Journal of
Plant Ecology

VOLUME 5, NUMBER 1,
PAGES 3–21

MARCH 2012

doi: 10.1093/jpe/rtr044

available online at
www.jpe.oxfordjournals.org

Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages

Robert K. Colwell[1]*, Anne Chao[2], Nicholas J. Gotelli[3], Shang-Yi Lin[2], Chang Xuan Mao[4], Robin L. Chazdon[1] and John T. Longino[5]

[1] Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA
[2] Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan
[3] Department of Biology, University of Vermont, Burlington, VT 05405, USA
[4] School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China
[5] Department of Biology, University of Utah, Salt Lake City, UT 84112, USA
*Correspondence address. Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT 06269, USA. E-mail: colwell@uconn.edu

Abstract

Aims

In ecology and conservation biology, the number of species counted in a biodiversity study is a key metric but is usually a biased underestimate of total species richness because many rare species are not detected. Moreover, comparing species richness among sites or samples is a statistical challenge because the observed number of species is sensitive to the number of individuals counted or the area sampled. For individual-based data, we treat a single, empirical sample of species abundances from an investigator-de ned species assemblage or community as a reference point for two es- numbers of sampling units.

Methods

The rst objective is a problem in interpolation that we address with classical rarefaction (multinomial model) and Coleman rarefaction (Poisson model) for individual-based data and with sample-based rarefaction (Bernoulli product model) for incidence frequencies. The second is a problem in extrapolation that we address with sampling-theoretic predictors for the number of species in a larger sample (multinomial model), a larger area (Poisson model) or a larger number of sampling units (Bernoulli product model), based on an estimate of asymptotic species richness. Although published methods exist for many of these objectives, we bring them together here with some new estimators under a uni ed statistical and notational framework. This novel integration of mathematically distinct approaches allowed us to link interpolated (rarefaction) curves and extrapolated curves to plot a uni ed species accumulation curve for empirical examples. We provide new, unconditional variance estimators for classical, individual-based rarefaction and for Coleman rarefaction, long missing from the toolkit of biodiversity measurement. We illustrate these methods with datasets for tropical beetles, tropical trees and tropical ants.

Important Findings

Surprisingly, for all datasets we examined, the interpolation (rarefaction) curve and the extrapolation curve meet smoothly at the reference sample, yielding a single curve. Moreover, curves representing 95% con dence intervals for interpolated and extrapolated richness estimates also meet smoothly, allowing rigorous statistical comparison of samples not only for rarefaction but also for extrapolated richness values. The con dence intervals widen as the extrapolation moves further beyond the reference sample, but the method gives reasonable results for extrapolations up to about double or triple the original abundance or area of the reference sample. We found that the multinomial and Poisson models produced indistinguishable results, in units of estimated species, for all estimators and datasets. For sample-based abundance data, which allows the comparison of all three models, the Bernoulli product model generally yields lower richness estimates for rare ed data than either the multinomial or the Poisson models because of the ubiquity of non-random spatial distributions in nature.

Keywords: Bernoulli product model • Coleman curve • multinomial model • Poisson model • random placement • species–area relation

Received: 20 July 2011 Revised: 15 October 2011 Accepted: 17 October 2011

INTRODUCTION

Exhaustive biodiversity surveys are nearly always impractical or impossible (a7s-188(richn(ss1m17 304]235 6ss6664-1ulliF (imC57)lTJsample-b(sa (356)7bl4nci)ence(354i3bldat(,(356)8TJth((355TJ)t

$$r^2_{area}(a) = \sum_{k=1}^{n} \left(1 - \left(1 - \frac{a}{A}\right)^k f_k - \tilde{S}_{area}(a)\right)^2 / S_{est}. \quad (7)$$

Coleman et al. (1982) provide an estimator for the variance of $\tilde{S}_{area}(a)$ conditional on the reference sample. We postpone specification of S_{est} for a later section.

Comparing the multinomial and Poisson models for interpolation

How different are the rarefaction estimates of species richness estimators under the multinomial and the Poisson models? From Equations (4) and (6), the estimates from the two models

$\hat{\text{var}}(\tilde{S}_{\text{area}}(A$

$$\hat{\mathrm{var}}(\tilde{S}_{\mathrm{sample}}(T+t^*)) = \sum_{i=1}^{T}\sum_{j=1}^{T} \frac{\partial \tilde{S}}{\partial Q_i}\frac{\partial \tilde{S}}{\partial Q_j}\,\hat{\mathrm{cov}}\,(Q_i, Q_j) \quad ; \qquad (19)$$

where $\hat{\mathrm{cov}}(Q_i, Q_j) = Q_i[1 - Q_i/(S_{\mathrm{obs}} + \hat{Q}_0)]$ for $i = j$ and $\hat{\mathrm{cov}}(Q_i, Q_j) = -Q_iQ_j/(S_{\mathrm{obs}} + \hat{Q}_0)$ for $i \neq j$. (For simplicity, we write $\tilde{S}$ for $\tilde{S}_{\mathrm{sample}}(T + t^*)$ in the above variance formula.)

Equations (18) and (19), above, both require an estimate of Q_0, the number of species present in the assemblage but not

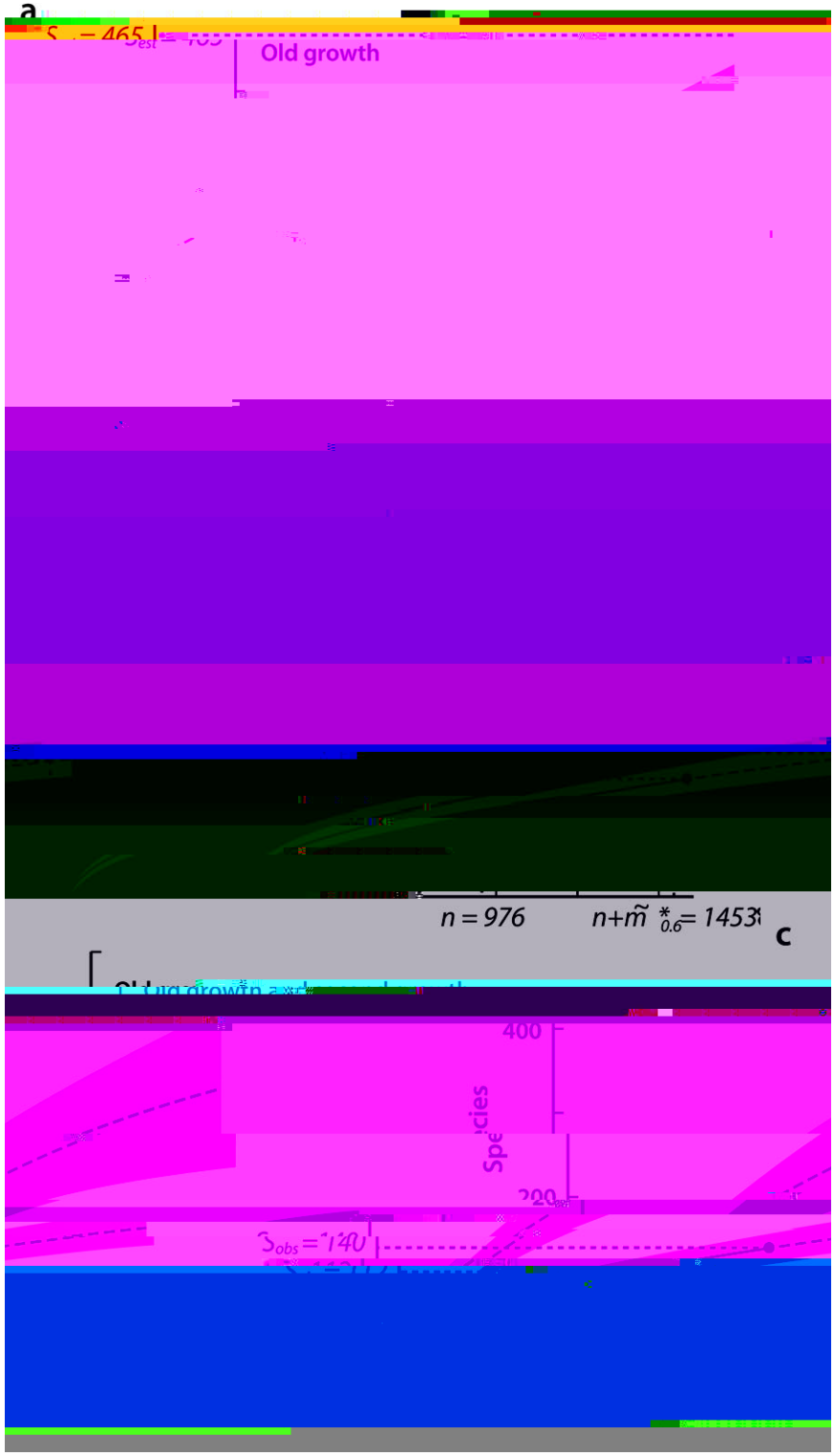

Figure 2: individual-based interpolation (rarefaction) and extrapolation from two reference samples (Þlled black circles) of beetles from southwestern Costa Rica (Janzen 1973a, 1973b), illustrating the computation of estimators from Fig. 1a for the multinomial model, with 95% unconditional conÞdence intervals. (a) Osa old-growth forest sample. (b) Osa second-growth forest sample (c) Comparison of the curves from the samples in (a) and (b). Based on observed richness, S_{obs}, the Osa second-growth assemblage (with 140 species in the reference sample) is richer in species than the Osa second-growth assemblage (with 112 species in the reference sample), but after rarefying the second-growth sample to 237 individuals to match the size of the old-growth sample (open black circle), the second-growth sample has only 70 species. Clearly the old-growth assemblage is richer, based on these samples.

sample (Fig. 2c, open point), using the multinomial model (Equation 4), the ordering of the two sites is reversed. The interpolated species richness for 237 individuals in the second-growth site is only 70, considerably less than primary site, with 112 species. Moreover, the 95% conÞdence intervals do not overlap (Fig. 2c).

Individual-based rarefaction of abundance data, like the interpolation analysis above, has been carried out in this way for decades. Here, we apply individual-based rarefaction and extrapolation to the same reference sample for the Þrst time. Applying the multinomial model (Equation 9) to the Janzen dataset to increase the sample size (number of individuals) in each site yields the extrapolated curves (broken line curves) for each site is shown in Fig. 2. Even though the mathematical derivations for interpolation and extrapolation are fundamentally different, the interpolation and extrapolation curves join smoothly at the single data point of the reference sample.

In Table 2a, using the multinomial model (classical rarefaction), we show for the Osa old-growth data ($S_{obs} = 112$, $n = 237$ in the reference sample): (i) values for the interpolated estimate $\tilde{S}_{ind}(m)$, for values of m from 1 up to the reference sample size of 237 individuals (Equation 4), along with the unconditional standard error (SE, Equation 5) values that are used to construct the 95% conÞdence intervals shown in Fig. 2a and c; (ii) the extrapolated estimate $\tilde{S}_{ind}(n + m^*)$ (Equation 9), where m* ranges from 0 to 1 000 individuals, along with the unconditional SE (Equation 10); and (iii) the number of additional individuals $\tilde{m}^*_g$ required to detect proportion g of the estimated assemblage richness (Equation 11), forg

For both samples, the unconditional variance, and thus the 95% conÞdence interval, increased with sample size. For extrapolation, the SE values are relatively small up to a doubling of the reference sample, signifying quite accurate extrapolation in this range. For the Osa old-growth site (Table 2a; Fig. 2a), the extrapolation is extended to Þve times of the original sample size in order to compare with the Osa second-growth curve. This long-range extrapolation (>3 3 the original sample size) inevitably yields very wide conÞdence intervals. For the Osa second-growth site (Table 2b; Fig. 2b), the extrapolation is extended only to double the reference sample size (not fully shown in Fig. 2b) yielding a quite accurate extrapolated estimate with a narrow conÞdence interval.

Based on Fig. 2, even though the Osa old-growth site extrapolation for large sample sizes exhibits high variance, the old-growth and second-growth conÞdence intervals do not overlap for any sample size considered. This implies that beetle species richness for any sample size is signiÞcantly greater in the old-growth site than that in the second-growth site for sample size up to at least 1 200 individuals.

Tropical beetles: individual-based rarefaction and extrapolation (Poisson model)

In addition to applying estimators based on the multinomial model, we also analysed the Janzen beetle dataset with estimators based on the Poisson model, including Coleman area-based

multinomial model (Equation 1). Moreover, the similarity applies not only to rarefaction (as previously noted by Brewer and Williamson 1994) but also to extrapolation. Figure 3 shows

between 500 and 1 600 individual s, based conservatively on non-overlapping conÞdence intervals. Due to the prevalence of rare species in old-growth tropical forests and widespread dispersal limitation of large-seeded animal-dispersed species, tree speeis richness is slow to recover during secondary succession and may require many decades to reach old-growth levels, ev en under conditions favorable to regeneration.

Tropical ants: sample-based rarefaction and extrapolation for incidence data (Bernoulli product model)

Longino and Colwell (2011) sampled ants at several elevations on the Barva Transect, a 30-km continuous gradient of wet forest on Costa RicaÕs Atlantic slope. For this example, we use results from Þve sites, at 50-, 500-, 1 070-, 1 500- and 2 000-m elevation, to illustrate sample-based rarefaction and extrapolation. The sampling unit consisted of all worker ants extracted from a 1-m^2 forest ßoor plot, applying a method called Ômini-Winkler extractionÕ. Because ants are colonial and the colony is the unit of reproduction, scoring each sampling unit for presence or absence of each species makes more sense than using abundance data (Gotelli et al. 2011). A sample-by-species incidence matrix was therefore produced for each of the Þve sites. The incidence frequency counts for the Þve sites appear in Table 6.

The results for sample-based interpolation and extrapolation from these Þve sites (at Þve elevations), under the Bernoulli product model, appear in Table 7 and Fig. 4b. For each of the Þve samples, Table 7 shows: (i) values for the interpolated estimate $\hat{S}_{sample}(t)$, under the Bernoulli product model (Equation 17), for values of t from 1 up to the reference sample size T for each elevation (T = 599, 230, 150, 200, 200 sampling units), along with the unconditional SE values (Colwell et al. 2004, their Equation 6) that are used to construct the 95% conÞdence intervals shown in Fig. 4b; and (ii) the

Figure 4: (a) individual-based interpolation (rarefaction) and extrapolation from three reference samplestee To4efer-a33220.3 (1-e).7 (aT62.5s97741 (y)ct3127 27mo4503913918614)1m o2

extrapolated estimate $\tilde{S}_{sample}(T + t^*)$, where t* ranges from 401 to 800 sampling units, to extrapolate all elevations to 1 000 sampling units (Equation 18), along with the unconditional SE (Equation 19).

DISCUSSION

In this paper, we developed a uniÞed theoretical and notational framework for modeling and analyzing the effects on observed species richness of the number of individuals sampled or the number of sampling units examined in the context of a single, quantitative, multispecies sample (an abundance reference sample) or a single set of incidence frequencies for species among sampling units (an incidence reference sample). We compared three statistically distinct models, one based on the multinomial distribution, for counts of individuals (Fig. 1a), the second based on the Poisson distribution, for proportional areas (Fig. 1b), and the third based on a Bernoulli product distribution, for incidence frequencies among sampling units (Fig. 1c).

For interpolation to samples smaller than the reference sample, these correspond to classical rarefaction (Hurlbert 1971), Coleman rarefaction (Coleman 1981) and sample-based rarefaction (Colwell et al. 2004). For the Þrst time, we have linked these well-known interpolation approaches with recent sampling-theoretic extrapolation approaches, under both the multinomial model (Shen et al. 2003) and the Poisson model (Chao and Shen 2004), as well as to methods for predicting the number of additional individuals (multinomial model, Chao et al. 2009) or the amount of additional area (Poisson model, Chao and Shen 2004) needed to reach a speciÞed proportion of estimated asymptotic richness. For the Bernoulli product model, we have developed new estimators, using a similar approach, for sample-based extrapolation (Fig. 1c). The fundamental statistics for all these estimators are the abundance frequency counts f_kÑthe number of species each represented by exactly $X_i = k$ individuals in a reference sample (e.g. Tables 1 and 4)Ñfor individual-based models, or the incidence frequency counts Q_kÑthe number of species that occurred in exactly $Y_i = k$ sampling units (e.g. Table 6)Ñfor sample-based models.

This novel integration of mathematically distinct approaches allowed us to link interpolated (rarefaction) curves and extrapolated curves to plot a uniÞed species accumulation curve for empirical examples (Figs 2 and 4). Perhaps the most surprising (and satisfying) result is how smoothly the interpolated and extrapolated moieties of the curve come together at the reference sample, in all examples we have investigated. The remarkable degree of concordance between multinomial and Poisson estimators (e.g. Fig. 3), not only for interpolation (as anticipated by Brewer and Williamson [1994] and Colwell and Coddington [1994]) but also for extrapolation (as Þrst shown here), was a second surprise, although the two models are closely related, as discussed earlier. We see little reason, for individual-based data, to recommend computing estimators based on one model over the other (although Coleman curves are computationally

less demanding than classical rarefaction), and no reason whatsoever to compute both.

The ability to link rarefaction curves with their corresponding extrapolated richness curves, complete with unconditional conÞdence intervals, helps to solve one of most frustrating limitations of traditional rarefaction: Ôthrowing awayÕ much

statistically different from the richness of a random sample of the same size drawn from the larger reference sample, Y

distributions with approximately equal variances, overlap or non-overlap of 84% conÞdence intervals (mean plus or minus

National Science Council (97-2118-M007-MY3 to A.C.); and the University of Connecticut Research Foundation (to R.L.C.).

ACKNOWLEDGEMENTS

We are grateful to Fangliang He and Sun Yat-sen University for the invitation to contribute this paper to a special issue of JPE and to an anonymous reviewer for helpful comments.
Con ict of interest statement None declared.

REFERENCES

Brewer A, Williamson M (1994) A new relationship for rarefaction. Biodiversity Conserv 3:373–9.

Burnham KP, Overton WS (1978) Estimation of the size of a closed population when capture probabilities vary among animals. Biometrika 65:625–33.

Chao A (1984) Non-parametric estimation of the number of classes in a population. Scand J Stat 11:265–70.

Chao A (1987) Estimating the population size for capture-recapture data with unequal catchability. Biometrics 43:783–91.

Chao A (2005) Species estimation and applications. In: Kotz S, Balakrishnan N, Read CB, Vidakovic B (eds). Encyclopedia of Statistical Sciences, 2nd edn. New York: Wiley, 7907–16.

Chao A, Colwell RK, Lin C-W, et al. (2009) Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90: 1125–33.

Chao A, Hwang W-H, Chen Y-C, et al. (2000) Estimating the number of shared species in two communities. Stat Sin 10:227–46.

Chao A, Lee S-M (1992) Estimating the number of classes via sample coverage. J Am Stat Assoc 87:210–7.

Chao A, Shen TJ (2004) Nonparametric prediction in species sampling. J Agric Biol Environ Stat 9:253–69.

Chazdon RL, Colwell RK, Denslow JS, et al. (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of NE Costa Rica. In: Dallmeier F, Comiskey JA (eds). Forest Biodiversity Research, Monitoring and Modeling: Conceptual Background and Old World Case Studies. Paris:

Mao CX (2007) Estimating species accumulation curves and diversity indices. Stat Sin17:761Ð74.

Mao CX, Li J (2009) Comparing species assemblages via species accumulation curves. Biometrics65:1063Ð7.

Norden N, Chazdon RL, Chao A, et al.(2009) Resilience of tropical rain forests: tree community reassembly in secondary forests. Ecol Lett 12:385Ð94.

Payton ME, Greenstone MH, Schenker N (2004) Overlapping conÞdence intervals or standard error intervals: what do they mean in terms of statistical signiÞcance. J Insect Sci3:34, http://insectscience.org/3.34 (13 November 2011, date last accessed).

Sanders H (1968) Marine benthic diversity: a comparative study. Am Nat 102:243.

Shen T-J, Chao A, Lin C- F (2003) Predicting the number of new species in further taxonomic sampling. Ecology84:798Ð804.

Simberloff D (1979) Rarefaction as a distribution-free method of expressing and estimating diversity. In: Grassle JF, Patil GP, Smith WK, Taillie C (eds). Ecological Diversity in Theory and Practice Fairland, MD: International Cooperative Publishing House, 159Ð76.

Shinozaki K (1963) Notes on the species-area curve. In: 10th Annual Meeting of the Ecological Society of JapanÞAbstract, p. 5. Ecological Society of Japan, Tokyo, Japan.

Smith EP, Stewart PM, Cairns J (1985) Similarities between rarefaction methods. Hydrobiologia120:167Ð70.

Smith W, Grassle F (1977) Sampling properties of a family of diversity measures. Biometrics33:283Ð92.

Solow A, Polasky S (1999) A quick estimator for taxonomic surveys. Ecology80:2799Ð803.

Ugland KI, Gray JS, Ellingsen KE (2003) The speciesÐaccumulation curve and estimation of species richness.J Anim Ecol72:888Ð97.