.

Holdridge, L. R., W. G. Grenke, W. H. Hatheway, T. Liang, and J. A. Tosi. 1971. Forest environments in tropical life zones. Pergamon Press, Oxford, UK.

Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. Ecology **52**:577–586.

Lindsay, B. G., and K. Roeder. 1992. Residual diagnostics for mixture models. Journal of the American Statistical Association **87**:785–794.

Longino, J., R. K. Colwell, and J. A. Coddington. 2002. The ant fauna of a tropical rainforest: estimating species richness three different ways. Ecology **83**:689–702.

Mao, C. X., R. K. Colwell, and J. Chang. 2004. Estimating species accumulation curves using mixtures. Technical report, Department of Statistics, University of California, Riverside, California, USA.

Mao, C. X., and B. G. Lindsay. 2003. Estimating the population size: heterogeneity, nonidentifiability and regularization. Technical Report, University of California, Riverside, California, USA.

Norris, J. L., and K. H. Pollock. 1996. Nonparameteric MLE under two closed capture–recapture models with heterogeneity. Biometrics **52**:639–649.

Rosenzweig, M. L. 1995. Species diversity in space and time. Cambridge University Press, Cambridge, UK.

Scheiner, S. M. 2003. Six types of species–area curves. Global Ecology and Biogeography **12**:441–447.

Simberloff, D. 1972. Properties of the rarefaction diversity measurement. American Naturalist **106**:414–418.

Smith, W., and J. Grassle. 1977. Sampling properties of a family of diversity measures. Biometrics **33**:283–292.

Soberón, M. J., and B. J. Llorente. 1993. The use of species accumulation functions for the prediction of species richness. Conservation Biology **7**:480–488.

Ugland, K. I., J. S. Gray, and K. E. Ellingsen. 2003. The species-accumulation curve and estimation of species richness. Journal of Animal Ecology **72**:888–897.