COMMUNITY ECOLOGY 9(1): 121-123, 2008
1585-8553/$20.00 © Akadémiai Kiadó, Budapest
DOI: 10.1556/ComEc.9.2008.1.14

FORUM

Discovering and rediscovering the sample-based rarefaction formula in the ecological literature

A. Chiarucci[1,2,4], G. Bacaro[1,2], D. Rocchini[1,2] and L. Fattorini[3]

[1]*Dipartimento di Scienze Ambientali "G. Sarfatti", Università di Siena, Via P.A. Mattioli 4, 53100 Siena, Italy*
[2]*TerraData environmetrics, Dipartimento di Scienze Ambientali "G. Sarfatti", Università di Siena,*
Via P.A. Mattioli 4, 53100 Siena, Italy
[3]*Dipartimento di Metodi Quantitativi, Università di Siena, P.za San Francesco 8, 53100 Siena, Italy*
[4] *Corresponding author. Phone: +390577232872, Fax: +390577232896, E-mail: chiarucci@unisi.it*

Keywords: Biodiversity, Ecological statistics, Sampling effort, Species accumulation curve, Species rarefaction curve, Species richness.

Abstract: Rarefaction has long represented a powerful tool for detecting species richness and its variation across spatial scales. Some authors recently reintroduced the mathematical expression for calculating sample-based rarefaction curves. While some of them did not claim any advances, others presented this formula as a new analytical solution. We provide evidence about formulations of the sample-based rarefaction formula older than those recently proposed in ecological literature.

Accumulation and rarefaction curves

In biogeography and community ecology, a widely-applied protocol to assess whether a site has been sufficiently sampled is based on curves showing the increase in the number of recorded species as the sampling effort increases. A curve approximately reaching an asymptote indicates that few or no species would be collected if sampling effort is further increased. On the other hand, a curve which sharply rises near its end should mean that many new species could be recorded by additional sampling effort. In a relatively new terminology, such curves are referred to as *species accumulation curves* (ACs), or, using an older jargon, *collectors' curves* (see Colwell and Coddington 1994). These curves are also widely adopted to describe species diversity patterns (e.g., Ricotta et al. 2002; Crist and Veech 2006).

In order to avoid misinterpretations, Gotelli and Colwell (2001) provided some unambiguous definitions regarding the curves adopted to describe the accumulation of species. According to Gotelli and Colwell (2001), given a collection of n individuals, an *individual-based* AC is the plot of S_i against i ($i = 1,\ldots,n$), where S_i represents the number of species observed among i individuals when they are pooled, one at a time, in a given order. Individual-based ACs require absolute *abundance data* for each species, in terms of the number of individuals. In many cases, however, data are collected by means of plots, transects or traps, each of them giving rise to a "sample" of individuals. Accordingly, given a collection of n such samples, a *sample-based* AC is built simply by successively pooling these samples rather than individuals. Thus, as pointed out by Gotelli and Colwell (2001), the key distinction between individual-based and sample-based ACs is the accumulation unit: an individual *vs.* a sample of individuals. Sample-based ACs have the advantage that only presence or absence of species needs to be detected or recorded (*incidence data*).

When constructing ACs, individuals or samples may be pooled in the order they are recorded (as in a time-series data) or in any other order. Obviously, the order in which individuals or samples are added affects the shape of the resulting curve (Gotelli and Colwell 2001; Ugland et al. 2003). For a given individual-based or sample-based set of data, there are n! possible ACs. To overcome this problem and provide a unique descriptor for any given data set, an order-free curve can be adopted. Given a collection of n individuals or samples, the *rarefaction curve* (RC) is the plot of $\overline{S}_i$ against i ($i = 1,\ldots,n$), where $\overline{S}_i$ represents the arithmetic mean of the S_i values arising from all the possible n! orderings. An *individual-based* or *sample-based* RC can be calculated, in accordance with the corresponding AC from which it derives. Thus, if G denotes the set of species observed in the collection of n individuals (or n samples), S_n denotes the total number of observed species and n_k denotes the number of individuals belonging to species $k \in \mathsf{G}$ (or the number or samples containing at least one individual of species $k \in \mathsf{G}$), then, from elementary combinatorial considerations, for both the sample- and individual-based curves the arithmetic mean of the S_i turns out to be

$$\overline{S}_i = S_n - \binom{n}{i}^{-1} \sum_{k\in \mathrm{G}} \binom{n-n_k}{i}, i = 1,\ldots,n \qquad (1)$$

Formally, expression (1) may be viewed as the expectation of S_i when i individuals or i samples are resampled by means of simple random sampling without replacement from the collection of the n individuals or samples. A different formulation was derived for cases in which the n individuals or samples, are resampled with replacement (see e.g., expression 30 in Hulberts 1971 paper for individual based curves). However, as stressed by Hulbert (1971), the two formulations produce similar results when the size of the considered population is large enough.

"Repetita iuvant"? Multiple rediscovering of sample-based rarefaction curves

The use of individual-based RCs became a well-known procedure, mostly after Sanders (1968). This author suggested the use of RCs on the basis of the intuition that, since the species richness in a collection of *n* biological units tends to increase with *n*, the effective comparison of species richness among different collections required all of them to be *reduced* to the same number of units (presumably that in the smallest collection). To this purpose, the author proposed a computational method, referred to as *rarefaction*, aimed to determine $\overline{S}_i$ for each $i = 1,\ldots,n$. Subsequently, Hurlbert (1971) and Simberloff (1972) independently noted that the Sanders method was incorrect, both arriving at the right expression (1).

The same elementary considerations leading to the individual-based RC can be used to derive expression (1) for the sample-based RC. Kobayashi (1974, p. 227), in the English-language Japanese journal *Researches on Population Ecology*, and then E. P. Smith et al. (1985, pp. 167-168), in the aquatic and marine biology journal *Hydrobiologia*, cited Shinozaki (1963, in a Japanese proceedings volume) as the author firstly deriving expression (1) for a collection of samples. Note that the same expression (or other equivalent forms) was independently derived in the marine biology literature also by Holthe (1975), Engen (1976) and later, in the statistical literature, by W. Smith et al. (1979, p. 188, in a book including conference proceedings). Unfortunately, these papers were ignored for long time and, as a consequence, the analytical formulation of sample-based RC was largely neglected in ecological papers as well. Hence, in most areas of ecology sample-based RCs were computed by means of randomisation methods in which a large set of sample orderings is randomly selected from the universe of all the orderings (see, e.g., early versions of the widely-applied Estimate*S* software; current versions use expression (1)). Even Gotelli and Colwell (2001) justified the use of randomization procedures to compute sample-based RCs, emphasizing the impossibility of deriving a closed expression for sample-based RCs. In fact, Gotelli and Colwell (2001, p. 383) stated that "*Because the sample-based rarefaction curve depends on the spatial distribution of individuals as well as the size and placement of samples…, it cannot be derived theoretically*". Recently, Ugland et al. (2003) and Koellner et al. (2004) reintroduced the analytical expression (1) for sample-based RCs in the ecological literature. While the latter research group did not claim any priority by presenting expression (1) as a result derived "*according to Hurlbert and Simberloff*" (Koellner et al., 2004, p. 544), Ugland et al. (2003, p. 889) claimed the derivation of an "*analytical method which gives exact cumulative numbers of species and so obviates the need for randomisation using Monte Carlo technique and curve fitting*". Later in the same paper, Ugland et al. (2003, p. 894 and 895) referred twice to it as "*our new method*", thus apparently claiming the first derivation of expression (1). However, the same analytical formulation of sample-based RC was already known at least 40 years earlier, as outlined above. In the same period, other authors (Colwell et al., 2004,; Mao et al., 2005) independently derived expression (1) as the unbiased moment estimator of the expected number of species detected by means of *i* samples, under the assumption that species detection occurs in accordance with a mixture of binomial densities (statistical details of the derivation are in Mao et al. 2005). Furthermore, Mao et al. 2005 also proposed a variance estimator which allows the construction of unconditional confidence intervals. It is also possible that other authors also presented the analytical derivation of expression (1) as their own result. On the other hand, at least a couple of ecological papers (Johnson and Patil, 1995; Ricotta et al., 2002) used the correct expression (1) for sample-based RC before its re-introduction by Ugland et al. (2003) and Colwell et al. (2004), and properly referred its derivation to Kobayashi (1974) and Engen (1976).

Of course, these multiple rediscoveries of the sample-based rarefaction formula (1) have been useful for present-day ecologists and biogeographers to know the proper way to calculate the average number of species detected by a given number of samples without using randomisation procedures: "*repetita iuvant*" according to the well known Latin locution! However, this redundancy is likely to determine instability and confusion about the credit to be given to the authors achieving this result, especially among ecologists working in different fields of ecology (e.g., marine and. terrestrial ecology) and those less familiar with the statistical literature. Moreover, even if expression (1) is now familiar to ecologists, the randomisation procedure is still offered by some software packages (e.g., PC-Ord version 5 - http://home.centurytel.net/~mjm/pcordwin.htm - even if this is done in the framework of species-area relations) and this is likely to create ambiguities and misunderstandings. Indeed, the randomisation procedure is completely needless, given the existence of expression (1). Of course, this is mostly a formal problem, since the results obtained with a large number of randomisations are virtually identical to those obtained with the analytical procedure.

Fortunately, the problem is less complex than it appears: sample-based RCs simply constitute order-free curves showing the increase in the number of recorded species as the number of samples increases from 1 to n. Consequently, sample-based RCs can be straightforwardly computed by expression (1) just as their individual-based counterparts, and this is now clear also in the ecological literature. Increasing importance has been attributed to the rarefaction formula in recent years, as a basic approach to many different community statistics. As an example, Crist and Veech (2006) used sampled-based rarefaction curves for computing the alpha, beta and gamma components of species diversity in a set of sampling units (the individual-based counterpart was derived by Olszewski 2004). Also, Ricotta (2004) used expression (1) as a starting point for deriving a parametric diversity index for combining species relative abundances with their taxonomic distinctiveness.

The only problem to be really clarified is the authorship that should be credited for introducing the analytical derivation of the sample-based RC (1). In this paper, we provided evidence that this is at least 40 years older than that reported in present-day ecological literature. It should be remarked that in almost all early cases, expression (1) was introduced in the framework of species area-relations and not for the specific purpose of rarefaction curves. In any case, the first introduction of this formula should be referred to a paper written in Japanese by Shinozaki (1963), or to Kobayashi (1974) if considering English-language literature only – at least until someone finds even earlier publications of this formula! As a concluding remark, this relatively short history highlights a problem that can be observed also with respect to other scientific concepts and methods, i.e. that often no, or too little, credit is given to the non Anglo-Saxon researchers, and especially to the research papers that were published in languages different from English. Geographic biases in the citation of scientific papers have been widely demonstrated in ecology as well as other scientific fields (see e.g., May, 1997; Paris et al., 1998; Wong and Kokko, 2005). Therefore, it is very important to recognise when possible the role of previous and non-cited authors in the progress of scientific knowledge. The rediscovery of parametric indices of diversity could be cited as an example of this problem in the ecological literature (see Ricotta, 2005 and Lövei, 2005 for a description of this story).

Acknowledgement. We would like to acknowledge R. K Colwell, who provided constructive comments on a previous version of this manuscript. We are indebted to J. Podani and two anonymous referees for valuable comments.

References

Colwell, R.K. and J.A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. *Philos. T. Roy. Soc. 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.

Crist, T.O. and J.A. Veech. 2006. Additive partitioning of rarefaction curves and species-area relationship: unifying alpha-, beta- and gamma-diversity with sample size and habitat area. *Ecol. Lett.* 9: 923-932.

Engen, S. 1976. A note on the estimation of the species-area curve. *J. Cons. Inter. Explor. Mer.* 36: 286-288.

Gotelli, N.J. and R.K. Colwell. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol. Lett.* 4: 379-391.

Holthe, T. 1975. A method for the calculation of ordinate values of the cumulative species-area curve. *J. Cons. Inter. Explor. Mer.* 36: 183-184.

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

Johnson, G.D. and G.P. Patil. 1995. Estimating statewide species richness of breeding birds in Pennsylvania. *Coenoses* 10: 81-87.

Kobayashi, S. 1974. The species-area relation I. A model for discrete sampling. *Res. Popul. Ecol.* 15: 223-237.

Koellner, T., A.M. Hersperger and T. Wohlgemuth. 2004. Rarefaction method for assessing plant species diversity on a regional scale. *Ecography* 27: 532-544.

Lövei, G.L., 2005. Generalised entropy indices have a long history in ecology – a comment. *Comm. Ecol.* 6: 245-247.

Mao, C.X., R.K. Colwell and J. Chang. 2005. Estimating the species accumulation curve using mixtures. *Biometrics* 61: 433-441

May, R.M. 1997. The scientific wealth of nations. *Science* 275: 793-796.

Olszewski, T. D. 2004. A unified mathematical framework for the measurement of richness and evenness within and among multiple communities. *Oikos* 104: 377-387.

Paris, G., G. De Leo, P. Menozzi and M. Gatto. 1998. Region-based citation bias in science. *Nature* 396: 210.

Ricotta, C. 2005. On parametric diversity indices in ecology. *Comm. Ecol.* 6: 241-244.

Ricotta, C., M.L. Carranza and G. Avena. 2002. Computing -diversity from species-area curve. *Basic Appl. Ecol.* 3: 15-18

Ricotta, C. 2004. A parametric diversity measure combining the relative abundances and taxonomic distinctiveness of species. *Divers. Distrib.* 10: 143-146.

Sanders, H.L. 1968. Marine benthic diversity: a comparative study. Am. Nat. 102: 243-282.

Shinozaki, K. 1963. Note on the species area curve. Proceedings of the 10th Annual Meeting of Ecological Society of Japan , 5 (in Japanese).

Simberloff, D. 1972. Properties of the rarefaction diversity measurement. *Am. Nat.* 106: 414-418.

Smith, W., J.F. Grassle and D. Kravitz. 1979. Measures of diversity with unbiased estimates. In: J.F. Grassle, G.P. Patil, W. Smith and C. Taillie (eds), *Ecological Diversity in Theory and Practice*. International Co-operative Publishing House, Fairland (MD). pp. 177-191.

Smith, E.P., P.M. Stewart and J. Cairns. 1985. Similarities between rarefaction methods. *Hydrobiologia* 120: 167-170.

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

Wong, B.B.M. and H. Kokko. 2005. Is science as global as we think? *Trends Ecol. Evol.* 20: 475-476.

Received April 2, 2008
Revised June 6, 2008
Accepted June 20, 2008