Journal of Animal Ecology 2003
72, 888–897

FORUM

The species–accumulation curve and estimation of species richness

KARL I. UGLAND, JOHN S. GRAY and KARI E. ELLINGSEN
Department of Biology, University of Oslo, PB 1064 Blindern, 0316 Oslo, Norway

Summary

1. One of the general characteristics of ecological communities is that the number of species accumulates with increasing area sampled. However, it is important to distinguish between the species–area relationship and species accumulation curves. The species–area relationship is concerned with the number of species in areas of different size irrespective of the identity of the species within the areas, whereas the species accumulation curve is concerned with accumulation rates of new species over the sampled area and depends on species identity.

2. We derive an exact analytical expression for the expectance and variance of the species–accumulation curve in all random subsets of samples from a given area. The analytical species accumulation curve may be approximated by a semilog curve. Both the exact accumulation curve and its semilog approximation are independent of the underlying species abundance distributions, but are influenced strongly by the distribution of species among the samples and the spatial relationship of the samples that are randomized.

3. To estimate species richness in larger areas than that sampled we take account of the spatial relationship between samples by dividing the sampled area into subareas. First a species–accumulation curve is obtained for randomized samples of all the single subareas. Then the species–accumulation curve for all combinations of two subareas is calculated and the procedure is repeated for all subareas. From these curves a new total species (T–S) curve is obtained from the terminal point of the subarea plots. The T–S curve can then be extrapolated to estimate the probable total number of species in the area studied.

4. Data from the Norwegian continental shelf show that extrapolation of the traditional species–accumulation curve gave a large underestimate of total species richness for the whole shelf compared with that predicted by the T–S curve. Application of non-parametric methods also gave large underestimates compared with actual data obtained from more extensive sampling than the data analysed here. Although marine soft sediments sampled in Hong Kong were not as variable as those from the Norwegian shelf, nevertheless here the new method also gave higher estimates of total richness than the traditional species–accumulation approaches.

5. Our data show that both the species–accumulation curve and the accompanying T–S curve apply to large heterogeneous areas varying in depth and sediment properties as well as a relatively small homogeneous area with small variation in depth and sediment properties.

Key-words: species accumulation, analytical expression.

Journal of Animal Ecology (2003) **72**, 888–897

Introduction

One of the primary goals of field studies in ecology is to estimate how many species of a given taxon (or all taxa) occur in an area. Initially many species are found as larger areas are sampled and a plot of accumulated number of species against area sampled rises steeply at first and then more slowly as the increasingly rare species are added. The species–accumulation curve may approach an asymptote for data sets of species that can be identified easily, such as of plants and breeding birds in Britain (Rosenzweig 1995) or tropical tree species in

Correspondence: John S. Gray, Department of Biology, University of Oslo, PB 1064 Blindern, 0316 Oslo, Norway.
E-mail: j.s.gray@bio.uio.no

50-ha plots (Hubble 2001) where it is possible to obtain a count of all the species present (Colwell & Coddington 1994). For other habitats (or taxa) one cannot expect to count all the species. Examples are estimating numbers of species of beetles in tropical forest tree canopies (Erwin 1988, 1991) or species inhabiting coral reefs or marine sediments. In such habitats all that can be done is to estimate total species richness and the sampling effort needed to obtain reliable estimates of this richness. Here we focus on fitting species–accumulation curves for one of these difficult habitats for which we have extensive data, namely marine soft sediments.

Gotelli & Colwell (2001) state that for patchy distributions the individual-based rarefaction curves 'inevitably overestimates the number of species that would have been found with less effort'. They suggest that it is preferable to use sample-based species–accumulation curves that take account of between sample heterogeneity, a view with which we entirely concur. Sample-based species–accumulation curves are plotted from samples taken randomly within a given area (Gotelli & Colwell 2001), and take into account the number of species and their identity, but no information of the distribution of individuals among species is utilized.

The order that samples are added to the species–accumulation curve affects the shape of the curve produced. This variation in the shape of the curve results from sampling error and heterogeneity among the species in the samples (Colwell & Coddington 1994). To overcome this problem various sample randomization procedures have been developed (Colwell & Coddington 1994; Gray *et al.* 1997; Colwell 2001; Gotelli & Entsminger 2001). The traditional method of plotting a species–accumulation curve starts by calculating and plotting the mean number of species (and its standard deviation (SD)) of the smallest sample size. Then all combinations of the next sample size are randomized and the mean cumulative number of species is calculated. This procedure is followed for all sample sizes. For the randomized sample data, once a curve has been obtained it can be used to estimate species richness. The traditional method is simply to extrapolate a parametric model for the species–accumulation curve to a larger area for which an estimate is needed. Here we develop an analytical method which gives exact cumulative numbers of species and so obviates the need for randomization using Monte Carlo techniques and curve fitting.

In the traditional method the curve for randomized data takes account of variance in number of species between samples, but does not take into account the fact that within the total area sampled there may be heterogeneity between subareas such that one subarea is species-rich, another subarea species-poor and intermediate subareas of moderate richness. By randomizing over all samples such heterogeneity is ignored. Here we take a different approach by recognizing that heterogeneity in species richness can occur within subareas and that this may have important consequences for estimating species richness.

Consideration of the covariance structure between species and between subareas leads to a largely unrecognized aspect of predicting numbers of species in large areas, namely with the addition of new subareas the new species–accumulation curve will not only cover a larger area, but will usually also lie above that for one subarea taken alone. It is the rate of increase of this new (and subsequent) species–accumulation curve as more subareas are combined which leads to the best estimate of total species richness, and this may be considerably higher than richness estimates from application of the traditional species–accumulation curve approach.

We first develop the analytical approach to the species–accumulation curve and then illustrate the importance of taking into account covariance structure of species between subareas.

The analytical species–accumulation (ASA) approach

In the following study we utilize two data sets. One is from the Norwegian continental shelf of benthic species occurring in five areas from a transect of 1960 km covering 809 species from a total of 101 samples, each of 0·5 m^2 (Ellingsen & Gray 2002). The other is from an extensive survey of the benthos of soft-bottom sediments over a spatial sampling scale of *c.* 1100 km^2 in the territorial waters of the Hong Kong SAR, China. The data are from 101 samples each of 0·5 m^2, and contain 386 species (Shin unpublished).

ANALYTICAL EXPRESSIONS FOR THE SPECIES ACCUMULATION CURVE, ITS VARIANCE AND SEMILOGARITHMIC APPROXIMATION

Arrhenius (1921) was the first to fit a model to data on the increase in number of species with increased size of the area sampled (see McGuinness 1984 for a good review of the history of the species–area curve). In this paper Arrhenius (1921) remarked explicitly that his power formula, $S = S(A) = dA^Z$, was empirical and should be regarded as an approximation whose existence was entirely dependent on agreement with data from lists of flora that he had obtained. Because his formula calculated the average number of species growing in an area, Arrhenius also posed the problem of establishing a stochastic model for species richness in a subarea a of a large area A. In order to relate area to species occurrence he assumed that any individual of any species has the same probability a/A of occurring in the subarea a. Now, if n_i is the number of individuals belonging to the i-th species, the probability that the i-th species is not found in the subarea is:

$$q_i(a) = \left(1 - \frac{a}{A}\right)^{n_i} \qquad \text{eqn 1}$$

Thus the probability that the i-th species grows in the subarea is

$$p_i(a) = 1 - q_i(a) = 1 - \left(1 - \frac{a}{A}\right)^{n_i} \qquad \text{eqn 2}$$

As 'the sum of these probabilities is the probable number of species on the area', Arrhenius (1921) obtained the following analytical expression for the expected species number in the subarea a:

$$E[S(a)] = p_1(a) + p_2(a) + \cdots + p_{S_{tot}}(a) = \sum_{i=1}^{S_{tot}} \left[1 - \left(1 - \frac{a}{A}\right)^{n_i}\right] = S_{tot} - \sum_{i=1}^{S_{tot}} \left(1 - \frac{a}{A}\right)^{n_i} \qquad \text{eqn 3}$$

It is important to note here that Arrhenius modelled the occupancy of each species as random placement of plants with replacement, but from a statistical viewpoint this expression rests on a limit theorem for the hypergeometric distribution; or as Feller (1950) states: 'For large populations there is practically no difference between sampling with and without replacement.'

The obvious generalization to sampling with replacement is simply to use the hypergeometric distribution rather than the binomial used by Arrhenius. Similar approaches have been taken more or less independently by several authors who examined the distribution of individuals and presence/absence pattern of species (e.g. Hurlbert 1971; Heck *et al*. 1975; Brewer & Williamson 1980; Coleman 1981; Ney-Nifle & Mangel 1999). All these proposed formulae may be regarded as variants of Arrhenius's (1921) original model.

In sharp contrast to the calculation of expectation, it is much more difficult to find analytical expressions for the variance. Heck, Belle & Simberloff (1975) suggested the formula for the variance of rarefaction curves based on individuals. The problem of finding a corresponding formula for rarefaction based on samples is more complex and is still unsolved. In their study of the presence/absence structure of the species in the assembly, Ney-Nifle & Mangel (1999) used the hypergeometric formula for the presence of species under sampling a cells without replacement. However, in order to obtain an expression for the variance, they derived Arrhenius' formula for $q_i(a)$ by passing to the limit when A is much larger than the number of cells and all the occupancies n_i are much lower than the total number of units (cells): $n_i << A$. Having made the transition to Arrenhius's (1921) model (sampling with replacement) Ney-Nifle & Mangel (1999) derived the variance of the species number in subarea a by regarding it as a sum of independent binomial trials (Feller 1950):

$$\mathrm{Var}\,[S(a)] = \sum_{i=1}^{S_{tot}} p_i(a)q_i(a) = \sum_{i=1}^{S_{tot}} \left[1 - \left(1 - \frac{a}{A}\right)^{n_i}\right]\left(1 - \frac{a}{A}\right)^{n_i} \qquad \text{eqn 4}$$

However, this procedure cannot be applied to accumulation curves because the assumption that all $n_i << A$ means that all species must be very rare in the samples. As such communities are practically never seen, this assumption should be avoided.

We now turn to rarefaction based on samples. Our problems are to find an analytical expression for (1) the average accumulation curve under all possible permutations of all the samples, i.e. the analogue to Hurlbert's (1971) formula for individual based accumulation curves, and (2) the corresponding variance, i.e. the analogue to Heck *et al*.'s (1975) formula. The advantage with such analytical expressions is that it will be no longer necessary to re-sample with Monte Carlo techniques.

It is important to note that the accumulation curve of species richness in samples is defined as the average number of species under all possible permutations of these samples. As such the samples are independent as they may occur anywhere in a random permutation. Thus the accumulation curve will simply be the average species number in a random collection of a samples from the existing A samples (i.e. a is one of the integers 1, 2, 3, … , A). When this is realized, Problem 1 becomes trivial. First we need a minor redefinition of the parameters: A is now the total number of samples, while a is the size of the random subsample, i.e. a is one of the integers 1, 2, … , A, and n_i is the number of samples where the i-th species is observed. Further, the number of observed species in all the A samples is denoted S_{obs}. Because the number of any species in a random subcollection has a hypergeometric distribution (Feller 1950), the arguments given by Arrhenius (1921) (i.e. using indicator functions whose expectation is the probability of occurrence) leads directly to an exact expression for the expected number of species:

$$E[S(a)] = \sum_{i=1}^{S_{obs}} \left[1 - \frac{\binom{A - n_i}{a}}{\binom{A}{a}}\right] = S_{obs} - \sum_{i=1}^{S_{obs}} \frac{\binom{A - n_i}{a}}{\binom{A}{a}} \qquad \text{eqn 5}$$

Note that this formula for the average accumulation curve under all possible permutations of the samples is an Arrhenius type of expectation under sampling without replacement, and therefore has the same form as the curves obtained by random placement of species over a grid as developed by (Hurlbert 1971; Coleman 1981; Williamson 1988; and Ney-Nifle & Mangel 1999). Expanding the binomial ratio (the probability of not being present in a randomly selected sample):

$$q_i(a) = \frac{\binom{A - n_i}{a}}{\binom{A}{a}} = \frac{\frac{(A - n_i)!}{a!(A - n_i - a)!}}{\frac{A!}{a!(A - a)!}} = \left(1 - \frac{a}{A}\right)\left(1 - \frac{a}{A - 1}\right)\cdots\left(1 - \frac{a}{A - (n_i - 1)}\right) \qquad \text{eqn 6}$$

we arrive at an expanded form for the average species number at each stage of the accumulation:

$$E[S(a)] = S_{obs} - \sum_{i=1}^{S_{obs}} q_i(a) = S_{obs} - \sum_{i=1}^{S_{obs}} \left(1 - \frac{a}{A}\right)\left(1 - \frac{a}{A - 1}\right)\cdots\left(1 - \frac{a}{A - (n_i - 1)}\right) \qquad \text{eqn 7}$$

From this expression we see that if the total number of samples is large in comparison with the occupancy number for each species, $n_i << A$, then:

$$q_i(a) \approx \left(1-\frac{a}{A}\right)\left(1-\frac{a}{A}\right)\cdots\left(1-\frac{a}{A}\right)=\left(1-\frac{a}{A}\right)^{n_i} \quad \text{eqn 8}$$

which is Arrenhius's (1921) expression. However, this transition is biologically unrealistic as many species are likely to be represented in more than 10–20% of the all the samples, so the requirement that n_i/A shall be close to zero is seldom fulfilled and must therefore be avoided.

For subsequent analysis we need a more compact form of the expression for the accumulation curve. Let R_k be the number of species that are observed in exactly k samples. Note that the sum of the R_ks is S_{obs}, so the frequencies, R_k/S_{obs} ($k = 1, \ldots, A$) form a probability distribution which contain information about how the species are represented over the various samples, and therefore we call this the 'representation distribution'. The species–accumulation curve is defined as the expected number of species, $E(S_a)$, in a random subset of $a = 1, \ldots, A$ samples. In Appendix I we show that the average number of observed species in any subset of samples may be expressed in terms of the species number S_{obs} and the representation distribution:

$$E[S(a)] = S_{obs}\left[1-\sum_{k=1}^{A-a}\left(1-\frac{a}{A}\right)\left(1-\frac{a}{A-1}\right)\cdots\left(1-\frac{a}{A-(k-1)}\right)\frac{R_k}{S_{obs}}\right] \quad \text{eqn 9}$$

The covariance structure between the species occurrences is complicated in the general case. Further, it is not permissible to reduce the problem to a binomial case by letting the sample size tend to infinity. Therefore the only possibility is to use combinatorial arguments in the development of an exact formula for the covariance. Because this is a purely statistical exercise, we put all the arguments and the final analytical expression into Appendix I.

Similarly, it is also complicated to find a compact approximation formula to the exact species accumulation curve. We have solved this by using a combination of linear regression, numerical analysis and approximation by infinite series. Because this is a purely mathematical exercise, we put all the arguments and the final analytical expression into Appendix II.

Results

Figure 1 shows the comparison of the species accumulation curves produced by the analytical formula and simulation by randomizing the samples with EstimateS (Colwell 2001). The two curves are almost exactly similar thus the new analytical method is to be preferred to the randomization method.

Figure 2 shows results for the standard deviations of the analytical formula and the randomization procedure (EstimateS) for the two areas studied. Initially, with small samples estimates of species richness vary considerably but after *c*. 20 samples both methods settle down, although of course the exact analytical method has more constant variance.

In order to approximate the accumulation curve we first derive an approximation to the representation distribution. This utilizes the empirical form of the representation distribution found in the Norwegian shelf (Fig. 3). Note that the observed representation distribution follows a straight line on a log–log scale ($R^2 = 0{\cdot}94$). Appendix II then shows how this implies a semilogarithmic form of the sample-based species–accumulation curve.

ESTIMATING TOTAL SPECIES RICHNESS IN AN AREA

The traditional way of estimating species richness of areas larger than that sampled is to randomize samples and plot the species–accumulation curve and then fit a

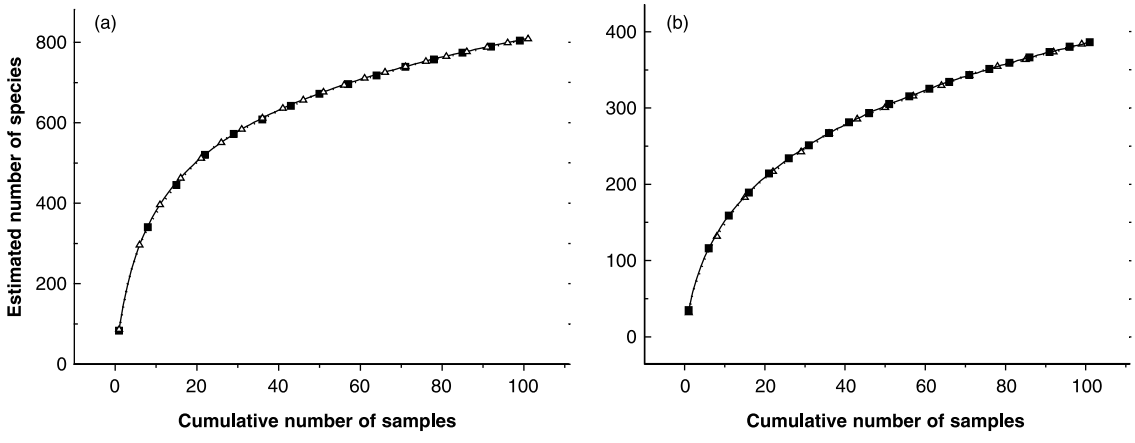

Fig. 1. Comparison of the species–accumulation curves produced by the analytical formulae and simulation by randomizing the samples with EstimateS. (a) The North Sea; (b) Hong Kong. Not all the data points are shown for clarity. △ = Analytical model; ■ = EstimateS.

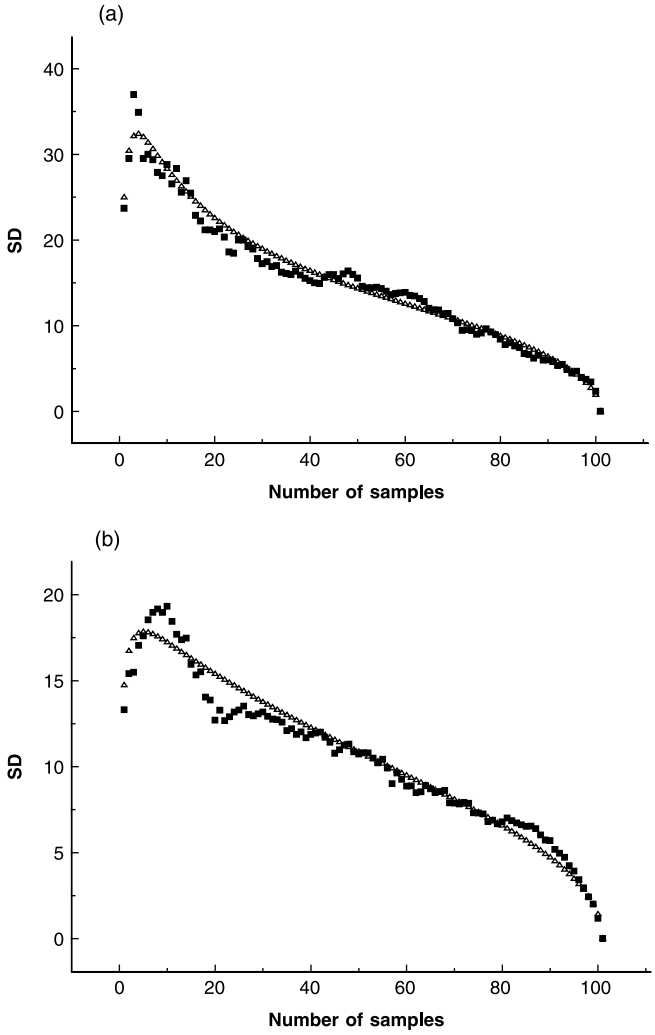

Fig. 2. Comparison of the standard deviations of the species–accumulation curves produced by the analytical formulae and simulation by randomizing the samples with EstimateS. (a) Norwegian continental shelf; (b) Hong Kong. △ = Analytical model; ■ = EstimateS.

Table 1. Summary of estimates of total species richness for the Norwegian continental shelf and Hong Kong SAR

Area	Species observed	Extrapolation of species–accumulation curve	Chao 2	ICE	T–S curve
Norwegian shelf	809	1192	1035	989	5403
Hong Kong SAR	386	780	549	530	2254

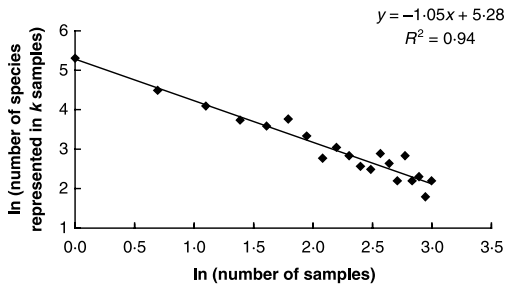

Fig. 3. The number of species, R_k, represented in $k = 1, \ldots, 20$ samples on a log–log scale in the pooled 101 samples from the Norwegian continental shelf.

model to the curve (see Bunge & Fitzpatrick (1993) and Colwell & Coddington (1994) for a review). For example, Colwell & Coddington (1994) used a hyperbolic model to extend the species–accumulation curve. Here we use the semilog estimate, for the areas covered by samples for the Norwegian shelf and Hong Kong (50·5 m^2 in both cases) are, respectively: Norwegian shelf 3773 species and Hong Kong 1745 species, Table 1.

To estimate the number of species in the whole area of the Norwegian shelf we have used the following method. The data consist of 101 samples from five large areas. We want to find the average species–accumulation

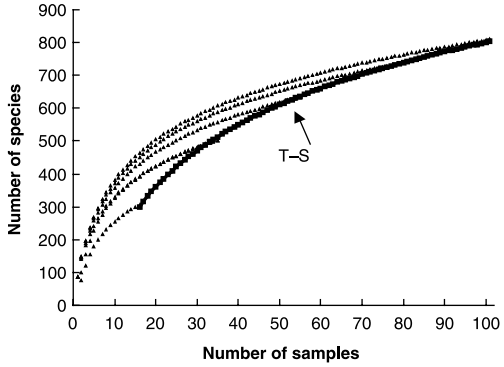

Fig. 4. The species–accumulation curves for all combinations of, respectively, one to five subareas surveyed on the Norwegian continental shelf. The total species projection curve (T–S) is a smooth curve through the average total number of species in all five combinations of the five areas.

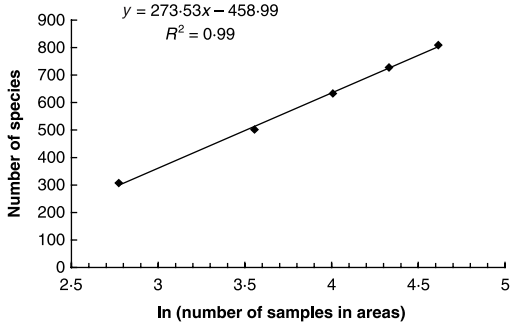

Fig. 5. Regression of the average number of species in all combinations of, respectively, one, two, three, four and five subareas from the Norwegian continental shelf against the logarithm of the number of samples in each of the subareas.

curves for the total area sampled (sum of the five subareas); that is, the average values when one, two, three, four or five subareas are drawn at random. The number of possible combinations are: five for one subarea, 10 for two subareas, 10 for three subareas, five for four subareas and just one possibility for the total area. Because the samples in the five subareas have the following number of samples, respectively, 16, 19, 20, 21 and 25, the sample allocated to the combination of one, two, three, four and five areas will be 16, 16 + 19 = 35, 16 + 19 + 20 = 55, 16 + 19 + 20 + 21 = 76 and all stations = 101. First calculate the species–accumulation curve for all combinations of one single subarea. This gives the bottom curve in Fig. 4 and terminates at 307 species. Thus 307 is the average of five ways to draw one subarea from five subareas. Then we constructed a curve for all combinations of two subareas and this gives 502 species, which is the average of 10 ways to combine two subareas from five subareas and so on, until finally 809 is the average of only one number (as there is only one way to combine all five subareas). The most interesting feature of our procedure is that the species–accumulation curve is steeper each time a new combination of subareas is included. Thus a separate curve is derived for successive random combinations of subareas one to five. We postulate that the reason for this increasing steepness is that the incorporation of samples from a new subarea will add relatively many new species and thereby increase the average number of species per combined sample.

In order to estimate the total species richness in a large area one needs to find a model that will estimate this upward shift of the species–accumulation curve. The rate of increase in the total species–accumulation curve (T–S) is the increment in the ratio between the current species number and logarithm of the number of samples, i.e. the rate of change of S/ln(c). When the subareas are pooled, the species number increases faster than the logarithm of number of samples, and therefore the species–accumulation curve is raised.

Figure 5 shows the species numbers from the terminal point of each of the subarea plots (307, 502, 633, 728 and 809) plotted against the logarithm of the corresponding number of samples (16, 35, 55, 76 and 101). Over 99% of the variability is explained and therefore justifies why we believe the extrapolation is a semilog function, here of the form:

$$S = 273{\cdot}5 * \ln(a) - 459 \quad (10)$$

Although we are acutely aware of the dangers of extrapolation beyond the data set, it is of interest to see whether or not our method gives realistic results. Our best estimate of species richness in a larger area than that sampled is obtained by use of equation 10. We predict that if a new area of the same size as that of the total area sampled is added, the total number of species will increase by 273·5 * ln(2) = 190, giving 999 species. In order to make an estimate of the total species richness of the continental shelf we make the following assumptions. The total area sampled from the Norwegian continental shelf is 50·5 m^2. One sample is the sum of five replicated grabs, the sampling units, each of 0·1 m^2. Here we assume that because the grab was on a wire of up to 434 m (the maximum depth sampled) the five sampling units are probably from an area of the order of 100 m^2. The samples themselves were collected from five large subareas covering in total *c*. 81 300 km^2. Assuming that one sample is representative of 100 m^2, an extrapolation yields 273·5 * ln(813 000 000) – 459 = 5152 species. Furthermore, assuming that the total shelf area is 2·5 times the area covered by the five subareas, the species richness on the whole continental shelf is estimated to be 5152 + 273·5 * ln(2·5) = 5403. As with the randomization procedure and curve-fitting procedure, as yet there are no means of estimating the variance of the extrapolated estimate of species richness.

Figure 6 shows an analysis of data from marine sediments from coastal areas of Hong Kong. The data are extensive, covering 101 samples of 0·5 m^2, a sampled area of 50·5 m^2 with a total of 386 species. We divided the data set into five subareas based on knowledge of

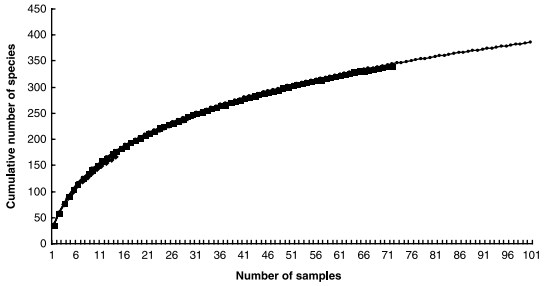

Fig. 6. Data from Hong Kong calculated as in Fig. 4 showing little variability between five subareas. Number of points reduced for clarity.

the hydrographic conditions and geography. The results are shown in Fig. 6. While there is clearly much less heterogeneity in species richness between subareas to that shown for the Norwegian continental shelf, nevertheless the T–S curve leads to higher predictions of richness compared with the traditional species–accumulation curve, Table 1. The total area sampled is 50·5 m^2 and we calculate the total area of sediment in coastal Hong Kong coastal to be 1088 km^2, giving an estimate of 2254 species. Thus, even though there is considerably less variability between subareas nevertheless our new method leads to a quite considerably increased estimate of species richness.

Non-parametric methods also have been used extensively to estimate total species richness of an area (see Colwell & Coddington's 1994 review). Here we have used just two, Chao 2 (Chao 1984) and ICE (Chao *et al.* 1993). Applying these methods to the continental shelf data (Table 1) the Chao 2 estimate gave 1035 ± 42 and the ICE estimate 989 species. For Hong Kong the estimates for Chao2 and ICE are 549 ± 43 and 530. However, neither of the two non-parametric methods approached an asymptote and the predicted number of species are much less than those predicted by the other methods.

Discussion

There has been much discussion as to whether the species–area relationship is, following Arrhenius 1921), a log–log relationship or a semilog relationship as proposed by Gleason (1922) (see Williams 1964; McGuiness 1984; Rosenzweig 1995; Connor & McCoy 1979, 2001 Martin 1981; and Gotelli & Colwell 2001 for discussions of this issue). Differences in goodness-of-fit have been attributed to differences in the size of areas studied. For example, Palmer (1990) analysed plant species in a hardwood forest and concluded that: 'an ecologist interested in comparing species richness can choose any estimator except the log–log model'. Yet he operated over only a small range of areas (2 m^2–15 ha). Krebs (2001), however, suggested that the log–log model applies between *c.* 4000 m^2 and $4 \cdot 10^6$ km^2. He & Legendre (1996) analysed a species-rich assemblage and concluded that 'the exponential model (semilog) is only appropriate for small sampling areas, the power model is the best for intermediate sampling areas and the logistic is the best for large-scale sampling … and that there is no model that is universally best, all depending on sampling scales'.

We believe that the scale of the study may not be relevant, but rather what is actually being measured. Rosenzweig (1995) acknowledges Watson's 1859 study as the first species–area plot showing numbers of species of plants in increasingly larger areas, beginning with parts of the county of Surrey, then South Thames, Southern England and Great Britain, and Watson found a log–log species–area relationship. Similarly, Arrhenius (1921) analysed numbers of species of plants on 13 islands of varying sizes in the Stockholm archipelago and derived his famous log–log relationship between species and area. In these studies the identity of the species was not taken into account. Thus the species–area relationship is concerned only with the number of species in the different sizes of area sampled. Species–accumulation curves, on the other hand, take account of the identity of the species and plot the rate of accumulation of the new species sampled as samples of identical size are pooled over the total area sampled. It is not surprising to us that studies of very small areas approximate a semilog curve (e.g. McGuinnes's 1984 study of species on rocks in the marine intertidal zone of up to only a maximum size of 1400 cm^2). In such cases the curves mainly measure rates of accumulation of species, but of course the samples are not randomized. Thus we contend that the species–area relationship is log–log and species–accumulation curves semilog.

The analytical method that we have developed does not depend on an underlying distribution of individuals among species and gives exact fits to data, which are almost identical to the randomization in Colwell's EstimateS software (Colwell 2001). We believe that this method for calculating exact species–accumulation curves has clear advantages over the randomization procedures and fitting parametric models, although obviously with a large number of randomizations the results will be extremely similar for the two methods. We provide the algorithms in an Excel spreadsheet (available for download at http://folk.uio.no/johnsg).

Extrapolation from without the bounds of a data set is always a dangerous procedure. However, estimates of species richness are often needed, for example for setting priorities for conservation. The data from the Norwegian continental shelf suggests that taking account of the heterogeneity of species richness in subareas leads to a much higher estimate of species richness via extrapolation. Colwell & Coddington (1994) distinguish between species–accumulation and species–area curves. They define species–accumulation curves as referring to data sets from local species assemblages in an area or habitat that is roughly homogeneous, both spatially and temporally, whereas species–area curves cover large-scale biogeographical patterns comprising explicitly heterogenous areas. Our data from the Norwegian continental shelf cover a large area (15° of

latitude), a wide range of sediment types and depths, yet the data clearly fit the semilog species–accumulation curve. The data discussed here from the Norwegian continental shelf are clearly from heterogenous subareas, yet the analytical species–accumulation curve for the complete data set (Fig. 4) (and the semilog approximation) fits the data extremely well. Thus application of a species–accumulation curve rather than a species–area plot as recommended by Colwell & Coddington (1994) seems appropriate. The Hong Kong data, on the other hand, are from environmentally homogeneous subareas and again the analytical species–accumulation curve fits and the relationship is semilogarithmic.

Table 1 shows that the traditional methods that have been used, such as the non-parametric methods and extrapolations from models fitted to randomized species–accumulation curves, give estimates which are in fact lower than the number of species known to occur on the Norwegian continental shelf. The number of species in the database of all samples taken on the Norwegian continental shelf from many more surveys than the few considered here is 2500 species. The estimates derived by the T–S curve for the Norwegian continental shelf is 5403 species and for Hong Kong SAR 2254 species. Because all surveys conducted on the Norwegian continental shelf have covered only a small fraction of the total area of the shelf, the estimate that we are likely to find *c*. double the number of species known seems reasonable. Interestingly, May (1992), in discussing Grassle & Maciolek's (1992) estimates of how many species were likely to be found in the deep sea, suggested that rather than their estimate of 10 million species he would guess that roughly double the known number of species at 500 000 are still to be found. Our finding that the T–S extrapolation gives an estimate of double the number of species found supports May's intuition! The most important aspect is that the traditionally used methods give large underestimates of species richness, which is a serious problem in relation to conservation of biodiversity. As they are gross extrapolations, our estimates of richness on the Norwegian continental shelf and in the Hong Kong SAR may also prove to be underestimates, but we are convinced that they are better than those produced with traditional methods.

One neglected aspect that needs to be considered is that in testing the estimates from extrapolations of the species–accumulation curves is that a number of 'typical' or 'classical' data sets have been used. These include Butler & Chazdon's (1998) seed-bank data used in the EstimateS software (Colwell 2001), Palmer's (1990) data on plant species in hardwood stands and Coleman *et al*.'s (1982) data on breeding birds in Pennsylvania–Ohio. A characteristic of all these studies is the small number of species found (32, *c*. 100, and a maximum 35 per island studied, respectively). With marine data sets we routinely operate with hundreds of species (Grassle & Maciolek 1992; Gray *et al*. 1997; Ellingsen & Gray 2002); yet only He & Legendre's (1996) analysis of tree data from a tropical forest in Malaysia with 783 species approaches the richness of the marine data. More attention needs to be given to species–accumulation relationships of species-rich assemblages. With the destruction of tropical forests, coral reefs and coastal habitats there is an urgent need to develop techniques that allow us to make comparative analyses and reliable predictions of the richness of these species-rich assemblages. We have demonstrated that our new method may be generally applicable to a wide variety of data and gives sound estimates of the species richness of larger areas than those sampled; however, more tests are clearly needed.

Acknowledgements

We thank two anonymous referees for extremely useful comments and criticisms of this article. Any remaining errors are our own. We also thank the Agriculture, Fisheries and Conservation Department of the HKSAR Government and Dr Paul Shin for allowing us to use his as yet unpublished data from Hong Kong.

References

Arrhenius, O. (1921) Species and area. *Journal of Ecology*, **9**, 95–99.

Brewer, A. & Williamson, M. (1980) A new relationship for rarefaction. *Biodiversity and Conservation*, **3**, 373–379.

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

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

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

Chao, A., Ma, M.C. & Yang, M.C.K. (1993) Stopping rules and estimation for capture debugging with unequal failure rates. *Biometrika*, **80**, 193–217.

Coleman, B.D. (1981) On random placement and species–area relations. *Mathematics and Bioscience*, **54**, 191–215.

Coleman, B.D., Mares, M.A., Willig, M.R. & Hsieh, Y.H. (1982) Randomness, area, and species richness. *Ecology*, **63**, 1121–1133.

Colwell, R.K. (2001) *Estimates: Statistical Estimation of Species Richness and Shared Species from Samples*, Version 6. User's guide and application published at: http://viceroy.eeb.uconn.edu/estimates.

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

Connor, E.F. & McCoy, E.D. (1979) The statistics and biology of the species-area relationship. *American Naturalist*, **113**, 791–833.

Connor, E.F. & McCoy, E.D. (2001) Species–area relationships. In: *Encyclopedia of Biodiversity*, Vol. 5. (ed. S.A. Levin), pp. 397–412. Academic Press, New York.

Ellingsen, K.E. & Gray, J.S. (2002) Spatial patterns of benthic diversity: is there a latitudinal gradient along the Norwegian continental shelf? *Journal of Animal Ecology*, **71**, 373–389.

Erwin, T.L. (1988) The tropical forest canopy: the heart of biotic diversity. In: *Biodiversity* (ed. E.O. Wilson), pp. 123–129. National Academy Press, Washington DC.

Erwin, T.L. (1991) How many species are there – revisited. *Conservation Biology*, **5**, 330–333.
Feller, W. (1950) *An Introduction to Probability Theory and its Application*, Vol. 1. Wiley, New York.
Gleason, H.A. (1922) On the relation between species and area. *Ecology*, **3**, 158–162.
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.
Gotelli, N. & Entsminger, G.L. (2001) *ECOSIM: Null Models Software for Ecology*, Version 60. Acquired Intelligence Inc. & Kesey-Bear, available at: http//: homepages.together.net/gentsim/ecosim.htm.
Grassle, J.F. & Maciolek, N.J. (1992) Deep-sea species richness: regional and local diversity estimates from quantitative bottom samples. *American Naturalist*, **139**, 313–341.
Gray, J.S., Poore, G.C.B., Ugland, K.I., Wilson, R.S., Olsgard, F. & Johannessen, Ø. (1997) Coastal and deep-sea benthic diversities compared. *Marine Ecological Progress Series*, **159**, 97–103.
He, F. & Legendre, P. (1996) On species–area relations. *American Naturalist*, **148**, 719–737.
Heck, K.L., Belle, G. & Simberloff, D. (1975) Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size. *Ecology*, **56**, 1459–1461.
Hubble, S.P. (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Monographs in Population Biology, no. 32. Princeton University Press, Princeton.
Hurlbert, S.H. (1971) the non-concept of species diversity: a critique and alternative parameters. *Ecology*, **52**, 577–586.
Krebs, C.J. (2001) *Ecology*, 5th edn. Benjamin Cummings, San Francisco.
Martin, T.E. (1981) Species–area slopes and coefficients: a caution on their interpretation. *American Naturalist*, **118**, 823–837.
May, R.M. (1992) Bottoms up for the ocean. *Nature*, **357**, 278–279.
McGuinness, K.A. (1984) Equations and explanations in the study of species–area curves. *Biological Review*, **59**, 423–440.
Ney-Nifle, M. & Mangel, M. (1999) Species–area curves based on geographic range and occupancy. *Journal of Theoretical Biology*, **196**, 327–342.
Palmer, M.W. (1990) The estimation of species richness by extrapolation. *Ecology*, **71**, 1195–1198.
Rosenzweig, M.L. (1995) *Species Diversity in Space and Time*. Cambridge University Press, Cambridge.
Watson, H.C. (1832) *Outlines of the Geographical Distribution of British Plant Communities*. Clarendon Press, Oxford.
Williams, C.B. (1964) *Patterns in the Balance of Nature*. Academic Press, New York.
Williamson, M. (1988) Relationships of species number to area, distance and other variables. *Analytical Biogeography* (eds A.A. Meyers & P.S. Giller), pp. 91–115. Chapman & Hall, London.

Received 30 Spetember 2002; accepted 11 April 2003

Appendix I

EXACT EXPRESSIONS FOR THE EXPECTANCE AND VARIANCE OF THE SAMPLE-BASED SPECIES–AREA CURVE

Let the data consist of a collection of $A = 100$ samples containg $S_{obs} = 800$ species with occurrence frequencies R_k = number of species represented in k of the samples ($k = 1, 2, 3, \ldots, 100$). Note that $R_1 + R_2 + \ldots + R_{100} = 800$. Suppose now that we select 'a' samples randomly from the collection of 100 samples and observe S_a species. The problem is to find $\mathrm{Var}(S_a)$.

It is convenient to make use of the indicator variables $I_1, I_2, \ldots, I_{800}$ defined as $I_i = 1$ if species number i is contained in one of the selected samples, and $I_i = 0$ if this species is not represented. Then $S_a = I_1 + I_2 + \ldots + I_{800}$, so:

$$E(S_a) = E(I_1) + E(I_2) + \ldots + E(I_{800})$$

(this equation was used by Arrhenius)

$$\mathrm{Var}(S_a) = E(S_a^2) - E(S_a)^2$$

where:

$$E(S_a^2) = \sum_{i=1}^{800} E(I_i^2) + 2\sum_{i<j} E(I_i I_j)$$

Thus, the problem is reduced to calculate the mean value and covariances of the indicator variables (n_i is the number of samples where the i-th species is observed):

$$E(I_i) = P\{I_i = 1\} = 1 - P\{I_i = 0\} = 1 - \binom{100 - n_i}{a} \Big/ \binom{100}{a}$$

$$E(I_i^2) = P\{I_i = 1\} = 1 - P\{I_i = 0\}$$

$$E(I_i I_j) = P\{I_i = 1, I_j = 1\} = 1 - P\{I_i = 0 \text{ or } I_j = 0\}$$

By expanding the binomial fraction we get in the general case:

$$E[S(a)] = S_{obs}\left[1 - \sum_{k=1}^{A-a}\left(1 - \frac{a}{A}\right)\left(1 - \frac{a}{A-1}\right)\cdots\left(1 - \frac{a}{A-(k-1)}\right)\frac{R_k}{S_{obs}}\right]$$

The term R_k/S_{obs} is recognized as the fraction of species which is represented in k samples, and may be regarded as a summary statistic of the community structure. It should also be remarked that the reason for stopping the summation at $A - a$ is simply that all species that are represented in more than $A - a$ samples must be represented in any subset of a samples.

With this expression for $E(S_a)$ we need only an expression for $E(I_i I_j) = P\{I_i = 1, I_j = 1\}$ to find an analytical expression for the variance of S_a. Again, it is easier to calculate the complementary event, $\{I_i = 0$ or $I_j = 0\}$, whose probability follows from the elementary addition law:

$$P\{I_i = 0 \cup I_j = 0\} = P\{I_i = 0\} + P\{I_j = 0\} - P\{I_i = 0 \cap I_j = 0\}$$

To calculate these probabilities we need the quantities: T_{10} = Number of samples in the collections where the i-th but not the j-th species is represented.

T_{01} = Number of samples in the collections where the *j*-th but not the *i*-th species is represented.
T_{00} = Number of samples in the collections where neither the *i*-th nor the *j*-th species is represented.

The only way that the *i*-th species can be lacking from the subset of a samples is when all samples are chosen from the $T_{00} + T_{01}$ samples that do not contain the *i*-th species. From this viewpoint it is realized that the probability that the a random samples contain both the *i*-th and the *j*-th species is given by:

$$E(I_iI_j) = P\{I_i = 1 \cup I_j = 1\} = 1 - \left[\frac{\binom{T_{00}+T_{01}}{a}}{\binom{c}{a}} + \frac{\binom{T_{00}+T_{10}}{a}}{\binom{c}{a}} - \frac{\binom{T_{00}}{a}}{\binom{c}{a}}\right]$$

Because $S_a^2 = I_1^2 + \ldots + I_{S_a}^2 + 2I_1I_2 + \ldots + 2I_{S_a-1}\,I_{S_a}$ the variance of the cumulative species number may be written

$$\mathrm{Var}(S_a) = \sum_{i=1}^{S_{obs}} E(I_i^2) + 2\sum_{i<j} E(I_iI_j) - [E(S_a)]^2$$

But $E(I_i^2) = E(I_i)$, so the sum of $E(I_i^2)$ is recognized as $E(S_a)$. Hence the general analytical expression for the variance of the number of species in a random subcollection of a samples is given as:

$$\mathrm{Var}(S_a) = E(S_a) - [E(S_a)]^2 + 2\sum_{1\le i<j\le S_{obs}} E(I_iI_j)$$

Appendix II

SEMI-LOGARITHMIC APPROXIMATION TO THE SAMPLE-BASED SPECIES–AREA CURVE

In the approximation of the accumulation curve we are most interested in the properties at the tail of the function $E(S(a))$, as contrary to our intuition, many of these curves seems to have no asymptote. Obviously, the behaviour at the tail depends on the form of the representation distribution over the subset of rare species. From our study of the Norwegian continental shelf data it is sufficient to consider the species that are represented in fewer than 20 samples; that is, the values from R_1 to R_{20} determines the asymptotic behaviour of $E(S_a)$. Figure 3 shows that the representations of this subset of species on the Norwegian continental shelf may be fitted surprisingly accurately on a logarithmic scale. Over 94% of the variability is explained and therefore indicates that a collection of A samples containing S_{obs} species has a representation distribution quite close to:

$$\frac{R_k}{S_{obs}} = \frac{k^{-1.05}}{\sum_{k=1}^{A} k^{-1.05}}$$

The numerical value of the denominator is:

$$\sum_{k=1}^{A} k^{-1.05} = 4.698244$$

so with the aid of linear regression, the accumulation curve may be approximated as:

$$E[S(a)] \approx S_{obs}\left[1 - \frac{1}{4.698244}\sum_{k=1}^{A-a}\left(1-\frac{a}{A}\right)\left(1-\frac{a}{A-1}\right)\cdots\left(1-\frac{a}{A-(k-1)}\right)\frac{1}{k^{1.05}}\right]$$

Thus the problem is now reduced to approximate the function:

$$L(a) = \sum_{k=1}^{A-a}\left(1-\frac{a}{A}\right)\left(1-\frac{a}{A-1}\right)\cdots\left(1-\frac{a}{A-(k-1)}\right)\frac{1}{k^{1.05}}$$

Numerical calculation shows that the function $-0.8999\ln(a) + 4.2006$ fits $L(a)$ with $R^2 = 0.9967$, implying a semi-logarithmic approximation:

$$E[S(a)] \approx S_{obs}\left[1 - \frac{1}{4.698244}(-0.8999\ln(a) + 4.2006)\right] = 0.19S_{obs} - 0.11S_{obs}\ln(a)$$

The underlying reason why the function $L(a)$ decreases as the logarithm of a may be seen by replacing $k^{-1.05}$ by k^{-1} and apply the Arrhenius approximation $1 - a/(A-k) \approx 1 - a/A$. The sum is then recognized as the integral of the geometric series with coefficient $(1 - a/A)$, and consequently $\ln(a)$ emerges in the approximation.