

A new local estimator of regional species diversity, in terms of ‘shadow species’, with a case study from Sumatra

Keith Rennolls*¹ and Yves Laumonier†

* Computing and Mathematical Sciences, University of Greenwich, Park Row, Greenwich, London SE10 9LS, UK

† CIRAD-Forêt, UPR 36, Campus International de Baillarguet, 34398 Montpellier, Cedex 5, France

(Accepted 10 December 2005)

Abstract: In a local biodiversity inventory the locally rare species are of particular importance. The main problem of sample-based inventories is that many species are so rare that they will not be observed. The observed frequencies of species in the sample provide an estimate of the species proportion in the population. This may be used to estimate the number of species which exist in the population, but which were not observed in the sample (shadow species). This non-parametric approach provides an unbiased estimate of the relative frequency distribution of the species in the population, which differs very significantly from the sample distribution, particularly for the rare species. The approach leads to a new and ecologically meaningful estimator of the Rényi–Hill generalized species diversity measure, which includes species abundance, the Shannon–Weaver and Simpson’s diversity measures, amongst others. The use of the estimator is illustrated on data from a biodiversity inventory of trees on a 3-ha forest sample plot in Sumatra.

Key Words abundance, coverage, diversity, expansion estimator, Rényi–Hill, shadow species, Shannon–Weaver, Simpson

INTRODUCTION

Biodiversity is a multi-dimensional concept that occurs in differing forms at a range of spatial scales and is defined variously according to the goals being addressed. Biodiversity indices abound and many indicators of sustainability have been proposed. There are continuing major international efforts to monitor biodiversity, and much attention is given to collecting data on biodiversity and sustainability indicators. Primary biodiversity and sustainability criteria and indicators used for monitoring by the Ministerial Conference on the Protection of Forests in Europe and the International Timber Trade Association are the number of observed species, and the number of species regarded as being under threat. Such sample-based indicators have to be treated with great care, since observed species counts can miss many very rare species.

Tropical forest ecology during most of the 20th century has been concerned with characterizing the structure of forests and developing an understanding of their dynamics (Ashton 1969, 1976; Hubbell 1979, Lieberman *et al.* 1996, May 1973, 1981; Newbery *et al.*

2000, Tilman 1994). Niche assembly theories have vied with dispersal assembly theories (Chase 2005), with the neutral theory of biodiversity adopted by Hubbell (1997, 2001) surprisingly explaining many qualitative aspects of forest biodiversity dynamics in terms of speciation and random population drift, without the use of niche assembly and dispersal assembly concepts. Unbiased estimators of species diversity from field data are an important basis for the building and the validation of a general and unified model of forest diversity and dynamics. Observed diversity-measure values from samples of differing sizes are not a satisfactory basis for such scientific considerations, because the raw estimates miss so much.

Even species diversity, an important but comparatively simple component of the biodiversity concept, has been defined in many ways (Gleason 1922, Hill 1973, Hurlbert 1971, Fisher *et al.* 1943, Magurran 1988, May 1975, Orłóci 1991, Peet 1975, Pielou 1975, Rényi 1961). Measures of species diversity are usually only meaningful in relation to the population or region under consideration. Species diversity measures evaluated on local sample data from the population are usually sample-size dependent, and hence are biased estimates of the corresponding population diversity parameters.

¹ Corresponding author. Email: k.rennolls@gre.ac.uk

In considering estimators of a regional species diversity measure it is essential to go beyond the simple evaluation of the diversity measures on the sample data and to devise estimators of population diversity which are not sample-size dependent.

There has been much work over a long period with the specific aim of developing species abundance estimators that eliminate the bias of the naïve sample-based estimator, i.e. the count of the number of species observed. The proportion of the population species distribution that is represented by the species in the sample (the coverage of that sample) was estimated by Good (1953) as $(1 - f_1/n)$ where f_1 is the number of singletons in the sample, and n is the sample size. This estimator was justified mathematically by using an empirical Bayes approach (Engin 1978), with an a priori assumption that all species have equal proportions; it is clear that Good's coverage estimate is at best a first approximation. Burnham & Overton (1978, 1979), and Otis *et al.* (1978) introduced and analysed jack-knife estimation methods for removing sample-size bias of species abundance estimators. These jack-knife estimators are purely technical and have no ecological content or interpretation. Similarly, Chao's moment-based estimators (Chao 1984, Chao & Lee 1992) are technical estimators based on various mathematical assumptions and they also have no ecological content or interpretation. Other methodological work in this area of abundance estimation use computationally intensive methods: bootstrap methods (Smith & van Belle 1984) and Bayesian and Monte Carlo Markov Chain (MCMC) methods (Rodrigues *et al.* 2001). Most of this work appears in the statistical literature and is highly technical. Some of these estimators are given in detail in the appendix to this paper and are used for comparative purposes in the case-study of this paper. While these abundance estimation techniques do achieve their aim of reducing bias, they seemingly do not involve any ecological insights.

In contrast to all the above-mentioned technical estimators, the use of species–area and species–sample-size models (Coleman 1981, Coleman *et al.* 1982, Condit *et al.* 1996, Goodman 1949, May 1975) is an intuitively plausible way of extrapolating observed species counts to the population scale, and the ecological interpretation is clear.

Many species diversity indices have been defined in the literature. It is often difficult for ecologists to choose which of the many diversity indices and estimators to use. Hence, it is highly desirable that any new estimator should provide both new ecological insight and be of general applicability.

This paper introduces and uses a new non-parametric probability-weighting method to scale up from the observed species frequencies to an estimate of the species abundance in the population (Chao & Shen 2003, Rennolls & Laumonier 1999a). This estimator, which we term the expansion estimator, is simple and

intuitively plausible, and has a clear ecological meaning. The approach estimates the number of unseen species (shadow species) corresponding to each observed species and hence provides in addition a new estimator of the population species distribution. In the case study, on forest tree data from Jambi, Sumatra, we compare the estimates obtained from this new estimator with those obtained from other estimators mentioned above.

Most of the publications mentioned above are concerned with obtaining unbiased estimators of global species abundance from a local sample. There has been comparatively little published on eliminating the sample-size dependency of estimators of other measures of species diversity. An exception is the sample-based estimate of the α parameter of the log-series distribution, which is often claimed to be sample-size independent (Fisher *et al.* 1943, Kempton 1979, May 1975).

The expansion-estimator is generalized to provide an estimator of the Rényi–Hill generalized species diversity, a diversity measure which includes species abundance, Shannon–Weaver, Simpson and Berger–Parker indices as special cases (Rennolls & Laumonier 1999a, 2000). Chao & Shen (2003) use a similar estimator but specifically for the estimation of the population Shannon–Weaver diversity index. We briefly demonstrate the use of new cover-adjusted α -diversity estimator on the case study data, as an extension of the results of Rennolls & Laumonier (2000).

ASSUMPTIONS AND TERMINOLOGY

Species may be regarded as a classification of the elements of the population, with a species label attached to each element of the population. We use j as an index/label to indicate species, with $j = 1, 2, \dots, S$, where S is the number of species in the population (i.e. the species abundance). Note that the term index is used in a different way in this section than it was in the previous section. Estimation of S from a sample of elements from the population is the species-abundance estimation problem. Even though the sampling situation considered is standard in ecology, and the estimators discussed in this paper are mathematically fairly simple, the notation can cause some confusion. Hence, we define our notation in detail.

We suppose that individuals in the population are indexed by i where $i = 1, \dots, N$, and N may be either finite or infinite. Each individual belongs to exactly one species, with the proportional numerical representation of species j in the population being p_j , $j = 1, \dots, S$.

For clarity of treatment we make the usual assumption that the sample is a simple random sample of n individuals ($i = 1, \dots, n$) drawn from the population. The fact that this assumption is almost never true in practice means the estimators that are derived from it are approximate.

n_j individuals of species j are observed in the sample, for $j = 1, \dots, s$, with

$$\sum_j^s n_j = n \tag{1}$$

The probability that the i th sample element is of species j is p_j (the multinomial model). Then $E(n_j) = E(\sum_{i=1}^n I_j(i)) = np_j$, where $I_j(i)$ is the indicator that the i th individual is of species j . Therefore, the usual maximum likelihood estimator of p_j is

$$\hat{p}_j = \frac{n_j}{n} \tag{2}$$

The coverage of the sample is defined to be

$$C = \sum_{j=1}^s p_j \tag{3}$$

Note that for notational simplicity we have used the same index, i , for the individual elements in both the population and the sample, and similarly for species j . We may suppose that the labelling of the elements in the population is such that the first n elements in the population correspond to the elements actually observed in the sample, and similarly for the labelling of species.

We introduce a third index, k ($= 0, 1, 2, \dots$) indicating the number of trees of a species that are observed in the sample. The k -value for species j is equal to n_j . The number of species observed in the sample with a count of k is then denoted by f_k , $k = 0, 1, 2, 3, \dots$. The frequency $f_0 \equiv S - s$ is the number of unobserved species. Species for which only one individual is observed in the sample ($k = 1$) are called singleton species, and f_1 is the number of singleton species observed in the sample. f_1 plays an important part in all estimators of the number of unobserved species, and hence the abundance S . The Appendix gives the estimating equations for the jack-knife and the Chao estimators, both of which make use of f_1 .

Generally, for the estimators given, variance estimation formulae can be found in the literature, or jack-knifing or boot-strap methods can be used. For clarity and brevity in our qualitative comparison of the estimators we do not consider variance issues in this paper.

THE COVER-ADJUSTED EXPANSION-ESTIMATOR OF ABUNDANCE

Intuitive justification

In many ecological species-diversity studies, it is found that many species are locally very rare in the population (MacArthur 1960, Preston 1962). Hence such rare species have a low chance of being observed in a relatively small sample. In the tropical rain forest case study of this paper, at least 43% of observed species have a stocking

density of 1 tree per 3 ha, or less, and each of these rare species constitutes less than 0.05% of the tree population (with dbh ≥ 10 cm).

The observation frequency of a species may be used to estimate the relative frequency of the species in the population, and hence the probability of a species being observed in a random sample can be calculated. An abundant species is almost certain to be observed in a random sample, and there is a high probability that such an abundant species will be observed many times. It is unlikely that there are equally abundant species in the population which were not observed. However, a rare species which is observed has a relatively high chance of not being observed in a random sample. This probability of non-observance maybe associated with species that are not observed in the actual sample, but would be observed on other samples. Such species, which were not observed, but which we argue must exist, we term 'shadow species'. This paper uses a probability-weighting method to obtain an estimator of the number of shadow species corresponding to each observed species. Hence a new estimator of the population species distribution results, as well as corresponding estimators of population species diversity measures (Chao & Shen 2003, Rennolls & Laumonier 1999a).

The expansion estimator of species abundance

For observed species j the usual maximum likelihood estimate of p_j is given by (2). Hence the probability of a species such as species j not being observed in the plot is estimated as

$$\hat{P}_0(j) = (1 - \hat{p}_j)^n \tag{5}$$

Hence the probability of species j being included in the sample, the inclusion probability π_j in sampling terminology, can be estimated as

$$\hat{\pi}_j = 1 - \hat{P}_0(j) = 1 - (1 - \hat{p}_j)^n \tag{6}$$

In the current context we may regard π_j to be a coverage probability for species j , ($C(j) \equiv \pi_j$). We postulate that there are other species, v_j in number, similar to species j in their chances of being selected, which were not observed in the sample. These are the shadow species to species j . We may write the ratio of the (single) observed species j to the shadow species count, v_j , as:

$$\begin{aligned} 1 : v_j &:: \pi_j : 1 - \pi_j \\ \Leftrightarrow \frac{v_j}{1} &= \frac{1 - \pi_j}{\pi_j} = \frac{1}{\pi_j} - 1 \\ \Leftrightarrow 1 + v_j &= \frac{1}{\pi_j} \end{aligned} \tag{7}$$

That is, we obtain an estimate of the total number of species (i.e. $1 + \hat{v}_j$) corresponding to observed species j by

expanding up the single species count by the factor $(1/\hat{\pi}_j)$. For example, if the inclusion probability for a particular species were 0.5, then we would say that there exist two species which correspond to the particular observed species: one, the observed species, and the other, the shadow species.

If we sum the individual-species expansion-estimators over all observed species, we obtain the new expansion estimator of abundance,

$$\begin{aligned} \hat{S} &= \sum_j^s \frac{1}{\hat{\pi}_j} = \sum_j^s (1 + \nu_j) \cdot 1 \\ &\equiv \sum_k \frac{f_k}{\hat{\pi}_k} \end{aligned} \tag{8}$$

where f_k is the number of observed species for which the observation frequency is $k = 1, 2, 3, \dots$. This estimator seems to have been first proposed as an abundance estimator by Rennolls & Laumonier (1999a). It is noted that this estimator has the form of the Horvitz–Thompson estimator (Cochran 1977, Horvitz & Thompson 1952, Särndal *et al.* 1992) of the total number of species; either (1) in terms of observed species, with species inclusion probabilities π_j , and the response variable being the unique species indicator, or (2) in terms of k -observed species groups.

We may write the first form of the abundance estimator given in (8) as

$$\hat{S} = \sum_j^s \frac{1}{\hat{\pi}_j} = \sum_j^s \frac{I_j}{\hat{\pi}_j} \equiv \sum_j^s \frac{I_j}{\hat{\pi}_j} \tag{9}$$

where I_j is the indicator function for species j being in the sample. Since $E(I_j) = \pi_j$, we have

$$\begin{aligned} E(\hat{S}) &= E\left(\sum_j^s \frac{I_j}{\hat{\pi}_j}\right) = \sum_j^s \frac{E(I_j)}{\hat{\pi}_j} = \sum_j^s \frac{\pi_j}{\hat{\pi}_j} \\ &= \sum_j^s 1 = S, \quad \text{if } \hat{\pi}_j = \pi_j \end{aligned} \tag{10}$$

That is, this expansion estimator of species abundance is unbiased, if the estimated species inclusion probabilities are unbiased.

However, it is noted that equation (9) is not strictly the Horvitz–Thompson estimator of classical sample survey theory since to be so the sample units would have to be species, and this is not the case. Also, I_j is not the kind of response variable required for the Horvitz–Thompson estimator of classical sample survey theory (Cochran 1977, Särndal *et al.* 1992). Chao & Shen (2003) make a slightly different presentation of this estimator for estimating Shannon–Weaver diversity.

Cover-adjusted expansion estimator of species abundance

The argument of the last section was that the i th observed species corresponded to a set of similar shadow species. However the extremely rare species will most likely not have been observed and hence the corresponding extremely rare shadow species cannot be estimated by the expansion estimator (9). We also note that (1) and (2) imply

$$\sum_{j=1}^s \hat{p}_j = 1 \tag{11}$$

whereas, by definition, we have

$$\sum_{j=1}^S p_j = 1 \text{ with } S \geq s \tag{12}$$

Hence the estimators of \hat{p}_j in (2) and $\hat{\pi}_j$ in (6) are overestimates, in general. Hence the estimated number of shadow species, from (8), will be under-estimates.

Good (1953) estimated the coverage, defined in (3), by

$$\hat{C} = \left(1 - \frac{f_1}{n}\right) \tag{13}$$

See also Engin (1978). The derivation of (13) by Good (1953) uses Bayesian methods with a prior distribution with equal proportions for each species, so at best this estimator maybe regarded as a first approximation. However, it may be used to adjust the bias of \hat{p}_j in (2) and $\hat{\pi}_j$ in (6) to give the following second stage cover-adjusted estimators:

$$\tilde{p}_j = \hat{C} \hat{p}_j = \hat{C} \frac{n_j}{n} \tag{14}$$

$$\tilde{\pi}_j = 1 - (1 - \tilde{p}_j)^n \tag{15}$$

$$1 + \tilde{\nu}_j = \frac{1}{\tilde{\pi}_j} \tag{16}$$

$$\tilde{S} = \sum_j^s \frac{1}{\tilde{\pi}_j} = \sum_j^s (1 + \tilde{\nu}_j) \cdot 1 \equiv \sum_k \frac{f_k}{\tilde{\pi}_k} \tag{17}$$

Ashbridge & Goudie (2000) use similar cover-adjusted estimators for mark-recapture data, as do Chao & Shen (2003) to estimate the Shannon–Weaver index.

THE COVER-ADJUSTED EXPANSION-ESTIMATOR OF THE RÉNYI-HILL GENERALIZED SPECIES DIVERSITY

The treatment in the previous sections introduced the expansion estimator of species abundance. The estimator has some virtue in terms of its simplicity and its ecological interpretability. Furthermore the concept of shadow species may be used to achieve a similar Horvitz–Thompson type of scaling from sample to population for

a very general diversity measure, the H_α of Rényi (1961), and equivalently for the N_α of Hill (1973).

Rényi’s generalized entropy: α -diversity

The Rényi generalized α -entropy H_α (which we call α -diversity), and equivalently the N_α of Hill (1973), are functional statistics for a population and are given by:

$$N_\alpha = \exp(H_\alpha) = \left(\sum_{j=1}^s p_j^\alpha \right)^{\frac{1}{1-\alpha}} : \alpha \geq 0 \quad (18)$$

This functional diversity index contains many of the standard diversity indices corresponding to particular values of α . For example, the species abundance $S = N_0$, the Shannon–Weaver index is $\lim_{\alpha \rightarrow 1}(H_\alpha) = -\sum_{j=1}^s p_j \log p_j$ and Simpson’s index of concentration is $\lambda = \sum_{j=1}^s p_j^2 = 1/N_2$. The Berger–Parker index is $N_\infty^{-1} = d$, the proportion of the most abundant species, regarded by May (1975) as one of the most informative indices.

Rennolls & Laumonier (1999b) made extensive use of H_α vs. α graphs to analyse the spatial distribution of the multi-dimensional structure of diversity of a tropical rain forest, and indeed the Berger–Parker index did turn out to be an important aspect of the diversity structure of their Batang Ule case study.

A cover-adjusted expansion-estimator of α -diversity

We introduce a new cover-adjusted expansion-estimator of population α -diversity, either H_α or N_α , through the following defining equation:

$$\tilde{N}_\alpha = \exp(\tilde{H}_\alpha) = \left(\sum_{j=1}^s (1 + \tilde{v}_j) \tilde{p}_j^\alpha \right)^{\frac{1}{1-\alpha}} = \left(\sum_{j=1}^s \left(\frac{\tilde{p}_j^\alpha}{\tilde{\pi}_j} \right) \right)^{\frac{1}{1-\alpha}} \quad (19)$$

This is a very compact estimator and has much potential for the diversity analysis of empirical data. \tilde{N}_0 is equivalent to \tilde{S} of equation (17). The population Shannon–Weaver diversity measure is estimated by

$$\lim_{\alpha \rightarrow 1} \tilde{H}_\alpha = \frac{-\sum_{j=1}^s \tilde{p}_j \ln \tilde{p}_j}{\tilde{C}} \quad (20)$$

This reflects the coverage correction in the estimator defined in equation (19). Use of the functional diversity statistics defined by equation (19) will allow comparisons of the diversity structure between regions which have samples of differing sizes, or from differing sample sizes collected in the same region at different times.

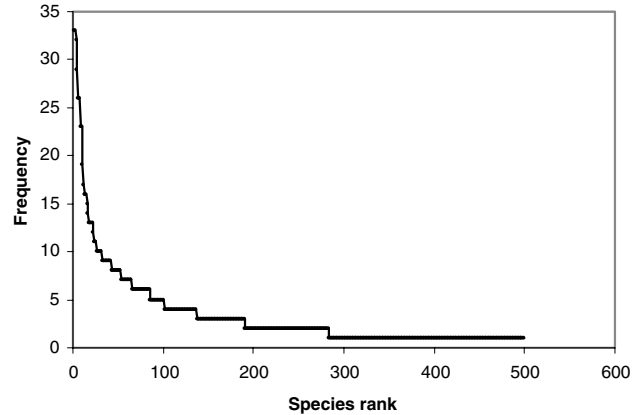


Figure 1. Species-frequency curve for trees in a 3-ha sample plot in Batang Ule, Jambi, Sumatra. The sample plot has dimensions of 300 × 100 m, with the longer dimension being oriented along the line of maximal topographic gradient. A total of 1897 trees of diameter greater than or equal to 10 cm were observed to fall into 497 identified species. Species are ranked according to how frequently they are observed, and then frequency is plotted against rank. Rank-1 species with frequency 196 is excluded, for clarity of presentation.

CASE-STUDY FROM SUMATRA

Batang Ule

In this section we demonstrate the use of the unbiased expansion-estimator of abundance. Data from a research site at Batang Ule, Jambi, Sumatra, were collected by the second author, his students and others (Laumonier 1997, Trichon 1996). The site is 3 ha in area, with dimensions of 300 × 100 m, with the longer dimension being oriented along the line of maximal topographic gradient. A total of 1897 trees of diameter greater than or equal to 10 cm were observed to fall into 497 identified species, with 13 trees which were assigned to two morphospecies groups. The species-frequency curve is shown in Figure 1, with the rank-1 species excluded (frequency = 196). Note the Berger–Parker index is $(196/1897)^{-1} = 9.68$, not a particularly informative number for a single plot, by itself.

Shadow species estimates

Table 1 shows the results from the shadow species calculations, using equations (14)–(17). The numbers of species with observation frequencies of 1, 2, 3, 4 and 5 were 216, 93, 53, 36 and 16 respectively (accounting for 83% of all species and 41% of observed trees). The estimated number of shadow species for each species with observation frequencies 1, 2, 3, 4 and 5 were 0.701, 0.205, 0.075, 0.030 and 0.012 respectively (using equation (16)). The coverage adjustment used in equation (16) takes place on a continuous scale, and

Table 1. Detailed calculations of the estimated number of shadow species in the population, for Batang Ule, 10 cm minimum dbh. The first two rows are from the sample data. The second two rows are from formulae (16) and (17).

Number of trees per species, k	1	2	3	4	5
Number of species, f_k	216	93	53	36	16
Shadow species/obs. species, v_k	0.701	0.205	0.075	0.030	0.012
Number of shadow species, $f_k \cdot v_k$	151.3	19.02	3.98	1.07	0.19

hence the estimates of the numbers of shadow species are real numbers (i.e. decimals). These values should be interpreted as estimated mean values. In reality, the count of shadow species that exist would take non-zero integer values.

Hence, the estimated numbers of shadow species from these five lowest frequency species classes are 151 (= 216×0.701), 19, 4, 1 and 0 respectively, rounded to the nearest integer. The more frequently occurring species contribute no extra shadow species. The total estimated number of shadow species is therefore 175. Hence, with 499 observed species, we arrive at the total species abundance estimate of 674.

The frequency distributions of the observed and shadow species, from Table 1, are shown in Figure 2. Note, in this figure, a species which is observed k -times is termed a 'k-ton' (since a single observance results in the species being called a singleton). The source for the relatively large increase in the abundance estimate from the observed

species count can be seen clearly; that is, the observation of a large number of singleton species leads to an estimate of a large number of corresponding shadow species. Of course, a general statement such as this could have been made from the coverage estimate of Good (1953), but not with an associated distribution of shadow species abundance estimates.

Rennolls & Laumonier (1999b) used the same Batang Ule data to fit species-area relationships (SARs). The SAR estimate of the total number of species for Batang Ule was 676 ($\text{dbh} \geq 10 \text{ cm}$), with a 95% confidence interval (631–720), virtually identical to the cover-adjusted expansion-estimate obtained here.

On the same data, the first and second order jack-knife abundance estimates are 715, and 1024 from estimators (A1) and (A2); Chao's moment-based estimate is 750, from estimator (A3); the conditions of estimator (A4) are not satisfied. The first-order jack-knife and Chao's estimate are not significantly different from the ecologically based cover-adjusted expansion and SAR estimates. Note that minimal tree diameter in plot-based tree sampling has a major effect on the population species abundance (as well as plot area) (Rennolls & Laumonier 1999b).

Graphical analysis of α -diversity

The aim of this section is to indicate briefly how the new estimator equation (19) may be used in a fairly complex diversity analysis situation. However we keep the presentation fairly concise and refer readers to Rennolls & Laumonier (2000) for more details and illustrations of the context. This earlier paper did not include the shadow species of the expansion-estimator, or the coverage correction included in this paper.

The 3-ha Batang Ule plot was actually divided into 30 subplots of dimensions $100 \times 10 \text{ m}$ (and hence the plots might be termed transects) arranged along an ecological gradient; from a valley, up a slope and onto a plateau. Each transect was perpendicular to the ecological gradient. Rennolls & Laumonier (2000) used graphs of the α -diversity for the subplots (Orlóci 1991) as 41-dimensional diversity vector statistics to compare the diversity structure between the 30 subplots, and over the plot as a whole. They found that whilst the majority of the plots retained their relative rank over the α -range, use of the k -means clustering method grouped together five contiguous subplots, numbered 14, 15, 16, 17 and 18, which exhibited anomalously low diversity measures with increasing α . Whilst these contiguous subplots were not distinguishable from other subplots in terms of species abundance, or the Shannon–Weaver measure of diversity, they were distinguishable in terms of their high α -diversity behaviour. That is, the Berger–Parker

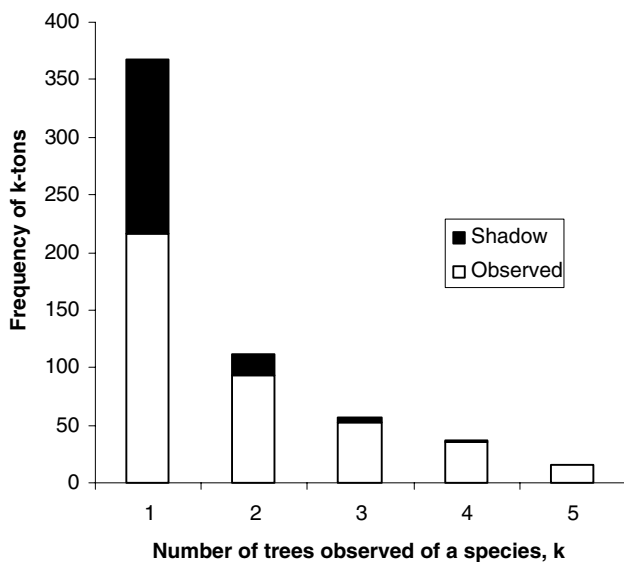


Figure 2. Frequency distribution of observed-rare and estimated-shadow species at Batang Ule. The abscissae (x-values) indicate the number of times a species is observed in the sample. The first white column (from the left) indicates the number of observed singletons; the black column above it is the estimated number of corresponding unobserved (shadow) species.

index shows its diversity discrimination power, in line with the expectations of May (1975). These contiguous subplots are those at the top of a slope, the ridge above the slope, and the first two subplots on the plateau next to the slope. Rennolls & Laumonier (2000) also investigated the use of factor analysis methods to analyse the diversity structure of the α -diversity plots. They found a two-dimensional ordination was adequate for diversity structure characterization at Batang Ule.

Analysis and ordination of cover-adjusted expansion-estimated α -diversity curves

The cover-adjusted expansion-estimated α -diversity curves for the 30 Batang Ule plots have been re-analysed using the estimation equation (19), and the results compared with those obtained previously. We summarize the results here, for the sake of brevity and clarity.

Exactly the same subplots, 14–18, are picked up by the same method of cluster analysis as previously. However, the estimated α -diversity curves for the subplots exhibit considerably more flexibility in shape than found in Rennolls & Laumonier (2000), possibly indicating a potential for more sensitive diversity structure analysis. In fact, factor analysis of the estimated α -diversity vectors reveals that the 30 subplots need a three-dimensional diversity space to represent their α -diversity structure, in contrast to the two-dimensional space that was found to be adequate previously.

DISCUSSION

The shadow-species concept and the expansion-estimator have been developed assuming that the sample plot observations are a random sample from the population. Plotkin *et al.* (2000) demonstrate that most of the tree species of the Smithsonian 50-ha plots display some degree of aggregation, due in part to the influence of local topography. Even so, the approach presented is useful in defining approximately unbiased estimators of a wide range of population diversity measures, as the jackknife and Chao estimators do for population species abundance.

So, what is achieved by this new estimator in terms of bias-correction that was not already available for the jack-knife and Chao estimators? The answer is four-fold.

First, the derivation of the cover-adjusted expansion estimator has clear intuitive appeal, as opposed to previous bias-correction estimators which were purely technical in nature. Also, the expansion estimator has been expressed in terms of the number of unobserved shadow species corresponding to each observed species, an ecologically meaningful concept.

Second, shadow species estimates have relevance to both conservation strategies operating at the global or

regional levels, and sustainability monitoring exercises operating at the local level. It is clear from Table 1 and Figure 1 for the case-study, that the estimated number of rare shadow species in tropical forests can be very high. Hence, both in terms of global conservation strategies, and local sustainability monitoring, it is important not just to rely on observed observation frequencies, but also to consider what the figures imply in terms of many shadow species that may not figure explicitly in the observed counts.

Third, the fact that estimated numbers of shadow species are naturally stratified by the frequency of the corresponding observed species means that each inferred shadow species has with it an associated spatial stocking level. Hence the observed species frequency distribution in the sample, $\{f_k\}$, has to be modified to $\{(1 + \nu_k)f_k\}$, where ν_k is the number of shadow species corresponding to a k -ton species. That is, the expansion-estimator approach estimates the population species-frequency distribution, and not only the population abundance. If ecological analysis of the species-frequency distribution is to be valid for the population concerned, then the appropriate species-frequency distribution needs to be used. The new estimator of the population species frequency distribution provided here could materially alter the interpretation of many of the previously published empirical analyses. Even the conclusions reached in the debate as to what might be the 'true' model for species frequency curves, i.e. geometric, log-normal or log-series, might be affected.

Fourth, the estimated population species-frequency distribution, and the way it changes, is a basic characterizing feature of a dynamic ecosystem, and only by having unbiased estimates of this distribution can we meaningfully hope to move to the construction and validation of a 'truly unified theory of biodiversity' (Hubbell 2001).

LITERATURE CITED

- ASHBRIDGE, J. & GOUDIE, I. B. J. 2000. Coverage-adjusted estimators for mark-recapture in heterogeneous populations. *Communications in Statistics: Simulation and Computation* 29:1215–1237.
- ASHTON, P. S. 1969. Speciation among tropical trees: some deductions in the light of recent evidence. *Biological Journal of the Linnean Society* 1:155–196.
- ASHTON, P. S. 1976. Mixed dipterocarp forest and its variation with habitat in Malayan lowlands: a re-evaluation of Pasoh. *Malayan Forester* 39:56–72.
- BURNHAM, K. P. & OVERTON, W. S. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika* 65:625–633.
- BURNHAM, K. P. & OVERTON, W. S. 1979. Robust estimation of population size when capture probabilities vary among animals. *Ecology* 60:927–936.

- CHAO, A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* 11:265–270.
- CHAO, A. & LEE, S. M. 1992. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association* 87:210–217.
- CHAO, A. & SHEN, T. S. 2003. Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environmental and Ecological Statistics* 10:429–443.
- CHASE, J. M. 2005. Towards a really unified theory for metacommunities. *Functional Ecology* 19:182–186.
- COCHRAN, W. G. 1977. *Sampling techniques*. (Third edition). John Wiley & Sons, New York. 428 pp.
- COLEMAN, B. 1981. Random placement and species-area relations. *Mathematical Biosciences* 54:191–215.
- COLEMAN, B., MARES, M., WILLIG, M. & HSIEY, Y. 1982. Randomness, area, and species richness. *Ecology* 63:1121–1133.
- CONDIT, R., HUBBELL, S. B., LAFRANKIE, J. V., SUUKUMAR, R., MANOKARAN, N., FOSTER, R. & ASHTON, P. S. 1996. Species-area and species individual relationships for tropical trees: a comparison of three 50-ha plots. *Journal of Ecology* 84:549–562.
- ENGIN, S. 1978. *Stochastic abundance models*. Chapman and Hall, London. 126 pp.
- FISHER, R. A., CORBET, A. S. & WILLIAMS, C. B. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* 12:42–58.
- GLEASON, H. A. 1922. On the relation between species and area. *Ecology* 3:156–162.
- GOOD, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* 40:237–264.
- GOODMAN, L. A. 1949. On the estimation of the number of classes in a population. *Annals of Mathematical Statistics* 20:572–579.
- HILL, M. O. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* 54:427–431.
- HORVITZ, D. G. & THOMPSON, D. J. 1952. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* 47:663–685.
- HUBBELL, S. P. 1979. Tree dispersion, abundance and diversity in a tropical dry forest. *Science* 203:1299–1309.
- HUBBELL, S. P. 1997. A unified theory of biogeography and relative species abundance and its application to tropical rain forest and coral reefs. *Coral Reefs* 16 Suppl.:S9–S21.
- HUBBELL, S. P. 2001. *The unified neutral theory of biodiversity and biogeography*. Monograph in Population Biology 32. Princeton University Press, Princeton. 375 pp.
- HURLBERT, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52:577–586.
- KEMPTON, R. A. 1979. The structure of species abundance and measurement of diversity. *Biometrics* 35:307–321.
- LAUMONIER, Y. 1997. *The vegetation and physiography of Sumatra*. Geobotany 22, Klüwer Academic Publishers, Dordrecht. 222 pp.
- LIEBERMAN, D., LIEBERMAN, M., PERALTA, R. & HARTSHORN, G. S. 1996. Tropical forest structure and composition on a large-scale altitudinal gradient in Costa Rica. *Journal of Ecology* 84:137–152.
- MACARTHUR, R. H. 1960. On the relative abundance of species. *American Naturalist* 45:25–36.
- MAGURRAN, A. E. 1988. *Ecological diversity and its measurement*. Chapman and Hall, London. 192 pp.
- MAY, R. M. 1973. *Stability and complexity in model ecosystems*. Princeton University Press, Princeton. 292 pp.
- MAY, R. M. 1975. Patterns of species abundance and diversity. Pp. 81–120 in Cody, M. L. & Diamond, J. M. (eds.). *Ecology and evolution of communities*. Harvard University Press, Cambridge.
- MAY, R. M. 1981. Patterns in multi-species communities. Pp. 197–227 in May, R. M. (ed.). *Theoretical ecology: principles and applications*. Blackwell, Oxford.
- NEWBERY, D. M., CLUTTON-BROCK, T. H. & PRANCE, G. T. (eds.) 2000. Changes and disturbance in tropical rainforest in south-east Asia. *Philosophical Transactions of the Royal Society* 354:1721–1897.
- ORLÓCI, L. 1991. *Entropy and information*. Volume 3 in the Ecological Computation Series, SPB Academic Publishing, the Hague. 72 pp.
- OTIS, D. L., BURNHAM, K. P., WHITE, G. C. & ANDERSON, D. R. 1978. Statistical inference from capture data on closed animal populations. *Wildlife Monographs* 62:1–135.
- PEET, R. K. 1975. Relative diversity indices. *Ecology* 56:496–498.
- PIELOU, E. C. 1975. *Ecological diversity*. John Wiley, New York. 385 pp.
- PLOTKIN, J. B., POTTS, M. D., LESLIE, N., MANOKARAN, N., LAFRANKIE, J. V. & ASHTON, P. S. 2000. Species-area curves, spatial aggregation, and habitat specialization in tropical forests. *Journal of Theoretical Biology* 207:81–99.
- PRESTON, F. W. 1962. The canonical distribution of communities and rarity, part I. *Ecology* 43:185–215, 431–432.
- RENNOLLS, K. & LAUMONIER, Y. 1999a. Analysis of species hyperdiversity in the tropical rain forests of Indonesia: the problem of non-observance. Pp. 19–23 in Sassa, K. (ed.). *Environmental forest science*. Proceedings of IUFRO Division 8 Conference, Oct. 1998, Kyoto University. Forestry Sciences Series 54. Kluwer, Dordrecht.
- RENNOLLS, K. & LAUMONIER, Y. 1999b. Species-area and species-diameter curves for three forest sites in Sumatra. *Journal of Tropical Forest Science* 11:784–800.
- RENNOLLS, K. & LAUMONIER, Y. 2000. Species diversity structure analysis at two sites in the tropical rain forest of Sumatra. *Journal of Tropical Ecology* 16:253–270.
- RÉNYI, A. 1961. On measures of entropy and information. Pp. 547–561 in Neyman, J. (ed.). *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Volume 1, University of California Press, Berkeley.
- RODRIGUES, J., MILAN, L. A. & LEITE, J. G. 2001. Hierarchical Bayesian estimation for the number of species. *Biometrical Journal* 43:737–746.
- SÄRNDAL, C.-E., SWENSSON, B. & WRETMAN, J. 1992. *Model assisted survey sampling*. Springer-Verlag, New York. 694 pp.
- SMITH, E. P. & VAN BELLE, G. 1984. Nonparametric estimation of species richness. *Biometrics* 40:119–129.
- TILMAN, D. 1994. Competition and biodiversity in spatially structured habitats. *Ecology* 75:2–16.
- TRICHON, V. 1996. *Hétérogénéité spatiale des structures en forêt naturelle de basse altitude à Sumatra*. Thèse Doctorat Université Toulouse III. 260 pp.

APPENDIX: THE ESTABLISHED ABUNDANCE ESTIMATORS

The jack-knife estimators

Burnham & Overton (1978, 1979), and Otis *et al.* (1978) used jack-knife estimation methods for removing the bias in the naïve sample-based estimate of species abundance. The first and second order jack-knife abundance estimators are respectively:

$$(i) \quad S_{J1} = s + \frac{n-1}{n} f_1 \tag{A1}$$

and

$$(ii) \quad S_{J1} = s + \frac{2n-3}{n} f_1 - \frac{2(n-2)}{n(n-1)} f_2 \tag{A2}$$

Chao's estimators

Chao's moment-based estimators (Chao 1984, Chao & Lee 1992) follow from complex and sophisticated mathematical analysis, and result in the following abundance estimators:

$$(i) \quad S_{Ch1} = s + \frac{f_1^2}{2f_2} \tag{A3}$$

and

$$(ii) \quad S_{Ch2} = \left\{ s + \left[\frac{f_1^2}{2f_2} \right] \left[1 - \frac{2f_2}{f_1} \right] \middle/ \left[1 - \frac{3f_3}{f_2} \right] \right\} \tag{A4}$$

provided $tf_1 > 2f_2$; $tf_2 > 3f_3$; $3f_1f_3 > 2f_2^2$ where t is the number of sample plots (Chao 1984). These estimators follow from various assumptions made for reasons of tractability.