Journal of Animal Ecology 2005
74, 375–386

Species richness estimators: how many species can dance on the head of a pin?

R. B. O'HARA

Department of Mathematics and Statistics, PO Box 68 (Gustaf Hällströmin katu 2b), FIN-00014 University of Helsinki, Finland

Summary

1. Several species richness estimators (two non-parametric, four based on rarefaction curves and two from fitting of abundance distributions) were compared by examining their performance in estimating the species richness for two moth data sets from the United Kingdom. Comparisons were also made using data simulated from the fitted abundance distributions.

2. The different species richness estimators gave different estimates. The non-parametric estimates and the rarefaction estimates were similar, but were smaller than the parametric estimates. When the simulated data were used, the only methods to give estimates near the true value was the parametric method using the distribution from which the data were simulated.

3. At present it is impossible to decide whether any of the estimation methods will give a realistic estimate, as not enough is known about the true numbers of species in communities. Until this is rectified, the most that can be hoped for is to obtain upper and lower bounds on species richness.

Key-words: ACE, Chao1, negative binomial, Poisson log-normal, taxon sampling curve.

Journal of Animal Ecology (2005) **74**, 375–386
doi: 10.1111/j.1365-2656.2005.00940.x

Introduction

One important component of ecological diversity is the number of species in a community. Although obtaining an estimate of this number would seem to be a vital part of the process of quantifying a community, such estimates are difficult to obtain. Censusing a community is usually unfeasible, so instead samples are taken from the community and some form of extrapolation is carried out to estimate the number of unobserved species (e.g. Colwell & Coddington 1994). This can be achieved either by taking a single sample and using the distribution of abundances to estimate the species richness, or by taking several samples and using the distribution of incidences in the different samples to get the estimate. This paper will be concerned with the former method, although much of the discussion will carry straight over to the latter.

A great many approaches to the problem of estimating species richness have been suggested in the statistical literature (reviewed in Bunge & FitzPatrick 1993). Only a few of these methods have seen use in ecology, and the choice of the methods seems to be due largely to the availability of software. The ideal approach to estimating species richness would be to use the data (in the form of the number of species observed once, twice, etc.) to estimate the distribution of species abundances, and from this estimate the number of species that were present but observed zero times. The problem here is that the distribution of abundances is unknown, so we have to either find a distribution that is flexible enough to take in all reasonable distributions, but for which the parameters can be estimated with a reasonable degree of precision, or we have to find the 'true' abundance distribution (or a good approximation to it). Neither of these seems likely. The problem that has to be faced is whether different distributions will give similar estimates of species richness, i.e. whether the estimation is robust to model misspecification.

One model of abundances that is often used is the log-normal distribution (Preston 1948). This is despite well-documented problems in fitting it to data (e.g. Hughes 1986). A better alternative is the Poisson log-normal distribution, which preserves the underlying concept (that species' abundances are log-normally distributed), but places a model of the sampling from

Correspondence: R. B. O'Hara, Department of Mathematics and Statistics, PO Box 68 (Gustaf Hällströmin katu 2b), FIN-00014 University of Helsinki, Finland. Tel: +358 9 191 23743; mobile: +358 50 5990540; fax: +358 9 191 22 779; e-mail: boh@rni.helsinki.fi

these distributions on top of this (Cassie 1962). An alternative distribution can be obtained by assuming that the abundances follow a gamma distribution. When random sampling is then assumed, the distribution of the numbers of species captured can be shown to follow a negative binomial distribution (Fisher, Corbet & Williams 1943; White & Bennetts 1996). Despite their statistical appeal, and their use in modelling diversity (e.g. Kempton & Taylor 1974), the use of these distributions in estimating species richness has been rare (for an exception see Peterson & Meier 2003).

Non-parametric estimates of species richness have also been developed and used. These do not make any explicit distributional assumptions about the species abundances, but instead use approximations based on statistics calculated from the data. Of the two methods commonly used, the first (historically speaking) is only a lower bound of the number of species (Chao 1984, 1987). The second (Chao & Lee 1992; Chao, Ma & Yang 1993) is a genuine estimate, which uses the estimate for the case where all species are equally abundant, and corrects this with a term based on the coefficient of variation, i.e. on the first two moments of the abundance distribution (Bunge & FitzPatrick 1993).

An alternative parametric approach to estimating species richness is to construct species sampling curves. These are plots of the number of species in samples of different sizes (Gotelli & Colwell 2001), and can be constructed either by sequentially adding a new sample, and re-counting the number of species (an accumulation curve), or by taking independent samples of varying sizes (a rarefaction curve) constructed by taking subsamples randomly of different sizes from the sample. The number of species in the subsamples is then plotted against the sample size to produce a taxon sampling curve (Gotelli & Colwell 2001). The species richness is then estimated by fitting an equation to the curve, and estimating its asymptote, i.e. how many species there would be if an infinite number of individuals were collected. Although it may not be apparent, this method is parametric. This can be seen if we consider that the distribution of abundances of species can be used to predict the number of species in a sample of size N, and then can clearly be used to predict the shape of the curve of the number of species as N is changed–, i.e. the species accumulation curve (Christen & Nakamura 2000).

All these methods assume (as they have to) that there is a clearly definable community. In practice this community is defined informally through empirical and taxonomic means, i.e. how the organisms are trapped, and which groups the researchers are able to identify. Neither aspect of this definition defines a community operationally, although both may be a reflection of an operational definition. This would weaken any connection between theoretical studies of community structure and the sort of data that are collected. For example, because samples tend to be of whatever species is trapped, the community that is sampled is usually an open one, including any immigrants or tourist species that are found (for an exception see Longino, Colwell & Coddington 2002).

There is a related question of whether the estimate is intended to estimate the number of species at the time of sampling, or whether it is meant to be a wider statement about the number of species that could be in the community. The estimation methods imply the latter definition, as they assume an infinite number of individuals in the community – in other words, they count any species that could be found in the community. The species richness that is being measured is then the total number of species in the region, where (depending on the dispersal abilities of the organisms) the region may have to be defined at the continental, or even global, scale. For field data, this is probably not the quantity required, although it is clearly the relevant one for museum collection data (e.g. Peterson & Meier 2003).

A more relevant quantity might be the number of species in a community at any one time, i.e. an answer to the question 'How many species are there here now?'. In principle this can be estimated using parametric methods. If the population size of the community (i.e. the number of individuals in the community) is known or can be estimated, then the predicted number of species at this population size can be calculated. Sampling curves are simply extrapolated out to the population size (rather than to infinity). Distribution-based estimates can be calculated by taking the difference between the number of species at an infinite population size and the predicted number of zeroes in a sample of size equal to the population size (i.e. the number of species not present). Obtaining an estimate of the number of individuals in a community is clearly a problem, although any estimates are probably fairly robust to bias.

Even when a community has been defined suitably, different estimates of species richness may arise from different abundance distributions (Colwell & Coddington 1994). It is certainly important that a fitted curve should fit the data, but little attention has been paid to this in the literature. For example, Colwell & Coddington (1994; their Fig. 1) show a species accumulation plot and the fitted hyperbolic curve. However, the curve clearly overestimates the number of species over the middle of the range of pooled samples. One effect of this is to make the estimate of species richness dependent on sample size. The authors note that their estimate of species richness depends on sample size, but do not link this to a lack of fit (the effect of the larger sample is to pull the fitted curve up, without these the curve is lower, so the predicted species richness is lower). If the curve fitted the data then the change in the estimate would not be directional, but rather would be pulled up and down by randomness in the data.

Several authors have attempted to address the question of how accurate species richness estimators are. A common comparison is of the estimated species richness with a 'true', known species richness at different re-sample sizes (e.g. Colwell & Coddington 1994; Walther

& Morand 1998; Longino, Colwell & Coddington 2002; Brose *et al*. 2003; Foggo *et al*. 2003; Peterson & Meier 2003). With the exception of Brose *et al*. (2003) and Peterson & Meier (2003), the 'true' species richness is defined as the number of species in the sample, which seems to be prejudging the issue. Even when data sets are screened to try to ensure a full coverage (e.g. Walther & Morand 1998; Brose *et al*. 2003; Foggo *et al*. 2003) the comparisons cannot be regarded as foolproof, as the extremely rare species will still be difficult to capture. It is then perhaps not surprising that the studies conclude that the early methods of Chao (1984, 1987) work best as these, like the observed number of species, are a lower bound. Several studies have looked at the behaviour of the estimators by taking re-samples of different sizes, and examining how the point estimates change with re-sample size (e.g. Walther & Morand 1998; Foggo *et al*. 2003). The criterion for a method being good has generally been that it approaches the observed richness quickly, but this is a property of the particular sample and not of the population from which it has been taken. A correct criterion should look at whether independent samples give similar estimates, and in particular that the 95% confidence interval for the estimates contains the true value 95% of the time.

This paper compares the different estimation methods, and investigates some of their properties. The principle question is whether the estimation methods give the same answers, and if not, can we use the data to discard any of the estimates? Investigating this latter point means examining the properties of the estimators. Both field data and data simulated from the fit of models to the field data are used. The former has the advantage that it is realistic, and so likely to be similar to other field data. However, it is not possible to assess the bias of the estimates as the true number of species is unknown. Hence, simulated data are also used, with the simulated data being drawn from the distributions that are fitted to the data. The true values are then known, and are reasonable if one assumes that the distributional assumptions are reasonable.

Methods

ANALYSES

We assume that individuals are sampled randomly from a community. Let S^* be the actual number of species in the community, and the number of species in our sample that are caught r times be n_r ($r = 0, \ldots, \infty$), so that $S^* = \sum_{i=0}^{\infty} n_r$. We define S_{obs} as the observed number of species (i.e. $S_{obs} = \sum_{i=0}^{\infty} n_r = S^* - n_0$). Our aim is to estimate S^*, from the subset n_r ($r = 1, \ldots, 8$), or equivalently to estimate n_0 from the same subset. Ten estimators will be used here – two non-parametric, four parametric methods based on the maximum likelihood approach and four species accumulation estimators.

All the estimators assume that for each individual sampled, there is a probability x_i ($i = 1, \ldots, S^*$) that the individual is of species i. These probabilities follow some distribution $f(x)$, which may not necessarily correspond to the abundances of the species in the community, as the catchabilities may vary between species.

The non-parametric estimators are derived by using the moment properties of the distribution f(x) to derive an estimate of S^*. In contrast, the sampling curve and maximum likelihood estimators work by fitting the distribution $f(x)$ to the data (although the actual form of $f(x)$ is not made explicit in the sampling curve methods).

Non-parametric

Chao (1984, 1987) developed a moment estimator of the lower bound of species richness, which uses the information in the number of species sampled once and twice. The point estimate is:

$$S_{Chao1} = S_{obs} + \frac{n_1^2}{2(n_2+1)} - \frac{n_1 n_2}{2(n_2+1)^2} \qquad \text{eqn 1}$$

with variance (Chao 1987)

$$V_{Chao1} = n_2\left(\frac{1}{4}\left(\frac{n_1}{n_2}\right)^4 + \left(\frac{n_1}{n_2}\right)^3 + \frac{1}{2}\left(\frac{n_1}{n_2}\right)^2\right) \qquad \text{eqn 2}$$

Although this is a lower bound rather than an estimate, it has been claimed that it works well as an estimator (Chao 1984; Foggo *et al*. 2003). Later, a proper estimator was developed (Chao & Lee 1992; Chao, Ma & Yang 1993), which is now known as ACE ('Abundance Coverage Estimator'). This method uses an estimator for the case where all species are equally probable, and then adds a correction based on the variance of the distribution (Bunge & FitzPatrick 1993). The estimator was derived originally by Chao & Lee (1992), using all the data. An alternative derivation was provided by Chao, Ma & Yang (1993), who also suggested that only rare species should be used in the estimation. If rare species are defined as those with a frequency less than or equal to t, and S_{COMMON} as the number of species with more than t individuals sampled, then the estimate can be written as:

$$S_{ACE} = S_{COMMON} + \frac{\sum n_i}{1-\frac{n_1}{\sum n_i}} + \frac{n_1}{1-\frac{n_1}{\sum n_i}}\left(\frac{\sum n_i}{1-\frac{n_1}{\sum n_i}}\frac{\sum i(i-1)n_i}{(\sum i n_i)(\sum i n_i - 1) - 1}\right) \qquad \text{eqn 3}$$

where the summations go from 1 to t. An approximation of the variance can be found using the following result (Chao & Lee 1992), where again the summations go from 1 to t:

$$V_{ACE} = \sum_i \sum_j \frac{\partial S_{ACE}}{\partial n_i}\frac{\partial S_{ACE}}{\partial n_j} Cov(n_i, n_j) \qquad \text{eqn 4}$$

This was solved symbolically with the R statistical package (Ihaka & Gentleman 1996) – the resulting

equation is large. This estimator is N_2 in Chao & Lee (1992). An alternative was proposed if the coefficient of variation is large, for the data here the alternative method gave estimates that were higher by two to five species.

Chao *et al.* (1993) used $t = 10$ on more or less arbitrary grounds, and this value has persisted. However, there is no formal argument for choosing any particular value, so the sensitivity of the estimate with regards to t is investigated by plotting the estimates for different values of t against t. The estimates with $t = 8$ are also presented, and called S_{ACE2}.

The estimators are not parametric, so there is no model from which to calculate residuals. As an alternative approach to assessing the estimates, subsamples of different sizes were taken from the data. If the estimators are good, then the subsamples should give rise to estimates that are distributed around the true (unknown!) value, without showing any trend. The standard errors should also decrease with increasing subsample size. Here 100 subsamples were taken, of sizes evenly spaced from 10% to 90% of the total sample size. The predicted number of species were then plotted against the observed number of species.

Sampling curves

Rather than using accumulation curves (which plot the addition of new species as sampling size increases), rarefaction curves were used here in order to avoid serial dependence between samples. For 1000 re-sample sizes s, evenly spaced between 10% and 90% of the total number of individuals (Table 1), s individuals were drawn from the sample without replacement and the number of species, $S(s)$, in the subsample was counted. The species richness was then estimated as the asymptote of a curve fitted to the rarefaction curve, i.e. estimating the species richness when an infinite number of individuals is caught. Two curves were used, the Michaelis–Menton curve and an exponential curve (Colwell & Coddington 1994). Both the Michaelis–Menton and the exponential curves are forced through the origin, which may adversely affect their fit to the data (e.g. Figure 1 in Colwell & Coddington 1994). In order to improve the fit, the curves were also fitted with an intercept. This is not realistic (as it implies that several species will be present in a sample size of zero), but may provide a reasonable approximation to the data over the range of the subsamples. The model for the Michaelis–Menton curve without the intercept is:

$$S(s) = \frac{S_{MM}}{B/s + 1} \qquad \text{eqn 5}$$

and with the intercept;

$$S(s) = S_{MMI} - \frac{A}{s + C} \qquad \text{eqn 6}$$

The equations for the exponential curve are:

$$S(s) = S_{exp}(1 - e^{-Ks}) \qquad \text{eqn 7}$$

for the model without the intercept, and

$$S(s) = S_{expI} - De^{-Ls} \qquad \text{eqn 8}$$

for the model with the intercept, where A, B, C, K, D and L are nuisance parameters. There are several ways of parameterizing these relationships. Using these forms makes the focal parameter, i.e. the species richnesses (S_{MM}, S_{MMI}, S_{exp} and S_{expI}) explicit. All four curves were fitted to the re-sampled data sets by non-linear least squares (Draper & Smith 1998: chapter 24). This requires the assumption that the variance is constant, which does not seem to be seriously violated (see the Results section).

It is not trivial to obtain an estimator of the standard error. The standard errors of the estimates from the re-sampled curves (as used by Colwell & Coddington 1994; Foggo *et al.* 2003; Keating & Quinn 1998 for example) are not the correct measures, as these are estimates of the error in the point estimate due to the re-sampling

Table 1. Summaries statistics and species richness estimates (approximate standard error). Definitions of the estimators are given in the text

	Fort Augustus	Barnfield	Barnfield (without most common species)
Number of individuals	4489	1496	1131
S_{obs}	158	130	129
S_{Chao1}	174·2 (10·5)	148·9 (9·76)	147·9 (9·76)
S_{ACE}	170·0 (6·33)	151·0 (5·68)	150·0 (5·64)
S_{ACE2}	222·1 (19·56)	416·4 (112·8)	230·3 (28·98)
Accumulation curves			
S_{MM}	167·6 (3·42)	157·3 (3·75)	157·7 (3·59)
S_{MMI}	177·8 (3·04)	168·9 (3·57)	166·8 (3·47)
S_{exp}	148·5 (5·78)	124·8 (4·68)	124·5 (4·45)
S_{expI}	158·6 (3·13)	135·6 (3·61)	134·2 (3·53)
S_{NB}	267 (78·2)	1495 (2831)	586 (693·0)
S_{PLN}	176 (10·3)	185 (29·1)	176 (24·7)
S_{NBNB}	392 (151·1)	397 (285·5)	279 (204·6)
S_{NBLN}	480 (275·4)	221 (163·3)	159 (62·4)

process, and they can be decreased simply by increasing the number of re-samples taken. Here the estimated standard deviation, i.e. the residual mean square, is used instead. This can be derived by viewing this as a prediction problem – how many species would we collect if we took a sample of infinite size? The variation about this value is then the same as the variation about any prediction, i.e. the standard deviation.

The fit of the models to the re-sampled data was assessed by plotting the residuals against the re-sample size. If the model fits well, then there should be no pattern in the residuals. Any pattern, such as curvature in the residuals, suggests that the model is not correctly fitting to the shape of the curve. An alternative would be to plot the residuals against the predicted values. The simple, monotonic, shape of the curves means that the resultant plots are similar to the plots against re-sample size, and hence these are not shown here.

Maximum likelihood

Developing a distribution for the number of captured individuals should be done in two steps (e.g. Fisher, Corbet & Williams 1943; Kempton & Taylor 1974). At the first step it is assumed that capture rates, x, arises from some probability distribution, f(x); possible distributions are described below. The second step assumes that r, the number of individuals of a species with capture rate x that are caught, follows a Poisson distribution with mean Qx, where Q is the fraction individuals in the population sampled. The unconditional distribution of the proportion of the species caught r times is then

$$p_r = \int_0^\infty \frac{(Qx)^r e^{-Qx}}{r!} f(x)dx \qquad \text{eqn 9}$$

If the total number of species is S^*, then the expected number of species sampled r times is $E_r = S^* p_r$. If the captures are independent, then the actual number then follows a Poisson distribution with this mean.

The Poisson sampling model restricts the sampling variance. If the actual variance is larger, the data can be said to be over-dispersed. This can occur either because the capture distribution or the sampling distribution is incorrect. One way of relaxing the Poisson sampling model is to assume that the number of individuals of a species caught, r, with capture rate x follows a negative binomial distribution with mean Qx, and an extra over-dispersal term (α). The unconditional distribution of the proportion of the species caught r times is

$$p_r = \int_0^\infty \frac{\Gamma(\alpha + r)}{\Gamma(\alpha) r!} \frac{(Qx)^r \alpha^\alpha}{(Qx + \alpha)^{\alpha + r}} f(x)dx. \qquad \text{eqn 10}$$

One interpretation of this is that the rate follows a gamma distribution with the shape parameter equal to α. For all of the models, the number of species, S^*, is estimated by finding the value of S^* and the parameters of $f(x)$ which maximize the log-likelihood.

Assessing confidence in the estimate is more difficult, because the estimates of S^* and the other parameters are not independent (Kempton & Taylor 1974). One solution to this is to calculate the *profile likelihood*. For each value of S^*, the likelihood at the ML estimates of the other parameters is used as the likelihood for that value of S^* (e.g. Clayton & Hills 1993: chapter 13). If the likelihood is of the form of a ridge, then the profile likelihood is a line along the top of this ridge. The likelihood can then be plotted against S^*, which would then be a projection of the line onto the S^* dimension. If S^* were continuous, the Hessian (i.e. the second derivative of the profile likelihood curve) could be calculated. As S^* is discrete a heuristic approximation is calculated instead. A quadratic curve is fitted to the likelihoods for the profile ML estimate and the two adjacent values (i.e. S^*, $S^* - 1$ and $S^* + 1$). This fit is perfect (as there are three points, and three parameters to estimate), and the quadratic term is equivalent to the Hessian for a continuous surface. The standard error can be estimated by taking the square root of the inverse of the negative value of this term. The estimate is precise only if the likelihood genuinely is a quadratic surface (as it would be if it were normally distributed), although it will be a good approximation if there are sufficient data.

Two abundance distributions were used. The gamma distribution was first used by Fisher, Corbet & Williams (1943). With abundance distribution and a Poisson sampling model, the number of individuals of a species that are caught follows a negative binomial distribution. The proportion of species caught r times is

$$p_r = \frac{\Gamma(k + r)}{r!\Gamma(k)} \left(\frac{m}{k}\right)^r \left(1 + \frac{m}{k}\right)^{-r-k} \qquad \text{eqn 11}$$

so that there are two nuisance parameters to estimate, k and m. k is the shape parameter of the distribution, when $k = 1$ the abundances have a mode at zero, and m is the mean abundance. This species richness estimate is denoted S_{NB}, and the overdispersed version (with a negative binomial sampling model) is denoted S_{NBNB}.

An alternative is to follow Preston's reasoning and assume that the abundances follow a log-normal distribution. Using the Poisson sampling model we get

$$p_r = \frac{1}{r!\sigma\sqrt{2\pi}} \int_0^\infty e^{-x} x^{r-1} \exp\left(-\frac{1}{2\sigma^2}(\log x - \mu)^2\right) \qquad \text{eqn 12}$$

where μ and σ^2 are the mean and the variance of the natural log of the abundances. This is the Poisson log-normal distribution, and gives an estimator S_{PLN}. This approach avoids the problems of fitting the log-normal directly (e.g. Cassie 1962). The over-dispersed version is denoted S_{NBLN}.

All the distributions can be fitted to the data numerically – at least part of the reason for the unpopularity

of the Poisson log-normal distribution seems to have been due to difficulties in evaluating the integral in eqn 12, but this is no longer a serious problem with modern computers. The fit of the models was examined by residual plots. As the raw residuals from a Poisson distribution are difficult to assess, the deviance residuals (McCullagh & Nelder 1989, p. 39) were plotted against r.

DATA

The data used here are a part of that used by Kempton & Taylor (1974), and comes from large-scale light trapping of macrolepidoptera moths in the United Kingdom. Two of their data sets are used in this paper – Fort Augustus (Invernesshire, northern Scotland, data from 1969) and Barnfield (Rothamsted, southern England, data from 1971). Barnfield was chosen as it seems typical of most of the data sets analysed, and the most extensive results are presented by Kempton & Taylor. Fort Augustus was used to give a contrast, Kempton & Taylor found that the log series did not provide a good fit to the data. As the data sets used by Kempton & Taylor were not available, they had to be reconstructed from the original data and the tables provided by Kempton & Taylor (1974). Any differences between their estimates and the ones found here may be due to this, or due to the way that Kempton & Taylor aggregated the data before fitting the abundance distributions.

An initial examination of the data shows that the Barnfield sample is dominated by a single species (the Setaceous Hebrew Character, *Xestia c-nigrum*). As this species is so much more common than any other species, it may have an effect on the other estimates. Therefore, the estimates were also calculated for the Barnfield data set with this species removed.

As a guide, about 850 macrolepidoptera have been recorded in the United Kingdom, although these include species which are either extinct or occasional migrants. No estimate of the true species richness is available, so there is no way of seeing which of the different estimates is closest to the true value. Instead, data sets were simulated with the negative binomial and Poisson log-normal distributions, using the ML estimates (including the ML estimates of the species richness, Table 1) for the two data sets. a total of 1000 simulated data sets were created, and for each simulated data set the different species richness estimates were calculated. The results were summarized by estimating the mode of the distribution of point estimates, as well as the 95% highest density confidence interval. This is the interval where 95% of the density lies, such that each point has a higher density than points outside the interval.

Results

The abundance distributions for the two data sets are shown in Fig. 1. Both samples show the typical pattern of a J-shaped curve, with the singletons being the modal class. The Fort Augustus sample has about three times as many individuals than Barnfield, but only a few more species (Table 1). The most common species in the Barnfield sample is about three times as common as the next most common species (365 individuals as against 118), and seems to be an extreme species. The difference between the two most common species in the Fort Augustus sample is much less (255 and 317 individuals). It is this difference that motivates the examination of the effect of the most common Barnfield species.

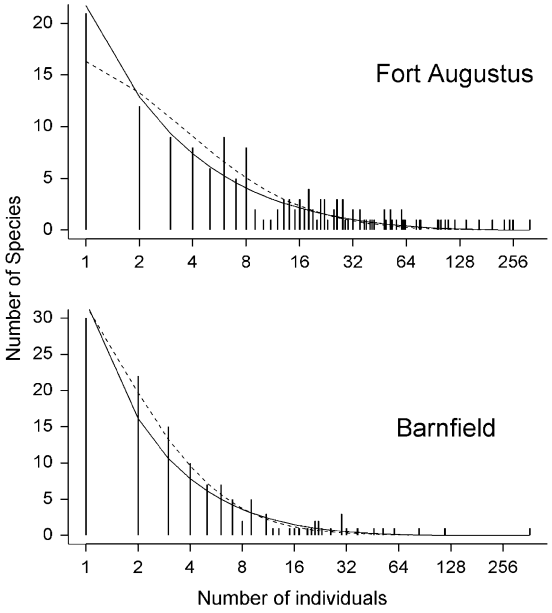

Fig. 1. Observed and fitted frequency distributions of species' abundances for two moth communities. Solid line: negative binomial. Dashed line: Poisson log-normal.

The different estimates of species richness are shown in Table 1. Most of the estimates are slightly higher than the number of species in the samples – for Fort Augustus this means that the methods suggest that between about five and 20 species are not in the sample; for Barnfield, between about 10 and 40 species are predicted as being present but not sampled. The exceptions to this come from ACE2, the negative binomial estimator and the two estimators based on the overdispersion model (all of which give much higher estimates of species richness), and estimates based on an exponential model fitted to the species accumulation curves, which with an intercept predict only slightly more species than present, and without an intercept manage to predict less species than sampled.

The predictions for the simulated data are shown in Fig. 2. In all cases, the only estimation method that gives a good estimate of the species number is the correct parametric estimate. The accumulation curve methods and the non-parametric methods all underestimate the true number of species. The negative binomial model gives the highest estimates, which considerably overestimates the number of species when the Poisson log-normal distribution is used and conversely, the Poisson log-normal distribution underestimates the species

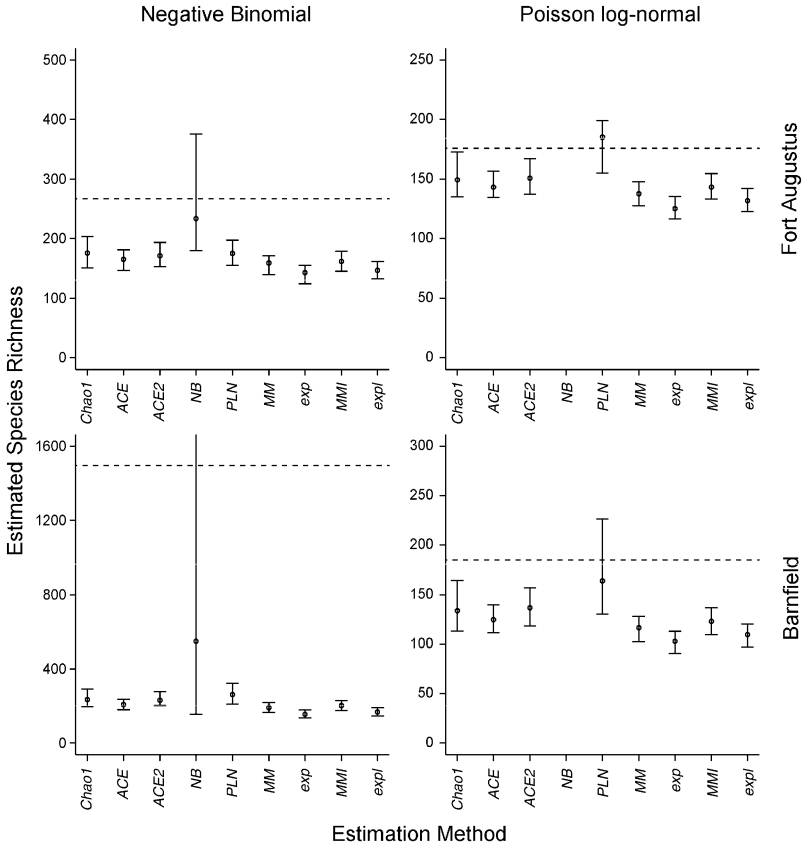

Fig. 2. Modes and 95% confidence intervals of point estimates of species richness for replicated simulated data from two distributions, and parameters estimated from data from moth abundances from Barnfield and Fort Augustus. Dotted line: true number of species (from simulation). Minima for negative binomial estimates for Poisson log-normal simulations: 514 for Barnfield, 304 for Fort Augustus. Definitions of the estimators are given in the text.

richness when the data are taken from a negative binomial distribution. It is worth noting that the order of the estimates is the same for all four simulated situations, and indeed the real data. This is despite the differences in the distributions and data sets that are used, and suggests that the patterns are due to the properties of the estimators themselves rather than the data.

It is curious that for Fort Augustus the ACEs estimates are smaller than the Chao1 estimates (Table 1). Chao1 is not strictly an estimate of the number of species, but an estimate of the minimum number of species (Chao 1984), so for these data sets the ACE appears to underestimate the species richness. However, using all the data in the ACE (i.e. ACE2) considerably increases the estimate. The effect of increasing the cut-off point (t) between rare and common species is initially to decrease, and then increase the estimate (Fig. 3), with little effect on the standard error. There seems no obvious point at which to place a cut-off, unless a lower bound rather than an estimate is required, in which case the smallest estimate should be preferred. The estimates follow the (re-)sample size (Fig. 4), which suggests strongly that the estimates are biased, and might be better used as lower bounds on species richness.

The accumulation curves and the residuals from the fitted curves are shown in Fig. 5. For all the fitted curves without an intercept term, the residuals are positive at low

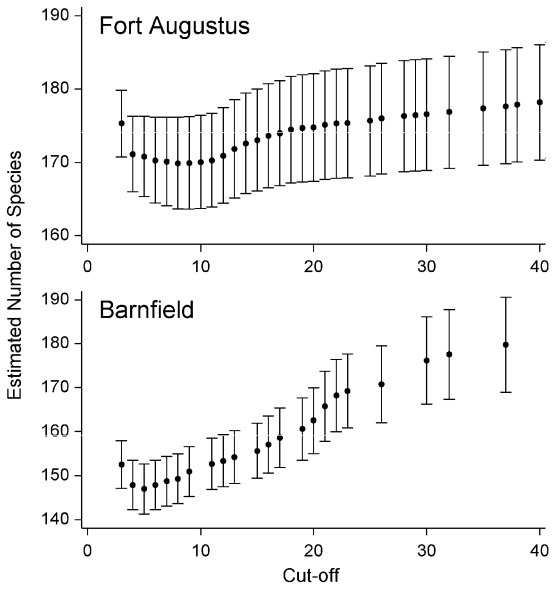

Fig. 3. Estimates of species richness from the ACE species richness estimator with different values of the common/rare cut-off point. Error bars are ± 1 standard error.

and high re-sample sizes, and negative at intermediate sizes. This is a clear indication that the models are not good fits to the data – if the models were correct, then there would be no structure in the residual plots. The

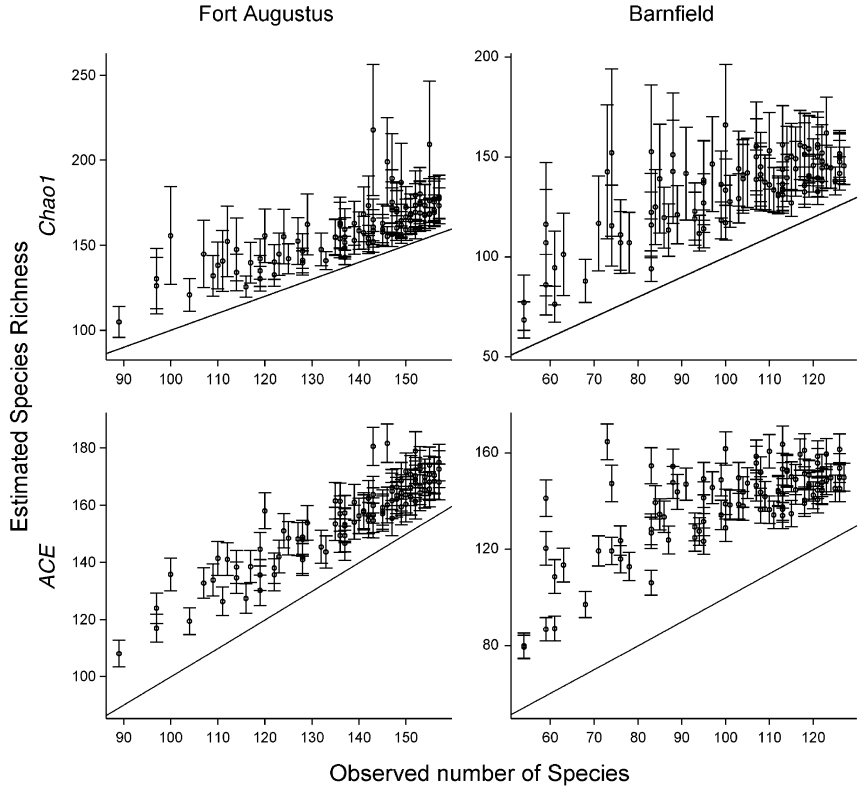

Fig. 4. Non-parametric estimates of species richness (± 1 standard error) from subsamples of different sizes from two moth communities. Solid lines are the 1 : 1 line.

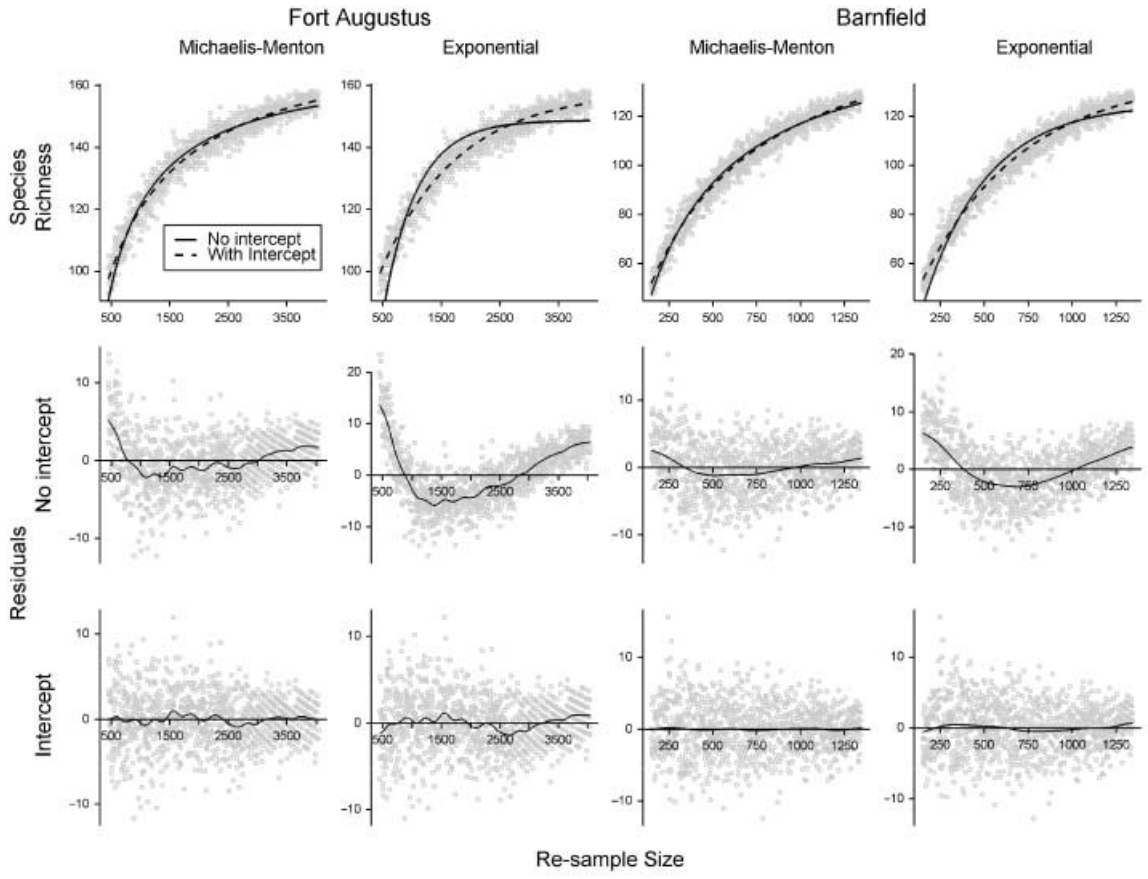

Fig. 5. Rarefaction plots (top) and residual plots of fitted sampling curves (middle: no intercept, bottom: with intercept) for data from two moth communities. Solid lines in middle and bottom rows: smoothed regression lines.

effect of this is that the accumulation curves are extrapolated so that the estimate of species richness is too low. The exponential curve fitted to the Fort Augustus data set is an extreme case, where the estimate is actually below the observed number of species.

When the intercept term is added the fits are much improved. The residual standard deviation (which is also the standard error of the species richness estimate, Table 1) is lower, and the residual plots show very little structure. The residuals do show a decrease in the variance as the re-sample size increases, although this does not seem to be large. This is opposite to the pattern assumed in the method developed by Raaijmakers (1987) that has commonly been used to fit the Michaelis–Menton model to species accumulation curves. The suggestion of Keating & Quinn (1998) that the variance will be largest at intermediate values is supported, as the sampling here does not include the lowest 10% of sample sizes, and the variance has to be zero at a sample size of 1. The residual variances for the two models with intercepts are very similar – indicating that both models fit to the data almost as well as each other. However, they still give different estimates of the species richness, with the exponential curve giving a smaller estimate than the non-parametric estimates, and the Michaelis–Menton curve giving estimates that are up to 20 species larger than the non-parametric estimates.

The profile likelihoods of the ML estimates are shown in Fig. 6. All of the profiles are positively skewed, even after a log transformation. The negative binomial distribution shows a maximum at a higher species richness (as already noted), but for Barnfield there is a large range of values with very similar likelihoods. The profiles for the Poisson log-normal distribution are more peaked, and have modes at smaller values. For Fort Augustus the maximum likelihoods are very similar, but for Barnfield the maximum is larger for the Poisson log-normal distribution; the likelihood ratio is almost 7000. The lack of fit is caused largely by the most common species – when this is removed, the likelihood ratio drops to 24·5. The effect of adding an overdispersal term (i.e. to change the sampling distribution from a Poisson to a negative binomial) is to massively flatten the profile likelihood. This is because the unexplained variation in the data is so large that the model allows for this excess variation. This also means that the standard errors are large (Table 1).

The fitted models are plotted in Fig. 1 and their residuals in Fig. 7. The residuals for Fort Augustus do not seem to show any pattern. In contrast, the residuals from the fit of the negative binomial distribution to the Barnfield data are sigmoid, being positive for lower abundances, and negative for higher abundances. This suggests that the lowest and highest values are pulling the fitted model away from the rest of the data. The effect of the most common species on the fit has already been examined. Overall, there is little difference between the fit of the two distributions to the Fort Augustus data, but the Poisson log-normal gives a better fit to the data from Barnfield.

Removing the most common species from the Barnfield data set has little effect on the non-parametric and accumulation curve estimates (Table 1). The effect on the ML estimates and the ACE2 estimates is then reduced considerably, in particular the estimate from the negative binomial distribution. This can be explained by the flatness of the profile likelihood – the estimated value is still well within the reasonable range for the full data set (Fig. 6). The effect of this species is reduced in the models with overdispersion, because the influence of outliers is reduced in these models (as a common species can be a common species with an additional effect of the extra variation allowed for in the overdispersion); however, there is still a reduction.

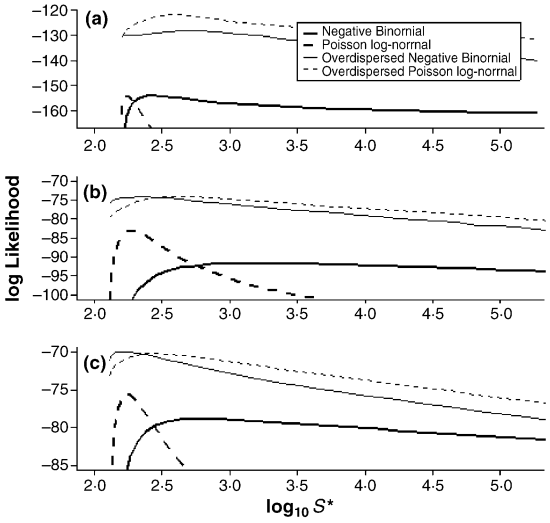

Fig. 6. Profile likelihoods for the estimated number of species in (a) Fort Augustus, (b) Barnfield and (c) Barnfield, with most frequent species omitted.

Discussion

The main conclusion to be drawn from the results is that different estimators give different estimates of species richness. When the data are simulated, only the distribution from which the data are simulated gives a correct estimate. Therefore, unless either the actual number of species is known (in which case the estimation is unnecessary), or if the form of the underlying sampling distribution is known the estimates cannot be relied upon. If a recommendation for which estimator to use has to be made, then either the negative binomial distribution or either of the distributions with overdispersion may be the most appropriate, simply because their standard errors seem to reflect most accurately the range of likely values.

Brose *et al.* (2003) recently examined the behaviour of incidence-based species richness estimators and showed that unless the abundance distribution was even, or the coverage (the proportion of species sampled) were high, the non-parametric estimators underestimated the true

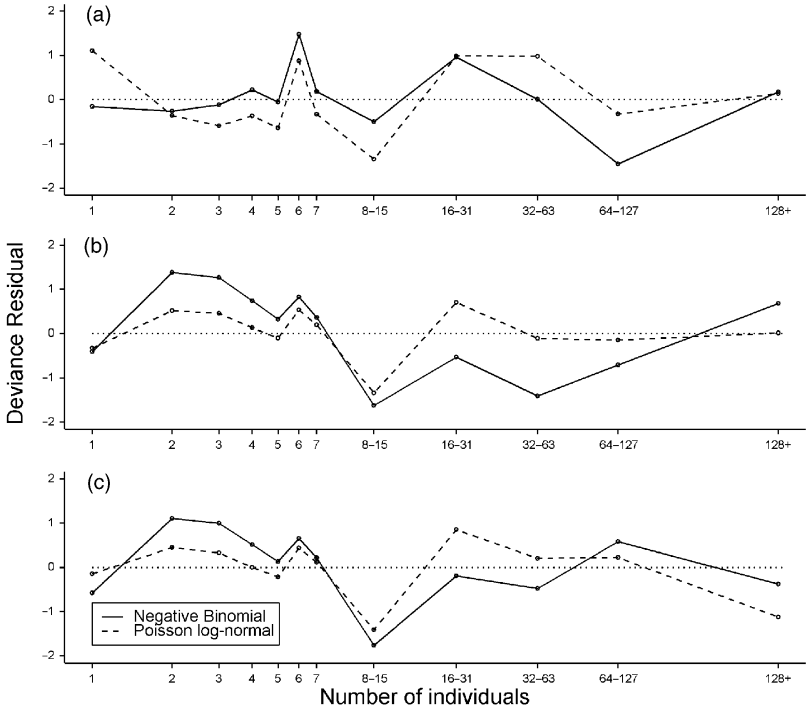

Fig. 7. Deviance residuals for fits of species abundance distributions to data on moth communities. Where the number of individuals is given as a range, the residuals were summed over that range. Dotted line: zero residual. (a) Fort Augustus, (b) Barnfield and (c) Barnfield, with most frequent species omitted.

species richness. They suggested a scheme for choosing an estimator based on taking an average estimated coverage from several methods. However, as shown here the estimates may all have an extremely large bias, so it will be difficult to trust the average estimate, and hence the chosen estimator. The problem, again, is that the only way to decide which estimator is good is to either know the correct answer in advance, or to know the underlying abundance distribution.

The non-parametric estimates and those based on extrapolating sampling curves all give similar estimates (Table 1, Fig. 2) with a similar ordering, regardless of the actual number of species. Their similarity is probably not a reflection of the true number of species, but is due to intrinsic differences between the properties of the estimators. This is apparent when the estimates for the simulated distributions are examined: regardless of the distribution, the same order is retained, with all estimates less than the true number.

It should perhaps not be a surprise that the non-parametric methods do not perform well – the derivation of Chao1 was as a lower bound (Chao 1984), and the ACE has also been shown to be an underestimate in simulations of negative binomial distributions (Chao & Lee 1992). In those simulations, the minimum value of k that was used was 1, whereas the highest estimates value here was 0·2 (for Fort Augustus; for Barnfield it was 0·024 and 0·076 when the most common species was removed). The two non-parametric methods used here were derived considering only the mean and variance of the empirical distribution. However, the distributions of the data used here are clearly positively skewed (Fig. 1), and this is probably a common feature of abundance distributions. Ignoring this may contribute to the bias.

The species curves methods assume an underlying abundance distribution (Christen & Nakamura 2000). However, the form of the abundance distribution may not be clear from the fitted curve, so any connections to ecological theory are obscure (although the exponential curve, eqn 7, can be derived from an even distribution of species). If the distribution is known then the estimator could be derived through the hierarchical approach used here for the gamma and log-normal abundance distributions, which should lead to a more efficient estimator (Bunge & FitzPatrick 1993). The exponential accumulation curve used here provides a clear example of this inefficiency, by providing impossible estimates.

The fit of parametric models can be checked using relatively simple techniques such as plotting residuals. If a model that has been fitted to data is unable to predict the data, then any predictions of new data should be viewed with extreme caution. Although the importance of these sorts of checks is emphasized in the statistical literature (e.g. Miller 1997), they are often missing from ecological analyses. Many estimates of species richness based on sampling curves may well be biased for this simple, and checkable, reason.

Other properties of the estimates are also important. Intuitively, the estimate of the number of unsampled species should depend largely on the number of rare species sampled. The Barnfield sample, however, tells a different story. The estimate from the negative binomial distribution shifts considerably when the most frequent species is removed. The ACE using all the data is also

sensitive to this species. A potential solution to this problem (at least for the parametric case) is to fit distributions which will allow flexibility in the upper tail of the distribution, while allowing the species richness to be influenced more by the rarer species. An alternative would be to follow the approach of Chao, Ma & Yang (1993), and use only the uncommon species in the estimation. However, it is unclear how uncommonness should be defined. Any cut-off will be arbitrary and placing it at the same abundance for all data sets does not seem reasonable, as it does not take into account variation in sampling effort. A better method might be to define common species as being the most frequent species that include a set proportion of the total number of individuals in the sample.

If a parametric model is to be used to estimate species richness, then the problem is choosing the right model. Finding such a model from ecological theory is difficult because the rate of capture of each species is a product of two quantities: the abundance of the species and the catchability, i.e. the rate of capture of each individual. If the catchabilities are identical for each species, then the fitted distribution will be an estimate of the abundance distribution. If the catchabilities are unequal, then the fitted distribution will be a biased estimate of the abundance distribution. If the catchabilities of each species can be estimated (for example, through a large-scale mark–recapture experiment), then the abundance distribution can be estimated. Even if the catchability can be modelled there is little good theory to guide the choice of which abundance distribution to use, so developments in this area (e.g. Hubbell 2001) seem necessary.

The large effects of the over-dispersion seen here are indicative of the two parameter models used not fitting well. The models with over-dispersion place all the extra variation into a random term, and so are the most flexible models. The cost of doing this is their large standard errors. An alternative would be to fit a three-parameter distribution, such as the generalized gamma distribution (Diserud & Engen 2000). Of course, this only moves the problem on one stage further: different three-parameter models will probably also give different estimates.

It should be clear from this discussion that it is difficult to estimate the number of species in a community. It appears that the only way to be sure that estimates are genuinely estimating species richness is the distribution of abundances of species in a community, and their catchabilties, is known. It seems unlikely that this will be possible, and hence unlikely that the bias of parametric estimates can be measured. Non-parametric estimators do not provide a solution as their bias is unbounded (Engen 1978). Estimating species richness therefore seems futile, as it is impossible to know how bad the estimates are. What, then, can be estimated? Chao1 is a lower bound to species richness and ACE seems to be an underestimate, so both of these can be regarded as providing lower limits to species richness. It is possible that an upper bound can be derived from ecological theory, by estimating the maximum number of species that a region can sustain, or if the empirical abundance distribution has an internal mode, in which case it might be expected that the number of unobserved species is less than the mode (but see Magurran & Henderson 2003 for an example of a bimodal data set). There seems little prospect for doing anything else than trying to provide tighter estimates of these bounds through refinements in ecological and statistical theory, with the hope that in the course of doing this a reliable estimator is discovered.

Acknowledgements

Many thanks to Ian Woiwod for providing the data, and to Johan Kotze and Tomas Roslin and three referees for their stimulating comments on the manuscript.

References

Brose, U., Martinez, N.D. & Williams, R.J. (2003) Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. *Ecology*, **84**, 2364–2377.

Bunge, J. & FitzPatrick, M. (1993) Estimating the number of species: a review. *Journal of the American Statistical Association*, **88**, 364–373.

Cassie, R.M. (1962) Frequency distribution models in the ecology of plankton and other organisms. *Journal of Animal Ecology*, **31**, 65–92.

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

Chao, A. (1987) Estimating the population size for capture–recapture data with unequal catchability. *Biometrics*, **43**, 783–791.

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., Ma, M.-C. & Yang, M.C.K. (1993) Stopping rules and estimation for recapture debugging with unequal failure rates. *Biometrika*, **80**, 193–201.

Christen, J.A. & Nakamura, M. (2000) On the analysis of accumulation curves. *Biometrics*, **56**, 748–754.

Clayton, D. & Hills, M. (1993) *Statistical Models in Epidemiology*. Oxford University Press, Oxford, UK.

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

Diserud, O.H. & Engen, S. (2000) A general and dynamic species abundance model, embracing the lognormal and the gamma models. *American Naturalist*, **155**, 497–511.

Draper, N.R. & Smith, H. (1998) *Applied Regression Analysis*, 3rd edn. Wiley, New York.

Engen, S. (1978) *Stochastic Abundance Models*. Chapman & Hall, London, UK.

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.

Foggo, A., Attrill, M.J., Frost, M.T. & Rowden, A.A. (2003) Estimating marine species richness: an evaluation of six extrapolative techniques. *Marine Ecological Progress in Series*, **248**, 15–26.

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

Hubbell, S.P. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, New Jersey, USA.

Hughes, R.G. (1986) Theories and models of species abundance. *American Naturalist*, **128**, 879–899.

Ihaka, R. & Gentleman, R. (1996) R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**, 299–314.

Keating, K.A. & Quinn, J.F. (1998) Estimating species richness: the Michaelis–Menton model revisited. *Oikos*, **81**, 411–416.

Kempton, R.A. & Taylor, L.R. (1974) Log-series and log-normal parameters as diversity discriminants for the Lepidoptera. *Journal of Animal Ecology*, **43**, 381–399.

Longino, J.T., Coddington, J. & Colwell, R.K. (2002) The ant fauna of a tropical rain forest: estimating species richness three different ways. *Ecology*, **83**, 689–702.

Magurran, A.E. & Henderson, P.A. (2003) Explaining the excess of rare species in natural species abundance distributions. *Nature*, **422**, 714–716.

McCullagh, P. & Nelder, J.A. (1989) *Generalized Linear Models*, 2nd edn. Chapman & Hall, London, UK.

Miller, R.G. (1997) *Beyond ANOVA*. Chapman & Hall/CRC, Boca Raton, Florida.

Peterson, F.T. & Meier, R. (2003) Testing species-richness estimation methods on single-sample collection data using the Danish Diptera. *Biodiversity and Conservation*, **12**, 667–686.

Preston, F.W. (1948) The commonness, and rarity, of species. *Ecology*, **29**, 254–283.

Raaijmakers, J.G.W. (1987) Statistical analysis of the Michaelis–Menton equation. *Biometrics*, **43**, 793–803.

Walther, B.A. & Morand, S. (1998) Comparative performance of species richness estimation methods. *Parasitology*, **116**, 395–405.

White, G.C. & Bennetts, R.E. (1996) Analysis of frequency count data using the negative binomial distribution. *Ecology*, **77**, 2549–2557.

Received 4 December 2003; accepted 25 September 2004