Variation in rank abundance replicate samples and impact of clustering

J.H. Neuteboom* and P.C Struik

Crop and Weed Ecology Group, Wageningen University, P.O. Box 430, NL-6700 AK Wageningen, The Netherlands

* Corresponding author: (e-mail: j.h.neuteboom@chello.nl)

Received 4 May 2005; accepted 12 December 2005

Abstract

Calculating a single-sample rank abundance curve by using the negative-binomial distribution provides a way to investigate the variability within rank abundance replicate samples and yields a measure of the degree of heterogeneity of the sampled community. The calculation of the single-sample rank abundance curve is used in combination with the negative-binomial rank abundance curve-fit model to analyse the principal effect of clustering on the species-individual (*S*-*N*) curve and the species-area curve. With the usual plotting of *S* against log *N* or log area, assuming that *N* is proportional to area, *S*-*N* curves and species-area curves are the same curves with only a shifted horizontal axis. Clustering results in a lower recorded number of species in a sample and stretches the *S*-*N* curve and species-area curve over the horizontal axis to the right. In contrast to what is suggested in the literature, we surmise that the effect of clustering on both curves will gradually fade away with increasing sample size. Since the slopes of the curves are not constant, they cannot be used as species diversity indices or site discriminant. *S*-*N* curves and species-area curves cannot be extrapolated.

Additional keywords: rank abundance curve, species-individual curve, species-area curve, spatial heterogeneity

Introduction

The negative-binomial rank abundance curve-fit model of Neuteboom & Struik (2005a) starts with fitting species abundance proportions. Next, the proportions from curve fit are re-converted into numbers of individuals by multiplying them with the total number of individuals in the sample. The resulting curve is a curve with species individual numbers on a continuous scale. In a second step, that curve is converted, using the Poisson-distribution, into a curve with discrete numbers of individuals for species, with the assumption that in case of replication these numbers are the species individual numbers expected in an average single sample. The resulting curve is called the 'single-sample rank abundance curve'. The number of species calculated from the single-sample curve is the number of species usually plotted in a species-individual curve (*S-N* curve) or species-area curve (Condit *et al.*, 1996; Neuteboom & Struik, 2005a).

However, using the Poisson-distribution assumes that the numbers of individuals within species in replicate samples follow a Poisson-distribution and thus that replication is from a homogeneous population. This assumption might approximately be true for very large samples but is unlikely to also hold for the much smaller sample sizes current in biodiversity research because species almost always occur in clusters of individuals.

In the first part of this paper we shall show that in samples taken as replicates from a spatially heterogeneous community, the pattern of distribution of the recorded numbers of individuals within species can easily deviate from the Poisson-distribution. Such patterns can be fitted per species with the negative-binomial distribution, which calculates a *k* value for clustering. The negative-binomial distribution with the fitted *k* values per species can be used as an alternative for the Poisson-distribution for calculating a single-sample rank abundance curve. The resulting curve appears to adequately describe the actual pattern of species abundance distribution in an average single sample in case of clustering.

Replicate samples are rarely analysed separately. Calculating a single-sample rank abundance curve by using the negative-binomial distribution provides a way to investigate the variability among rank abundance replicate samples and could yield a measure for the degree of homogeneity or heterogeneity of the sampled community.

In the second part of the paper we shall use the calculation of the single-sample rank abundance curve in combination with the negative-binomial rank abundance curve-fit model (Neuteboom & Struik, 2005a) to explore the principal effect of clustering on the *S-N* curve and the species-area curve. In an *S-N* curve the number of species is plotted against the total number of individuals (*N*) in the sample; in a species-area curve the number of species is plotted against the area of the sample. Assuming (1) that species and individuals are counted in sampling quadrats, and (2) that the number of individuals is proportional to area, both curves are related via $N = area \cdot d$, where d is the general density equal to the totalized densities of all species. With sample size plotted as usual on log-scale the relation between N and area may be written as $\log N = \log area + \log d$, which means that under the given assumptions the *S-N* curve and species-area curve represent the same curve with only a shifted horizontal axis.

Clustering must inevitably have an effect on the *S-N* curve and the species-area curve since both always start with species numbers in small samples. Clustering will lead to a lower presence frequency of a species and to the missing of species in samples (Greig-Smith, 1983). Since the number of species in a sample equals the sum of the probabilities of presence of all species present in the system (Coleman, 1981), clustering is expected to result in a lower number of species, and in case of replicates in a lower mean number of species per sample. So clustering is expected to result into lower estimated species richness in an *S-N* curve or species-area curve.

In *S-N* curves and species-area curves the number of species (*S*) or the logarithm of the number of species (log *S*) is plotted against the logarithm of the total number of individuals (log *N*), or against the logarithm of the area of the sample (log *area*). According to the negative-binomial rank abundance curve-fit model of Neuteboom & Struik (2005a), in case of an *S*-log *N* curve, the theoretically expected number of singleton species ($E(S(1))$) is the slope of the curve. However, that applies to communities with Poisson-distributed individuals. Whether this also holds for communities with strong clustering can be checked by assessing the slope as the tangent at consecutive points along the curve.

The relation between number of species and sample size, along with the rank abundance relation, is a central theme in community ecology. He & Legendre (2002) also analysed the principal effect of clustering on the species-area relation, using the negative-binomial distribution. However, they calculated theoretical species-area curves from the probabilities of presence or absence of species from samples. Their approach and conclusions will be compared with ours. The shape of the species-area curve we predict for clustering is different. Below we shall primarily focus on the *S*-*N* curve as the key to the species-area curve.

Appendix 2 of Neuteboom & Struik (2005a) contains a glossary with relevant terms, parameters and symbols that are also used in this paper. However, in the present paper x is used in a different way (see Equations 4, 6 and 7). Additional terms, parameters and symbols used in this paper, but not listed in Neuteboom & Struik (2005a), are given in Appendix 3.

Theory

The number of species plotted against the total number of individuals in an *S*-*N* curve is the number of species in a single sample or, in case of replicate samples (Condit *et al.*, 1996), the mean number of species per sample. That number can be theoretically derived from the rank abundance curve for the average numbers of individuals per species (the 'average' rank abundance curve), using a function describing the pattern of distribution of the numbers of individuals within species over replicates. We use the Poisson-distribution and, for detecting clustering of individuals within species, the negative-binomial distribution with separately fitted 'k values' per species (see below). Both probability functions calculate, from the sample means for the numbers of individuals per species, the contributions per species to the theoretically expected numbers of species with $n = 1, 2, 3 ...., j$ individuals and the total number of species in an average single sample.

Consecutively, the Poisson-distribution, the negative-binomial distribution (distinct from the negative-binomial rank abundance curve fit model (Neuteboom & Struik, 2005a)), the procedure for calculating the expected numbers of species with $n = 1, 2, 3,..., j$ individuals in an average single sample, and the procedure for generating a single-sample rank abundance curve are discussed below.

In the negative-binomial rank abundance curve-fit model (Neuteboom & Struik, 2005a) the equations of the negative-binomial distribution are used for curve fitting with m and k (the two parameters of the distribution, see later) in a totally different role as pure curve-fit coefficients without any further statistical meaning. In order to avoid confusion, they will be referred to in that role in the second half of this paper, as μ and κ, respectively.

Poisson-distribution

In case of a random distribution of individuals the expected relative frequencies $f(n)$ of samples containing $n = 0, 1, 2, 3, \ldots, j$ individuals will approximate a Poisson-series, expressed by:

$$e^{-m} \left(1, m, \frac{m^2}{2!}, \frac{m^3}{3!} \ldots\ldots \frac{m^n}{n!}\right) \quad (1)$$

where m is the mean number of individuals per sample and e the exponential base, 2.7183.

Negative-binomial distribution

The negative-binomial distribution is described by two parameters: the mean m and an exponent k. Parameter k is a measure of the degree of clustering, often referred to as the dispersion parameter. The expected relative frequencies of sampling units containing $n = 0, 1, 2, 3, \ldots, j$ individuals are given by (Davies, 1971):

$$f(n) = [(k + n - 1)! \,/\, n!(k - 1)!]\ (m/k)^n \,/\, [1 + (m/k)]^{(k+n)} \quad (2)$$

Methods for fitting k are discussed by Southwood (1978). Davies (1971) gives a computerized calculation of k, based on the maximum likelihood method of Fisher (1953). In Davies' (1971) programme the expected relative frequencies of samples containing $n = 0, 1, 2, 3, \ldots, j$ individuals are calculated by first calculating the expected relative frequency in the first class for $n = 0$ individuals from:

$$f(0) = 1 \,/\, [1 + (m \,/\, k)]^k \quad (3a)$$

The relative frequencies in the second and higher classes for $n = 1, 2, 3 \ldots, j$ individuals (note that j is the maximum number of individuals in the sample) are then derived from:

$$f(n) = f(n - 1)\ m\ (k + n - 1) \,/\, ((m + k)\ n) \quad (3b)$$

As stated before, low k values simulate clustering, high k values indicate that the frequency distribution approximates a Poisson-distribution. In Davies' (1971) programme for fitting the negative-binomial distribution the condition is built in that the variance has to be greater than the mean. In our application of the programme, where we fit a k value for each single species, we use as a criterion for fitting the negative-binomial distribution that the ratio variance : mean has to be larger than 1.2. If that criterion is not met, the relative frequencies are calculated from the Poisson-distribution.

Calculation of the expected numbers of species with n = 1, 2, ..., j individuals in an average sample from the mean numbers of individuals per species in replicate samples

First, the expected relative frequencies $f(n)$ of sample-units containing $n = 1, 2, 3, \ldots, j$ individuals are calculated for each of the S species in the lumped sample of all replicates. For species with Poisson-distributed individuals the relative frequencies are calculated from the Poisson-distribution. For negative-binomial-distributed species the relative frequencies are calculated from Equations 3a and 3b, using the means (m) and the fitted negative-binomial k values per species. The expected numbers of species $E(S(n))$ with $n = 1, 2, 3, \ldots, j$ individuals in an average sample are given by the totalized relative frequencies $f(n)$ of all species. That is, for each n the number of species expected in an average single sample is:

$$E(S(n)) = \sum_{x=1}^{x=S} f(n))_x \qquad (4)$$

The total number of species expected $E(S)$ is:

$$E(S) = \sum_{n=1}^{n=\infty} E(S(n)) \qquad (5)$$

The philosophy behind the calculations in Equation 4 is that each species (x) contributes to the expected number of species ($E(S(n))$) for each of the theoretically possible numbers of individuals (n). Given the steadily decreasing sample means for the abundances of the sequential species (m in Equation 3b), in case of a sufficient number of replicates (see Appendix 2) these contributions finally become so small that for each n the expected number of species $E(S(n))$ is (almost) fixed. In that case also $E(S)$ in Equation 5 is (almost) fixed.

$E(S)$ can also be derived from the probabilities of absence [$f(0)$ for $n = 0$] of all species. The probability of absence of a species can be calculated from the Poisson-distribution [$f(0) = e^{-z}$] as well as from the negative-binomial distribution (Equation 3a). The probability that a species will be present with at least one individual is $1 - f(0)$. According to Coleman (1981) the expected number of species in a sample ($E(S)$) equals the sum of the probabilities of presence of all species, or:

$$E(S) = \sum_{x=1}^{x=S} (1 - f(0)_x) \qquad (6)$$

Equations 4 and 5 or Equation 6 can also be used to calculate $E(S)$ from the averages from curve fit instead of the averages from sampling for the numbers of individuals per species. Using the negative-binomial rank abundance curve-fit model (Neuteboom & Struik, 2005a) and substituting '$x = S$' in the summation sign of Equation 4 or Equation 6 by '$x = \infty$' enables the calculation of $E(S)$ for an infinite number of species from curve fit. However, no negative-binomial k values are available for the species from extrapolation. For cases of clustering, this problem could be solved by using, analogous to He & Legendre (2002) and Plotkin & Muller-Landau (2002), one negative-binomial k with best fit as a common k for all species. The estimation of a common negative-binomial k for all species will be discussed below in an application to observational data. $E(S)$ can always be calculated from curve fit means for the theoretical case that all species inclusive of those from extrapolation are assumed to have Poisson-distributed individuals. By using the negative-binomial rank abundance curve-fit model, for theoretical purposes also the numbers of species expected for smaller and extrapolated larger sample sizes can be calculated.

In case P is the maximum number of species present in the sampled system the expected number of species for consecutive sample sizes can be calculated from (Coleman, 1981; note that $x = S$ in the sum sign of Equation 6 is $x = P$):

$$E(S) = P - \sum_{x=1}^{x=P} f(0)_x \qquad (7)$$

Procedure for generating a rank abundance curve from the numbers of species expected for consecutive numbers of individuals in an average single sample; the single-sample rank abundance curve

A single-sample rank abundance curve is made by plotting the numbers of individuals n of an average single sample (Equation 4) on log-scale (vertical axis) against the accumulated expected numbers of species ($E(S(n))$) as species sequence (horizontal axis). The first number in the plotting is the highest theoretical number of individuals with an arbitrarily chosen lowest expected relative frequency of 10^{-3}. The values for species sequence on the horizontal axis are values on a continuous scale.

In case individuals within species actually behave as being Poisson-distributed, the single-sample rank abundance curves for Poisson- and negative-binomial-distributed individuals will both describe the same pattern of species abundance distribution as the 'average' rank abundance curve and thus coincide with that curve. In case of clustering, the theoretical curve for negative-binomial-distributed individuals will deviate.

The expected numbers of species for consecutive numbers of individuals in the single-sample rank abundance curve can be transformed into a frequency distribution with numbers of species in $\log_2$-classes of numbers of individuals per species according to the method of Preston (1948). The frequency distribution can be tested versus the frequency distribution of the actual mean numbers of species classified in $\log_2$-classes of numbers of individuals by way of a χ^2-test.

Analysis of clustering using observational data

Four data sets were used to compare the calculated single-sample rank abundance curves for the case of Poisson- and negative-binomial-distributed numbers of individuals within species in replicate samples with the rank abundance curve for the mean numbers of individuals per species per sample (the 'average rank abundance curve'). To assess whether the numbers of individuals within species in replicate samples are negative-binomial-distributed, the *m* values and separately fitted *k* values per species are used. Species yielding a *k* value larger than 40 were considered to have Poisson-distributed numbers of individuals in replicate samples.

Data sets

Three of the four data sets are from Coleoptera research (Col1, Col2, Col3). The fourth data set is from Patrick (1968) with numbers of diatoms in 8 equally treated compartments (replicates) in part of the stream of Darby Creek, Pennsylvania.

The Col1, Col2 and Col3 data sets apply to Coleoptera (beetle) specimens counted in 5, 5 and 8 rectangular replicate plots, respectively, in grasslands on different localities in Zuid Limburg, The Netherlands. Specimens were collected using pitfall trapping at regular intervals.

Patrick's (1968) experiment was set up in the framework of a study of the effects environmental factors may have on the structure of diatom communities. First aim of the experiment was to determine the degree and kind of variability one might expect in the structure of a community under very similar ecological conditions. Two experiments were conducted, one in 1965 (8 replicates) and one in 1966 (4 replicates). The data analysed in this paper are from the experiment in 1965.

Analysis

Figure 1 summarizes the analysis of the Col1-data (for the detailed data see Appendix 1). Plotted in Figure 1a are the log-numbers of individuals per species against species sequence (species in sequence from most to least abundant) for the total (□1) and average (□2) of 5 replicate samples, and for each of the single replicates (*H*1–*H*5). Plotted also are the points (o) derived from the mean species frequencies per sample in $\log_2$-classes of numbers of individuals per species. The scaling of the horizontal and vertical axes for $\log_2$-class-data is explained in Appendix 2. Figure 1b shows the plotting of the calculated negative-binomial *k* values and the mean numbers of individuals per species per replicate sample. Species that satisfied the Poisson-distribution (variance not significantly greater than the mean, see Appendix 1), are plotted with a negative-binomial *k* of 40. The *k* values and mean numbers of individuals per species were not correlated.

The points marked by □1 and □2 in Figure 1c are the same as in Figure 1a. The curves 3 and 4 show for an average single sample the theoretically expected numbers of individuals of sequential species with the assumption of Poisson- (curve 3) and negative-binomial-distributed (curve 4) numbers of individuals within species, respectively. Curve 4 is calculated using the negative-binomial *k* values per species. Figure 1d summarizes the data from Figures 1a (o) and 1c (curves 3 and 4). The mean species frequencies in $\log_2$-classes (real and fitted frequencies) are given in the histograms in Figures 2a (frequencies fitted with the negative-binomial distribution from the *k* values per species) and 2b (frequencies fitted with the Poisson-distribution).

The Poisson-curve (Figure 1c, curve 3) almost exactly follows the course of the mean numbers of individuals per species (□2). However, the mean species frequencies in $\log_2$-classes (o) are significantly better described by the curve for negative-binomial-

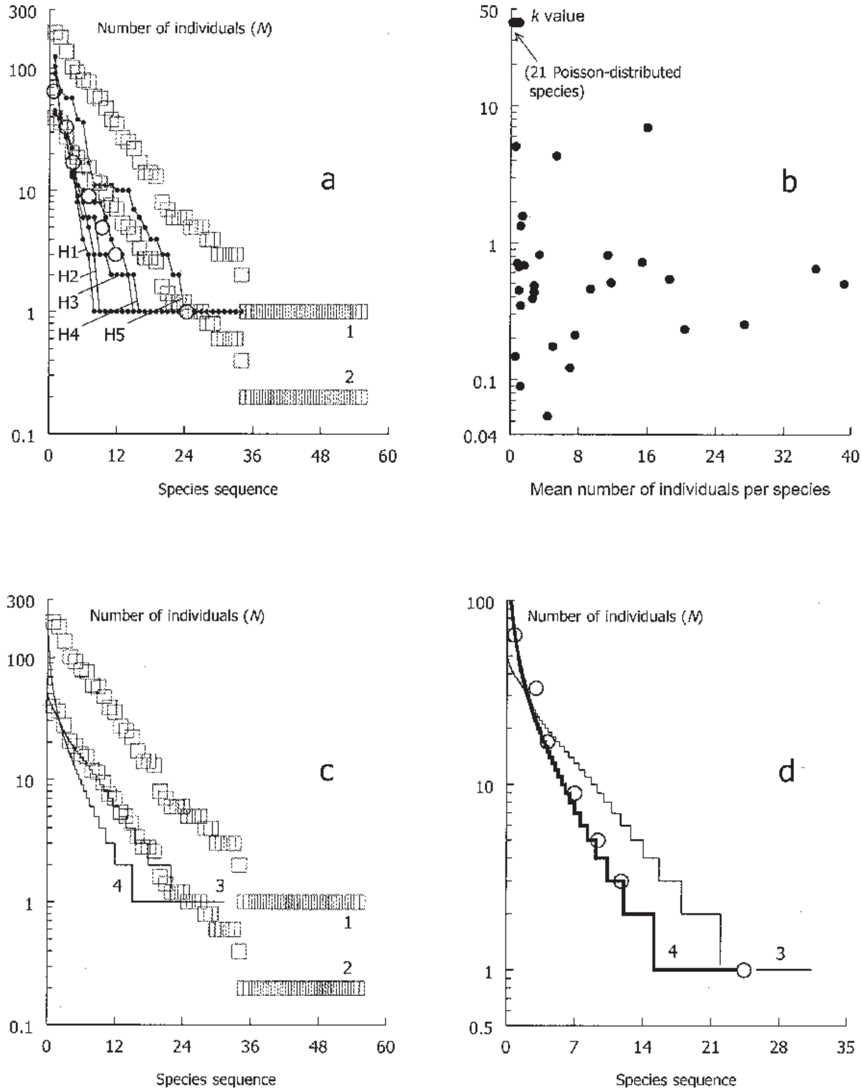

Figure 1. Data set Col1, Coleoptera species. (a) Plotted against species sequence are the total (□1) and mean (□2) numbers of individuals per species in 5 replicate samples, the numbers of individuals per species in single replicates (H1–H5), and the numbers of individuals of the consecutive $\log_2$-classes of numbers of individuals per species in an average sample (○). (b) Negative-binomial *k* values (vertical axis) and mean number of individuals per species per replicate sample (horizontal axis). Note that 'Poisson'-distributed species are plotted with a *k* value of 40. (c) □1 and □2 as in Figure 1a; curves 3 and 4 show the expected numbers of individuals per species in an average single sample. Curve 3 refers to Poisson-distributed numbers of individuals within species in replicate samples. Curve 4 refers to negative-binomial-distributed numbers of individuals as calculated from the negative-binomial *k* values per species. (d) Summary of Figures 1a and 1c.

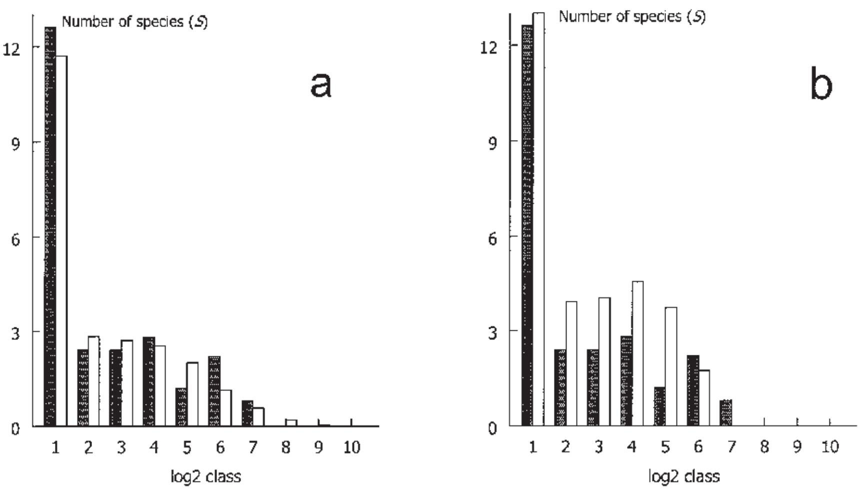

Figure 2. Data set Col1, Coleoptera species. Frequency distributions of the mean numbers of species per replicate sample in consecutive $\log_2$-classes of numbers of individuals per species. Black bars: mean numbers actually found from sampling. White bars: mean numbers expected, (a) for negative-binomial *k* values per species, (b) for Poisson- distributed numbers of individuals within species.

distributed individuals within species (curve 4, Figure 1d). Curve 4 is the single-sample rank abundance curve of which the accumulated number of species could be plotted in a species-individual (*S-N*) curve. Clustering, as indicated by the better fit with the negative-binomial distribution (curve 4 vs. curve 3, and Figure 2a vs. 2b), results in lower and higher numbers of species in the classes with low and high numbers of individuals, respectively, and in a lower accumulated total number of species. It is striking that only five replicate samples already suffice for the calculation of workable negative-binomial *k* values per species. This could be explained from the large number of species and from the fact that too high and too low estimated *k* values for species (Figure 1b) compensate each other.

In the four data sets summarized in Table 1, best fit according to χ^2 (lowest values) was found for negative-binomial-distributed numbers of individuals within species. However, the difference in the expected numbers of species compared with those expected with the assumption of Poisson-distributed individuals was relatively small for the Col3-data. The same holds for the diatom data for the numbers of species in $\log_2$-classes exclusive of those in the first class with 1 and 2 individuals (shown by the second data row for each data set in Table 1). The numbers of individuals in that range comprised 99.8% of all individuals in the total of 8 replicate samples. The low variation in the diatom replicate samples also appears from the similarity of curves 3 and 4 in Figure 3a and from the frequency distributions in Figures 3c and 3d if the species frequencies in $\log_2$-classes are considered exclusive of those in the first class. The diatom data consisted of 8 replicates of steady state conditions that had been kept as

Table 1. Sampling data on numbers of individuals and numbers of species in replicate samples of four data sets, and theoretical calculations on expected numbers of species ($E(S)$) in an average single sample. Three of the four data sets are from Coleoptera research (Col1, Col2 and Col3), one data set is on diatom species (Diatoms; Patrick, 1968; experiment 1965). Numbers of species expected in an average single sample were calculated in three ways: (1) using the fitted negative binomial *k*-values per species, resulting in $E(S)_1$, (2) using the terms of the Poisson distribution, resulting in $E(S)_2$, and (3) using the negative-binomial distribution with one common *k* for all species, resulting in $E(S)_3$.
Note that *N* is the total number of individuals of all species, *S* the number of species in the total sample of *H* replicates, N_{avg} the average number of individuals-, and S_{avg} the average number of species per replicate sample. The χ^2-values were calculated for the average numbers of species expected versus the average numbers of species actually found in $\log_2$-classes of numbers of individuals per species in an average sample. Upper row per data set: species in all $\log_2$-classes; lower row per data set: species in all $\log_2$-classes minus the first class with 1 and 2 individuals. The total number of $\log_2$-classes is given in parentheses.

	Data from sampling					Numbers of species expected ($E(S)$) in an average single sample						
	N	*S*	*H*	N_{avg}	S_{avg}	From *k* values per species		From Poisson; all species		From one common *k*; all species		
						$E(S)_1$	χ^2	$E(S)_2$	χ^2	*k* from best fit	$E(S)_3$	χ^2
Col1	1323	55	5	264	24.4	23.7	1.08 (6)	31.3	4.60 (6)	0.50	23.6	1.42 (6)
					11.8	11.7	1.02 (5)	17.9	4.55 (5)		12.8	1.12 (5)
Col2	1231	43	5	246	20.8	20.4	0.18 (4)	24.5	1.64 (4)	1.10	20.9	0.17 (4)
					9.0	9.1	0.07 (3)	11.4	1.55 (3)		9.7	0.12 (3)
Col3	4635	47	8	579	23.3	22.7	0.24 (6)	23.9	1.11 (6)	2.25	22.9	0.16 (6)
					14.0	14.1	0.17 (5)	14.9	1.11 (5)		13.8	0.15 (5)
Diatoms	36683	182	8	4585	106.0	106.8	1.49 (9)	118.6	6.58 (9)	7.42	116.8	5.52 (9)
					79.4	76.0	0.93 (8)	79.5	2.62 (8)		78.1	1.72 (8)

similar as possible. That is, the data set was a case with an expected low variation.

The degree of clustering of individuals can be reasonably expressed by one *k* value with best-fit, i.e., a common negative-binomial *k* for all species (Table 1). A low *k* value expresses clustering, high *k* values indicate a distribution towards Poisson. With *k* values ≥ 10 the distribution almost equals a Poisson-distribution. The single-sample rank abundance curve calculated from the average-rank abundance curve on the basis of a common *k* with best fit for all species is presented for the Col1-data in Figure 4a. The frequency distribution of the numbers of species in $\log_2$-classes derived from it is presented in Figure 4b. The single-sample rank abundance curve (Figure 4a) reasonably agreed with curve 4 in Figure 1d, which was calculated from the separately fitted *k* values per species, and yielded a comparable reasonable fit of the average numbers of species in $\log_2$-classes of numbers of individuals (Figures 4a and 4b). Fitting a common negative-binomial *k* for all species is not unrealistic since species may force clustering upon each other.

Clustering and species-individual curve

As stated before, the expected numbers of species in a theoretical single sample can also be calculated using the *fitted* averages from curve fit instead of the *actual* averages

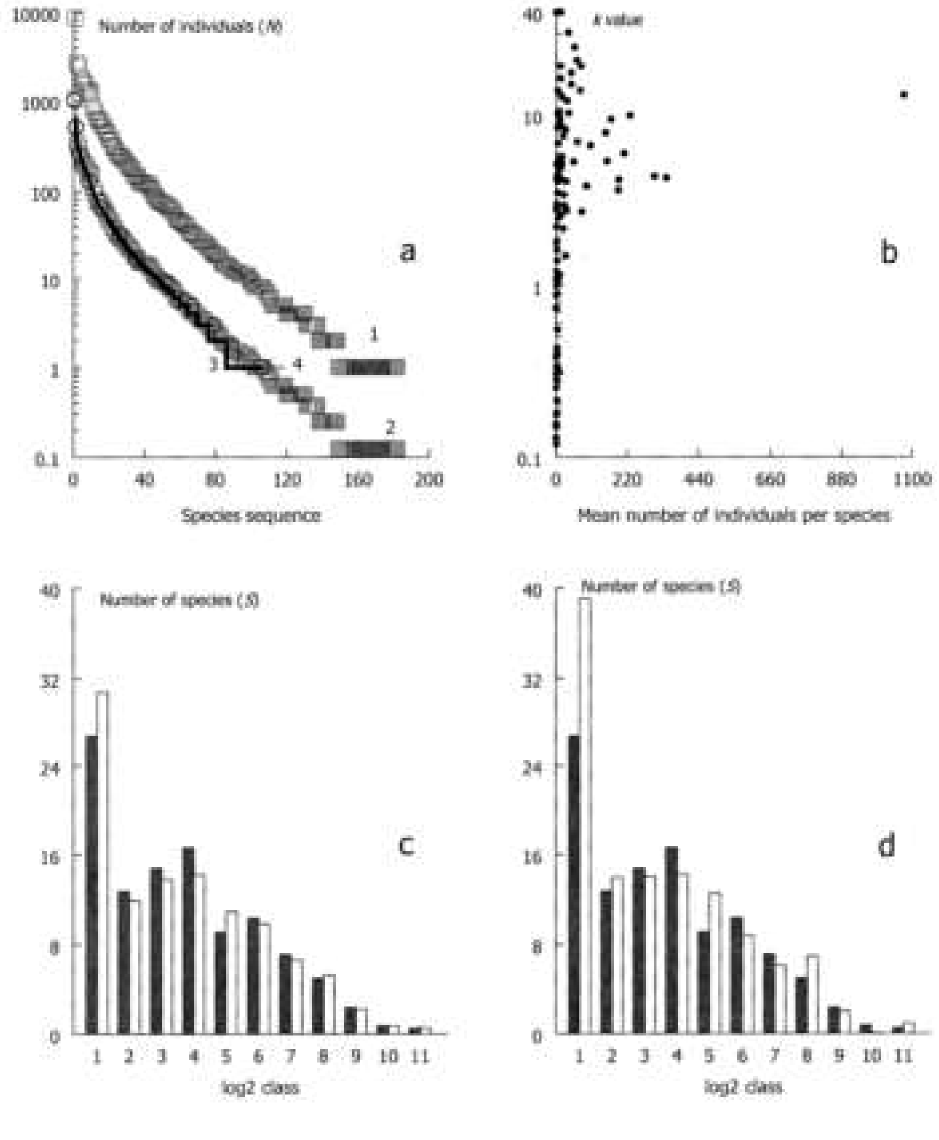

Figure 3. Number of diatoms in 8 replicate samples, Darby Creek, Pennsylvania, data from Patrick (1968), (experiment 1965). For legend of markers and curves in (a) and (b), see Figures 1a, 1c and 1d. For legend of the black and white bars in (c) and (d), see Figures 2a and 2b.

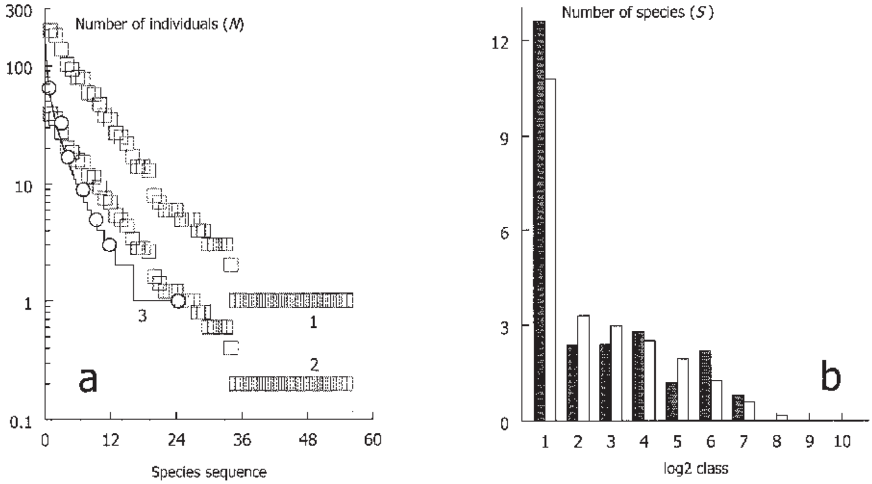

Figure 4. Data set Col1, Coleoptera species. (a) Total (□1) and mean (□2) number of individuals per species in 5 replicate samples, and numbers of individuals of the consecutive $\log_2$-classes of numbers of individuals per species in an average sample (○), plotted against species sequence. Curve 3 is the single-sample rank abundance curve calculated from the average numbers of individuals per species on the basis of a best fitting common *k* for all species. (b) Frequency distributions of the mean numbers of species per replicate sample in consecutive $\log_2$-classes of numbers of individuals per species. Black bars: mean numbers of species actually found from sampling. White bars: mean numbers of species expected on the basis of a best fitting common negative-binomial *k* for all species.

from sampling for the numbers of individuals per species. The procedure of calculations is discussed in Figures 5a and 5b. We applied that type of calculations to produce Figure 6a. Presented in Figure 6a are three *S-N* curves. Each point ($E(S)$) in each curve was calculated individually from the rank abundance curve fitted to the Col1-data, using the negative-binomial rank abundance curve-fit model (Neuteboom & Struik, 2005a). The negative-binomial curve-fit model (parameters μ, κ and c) fits species abundance proportions. These can be extrapolated for an infinite number of species (Figure 5a) and transformed into numbers of individuals by multiplying them with the total number (N) of individuals for any given sample size [see Figure 5b, curve 1 for $N = 4096$ (2^{12})]. The resulting numbers of individuals are conceived as the mean numbers of individuals per species per sample. From these, using the Poisson-distribution or the negative-binomial distribution, the expected numbers of species with 1 ($E(S(1))$), 2, 3, …., etc. individuals and the accumulated total number of species expected ($E(S)$) in an average single sample can be calculated (Figure 5b, curve 2). For the calculation we used Equations 4 and 5 for an infinite number of species. That is, '$x = R$' in the sum-sign of Equation 4 is substituted by '$x = \infty$'. $E(S(1))$ is for Poisson-distributed species the theoretical slope of the *S-N* curve at any given N, which can be checked by also calculating the

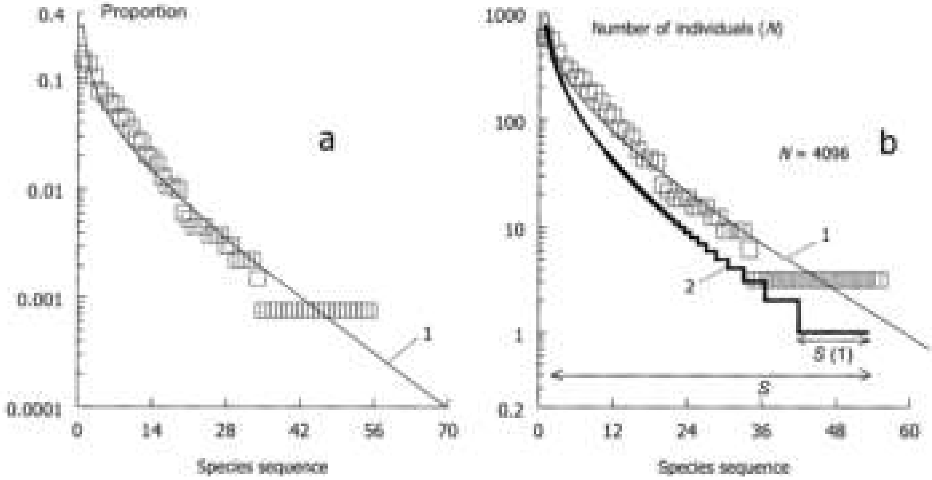

Figure 5. Example of calculations, Coleoptera species Col1. (a) Abundance proportions (□) and fitted rank abundance curve (curve 1). The curve is fitted with the negative-binomial rank abundance curve fit model and extrapolated for an infinite number of species. Parameter values of the model (Neuteboom & Struik, 2005a): $\mu = 5.968$, $\kappa = 0.456$ and $c = 0.98239$. (b) Abundance proportions (□) and proportions from curve fit (curve 1) transformed into numbers of individuals by multiplying them with a value for the total number of individuals in the sample of $N = 4096$ (2^{12}). Curve 2 is the rank abundance curve of a theoretical single sample calculated from the fitted 'total' rank abundance curve (curve 1) on the basis of a common negative-binomial k for all species of 0.50 (Col1-data, Table 1), using Equations 4 and 5. The theoretically expected number of species ($E(S) = 52.273$) and the expected number of singleton species ($E(S(1))$ = 11.089) as calculated from the single-sample rank abundance curve are plotted for $N = 4096$ in the curves numbered as '2' in Figures 6a (S-N curve) and 6b, respectively.

slope directly from the tangent at consecutive positions along the curve. Values calculated for $E(S(1))$ and tangents should be the same (Neuteboom & Struik, 2005a). The calculation of the slope as tangent is explained in the caption of Figure 6.

The numbers of species in Figure 6a were calculated starting with $N = 1$ and subsequently by doubling N. That is, $N = 1, 2, 4, 8, 16$, ..., etc., up to and including $N =$ 65,536 (2^{16}). $E(S)$-values for curve 1 were calculated from the fitted rank abundance curve with the assumption of Poisson-distributed numbers of individuals within species in replicate samples. The $E(S)$-values of curve 2 were calculated assuming clustering of individuals within species with one common negative-binomial k value for all species for all sample sizes (the common k of the Col1-data, $k = 0.50$ in Table 1). The $E(S)$-values of curve 3 were calculated with the assumption of one common negative-binomial k for all species that gradually increases with increasing sample size. An increasing k with increasing sample size seems realistic since the effect of clustering on the S-N curve is expected to gradually decrease and to finally fade away with increasing sample size. The arguments for this were discussed by Neuteboom & Struik (2005a).

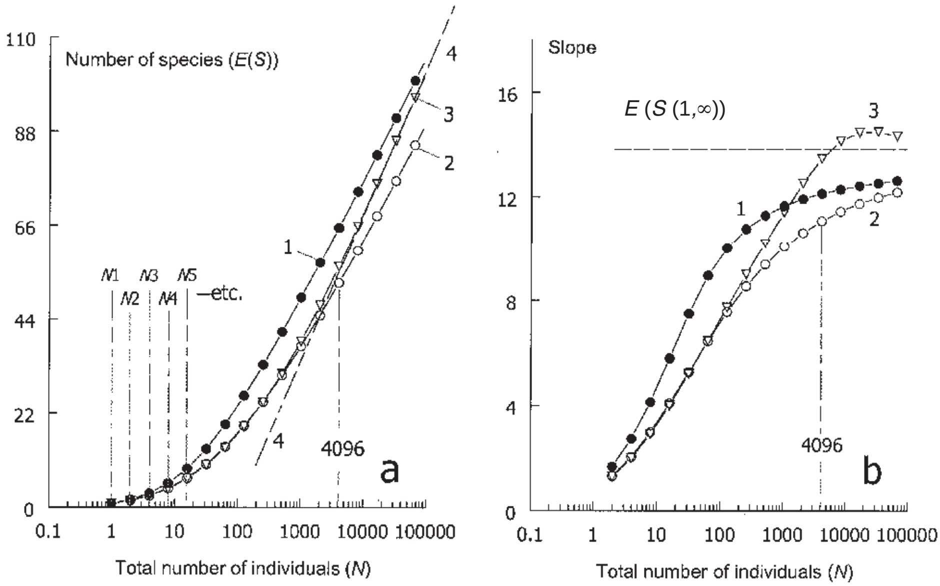

Figure 6. (a) Expected number of species $E(S)$ plotted against the total number of individuals N. $E(S)$ calculated for each N from the rank abundance curve fitted by the negative-binomial rank abundance curve fit model to the data of Col1. Curves based on different assumptions: Curve 1 (case 1): Poisson-distributed numbers of individuals within species in replicate samples. Curve 2 (case 2): negative-binomial-distributed numbers of individuals within species in replicate samples with one common negative-binomial k for all species for all sample sizes of 0.50 (Col1-data Table 1). Curve 3 (case 3): negative-binomial-distributed numbers of individuals within species in replicate samples with a common negative-binomial k for all species that gradually increases with increasing sample size (see text). (b) Slope values of the S-N curves 1, 2 and 3 in (a). The slope values are calculated as the tangent at consecutive points along the curves. The sample sizes N are numbered $N1$, $N2$, $N3$, ..., Nn in (a). The tangent of the slope is approached in all cases by calculating it for $N2$, $N3$, ... Nn from $(E(S_{N3})-E(S_{N1}))/[\ln(N3)-\ln(N1)]$, $(E(S_{N4})-E(S_{N2}))/[\ln(N4)-\ln(N2)]$, etc. Curve 4 in (a) is explained in text. The number '4096' in (a) and (b) is N (sample size) for which $E(S)$, $E(S(1))$ and slope were calculated in the example in Figure 5b.

We let k (k_A) increase using an equation of Plotkin & Muller-Landau (2002), empirically derived from abundance data of single species: $k_A = \beta A^{\zeta} + \delta$, where A is the size of the area sampled, and β δ (the value for k_A at $A = 0$) and ζ are constants. We replaced A by N (i.e., $k_A = \beta N^{\zeta} + \delta$) and set β at 0.002923, δ at 0.33 and ζ at 0.545. For $N = 1$, $k = 0.44$. For $N = 256$ (2^8), the calculated k approximated the common k of the Col1-data, i.e., $k = 0.50$ for the average sample size of $N = 256$. Our values for β δ and ζ are arbitrary but could be realistic.

Figure 6a shows that compared with the curve for Poisson-distributed individuals (case 1, curve 1), clustering of individuals within species stretches the S-N curve over the horizontal axis to the right (cases 2 and 3, curves 2 and 3). The slope calculated as

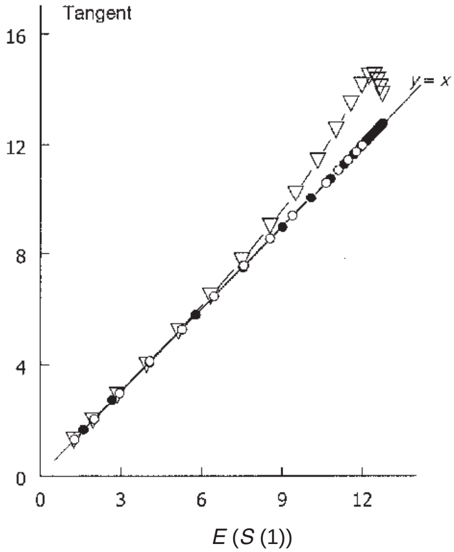

Figure 7. Slope of the *S-N* curve determined as tangent, plotted against $E(S(1))$ as value for the slope calculated from the rank abundance curve. Tangent- and $E(S(1))$-values of the *S-N* curves 1, 2 and 3 in Figure 6a: (•) Poisson-distributed individuals within species (case 1); (o) negative-binomial-distributed numbers of individuals within species in replicate samples with one negative-binomial *k* for all species for all sample sizes of 0.50 (case 2); (▽) all species with negative-binomial-distributed numbers of individuals in replicate samples characterized by one common negative-binomial *k* for all species that gradually increases with increasing sample size (case 3, see text).

tangent at consecutive points along the curve (the consecutive values of *N*) is presented in Figure 6b. Figure 7 shows the relation between the slope of the curve calculated as tangent and the theoretical slope calculated as $E(S(1))$ from the consecutive single-sample rank abundance curves. For Poisson-distributed species (case 1) the slope as tangent equals $E(S(1))$ (curve 1 in Figure 6b, and Figure 7). The same holds for clustering if simulated with one common negative-binomial *k* for all species that remains constant for all sample sizes (curve 2 in Figure 6b, and Figure 7). In other words, clustering with a constant common negative-binomial *k* for all species independent of sample size (case 2) did not change the meaning of $E(S(1))$ as the slope of the curve. However, for case 3 with a common negative-binomial *k* for all species that increases with sample size, the slope calculated as tangent gradually exceeded the value of $E(S(1))$ (Figure 6b, curve 3, and Figure 7). So in that case $E(S(1))$ does not represent the slope of the *S-N* curve anymore. We believe, with the arguments given, that curve 3, reflecting a decreasing effect of clustering with increasing sample size (Figure 6a), is the most likely type of curve for a community with clustering.

Neuteboom & Struik (2005a) also discussed the $E(S(1,\infty))$, the ultimate value of the slope for an infinitely large sample and independent of clustering. The three curves in

Figure 6b are expected to finally end up in one and the same value, the value for $E(S(1,\infty))$.

Discussion and conclusions

Neuteboom & Struik (2005a) explained that with abundance plotted on log-scale, the usual rank abundance curve obtained by totalizing the numbers of individuals per species in a series of replicate samples (the 'total' rank abundance curve) is the same as the curve for the average species individual numbers per sample (the 'average' rank abundance curve), with only a level difference. Both curves are in principle independent of clustering. The rank abundance curve from which the number of species is calculated for plotting in an *S-N* curve or species-area curve, is the curve of an average single sample (the 'single-sample' rank abundance curve), and strongly dependent on clustering. The present paper offers a technique for detecting a deviation from Poisson-distributed numbers of individuals within species in rank abundance replicate samples and shows the expected effect clustering will have on the shape of the single-sample rank abundance curve and on the shape of the *S-N* curve. The single-sample rank abundance curve will be steeper with a lower accumulated number of species (Figures 1c and 1d, curve 4 vs. curve 3). Clustering stretches the *S-N* curve over the horizontal axis to the right (Figure 6a). However, the effect of clustering is expected to gradually decrease and to finally fade away with increasing sample size (Neuteboom & Struik, 2005a). The most likely *S-N* curve for a community with clustering is therefore a curve of the type of curve 3 in Figure 6a, i.e., a curve that starts with a smaller slope and is temporarily steeper compared with the curve expected for Poisson-distributed species. With the assumption that number of individuals is proportional to area, the same would apply to the species-area curve.

One could question whether the suggestion of Poisson-distributed individuals for the rare species in the tail of the curve as shown for the Col1-data in Appendix 1, is realistic. In fact that is not so important since in the Col1-data the sample means over 5 replicates for the numbers of individuals of these species (*m* in Equations 3a and 3b) were so low that they hardly contributed to the theoretically expected number of species in a single sample. All in all, the method presented seems to work for getting an idea of the amount of variation in rank abundance replicate samples. The fitted common *k* for all species could be used as a measure of the degree of homogeneity or heterogeneity of the sampled community.

The species-area curve has an important application in conservation ecology where it is used to calculate the reduction in species richness that would occur if the area of a sanctuary were to be reduced. This reduction can be calculated with a variety of curve-fit models. A model frequently used in botanical research is Gleason's (1922) model, which fits the more or less linear part of the curve for large samples with *S* plotted against log area as a straight line with a slope and an intercept. In theory, the fitted curve could also be used to calculate, by means of extrapolation, the number of species for larger samples. However, curve 4 in Figure 6a shows that with clustering, after further extrapolation, species richness would clearly be overestimated. So extrapolating a species-area curve may be misleading.

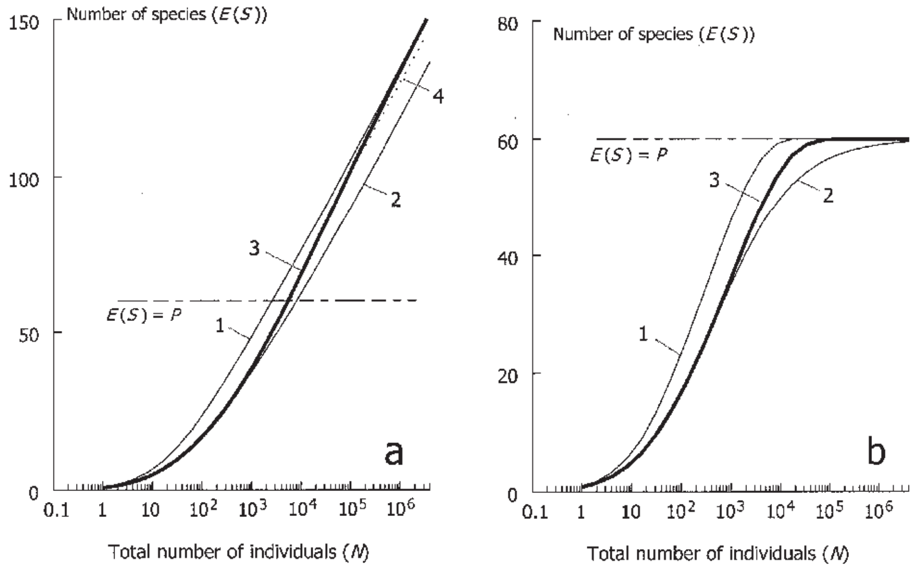

Figure 8. (a) *S-N* curves 1, 2 and 3 from Figure 6a, extrapolated for larger samples, using Equations 4 and 5 or Equation 6 for calculating *E*(*S*). (b) *S-N* curves 1, 2, and 3 from Figure 6a with *E*(*S*) calculated using Equation 7 with *P* for a finite number of species in the sample system, see text. The *S-N* curves presented also represent the shape of species-area curves, since with sample size (*N* or area) plotted on log scale, *S-N* curves and species-area curves are the same curves with only a shifted horizontal axis.

We have generalized the ultimate shape of the *S-N* curve that we expect for clustering by extending curve 3 (Figure 6a) for larger samples: curve 3 in Figure 8a. Like in Figure 6a the curve starts with a smaller slope and is, after a while, steeper than the curve expected for Poisson-distributed species (Figure 8a, curve 1). However, since the effect of clustering is expected to further decrease with increasing sample size, the curve for clustering (curve 3) may ultimately coincide with the curve for Poisson-distributed species. On the other hand we cannot be sure. Suppose that in the ever-increasing sampling area all new incoming rare species keep showing some degree of clustering. In that case curve 4 in Figure 8a may be more realistic. Curve 3 is to be expected in case the rarest species would have approximately Poisson-distributed individuals.

He & Legendre (2002) also predicted a species-area curve for clustering that, compared with the curve for Poisson-distributed species, is stretched over the horizontal axis to the right. Nevertheless the type of curves they predict is different (Figure 8b). To simulate clustering, He & Legendre (2002) also used the negative-binomial distribution with a common negative-binomial *k* for all species that increases with sample size. However, they calculated *E*(*S*) from the probabilities of absence of species from samples [the *f*(0) in our Equation 3a]. Values for *m* they derived from a special algorithm or from discrete rank abundance models.

The types of curve He & Legendre (2002) predict for clustering (curves of the type of curves 2 and 3 in Figure 8b) finally coincide with the curve for Poisson-distributed species (curve 1) at a maximum number of species. However, that is due to two assumptions: (1) a finite number (P) of species present in the sampled system by finally using Equation 7, and (2) the assumption that that number is the number of species counted in the total community or largest sampling quadrat.

We believe that the second assumption is not always realistic. In our approach we do not need to know P, because we calculate $E(S)$ for each sample size from the rank abundance curve fitted by the negative-binomial rank abundance curve-fit model (Neuteboom & Struik, 2005a). That model results in proportions or individual numbers from curve fit for, in principle, an infinite number of species. Given the continuously decreasing abundances of the species from curve fit and $x = \infty$ in the sum-sign of Equation 4 or Equation 6, the contributions per species to the theoretically expected numbers of species with consecutive numbers of individuals in a single sample finally become so small that for each sample size the number of species is fixed (Equation 5). In the analysis of He & Legendre (2002) the effect of clustering on the slope and the course of the species-area curve is obscured by the effect of P, a maximum number of species present in the sampled system.

He & Legendre's (2002) analysis of the species-area relation based on the absence or presence frequencies of species in sampling quadrats, fits in an approach discussed before by He & Gaston (2000). They consider that a lot of data on the commonness and rarity of species is available from their grid cell occupancy on species occurrence maps (geographical observation grids), and believe that in some way these data could be used for making population estimates of species. Gaston *et al.* (1998), Hartley (1998), Kunin (1998), Kunin *et al.* (2000) and Witte & Torfs (2003) also discussed aspects of that approach. We do not further discuss the details of that approach here, as it is beyond the scope of this paper.

Starting point in our analysis is the variation in species individual numbers within rank abundance replicate samples. We use that variation to construct, at least for one sample size, a single-sample rank abundance curve and calculate from that the total number of species for plotting in an S-N curve or species-area curve. The single-sample curve can be characterized by one common negative-binomial k for all species. A common k could in principle also be estimated for consecutive smaller and larger sample sizes. However, rank abundance data from differently sized sampling units are usually not available. We let the common k increase linearly with increasing sample size for calculating the expected numbers of species in larger samples via extrapolation using the negative-binomial rank abundance curve fit model. However, whether the types of S-N curves and species-area curves we predict for clustering are realistic should be further checked on the basis of observational data. For this checking we shall use a computer programme for artificial sampling (Neuteboom & Struik, 2005b).

A provisional conclusion is that the slope of the S-N curve and species-area curve is unreliable as site discriminant and species-diversity index. First, because even for Poisson-distributed species the slope is not constant (Neuteboom & Struik, 2005a), and second, because of clustering. Different from what is sometimes assumed, S-N curves and species-area curves cannot be extrapolated to calculate numbers of species for larger samples. Clustering is an unpredictable factor of which we never know where in the *S-N* curve or species-area curve its effect will cease.

However, if rank abundance data are available and reasonably fitted by the negative-binomial rank abundance model, $E(S(1,\infty))$ can always be calculated for an infinitely large sample as site discriminant and species diversity index. As stated before (Neuteboom & Struik, 2005a), $E(S(1,\infty))$ can also be calculated for data sets in which the abundances of species are available only in terms of proportions.

Acknowledgements

We are grateful to H. Turin, Stichting Faunistisch Onderzoek Carabidae (Loopkever-stichting), for providing the data set on the Coleoptera in a grassland on the Pietersberg (Zuid Limburg, The Netherlands) and to Dr L. Hemerik, Biometris, Wageningen University, for advice on mathematical aspects.

References

Coleman, B.D., 1981. On random placement and species-area relations. *Mathematical Biosciences* 54: 191–215.

Condit, R., S.P. Hubbell, J.V. Lafrankie, R. Sukumar, N. Manokaran, R.B. Foster & P.S. Ashton, 1996. Species-area and species-individual relationships for tropical trees: a comparison of three 50-ha plots. *Journal of Ecology* 84: 549–562.

Davies, R.G., 1971. Computer Programming in Quantitative Biology. Academic Press, London, 492 pp.

Fisher, R.A., 1953. Note on the efficient fitting of the negative binomial. *Biometrics* 9: 97–200.

Gaston K.J., T.M. Blackburn & J.H. Lawton, 1998. Aggregation and interspecific abundance-occupancy relationships. *Journal of Animal Ecology* 67: 995–999.

Gleason, H.A., 1922. On the relation between species and area. *Ecology* 3: 158–162.

Greig-Smith, P., 1983. Quantitative Plant Ecology (3rd edition). Blackwell Scientific Publications, Oxford, 359 pp.

Hartley, S., 1998. A positive relationship between local abundance and regional occupancy is almost inevitable (but not all positive relationships are the same). *Journal of Animal Ecology* 67: 992–994.

He, F. & K.J. Gaston, 2000. Estimating species abundance from occurrence. *American Naturalist* 156: 553–559.

He, F. & P. Legendre, 2002. Species diversity patterns derived from species area models. *Ecology* 83: 1185–1198.

Kunin, W.E., 1998. Extrapolating species abundance across spatial scales. *Science* 281: 1513–1515.

Kunin,W.E., S. Hartley & J.J. Lennon, 2000. Scaling Down: On the challenge of estimating abundance from occurrence patterns. *American Naturalist* 156: 560–566.

Neuteboom, J.H. & P.C. Struik, 2005a. A versatile curve-fit model for linear to deeply concave rank abundance curves. *NJAS – Wageningen Journal of Life Sciences* 53: 167–197.

Neuteboom, J.H. & P.C. Struik, 2005b. *In silico* sampling reveals the effect of clustering and shows that the log normal rank abundance curve is an artefact. *NJAS – Wageningen Journal of Life Sciences* 53: 223–245.

Patrick, R., 1968. The structure of diatom communities in similar ecological conditions. *American Naturalist* 102: 173–183.

Plotkin, J.B. & H.C. Muller-Landau, 2002. Sampling the species composition of a landscape. *Ecology* 83: 3344–3356.

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

Southwood, T.R.E., 1978. Ecological Methods. Chapman and Hall, London, 524 pp.

Witte, M.W. & J.J.F. Torfs, 2003. Scale dependency and fractal dimension of rarity. *Ecography* 26: 60–68.

Appendix 1

Numbers of individuals of Coleoptera species in 5 replicate samples taken at Pietersberg, The Netherlands. Species ranked in order of mean number of individuals per sample (z)

Coleoptera species	z	Sample no.					k [1]
		1	2	3	4	5	
Calathus fuscipes	39.2	45	42	103	0	6	0.4952
Pterostichus madidus	35.8	1	27	57	3	91	0.6361
Anchomenus dorsalis	27.4	1	1	11	124	0	0.2526
Harpalus latus	20.4	0	0	57	3	42	0.2325
Amara convexior	18.6	0	8	65	6	14	0.5371
Carabus violaceus purpura	16.0	8	13	10	22	27	6.9265
Amara lunicollis	15.4	1	1	38	28	9	0.7076
Nebria brevicollis	11.8	1	39	10	9	0	0.5046
Harpalus rubripes	11.4	33	11	7	0	6	0.7994
Harpalus rufipes	9.4	33	1	2	11	0	0.4553
Calathus melanocephalus	7.6	1	0	36	0	1	0.2113
Trechus quadristriatus	7.0	0	1	0	34	0	0.1217
Notiophilus palustris	5.4	1	5	10	8	3	4.3020
Harpalus tardus	5.0	0	0	17	8	0	0.1747
Carabus coriaceus	4.4	22	0	0	0	0	0.0541
Pterostichus vernalis	3.4	1	1	11	4	0	0.8109
Ophonus puncticeps	2.8	0	1	11	0	2	0.4806
Bembidion lampros	2.8	0	1	3	10	0	0.4322
Ophonus melletii	2.6	1	1	11	0	0	0.3867
Abax parallelepipedus	1.6	0	1	0	1	6	0.6762
Amara communis	1.4	4	0	1	0	2	1.6740
Pterostichus melanarius	1.2	1	0	4	0	1	0.3450
Amara similata	1.2	0	0	6	0	0	0.0894
Parophonus maculicornis	1.2	0	1	5	0	0	1.3339
Pterostichus niger	1.0	0	3	0	0	2	0.4439
Panagaeus bipustulatus	1.0	1	0	2	0	2	Poiss.[2]
Amara familiaris	1.0	0	0	4	1	0	0.6594
Amara nitida	0.8	3	0	1	0	0	0.7069
Leistus ferrugineus	0.8	0	0	1	0	3	0.7069
Agonum muelleri	0.6	0	1	0	2	0	5.0398
Amara aenea	0.6	0	0	3	0	0	Poiss.
Trechus obtusus	0.6	0	1	1	1	0	0.1476
Bembidion obtusum	0.6	0	1	0	2	0	5.0398
Nebria salina	0.4	1	0	0	0	1	Poiss.
Amara consularis	0.2	0	0	0	1	0	Poiss.

Coleoptera species	z	Sample no.					k [1]
		1	2	3	4	5	
Philorhizus melanocephalus	0.2	0	0	0	1	0	Poiss.
Amara montivaga	0.2	0	0	1	0	0	Poiss.
Paradromius linearis	0.2	1	0	0	0	0	Poiss.
Calathus micropterus	0.2	0	0	0	1	0	Poiss.
Amara aulica	0.2	0	0	1	0	0	Poiss.
Bradycellus verbasci	0.2	0	0	0	0	1	Poiss.
Amara curta	0.2	0	0	1	0	0	Poiss.
Ophonus nitidulus	0.2	0	0	0	1	0	Poiss.
Amara plebeja	0.2	0	0	1	0	0	Poiss.
Amara eurynota	0.2	1	0	0	0	0	Poiss.
Anisodactylus binotatus	0.2	0	0	0	1	1	Poiss.
Amara ovata	0.2	0	1	0	0	0	Poiss.
Laemostenus terricola	0.2	0	1	0	0	0	Poiss.
Asaphidion flavipes	0.2	0	1	0	0	0	Poiss.
Bembidion tetracolum	0.2	0	0	0	1	0	Poiss.
Dyschirius globosus	0.2	0	0	1	0	0	Poiss.
Notiophilus biguttatus	0.2	0	0	1	0	0	Poiss.
Syntomus truncatellus	0.2	1	0	0	0	0	Poiss.
Notiophilus rufipes	0.2	0	0	0	1	0	Poiss.
Notiophilus substriatus	0.2	0	0	1	0	0	Poiss.
Total number of individuals		162	164	494	284	219	

[1] *k* = calculated negative-binomial *k*-value per species.
[2] Poiss. = species with variance not greater than the mean. These species were treated as Poisson-distributed.

Appendix 2

Example (Col1) of scaling horizontal axis for mean numbers of species per replicate in log_2-classes of numbers of individuals per species. [Open circles (○) in Figures 1a and 1d]

Log_2-class	Number of individuals per species	Value along vertical axis	Mean number of species per sample	Mean cumulative number of species (horizontal axis for species sequence)
1	1–2	1	12.6	24.4
2	3–4	3	2.4	11.8
3	5–8	5	2.4	9.4
4	9–16	9	2.8	7.0
5	17–32	17	1.2	4.2
6	33–64	33	2.2	3.0
7	65–128	65	0.8	0.8

Appendix 3

Symbols used in this paper and not listed in Appendix 2 of Neuteboom & Struik (2005a)

A	Size of the area sampled.
Col	Coleoptera data set.
d	General density equal to the totalized densities of all species.
k_A	Value for k of the negative-binomial distribution dependent on area.
N_{avg}	Mean number of individuals per replicate sample.
S_{avg}	Mean number of species per replicate sample.
$E(S)_1$	Number of species expected in an average single sample as calculated from the separately fitted negative-binomial k values per species for the numbers of individuals distribution over replicate samples.
$E(S)_2$	Number of species expected in an average single sample as calculated from the Poisson-distribution.
$E(S)_3$	Number of species expected in an average single sample as calculated on the basis of one common negative-binomial k for all species.
β, δ and ζ	Constants in the Plotkin & Muller-Landau equation for the relation between k of the negative-binomial distribution used as best fitting common k for all species and size of the area sampled.