A versatile curve-fit model for linear to deeply concave rank abundance curves

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

A new, flexible curve-fit model for linear to concave rank abundance curves was conceptualized and validated using observational data. The model links the geometric-series model and log-series model and can also fit deeply concave rank abundance curves. The model is based – in an unconventional way – on the negative-binomial distribution and calculates (like the log-series model) a species-diversity index. The index is defined as the expected number of singleton species (species present with one individual) in an infinitely large sample. The new model could satisfy the need for more flexible curve-fit models with which differences and changes in the shape of the rank abundance curve can be more accurately investigated. The common rank abundance curve-fit models are lacking that flexibility.

Additional keywords: species-individual curve, species-area curve, geometric-series model, log-series model, species-diversity index

Introduction

In biodiversity research, a usual way of graphical presentation of data includes the abundance of species plotted on a log scale against the species' rank, in order from most abundant to least abundant species (see e.g. Magurran, 1988; see also Figure 1). Many species abundance relations thus presented show a linear or concave descending curve. Concave curves are common, for example, in entomological research (Taylor, 1978) and are preferably fitted with the log-series model (Fisher *et al*., 1943). However, the quality of fit is not always satisfactory. Reason for an imperfect fit is often the rigid shape of the always shallow log-series rank abundance curve. Wilson (1991) used the Zipf model and the Zipf-Mandelbrot model for fitting concave rank abundance relations. However, also

with these two models the quality of fit is often unsatisfactory.

In this paper, we shall show that for fitting concave rank abundance relations the negative-binomial distribution has the potential of an ideal curve-fit model if it is used in an unconventional way by taking the relative frequency for zero individuals $f(0)$ as quantifier for the abundance proportion of the first (dominant) species, the $f(1)$ for the abundance proportion of the second species, the $f(2)$ for the abundance proportion of the third species, etc. The resulting model can fit with only two iterable parameters and a third parameter that follows from a level difference (see later), a variety of rank abundance curves ranging in shape from linear (geometric-series model) via shallow-concave (log-series model) to deeply concave. When used in combination with the Poisson-distribution (see below), the model can even calculate a species-diversity index. The new model links the geometric series model and log-series model and could solve the problem discussed by Hughes (1986) that there are no statistical models that can fit deeply concave rank abundance curves.

Curve fitting with the new model is done in two steps. In the first step, species abundance proportions (see below) are fitted. 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. This results in a rank abundance curve with species individual numbers on a continuous scale. In the second step that curve is converted, using the Poisson-distribution, into a curve with discrete numbers of individuals for species. This second step assumes that the species individual numbers in replicate samples follow a Poisson-distribution and thus that replication is from a homogeneous population. The same assumption is made in Fisher's log-series model (Fisher *et al.*, 1943) and might be valid for very large samples, as will be discussed later. The assumption is needed for calculating the proposed species-diversity index defined as the number of species with one individual ('the number of singleton species') in an infinitely large sample. Below, the model will be explained in detail and validated using observational data from literature.

This paper is the first of a series of three papers in which we present new insights into species-abundance relations and species-area relations. In the second paper (Neuteboom & Struik, 2005a), the validity of the assumption of Poisson-distributed numbers of individuals within species in replicate samples will be further investigated. The assumption might approximately be true for very large samples but is unlikely to also hold for the relatively small sample sizes current in biodiversity research because species almost always occur in clusters of individuals. Clustering can have strong implications for the shape of both the species-individual curve (*S-N* curve) and the species-area curve in which the number of species is plotted against the size of the sample (sample size expressed as total number of individuals or area of the sample, respectively). In the third paper (Neuteboom & Struik, 2005b), the effect of clustering on rank abundance curves, *S-N* curves and species-area curves is further investigated using a computer programme for *in silico* sampling. In that paper we also shall explain why sigmoid rank abundance curves (log-normal and broken-stick types of curves) are deceptive as descriptors of the species abundances in a community as the rank abundance curve is basically concave.

The new curve-fit model could satisfy the need for more flexible models with which

differences in the shape of the rank abundance curve can be investigated more accurately. The common rank abundance curve-fit models are lacking that flexibility.

Existing curve-fit models

Before discussing the new curve-fit model based on the negative-binomial distribution, we discuss some models that are currently used for fitting linear to concave curves, i.e., the geometric-series model and log-series model. We shall link our new model to these existing models.

Appendix 2 contains a glossary with terms, parameters, acronyms and symbols used in this paper.

Geometric-series model

According to the geometric-series model, the abundance proportion $p_{R,geom}$ of a species with rank R is calculated from:

$$p_{R,geom} = q\ (1 - q)^{\ (R-1)} \tag{1}$$

where q is the abundance proportion of the first (dominant) species.

Log-series model

According to Fisher's log-series model (Fisher *et al.*, 1943) the expected number of species with n individuals in a single sample ($E(S(n))$) is calculated from:

$$E(S(n)) = (\alpha / n)\ x^n \qquad\qquad n = 1, 2, 3, \ldots, j \tag{2}$$

in which α ($\alpha > 0$) and x ($0 < x < 1$) are constants.

Parameters α (called the species-diversity index) and x (a 'constant' approaching unity with increasing sample size) are related to the number of species (S) and the total number of individuals in the sample (N) according to:

$$S = -\alpha \ln(1 - x) \tag{3}$$

$$N = \alpha x / (1 - x) \tag{4}$$

α is calculated after x is found by iteration from:

$$S / N - [(1 - x) / x][-\ln(1 - x)] = 0 \tag{5}$$

From Equation 2 it follows that for an infinitely large sample ($x = 1$), α equals the number of singleton species.

Negative-binomial rank abundance curve-fit model

Outline of the model

The new rank abundance curve-fit model is based – in an unconventional way – on the parameters of the negative-binomial distribution. First, proportions are fitted and thus expected abundance proportions for consecutive species are calculated from curve fitting. These proportions can be multiplied by the total number (N) of individuals in the sample to obtain the expected abundances of species in terms of numbers of individuals. These numbers are numbers on a continuous scale. Using the Poisson-distribution these numbers can be converted into discrete numbers of individuals. The resulting discrete 'Poisson-curve' or 'single-sample rank abundance curve' and the log-series curve have in common (Fisher *et al.*, 1943) that, with extrapolation, the calculated expected number of singleton species (species present in the sample with one individual) approaches a constant value when the total number of individuals becomes very large. We shall show why. It is the basis for calculating the number of singleton species for an infinitely large sample as a new species-diversity index. It can be shown that the number of singleton species is the slope of the species – individual (S-N) curve in which, like in Fisher's log-series model (Fisher *et al.*, 1943), the number of species (S) is plotted against the logarithm of the size (N) of the sample.

Equations

In this section we first discuss the negative-binomial distribution as the basis of our rank abundance curve-fit model. Discussed is also how to generate a 'Poisson-curve' (single-sample rank abundance curve) for graphical presentation.

Negative-binomial distribution

The negative-binomial distribution can be used for fitting the pattern of distribution of the numbers of individuals within species over replicate samples. The distribution is described by two parameters, the mean number of individuals per sample m and the exponent k. k is a measure of the degree of clustering and is often referred to as the 'dispersion parameter'. The expected relative frequencies ($f(n)$) of sampling units containing n = 0, 1, 2, 3 ...., j individuals are calculated with the following equation (Davies, 1971):

$$f(n) = [(k + n - 1)! \,/\, n!(k - 1)!]\,(m/k)^n \,/\, [1 + (m/k)]^{(k+n)} \qquad (6)$$

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

$$f(0) = 1 \,/\, [1 + (m \,/\, k)]^k \qquad (7a)$$

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

$$f(n) = f(n-1)\ m\ (k+n-1)\ /\ ((m+k)n) \qquad (7b)$$

Use of the parameters of the negative-binomial distribution for a curve-fit model, curve-fit coefficients m and k

By varying m and k a large diversity of curve types can be generated of which, plotted on a log scale, the calculated relative frequencies can very satisfactorily describe the course of the abundance proportions of species in sequence of abundance. The relative frequency $f(0)$ is used for the purpose of fitting the abundance proportion p_1 of the first (dominant) species, the relative frequency $f(1)$ to fit the abundance proportion p_2 of the second species, the relative frequency $f(2)$ to fit the abundance proportion p_3 of the third species, etc., or (R is species rank):

$$p_R = f(R-1) \qquad (R = 1, 2, 3, ..., 0) \qquad (8a)$$

or, adapted for the case of curve fitting:

$$p_R = f(R-1)\ /\ c \qquad (R = 1, 2, 3, ...0) \qquad (8b)$$

Parameter c will be further explained in the validation of the model, using observational data. In many curve-fit cases, c will turn out to be approximately 1, which means that often c has little effect. For the moment we shall therefore ignore c in the further theoretical considerations. Parameter c does not affect the essentials of the paper, such as the calculation of the $E(S(1,\infty))$ as site discriminant and species diversity index (see below).

However, the way in which they are used in the negative-binomial rank abundance curve-fit model, parameters m and k have fully lost their meaning as 'mean' of a series of samples and 'dispersion factor', respectively. They are degraded in the new curve-fit model to pure curve-fit coefficients without any further statistical meaning and should therefore be given a different name. In order to distinguish them from the original m and k in the negative-binomial distribution we shall refer to them in the remaining of this paper as μ and κ, respectively.

Abundance in terms of proportions and numbers of individuals

The nature of the negative-binomial curve-fit model makes that the fit is on proportions. However, the abundance proportions observed and those expected from curve fit (p_R) can be converted or re-converted into numbers of individuals by multiplying them by the total number (N) of individuals of all species in the sample. That is, the number of individuals (z_R) from curve fit for each species is:

$$z_R = p_R N \qquad (9)$$

Relation to the geometric series curve-fit model

One reason why the negative-binomial distribution works so well as the basis for a curve-fit model, is that in many data sets the abundance proportion of the dominant species has an extremely high value, just like the relative frequency $f(0)$ from the negative-binomial distribution for $k < 1$. Using $f(0)$ for p_1 has the further advantage that the abundance proportions of all species from curve fit add up to 1, such that the linearly declining curve found for $k = 1$ automatically reflects a geometric-series distribution (see above). That in turn means that the negative-binomial curve-fit model can be used for testing whether observational data follow a geometric-series distribution. Even the q-value of the geometric series can be calculated because it equals the value of the calculated first proportion. That is, from Equation 7a for $k = 1$ it follows that (see also above):

$$q = 1 / (1 + m) \qquad (10)$$

The flexibility of the negative binomial curve-fit model is illustrated in Figure 1 by a number of curves calculated for different values of κ and μ.

Calculation of the expected numbers of species in a single sample

The numbers of individuals per species from curve fit are numbers on a continuous scale, from which – using the Poisson-distribution – the expected numbers of species

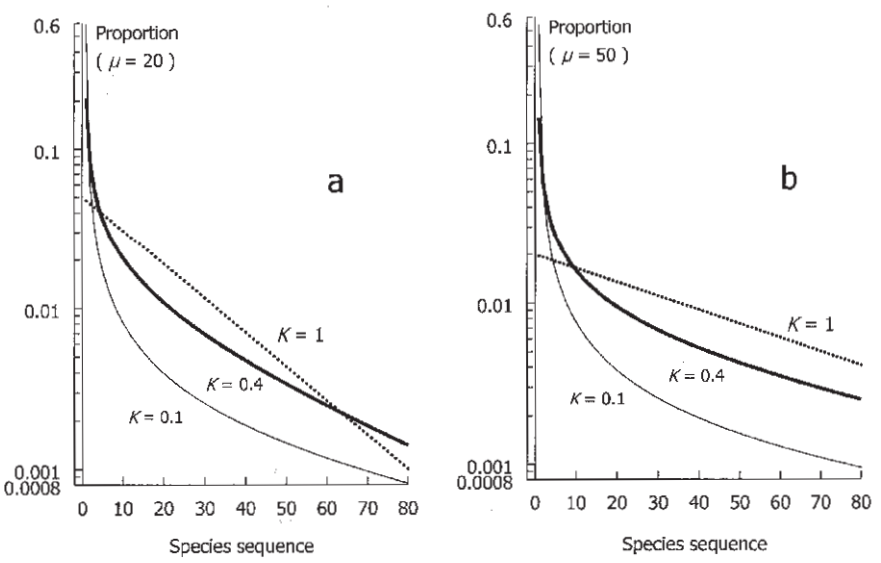

Figure 1. Curves demonstrating the flexibility of the negative-binomial curve-fit model. Parameter κ determines whether the curve is concave ($\kappa < 1$) or linear ($\kappa = 1$). Curves can even be convex ($\kappa > 1$; not shown because not used). Parameters μ and κ determine together the steepness and the degree of concavity of the curve and the height of the abundance proportion of the first (dominant) species. (a): curves for $\kappa = 1$, $\kappa = 0.4$ and $\kappa = 0.1$, for $\mu = 20$. (b): curves with the same values of κ, for $\mu = 50$.

with $n = 1, 2, 3, ..., j$ individuals in a theoretical single sample can be calculated. That is, the expected number of species S with n individuals in a single sample is:

$$E(S(n)) = \sum_{R=1}^{R=\infty} \frac{e^{-z_R} (z_R)^n}{n!} \qquad (11)$$

The expected total number of species in a single sample is:

$$E(S) = \sum_{n=1}^{n=\infty} E(S(n)) \qquad (12)$$

For each expected number of species $E(S(n))$ (Equation 11) its contribution to the expected total number of individuals, $E(N(n))$, can be calculated from:

$$E(N(n)) = E(S(n))n \qquad (13)$$

The expected total number of individuals of all species $E(N)$ in a single sample is:

$$E(N) = \sum_{n=1}^{n=\infty} E(N(n)) \qquad (14)$$

For rank abundance curves fitted with Equation 8a, N must equal the actual total number of individuals from sampling.

The expected total number of species ($E(S)$) can also be derived directly from the probabilities of absence of the consecutive species in a single sample. The probability of absence of a species is e^{-z_R}. That means that the probability that a species will be present with at least one individual is $1 - e^{-z_R}$. The expected total number of species in a sample ($E(S)$) is the sum of the probabilities of presence of all species (Coleman, 1981), or:

$$E(S) = \sum_{R=1}^{R=\infty} (1 - e^{-z_R}) \qquad (12)$$

Effect of sample size and calculation of the number of singleton species

Varying N in the equations (N in Equation 9) makes it possible to investigate the theoretical relation between the number of species and the size of the sample, and to calculate $E(S(1))$ (the expected number of singleton species $E(S(n))$ for $n = 1$ in Equation 11) for larger samples. Below we shall show that $E(S(1))$ approaches a constant value for large N. For Poisson-curves derived from the extrapolated continuous rank abundance curve fitted by the negative-binomial curve-fit model for $\kappa = 1$ (a geometric-series rank abundance curve) it even becomes exactly constant. For that type of curve it can be mathematically derived that the number of singleton species is the slope of the number of species - number of individuals curve (S-N curve), with number of individuals (horizontal axis) on a logarithmic scale (see below).

Procedure to generate a theoretical rank abundance curve for expected numbers of individuals of sequential species; 'Poisson-curve' and log-series rank abundance curve

The philosophy behind the calculations in Equation 11 is that each z contributes to the expected number of species ($E(S(n))$) for each of the theoretically possible numbers of individuals n. As z steadily decreases, the contributions finally become so small that for each n, $E(S(n))$ is fixed. In that case also the expected number of species ($E(S)$) calculated from Equation 12 or Equation 15 is fixed. For graphical presentation of a 'Poisson'-rank abundance curve, the numbers n are plotted on a logarithmic scale (vertical axis) against the accumulated expected numbers of species $E(S(n))$ as species sequence (horizontal axis). The first n in the plotting is the highest theoretical plant number with an arbitrarily chosen lowest expected relative frequency of 10^{-3}.

Since expected numbers of species for consecutive numbers of individuals are also calculated from the log-series model (Equation 2), exactly the same procedure can in principle be followed for generating a log-series rank abundance curve.

$E(S(1))$ and sample size; calculation of $E(S(1))$ for an infinitely large sample, $E(S(1,\infty))$

The expected number of singleton species $E(S(1))$for large samples is constant only in case of a linear rank abundance relationship fitted by the negative-binomial curve-fit model with $\kappa = 1$. This is a geometric-series rank abundance curve. For concave rank abundance curves (continuous curves) fitted by the negative binomial with $\kappa < 1$, $E(S(1))$ only approaches constancy for large samples. $E(S(1))$ is the tangent of the slope of the number of species - log number of individuals curve (S-log(N) curve).

We shall address the following questions: (1) why does $E(S(1))$ become constant for large samples, and (2) how to calculate $E(S(1))$ for an infinitely large sample for cases with $\kappa < 1$?

Geometric-series rank abundance curve

Figure 2a shows for a fictitious linear rank abundance curve (curve 1) the calculated numbers of individuals of sequential species ranked from most to least abundant. The curve is generated from a curve on proportions for $\kappa = 1$ and $\mu = 9$. This is a geometric-series rank abundance curve for $q = 0.1$ (Equation 10). The ultimate numbers of individuals per species were obtained by multiplying the calculated proportions by a total number of individuals of $N = 100$. The calculated numbers are numbers on a continuous scale. Curve 2 in Figure 2a shows the course of the contributions per species to the expected number of singleton species $E(S(1))$ as calculated from the term $e^{-z}z$ (Equation 11 for $n = 1$). Curves 1 and 2 coincide in the range of rare species due to the fact that for very small values of z (z approaching zero), the term $e^{-z}z$ (the contribution per species) becomes equal to z.

In case of a geometric-series relationship, the abundance proportions of the consecutive species differ by a constant factor $(1 - q)$. That is, for each species S_R it holds that its abundance proportion is a factor $(1 - q)$ smaller than that of its predecessor S_{R-1} (Equation 1). The consequence is that from a certain point onwards (the species sequence number

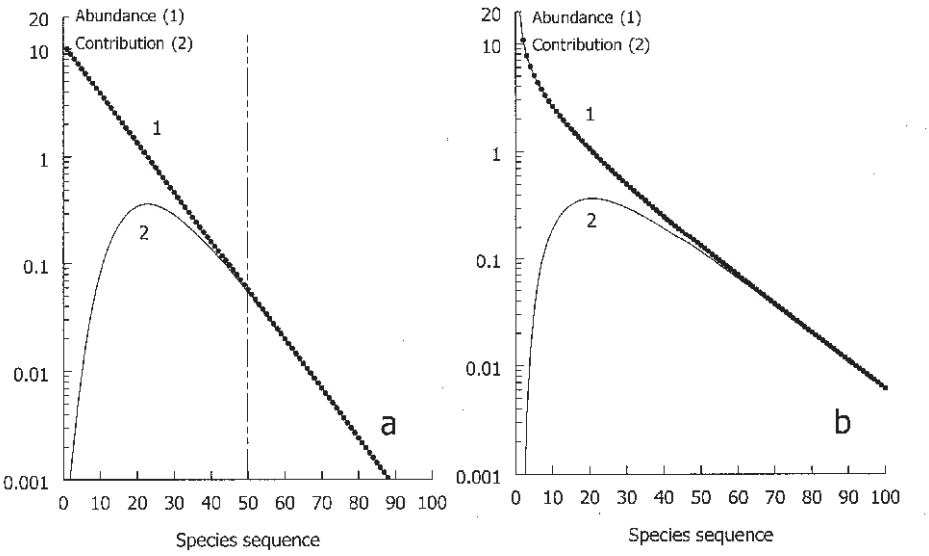

Figure 2. Curve 1: Rank abundance curve calculated from the negative-binomial curve-fit model for the abundances per species. Curve 2: Contribution per species to the expected number of singleton species. The abundances per species in (a) were calculated for $\mu = 9$, $N = 100$ and $\kappa = 1$, those in (b) were calculated for $\mu = 9$, $N = 100$ and $\kappa = 0.2$.

where curves 1 and 2 coincide) for each N always the same series of sequential z-values will be calculated, so that the number of singleton species becomes constant. The z-values contributing (according to the Poisson-distribution) to the number of singleton species are approximately values smaller than or equal to 6 or, to be on the safe side, values smaller than or equal to 30.

In the concrete procedure for calculating the expected number of singleton species $E(S(1))$ for an infinitely large sample ($E(S(1,\infty))$), first the consecutive values of z of R' fictitious sequential species are calculated from:

$$z_{R'} = z_o \, (1 - q)^{(R' - 1)} \qquad (16)$$

where z_o is the start-value for the abundances per species, which thus for safety is set at 30.

Next, the contributions $e^{-z} z$ per species R' to $E(S(1))$ (with $z = z_{R'}$) are calculated and finally these contributions are totalized in a comparable way as in Equation 11 for $n = 1$. The calculation of $z_{R'}$-values for R' ranging from 1 to infinite stops when the sum of the contributions is constant at an arbitrarily chosen precision of 10^{-6}.

Concave rank abundance curve

Curves 1 and 2 in Figure 2b show the course of the numbers of individuals (curve 1) and the contributions per species to the expected number of singleton species (curve 2) in an average sample for a fictitious rank abundance curve calculated from the negative-bino-

mial curve-fit model for $\mu = 9$ and a very low value for κ of 0.2. This is a deeply concave rank abundance curve. Also here, the curves seem to coincide. However, the factor difference in abundance between the sequential species, given by the term $[\mu / (\kappa + \mu)]$ $(\kappa + n - 1)/n$ in Equation 7b, will only approach constancy if n is so large that the right hand part $[(\kappa + n - 1)/n]$ almost equals 1. The number of singleton species for an infinitely large sample $E(S(1,\infty))$ can be calculated by taking that part equal to 1. That means that for infinitely large n, the multiplication factor for the sequential abundances per species defined in Equation 16 as $(1 - q)$ is now a constant equal to:

$$v = \mu / (\kappa + \mu) \tag{17}$$

The calculation of $E(S(1,\infty))$ can be summarized in one equation:

$$E(S(1, \infty)) = \sum_{1}^{R' = \infty} [z_o v^{(R'-1)} \exp\{-z_o v^{(R'-1)}\}] \tag{18}$$

where z_o is set at 30.

Properties of the *S-N* curve derived from a geometric series rank abundance curve

For the linear rank abundance curve calculated for $\kappa = 1$ and $\mu = 9$ (a geometric series curve for $q = 0.1$; Figure 3a) Table 1 shows that as soon as with increasing sample size (N), $E(S(1))$ from the discrete Poisson-curve becomes constant, there are fixed ratios between the numbers of species with 1, 2, ..., 6 individuals. Presented in the table are the calculated expected numbers of species with 1, 2, ..., 6 individuals [$E(S(1))$, $E(S(2))$,...., $E(S(6))$] of Poisson-curves calculated for the total numbers of individuals in the sample N = 25, 50, 100, 200, 400 and 800, respectively. The table shows that the expected numbers of species $E(S(2))$, $E(S(3))$, $E(S(4))$, ..., $E(S(n))$ finally end up in a constant ratio to $E(S(1))$ of exactly 1/2, 1/3, 1/4, ..., $^1/n$, respectively (see also Figure 3c). That means that for a large sample (large N) the total expected number of species ($E(S)$), is:

$$E(S) = E(S(1)) + E(S(2)) + E(S(3)) + ... + E(S(n))$$

or

$$E(S) = E(S(1)) \ \ (1 + \tfrac{1}{2} + \tfrac{1}{3} + ... + \tfrac{1}{n})$$

or

$$E(S) = E(S(1)) \sum_{n=1}^{n=h} (1 / n) \tag{19}$$

The expected total number of individuals ($E(N)$) is:

$$E(N) = E(S(1)) \sum_{n=1}^{n=h} (1 / n)n$$

or

$$E(N) = E(S(1))\ h \tag{20}$$

where h is the expected number of individuals of the most abundant (dominant) species.

Table 1. Numbers of species expected from Poisson-curves calculated from a rank abundance curve with $\kappa = 1$ and $\mu = 9$ for different numbers of individuals N. Calculated are the expected total numbers of species $E(S)$ and the expected numbers of species with 1 [$E(S(1))$], 2 [$E(S(2))$], 3 [$E(S(3))$], 4 [$E(S(4))$], 5 [$E(S(5))$] and 6 [$E(S(6))$] individuals. Presented are also values for the slope of the S-N curve calculated as tangent from the numbers of species found at consecutive values of N. For example: the slope at $N = 50$ is calculated as $\{E(S_{N100}) - E(S_{N25})\} / \{\ln(N_{100}) - \ln(N_{25})\}$, at $N = 100$ as $\{E(S_{N200}) - E(S_{N50})\} / \{\ln(N_{200}) - \ln(N_{50})\}$, etc. Calculated are also the ratios $E(S(1))/E(S(1))$, $E(S(2))/E(S(1))$, $E(S(3))/E(S(1))$ ..... $E(S(6))/E(S(1))$, etc., see text.

	Number of individuals (N)					
	25	50	100	200	400	800
$E(S)$	14.87243	21.26184	27.83285	34.41164	40.99045	47.56927
Slope		9.348963	9.485577	9.491207	9.491222	9.491222
$E(S(1))$	8.812036	9.442932	9.490982	9.491222	9.491222	9.491222
$E(S(2))$	3.509340	4.593649	4.744217	4.745611	4.745611	4.745611
$E(S(3))$	1.551079	2.837092	3.158299	3.163740	3.163741	3.163741
$E(S(4))$	0.643787	1.830141	2.356743	2.372803	2.372805	2.372805
$E(S(5))$	0.241438	1.149816	1.859970	1.898233	1.898244	1.898244
$E(S(6))$	0.081243	0.681900	1.505057	1.581826	1.581870	1.581870
$E(S(1))/E(S(1))$	1	1	1	1	1	1
$E(S(2))/E(S(1))$	0.398244	0.486464	0.499866	0.500000	0.500000	0.500000
$E(S(3))/E(S(1))$	0.176018	0.300446	0.332768	0.333333	0.333333	0.333333
$E(S(4))/E(S(1))$	0.073058	0.193811	0.248314	0.250000	0.250000	0.250000
$E(S(5))/E(S(1))$	0.027399	0.121765	0.195972	0.199999	0.200000	0.200000
$E(S(6))/E(S(1))$	0.009220	0.072213	0.158578	0.166662	0.166667	0.166667

The S-N curve 1 in Figure 3d shows the result of calculations with Equations 19 and 20. For large N, the curve appears to perfectly fit the points (open squares) for S and N calculated from the Poisson-curves (the values in Table 1 and some more from supplementary calculations). Presented in Table 1 are also the slope-values of the S-N curve calculated as tangents from the single S- and N-values of the Poisson-curves. That is, the slope at $N = 50$ calculated as tangent (best estimate) from $[E(S_{N100}) - E(S_{N25})] / [\ln(N_{100}) - \ln(N_{25})]$, the slope at $N = 100$ calculated as tangent from $[E(S_{N200}) - E(S_{N50})] / [\ln(N_{200}) - \ln(N_{50})]$, etc. These slope-values are all very similar to the values of $E(S(1))$ indeed, which thus confirms that $E(S(1))$ is the slope of the S-N curve.

Equation 19 explains why the curves for the expected numbers of species with at least 2, or at least 3, or at least i individuals against log N have the same slope as the curve for the total number of species (Figures 3b). The series in Equation 19 is the harmonic series, which apparently follows from the specific properties of the Poisson-distribution.

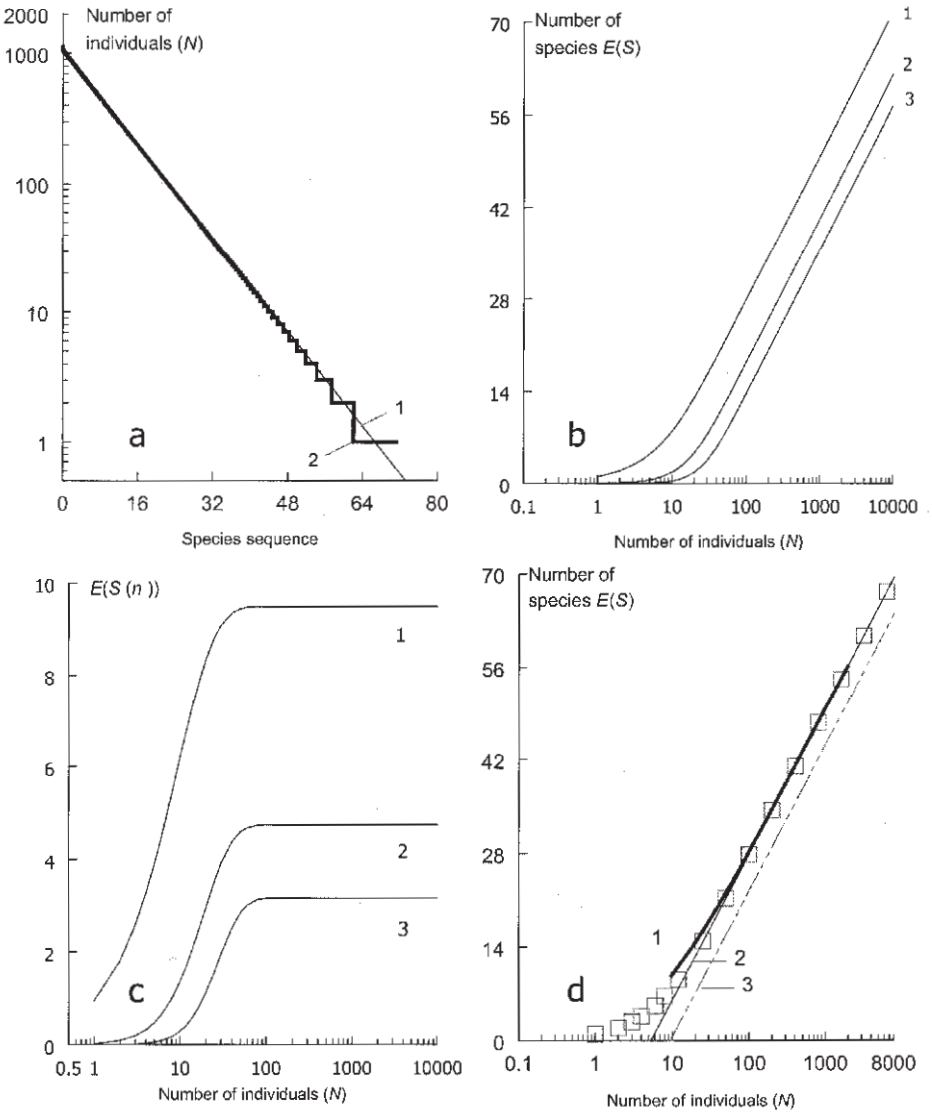

Figure 3. (a) Negative-binomial rank abundance curve (curve 1) and Poisson-curve (curve 2) calculated for $\kappa = 1$, $\mu = 9$ and $N = 10{,}000$. (b) Total number of species (curve 1), number of species with at least 2 individuals (curve 2) and number of species with at least 3 individuals (curve 3) versus log N. (c) Number of singleton species [$E((S(1))$, curve 1], number of species with 2 individuals [$E(S(2))$, curve 2] and number of species with 3 individuals ([$E(S(3))$, curve 3] versus log N. (d) Total number of species [$E(S)$] versus log N. □: data from the single Poisson-curve calculations. Curve 1 is based on calculations using the values in Table 1 (see text) and on supplementary calculations; curves 2 and 3 were calculated from Equation 22, including γ and stripped of the effect of γ, or from Equation 23 including w and stripped of the effect of w, respectively. For the equations see text.

An alternative expression for Equation 19

For very large h the progression $\sum_{n=1}^{n=h}(1/n)$ may be approached by (Abramowitz & Stegun, 1995):

$$\sum_{n=1}^{n=h}(1/n) = \ln(h) + \gamma \qquad (21)$$

which means that Equation 19 can be written as (see also Equation 20):

$$E(S) = E(S(1))[\ln(\frac{E(N)}{E(S(1))}) + \gamma] \qquad (22)$$

where γ is the Euler-Mascheroni constant, equal to 0.577216.

The curve calculated from Equation 22, is the linear curve 2 in Figure 3d. Curve 3 in the same figure shows the linear curve that would have resulted when leaving out γ. Parameter γ causes that the linear part of the *S-N* curve is shifted to the left.

Equation 22 may also be written as:

$$E(S) = E(S(1))[\ln(\frac{E(N)}{E(S(1))}) + \ln(e^{\gamma})] = E(S(1))\ \ln(\frac{E(N)}{E(S(1))}\ e^{\gamma})$$

or as

$$E(S) = E(S(1))[\ln(\frac{E(N)}{E(S(1))})\ w), \qquad (23)$$

in which $w = e^{\gamma} = 1.781073$.

S-N relationships for concave rank abundance curves

For concave rank abundance curves (Figure 4a) calculated from the negative-binomial model for $\kappa < 1$, the numbers of species with 1, 2, 3, ..., n individuals ultimately only approach constant ratios of also 1, 1/2, 1/3, 1/4, ..., $^1/n$ for large N (Figure 4b). $E(S(1))$, which in principle can be calculated (via the procedure of Poisson-curve calculations) for each sample size (N), resulting in a specific value for w, is the slope of the *S-N* curve at consecutive points along the curve.

The curves in Figure 4 were calculated from species abundance proportions calculated from the negative-binomial curve-fit model for $\kappa = 0.5$ and $\mu = 9$. The fitted continuous rank abundance curve (curve 1) and the discrete Poisson-curve (curve 2) in Figure 4a were calculated for $N = 10{,}000$. The value for w iterated from Equation 23 for $N = 16{,}384$ (2^{14}), with $E(S) = 109.599$ and $E(S(1)) = 17.008$, is 0.652766. $w < 1$ causes a shift of the approximately linear part of the *S-N* curve to the right (Figure 4c; compare curve 2 (inclusive-) with curve 3 (exclusive of the effect of w)). The straight line 4 in Figure 4b represents the value of $E(S(1))$ for an infinitely large sample [$E(S(1,\infty)) = 18.495$].

Both $E(S(1))$ and w are dependent on N and become constant in the range of large samples. $E(S(1))$ is linearly related to μ for given N and given κ (Figure 5), w is curvilin-

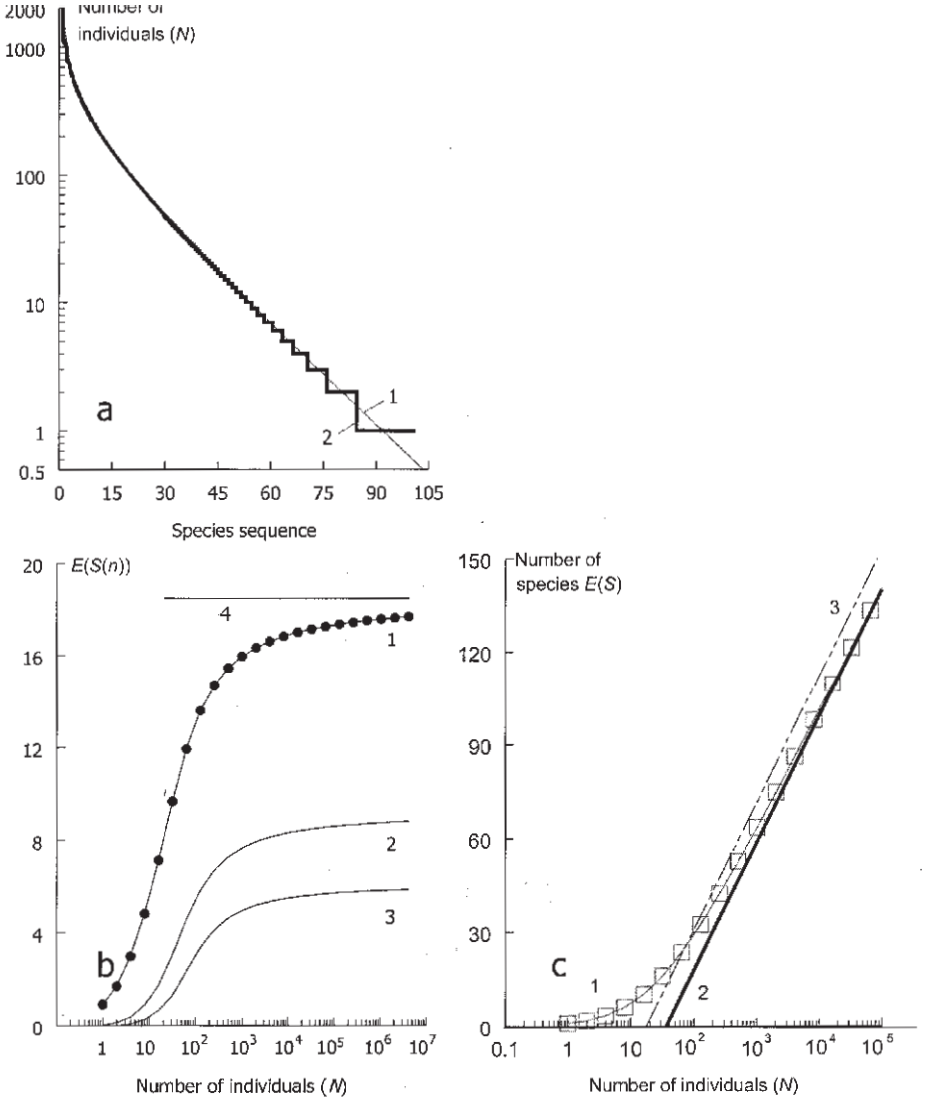

Figure 4. (a) The same type of curves as in Figure 3a but with $\mu = 9$, $\kappa = 0.5$ and $N = 10{,}000$. (b) The same type of curves [curve 1, $E(S(1))$, curve 2 $E(S(2))$ and curve 3 $E(S(3))$] as in Figure 3c. Curve 4 shows the number of singleton species for infinitely large N, $E(S(1,\infty))$. (c) $E(S)$ versus log N curves (see Figure 3d and text). Curves 2 and 3 were calculated from Equation 23 (see text), including w and stripped of the effect of w, respectively. The black dots in (b) and the open squares in (c) are the data from the single Poisson-curve calculations.

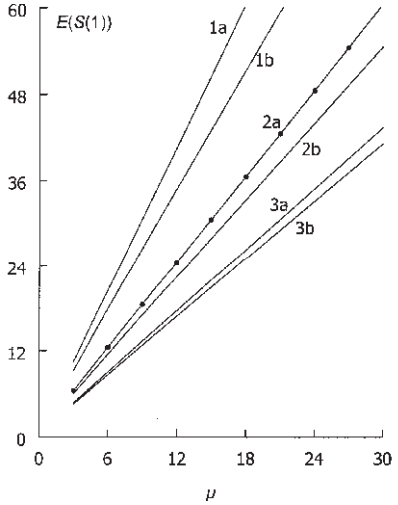

Figure 5. Relation between $E(S(1))$ and μ for the following cases: $\kappa = 0.3$ and N infinitely large (curve 1a), $\kappa = 0.3$ and $N = 16{,}384$ (curve 1b), $\kappa = 0.5$ and N infinitely large (curve 2a), $\kappa = 0.5$ and $N = 16{,}384$ (curve 2b), $\kappa = 0.7$ and N infinitely large (curve 3a), and $\kappa = 0.7$ and $N = 16{,}384$ (curve 3b). The dots in curve 2a are from the single Poisson-curve calculations.

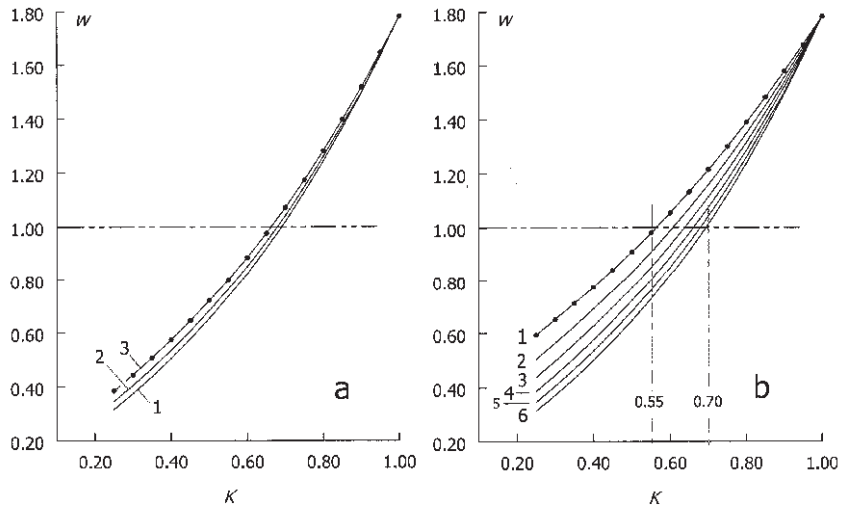

Figure 6. (a) Relation between w and κ for given values of μ, and N constant at 16,384; cases $\mu = 9$ (curve 1), $\mu = 18$ (curve 2) and $\mu = 36$ (curve 3). (b) Relation between w and κ for different given values of N, and μ constant at 18; cases $N = 1024$ (curve 1); $N = 2048$ (curve 2); $N = 4096$ (curve 3); $N = 8192$ (curve 4); $N = 16{,}384$ (curve 5) and $N = 32{,}768$ (curve 6). The dots in curve 3 in (a) and in curve 1 in (b) are based on the single Poisson-curve calculations. The curves through the calculated points in both figures were fitted by a 6-term polynomial multiple linear regression.

early related to κ for given N (Figure 6a) and given μ (Figure 6b). The curves in Figures 6a and 6b were fitted using a 6-term polynomial multiple linear regression.

Relation to log-series α

For $w = 1$, Equation 23 looks very similar to the log-series equation for S (Fisher *et al.*, 1943):

$$S = \alpha \ln(1 + \frac{N}{\alpha}) \tag{24}$$

and in case of a very large sample it should almost equal it because in that case the value '1' in the equation can be neglected. Calculations showed that this is true. From a rank abundance curve for $\kappa = 0.679$, $\mu = 18$, and $N = 16{,}384$ (2^{14}), i.e., a combination satisfying $w = 1$, we calculated values for $E(S)$ and $E(S(1))$ of 165.5 and 25.62, respectively. The same values for $E(S)$ and N yielded a log-series $E(S(1))$ and log-series α (using Equations 2–5) of 25.62 and 25.61, respectively, with an x from iteration (Equation 5) of 0.998439. The value for κ was calculated from curve 5 ($\mu = 18$, $N = 16{,}384$) in Figure 6b at the value for $w = 1$.

κ for $w = 1$ slightly decreases with decreasing sample size (N) and increasing μ but usually will not reach values lower than 0.5 (Figure 6b), which confirms that the log-series model can only fit relatively shallow rank abundance curves. Parameter w causing the part of the *S-N* curve for large N to shift along the horizontal axis to the left (Figure 3d) or to the right (Figure 4c) could be called the 'shifting' factor.

For further explanation of the similarities and differences between the parameters of the *S-N* curve according to Fisher's log-series model and our negative-binomial curve-fit model we have created Figure 7. The *S-N* curve in Figure 7a is generated from rank abundance curves for $\kappa = 0.679$, $\mu = 18$ and consecutive values for N. The 'Poisson'- rank abundance curve from which $E(S)$ for $N = 16{,}834$ was calculated (curve not presented) is a curve satisfying $w = 1$, and thus a log-series rank abundance curve. The linear curve 2 in Figure 7a is calculated from Equation 23, using the 'Poisson'-$E(S(1))$ for $N = 16{,}834$. Figure 7b shows the course of the 'Poisson'-$E(S(1))$ (curve 1) and the level of $E(S(1,\infty))$ (curve 2). It also shows the course of the log-series $E(S(1))$ (curve 3) and log-series α (curve 4) as calculated from the *S-N* curve in Figure 7a by iterating x for each of the *S-N* combinations, using Equation 5. Moreover, it shows the log-series α (curve 5) as best fitting value from iteration using Equation 24 in a direct log-series fit of the entire *S-N* curve in Figure 7a. Curve 6 finally shows the course of the log-series $E(S(1))$ as calculated by iterating x using Equation 5 for each *S-N* combination of the log-series *S-N* curve created from that α.

From the fact that curves 1, 3 and 6 almost coincide it may be concluded that the $E(S(1))$ we calculated has the same meaning as the $E(S(1))$ of the log-series model. $E(S(1))$ is the tangent of the *S-N* curve and depends on the size of the sample for which it is calculated. Fisher's log-series α soon becomes independent of sample size. The $E(S(1,\infty))$ of the new model is the extrapolated $E(S(1))$ for an infinitely large sample, calculated from Equation 18. Since it can be calculated for each fitted concave rank abundance curve independent of the given size of the sample, it could replace log-series α as a more

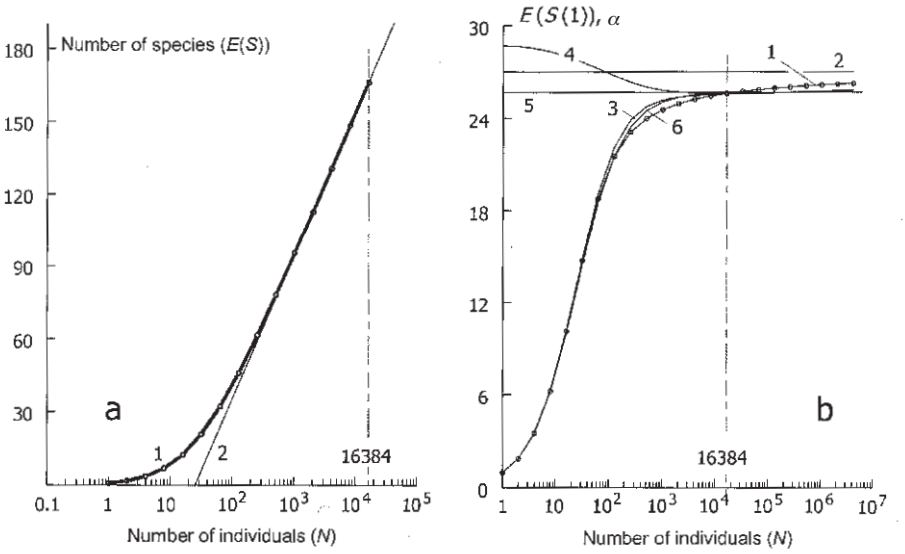

Figure 7. (a) Number of species ($E(S)$) versus the total number of individuals (N) (curve 1) with values for $E(S)$ calculated from 'Poisson'- rank abundance curves for $\kappa = 0.679$ and $\mu = 18$ by varying N. The 'Poisson'- rank abundance curve yielding the $E(S)$ for $N = 16{,}834$ (curve not shown) is a curve satisfying $w = 1$, and thus a log-series rank abundance curve. Curve 2 was calculated from Equation 23 using the 'Poisson'-$E(S(1))$ for $N = 16{,}834$. (b) The 'Poisson'-$E(S(1))$ (curve 1), the level of $E(S(1,\infty))$ (curve 2), the log-series $E(S(1))$ (curve 3), log-series α (curve 4) calculated from the S-N curve in (a) by iterating x for each S-N combination using Equation 5, log-series α (curve 5) as best fitting value using Equation 24 in a direct log-series fit of the entire S-N curve in (a), and the log-series $E(S(1))$ (curve 6) calculated by iterating x for each S-N combination of the latter curve, using Equation 5. For the equations see text.

suitable site discriminant. It has the additional advantage that it can also be calculated in case the species abundances are given in terms of proportions because, as follows from Equation 18, no N and no S are needed for its calculation.

Procedure of curve fitting and criteria for goodness of fit

Curve fitting with the new model is on proportions, which means that in case the abundances of species are given in terms of numbers of individuals, these numbers first have to be converted into proportions. Best fitting values for κ and μ are found by iteration, using the method of the least squares of the deviations between the points from observation (species abundances on log scale) and those on the fitted curve. Wilson (1991) used this method for fitting rank abundance data to the log-normal model and the broken-stick model. However, he calculated the least sum of squares as criterion for best fit. We prefer to use the least *mean* square of the deviations (mean least square deviance, D_{lsq}) because that enables a better comparison of data sets with different sample sizes. Best fitting value for c in case of Equation 8b can also be found by iteration. However, its

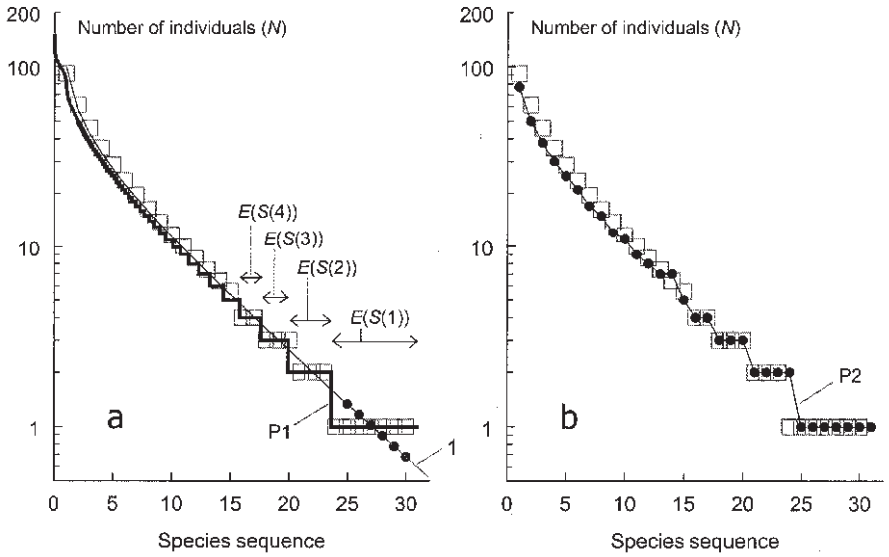

Figure 8. (a) Rank abundance curves. Curve 1 is the continuous curve fitted through the data (□) by the negative-binomial curve-fit model. P1 is the Poisson-curve derived from curve 1 with discrete numbers of individuals versus the accumulated species frequencies, i.e., the accumulated numbers of species calculated for consecutive values of n in Equation 11, as species rank. The calculated numbers of species with 1, 2, 3 and 4 individuals are indicated in the graph as $E(S(1))$, $E(S(2))$, $E(S(3))$ and $E(S(4))$, respectively. The six black dots on the continuous curve indicate the values for species 25 up to and including 30. (b) P2 is the Poisson-curve with discrete numbers of individuals for discrete sequential species (for explanations see text). The black dots on the curve are calculated through interpolation (see text). The P2-curve can be used for calculating a mean least square deviance versus the data from observation.

value can be determined much faster from a level difference, as will be explained below in the validation of the model with observational data.

A mean square deviance versus the data from observation can also be calculated for the Poisson-curve derived from the continuous curve fitted by the negative-binomial curve-fit model and for a theoretical rank abundance curve generated from the log-series model. However, to that end first the continuous scale of the horizontal axis with totalized expected species frequencies (Equation 12) has to be converted into a scale with discrete sequential species. This is done by interpolation. The procedure is that the frequencies are totalized, starting with the expected number of species for the highest theoretical plant number (for safety reasons set at n = 10,000 in Equation 11). As soon as the totalized frequency is 1, the numbers of individuals belonging to the last and to the second last added frequency are read from an array. The expected number of individuals for species 1 (the dominant species) is assumed to lie between these two values, and can thus be calculated by interpolation. Totalizing the frequencies is continued until the totalized frequency is 2 (enabling the calculation of the best estimate of the expected number of individuals for species 2), etc. Since the sum of all frequencies can end up in a value with a digit behind the decimal point, the ultimate expected number of species

Table 2. Analysis of four data sets: (1) numbers of moths (Taylor & French, 1974); (2) numbers of Coleoptera in a pasture, Zuid Limburg, The Netherlands; (3) numbers of diatom species counted in 4 experimental boxes in part of the flow of Darby Creek, Pennsylvania, USA (Patrick, 1968; experiment 1966); (4) species abundance of arboreal arthropods associated with an Australian rainforest tree (Basset & Kitching, 1991). Presented are the data (*N*, *S* and *S*(1)) directly from sampling (Sampling data), parameter values (κ, μ and *c*) of the fitted negative-binomial curve, values calculated from the Poisson-curve derived from it (*Rcalc-N*, *E*(*S*) and *E*(*S*(1))), and Fisher's α from a direct log-series fit. *N*, *S* and *S*(1) are the total number of individuals, the total number of species and the number of singleton species from sampling, respectively. D_{lsq} is the mean least square deviance between the fitted curve and the points from observation. *Rcalc-N*, *E*(*S*) and *E*(*S*(1)) are the re-calculated total number of individuals, the expected total number of species and the expected number of species with one individual as calculated from the Poisson curve, respectively. $E(S(1,\infty))$ is the expected number of singleton species in an infinitely large sample. Curves were fitted with parameter *c* set at 1 (Equation 8a, see text) and *c* calculated as level difference between the curve fitted by κ and μ and the points from observation (Equation 8b, see text).

Data set	Sampling data			Negative-binomial curve				Poisson curve						Log-series curve	
	N	*S*	*S*(1)	κ	μ	*c*	D_{lsq}	*Recalc-N*	*E*(*S*)	D_{lsq}	*S*-range	*E*(*S*(1))	$E(S(1,\infty))$	α	D_{lsq}
Moths	4046	89	22	0.493	9.147	1	0.0530	4046	87.8	0.0405	1–88	17.1	19.06	16.1	0.0711
				0.487	9.068	0.989	0.0530	4091	87.8	0.0401		17.1	19.10		
Coleoptera	1231	42	10	0.195	3.251	1	0.0854	1231	44.8	0.1011	1–42	12.78	17.15	8.41	0.3739
				0.199	3.298	1.016	0.0854	1212	44.8	0.1013		12.76	17.10		
Diatom spp.	57539	180	27	0.133	5.380	1	0.0496	57539	187.8	0.0491	1–180	34.22	40.98	23.0	0.8270
				0.109	4.487	0.827	0.0488	69615	188.2	0.0477		34.43	41.50		
Arthropods	22940	773	271	0.182	54.950	1	0.0877	22940	776.6	0.0764	1–773	223.9	302.6	154.3	0.5682
				0.090	29.642	0.524	0.0812	17759	785.9	0.0678		232.5	330.9		

indicated by the last species is determined by rounding off. Examples of Poisson-curves with numbers of species (horizontal axis) on a continuous scale and on a scale with discrete numbers of species are given in Figures 8a and 8b, respectively.

To compare the quality of fit of different models the calculated mean squares of the deviations can be used in an F-test. Another possibility is to use a χ^2-test applied to the frequency distributions of the actual and expected numbers of species in $\log_2$-classes of numbers of individuals according to the method of Preston (1948). However, as was explained by Hughes (1986) and Wilson (1991), a χ^2-test leads to loss of information and is therefore much less accurate.

Model validation

Parameter *c*

With Equation 8a we try to find the best fitting curve with abundance proportions from curve fit that add up to 1. With Equation 8b the best *shape* of the curve is fitted with abundance proportions that after correcting for a level difference on log scale between the fitted curve and the points from observation can add up to a value larger or smaller than 1. The consequence is that after re-conversion of the abundance proportions from curve fit into numbers of individuals (Equation 9) the re-calculated total number of individuals can be higher or lower. However, that is no problem as it only suggests that the actual numbers of individuals per species in the sampled plant community are higher or lower indeed, which is quite well possible. Adding c improves the quality of fit as is demonstrated in Table 2 where parameter values are presented of curves fitted with both Equation 8a ($c = 1$) and Equation 8b.

Adding parameter c does not invalidate the calculation of $E(S(1,\infty))$ as site discriminant. As explained before, the latter is based on the moment in the fitted curve where the subsequently calculated species abundance proportions start to differ by a constant factor $v = \mu / (\kappa + \mu)$, see Equation 17. That moment can be calculated for each asymptotically ending concave rank abundance curve, regardless of whether or not the curve is corrected for a level difference versus the points from observation.

Since parameter c (Equation 8b) is the factor that corrects for the level difference between the abundance proportions on log scale [Note that $\log(p_R) = \log(f(R-1) - \log(c)$], its value can be determined as such; c can be calculated from the mean of the relative frequencies (λ_f) calculated from the submitted values of μ and κ at the subsequent steps of the iteration and the mean of the abundance proportions (λ_p) from data according to:

$$c = \lambda_f / \lambda_p \tag{25}$$

Correction for level difference is carried out at each step of the iteration before calculating the mean least square of the deviances (D_{lsq}) between the points from observation and those from the fitted curve. Determining c that way is much faster than finding the best fitting value for it by treating it as an independent third iterable variable.

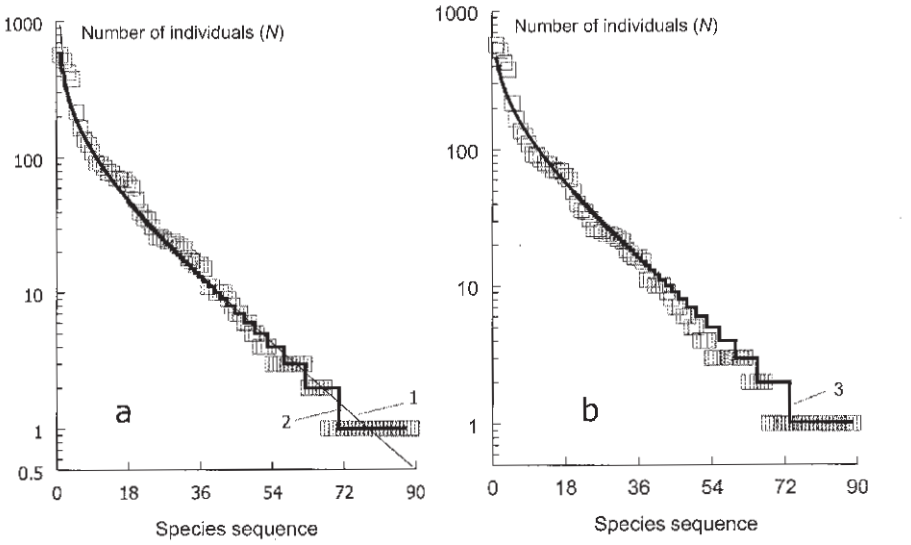

Figure 9. Numbers of moths caught in light-traps using Robinson mercury-vapour lamps. (a) □: data from sampling. Continuous curve fitted by the negative-binomial curve-fit model (curve 1) using Equation 8b and Poisson-curve (curve 2) derived from it. (b) Curve 3 is from a direct log-series fit to the data. Data from Taylor & French (1974).

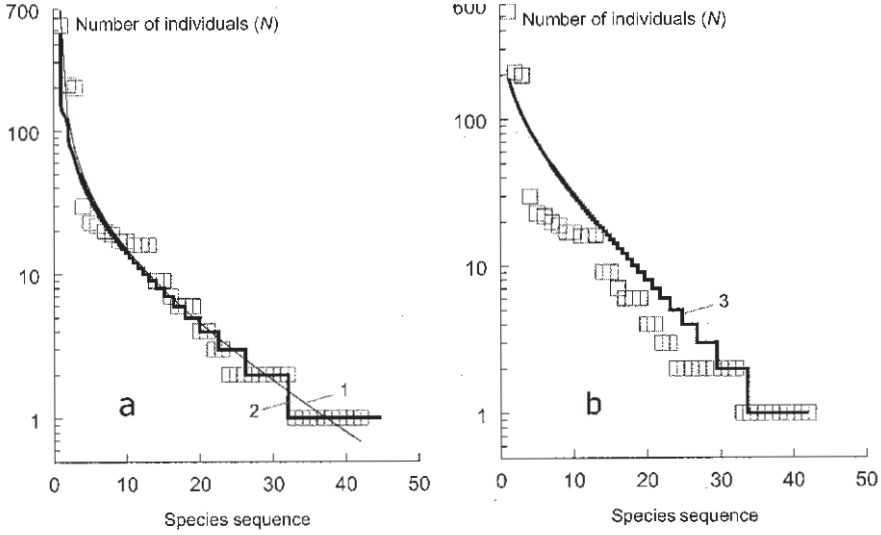

Figure 10. Numbers of individuals versus species rank for Coleoptera specimens trapped in five rectangular plots in a pasture on 'Pietersberg', Zuid Limburg, The Netherlands. □: data from sampling. (a) Continuous curve fitted by the negative-binomial curve-fit model (curve 1) using Equation 8b (see text), and Poisson-curve (curve 2) derived from it. (b) Curve 3 is from a direct log-series fit to the data.

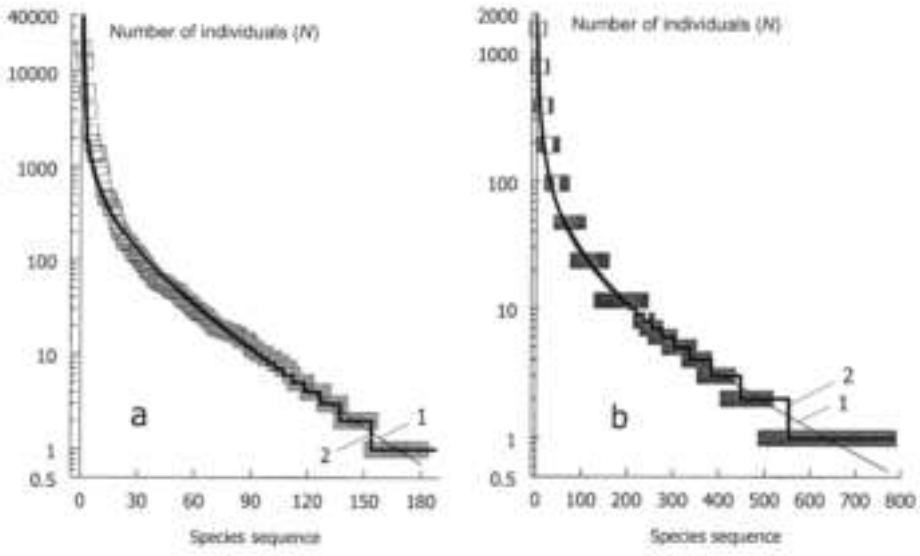

Figure 11. Rank abundance curves. (a) Diatom specimens counted in four experimental boxes placed in part of the flow of Darby Creek, Pennsylvania, USA (Patrick, 1968; experiment 1966). (b) Arboreal arthropods associated with an Australian rainforest tree (Basset & Kitching, 1991). □: data from sampling. Continuous curves fitted by the negative-binomial curve-fit model (curve 1), using Equation 8b (see text), and Poisson-curve (curve 2) derived from it.

Fitting data from observations

Four sets of rank abundance data with clear indications for a concave curve were used to test the negative-binomial curve-fit model: (1) a data set (Figure 9) from Taylor & French (1974) (numbers of moths obtained from light-trap samples using Robinson mercury-vapour lamps), (2) a data set (Figure 10) from Coleoptera (ground beetles) research (Coleoptera specimens counted by pitfall trapping in 5 rectangular plots in a grassland on 'Pietersberg', Zuid Limburg, The Netherlands), (3) a data set (Figure 11a) from Patrick (1968) (diatom specimens counted in 4 experimental boxes placed in part of the flow of Darby Creek, Pennsylvania, USA, experiment-1966), and (4) a data set (Figure 11b) from Basset & Kitching (1991) (species abundance of arboreal arthropods associated with an Australian rainforest tree).

Parameter values and statistics of the fitted curves using both Equation 8a ($c = 1$) and Equation 8b (c iterated or determined as level difference) are summarized in Table 2. Given are also Fisher's α and D_{lsq} from a direct log-series fit. In the moth case (Taylor & French, 1974) the log-series model fitted the data reasonably well (Figure 9b; D_{lsq} = 0.0771). However, the curve fitted by the new model was slightly better (Figure 9a; values for D_{lsq} 0.0530 and 0.0405 (Equation 8a) or 0.0530 and 0.0401 (Equation 8b), respectively); log-series $\alpha = 16.1$ versus $E(S(1)) = 17.1$ (Poisson-curve) and $E(S(1,\infty)) = 19.1$. As illus-

trated for the Coleoptera data in Figure 10b, the log-series model could not fit the data of the three remaining data sets. All three (all cases with a low value for κ) were well fitted by the new curve-fit model [see Figures 10a, 11a and 11b and the relatively low values for D_{lsq} in Table 2 varying from 0.0488 to 0.0877 (continuous curve) or from 0.0477 to 0.1013 (Poisson-curve)]. In all four data sets (Table 2) the Poisson-curve resulted in a reasonable estimate of the actual number of species in the sample. The re-calculated total numbers of individuals (*Recalc-N*) from curve fit using Equation 8a ($c = 1$) equalled the actual total numbers of individuals in the sample. This indicates that in Equation 11 a sufficient number of species (R) and in Equation 12 a sufficient number (n) of Poisson-terms had been included in the calculations. Along with sometimes relatively large differences in fitted values for μ and κ, relatively small differences were found between the values for $E(S(1))$ and $E(S(1,\infty))$ when Equations 8a and 8b were used. However, in case of differences, best fit was always obtained with Equation 8b. This equation with parameter c should always be applied, as is illustrated for a specific case in Appendix 1.

Discussion and conclusions

With the log-series model (Fisher *et al.*, 1943), rank abundance data are fitted by iterating only one parameter (x in Equation 5). With the new model, curves are fitted with two iterable parameters (κ and μ, see k and m in Equations 7a, b). This could explain why, even in cases with an acceptable log-series fit, the fit obtained with the new model is nearly always better.

The new model links the geometric-series model and log-series model and can also fit deeply concave rank abundance curves. Moreover, it can calculate a species-diversity index (the number of singleton species in an infinitely large sample, $E(S(1,\infty))$. The new model could solve the problem discussed by Hughes (1986) that deeply concave curves cannot be fitted by any of the existing rank abundance models. The log-series model (Fisher *et al.*, 1943) is not suitable because that model can only fit shallow rank abundance curves whereas the Zip-Mandelbrot model used by Wilson (1991) as another alternative for concave curves is inaccurate. Hughes (1986) suggested using his iterative dynamics model as a more flexible alternative for curve fitting. However, that model is too complex and lacks the ability to calculate a species-diversity index. Moreover, it is not really a statistical model and cannot fit proportions.

It may be questioned whether the basic assumption of the model of Poisson-distributed numbers of individuals within species in replicate samples is realistic because clustering is likely to occur always. On the other hand, the model may be realistic for very large samples because whether or not clustering will have an effect on the recorded numbers of individuals of a species in a sample depends on the cluster size to sample size ratio. That is, a species with strong clustering can be completely dominant or be totally absent in small samples whereas in a very large sample its recorded number of individuals will probably hardly differ from that of an equally abundant species with Poisson-distributed individuals. With the further, reasonable assumption that the number of species in the sampled community is finite, the calculation of the $E(S(1,\infty))$ for an infinitely large sample as site discriminant seems therefore justified. Adding the

condition of a finite number of species is needed as in that case it may be assumed that in an infinitely large sample of even the rarest species a sufficient number of clusters will be present to make that the number of individuals is almost fixed. Kempton & Taylor (1974) used the Poisson-distribution for transforming a continuous log-normal rank abundance curve into a curve with discrete numbers of individuals for species (the 'Poisson'- log-normal). And as stated before, also Fisher's log-series model (Fisher *et al.*, 1943) is based on the assumption of Poisson-distributed numbers of individuals in replicate samples.

Table 2 shows that despite the possible effects of clustering with small samples, even for the Coleoptera data set with a total of only 1231 individuals, the expected number of species ($E(S) = 44.8$) re-calculated from the Poisson-curve still reasonably agreed with the actually counted number of species ($S = 42$). The latter number was the total of five replicate samples with a fairly large variation. Possibly due to the bulking of randomly distributed replicate samples to one large sample, the majority of all occurring species can still be caught. However, using that value of S in an S-N curve is one option. In vegetation science the numbers of species are usually counted in series of nested expanding quadrats for making an S-N curve or *species-area* curve. The S plotted that way against N or against area in case of replicates is the S of an average single sample of increasing size (Condit *et al.*, 1996). The way in which S is calculated will be discussed in a next paper (Neuteboom & Struik, 2005a). S calculated as the mean number of species per sample is strongly dependent on clustering.

A rank abundance curve is usually made by totalling the numbers of individuals per species in a series of replicate samples. With numbers of individuals plotted on log scale against species sequence, the resulting curve is the same as the curve for the average species individual numbers per sample, with only a level difference. Since the mean number of specimens of a species is not affected by clustering, the curve for the average- (the 'average' rank abundance curve) and thus the curve for the total number of individuals per species (the 'total' rank abundance curve) is not affected by clustering. This means that the continuous rank abundance curve fitted by the negative-binomial curve-fit model represents in principle a curve that is free from clustering. Clustering could lead to a larger standard error. However, taking more samples can reduce that error.

Basset & Kitching (1991) stated that they could not fit their data with the log-series model. The same applies to the presented Patrick-data (experiment 1966) (Table 2). Patrick (1968) stated that the structure of diatom communities simulates a log-normal curve and for large samples a truncated log-normal. In our opinion the Basset & Kitching data (Figure 11b) and the Patrick data (Figure 11a) represent deeply concave rank abundance curves with a high species diversity as expressed by high values of $E(S(1,\infty))$.

By taking the relative frequency $f(0)$ for the abundance proportion of the first (dominant) species, the $f(1)$ for the second species, etc., we use in fact the negative-binomial distribution as a series, like the geometric series, in which after extrapolation the proportions of all species add up to 1. Since in the first instance proportions are fitted (even in case the abundances of species are given in numbers of individuals), the model can be used for fitting numbers of individuals as well as abundance proportions of species. As stated before, the model can calculate $E(S(1,\infty))$ for both cases without the need to know N or S.

Application of the equation for the *S-N* curve developed from the new model (Equation 23) shows that depending on the concavity of the rank abundance curve of the sampled community, the part of the *S-N* curve for large *N* is shifted along the x-axis to the left (Figure 3d) or to the right (Figure 4c). That effect is quantified in a 'shifting' parameter *w*. The log-series model applies to communities with a shallow rank abundance curve for which $w = 1$ in the new equation for the *S-N* curve. Fisher's log-series equation for the *S-N* curve (Equation 24) is lacking a shifting parameter and is therefore incomplete. A final conclusion could be that a species-diversity index (Fisher's α) as an integral property of the rank abundance and *S-N* curve is a specific property of the log-series model and only valid for communities with a typical shallow log-series rank abundance curve.

Acknowledgements

We are grateful to H. Turin, Stichting Faunistisch Onderzoek Carabidae (Loopkeverstichting), 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

Abramowitz, M. & I.A. Stegun, 1965. Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables. United States Department of Commerce, National Bureau of Standards, New York, 1045 pp.

Bliss, C.I. & R.A. Fisher, 1953. Fitting the negative binomial distribution to biological data. *Biometrics* 9: 176–200.

Basset, Y. & R.L. Kitching, 1991. Species number, species abundance and body length of arboreal arthropods associated with an Australian rainforest tree. *Ecological Entomology* 16: 391–402.

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: 197–200.

Fisher, R.A., A.S. Corbet & C.B. Williams, 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. *Journal of Animal Ecology* 12: 42–58.

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

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

Magurran, A.E., 1988. Ecological Diversity and its Measurement. Princeton University Press, Princeton,, 192 pp.

Neuteboom, J.H. & P.C. Struik, 2005a. Variation in rank abundance replicate samples and impact of clus-

tering. *NJAS – Wageningen Journal of Life Sciences* 53: 199–221.

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 (924): 173–183.

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.

Taylor, L.R., 1978. Bates, Williams, Hutchinson – a variety of diversities. In: L.A. Mound & N. Warloff (Eds), Diversity of Insect Faunas. 9th Symposium of the Royal Entomological Society. Blackwell, Oxford, pp. 1–18.

Taylor, L.R. & R.A. French, 1974. Effects of light trap design and illumination on samples of moths in an English woodland. *Bulletin of Entomological Research* 63: 583–594.

Wilson, J.B., 1991. Methods for fitting dominance / diversity curves. *Journal of Vegetation Science* 2: 35–46.

Appendix 1

Details on parameter *c* in Equation 8b

For curve fitting, the species abundances are first transformed into proportions adding up to 1. Things may especially go wrong with Equation 8a if a substantial part of the low-abundant and rare species is missing. Figures 2.1a–2.1c illustrate this. Given in Figure 2.1a is a Poisson-rank abundance curve created by the negative binomial curve-fit model for $\mu = 3$, $\kappa = 0.3$ and a total number of individuals in the sample of $N = 12{,}000$ (curve 1). Suppose that the curve is based on real data (sample 1), and next that only species $S1$–$S5$ are present in the sample with 5845, 1594, 942, 656 and 492 individuals, respectively (curve 2, sample 2). The total number of individuals N is 9529, and the abundance proportions of species $S1$–$S5$ calculated on that basis are 0.613, 0.167, 0.099, 0.069 and 0.052, respectively (Figure 2.1b, curve 2). These proportions are logically higher than the respective proportions 0.487, 0.133, 0.078, 0.055 and 0.041 for species $S1$–$S5$ in sample 1 (Figure 2.1b, curve 1) while the fitted rank abundance curve with $\mu = 1.795$ and $\kappa = 0.436$, using Equation 8a, is unsatisfactory (Figure 2.1c, curve 2a). Curve fitting with Equation 8b results in a perfect fit with the original $\mu = 3$ and $\kappa = 0.3$ and a value for c of 0.795 (Figure 2.1c, curve 2b). However, the abundance proportions from curve fit calculated for an infinite number of species now add up to a value larger than 1 (sum of proportions = 1.258). The latter is no problem because multiplying the proportions from curve fit by the lower N of 9529 results in the original numbers of individuals for species $S1$–$S5$ with a re-calculated N from curve fit of 12,000. As stated before, including parameter c in Equation 8a does not affect the validity of calculating an $E(S(1,\infty))$ as site discriminant.

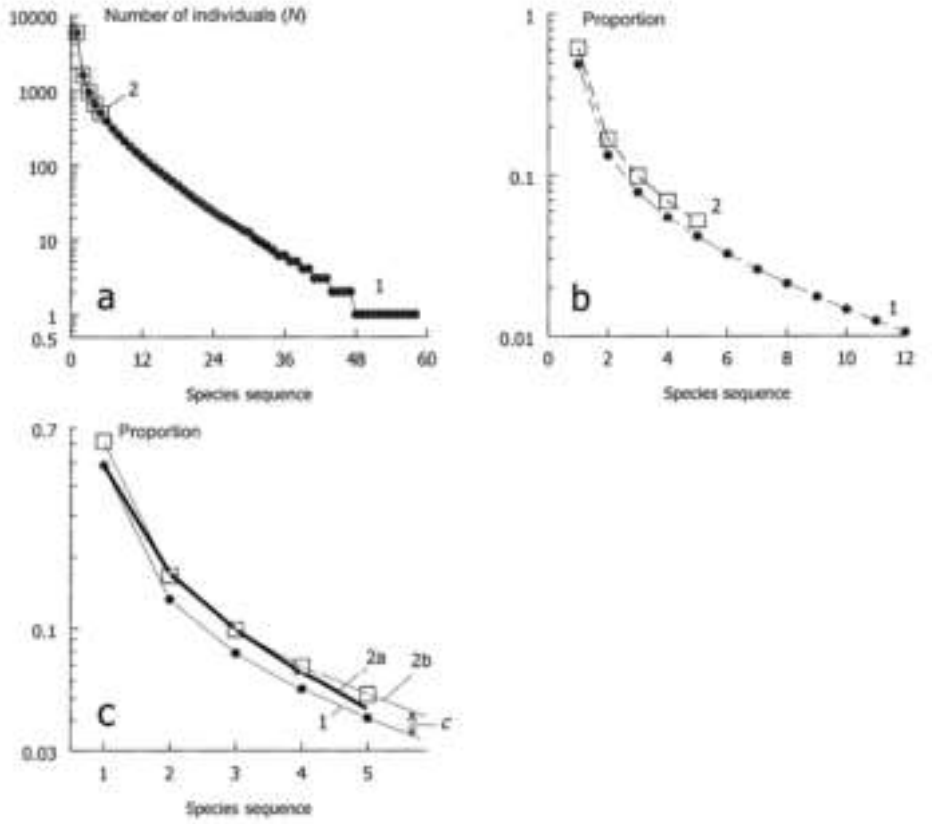

Figure 2.1. (a) Sample 1 (dots, curve 1); species individual numbers in a fictitious sample with 58 species. The numbers of individuals are the numbers from a Poisson-rank abundance curve created by the negative-binomial rank abundance curve-fit model for $\mu = 3$, $\kappa = 0.3$ and $N = 12{,}000$. Sample 2 (open squares, curve 2); a sample with only the species $S1–S5$ with the same numbers of individuals per species. (b) The numbers of individuals per species in samples 1 and 2 transformed into proportions totalling 1. (c) Fitted rank abundance curves using Equation 8a for sample 1 (curve 1) and sample 2 (curve 2a), and Equation 8b with parameter c for sample 2 (curve 2b). Equation 3b fits best. For the equations see text.

Appendix 2

Glossary of terms, parameters and symbols frequently used

Clustering	Phenomenon that species occur in clusters of individuals.
In silico sampling	Virtual sampling making use of computer software.
Poisson-curve	Synonym for the single sample rank abundance curve derived from the average rank abundance curve using the Poisson distribution.
Rank-abundance curve	Curve fitted through the relation between species abundance and species rank.
Total rank abundance curve	Rank abundance curve with the per species accumulated numbers of individuals in a series of replicate samples plotted on log scale against species rank.
Average rank abundance curve	Rank abundance curve with the sample means (means per replicate sample) for the numbers of individuals per species plotted on log scale against species rank.
Single sample rank abundance curve	Rank abundance curve with the species individual numbers theoretically expected in an average single sample plotted on log scale against species rank.
Singleton species	Species present in the sample with one individual.
Species-diversity index	Index expressing species richness of the sampled system in one value. Examples are Fisher's α and $E(S(1,\infty))$ of the negative-binomial rank abundance curve fit model.
Species-individual curve	Curve with the number of species plotted against the logarithm of the total number of individuals of all species in the sample. Synonyms are *S*-*N* curve and *S*-log(*N*) curve.
Species-area curve	Curve with the number of species plotted against the logarithm of the area of the sample.
c	Constant in the negative-binomial curve-fit model.
D_{lsq}	Mean least square deviance as calculated versus the actual data from sampling.
e	Exponential base (2.7183)
$E(N)$	Expected total number of individuals in an average single sample.
$E(N(n))$	Contribution of the expected number of species with n individuals to the expected total number of individuals in an average single sample.
$E(S)$	Expected total number of species in an average single sample.
$E(S(1,\infty))$	Expected number of singleton species in an infinitely large sample.
$E(S(n))$	Expected number of species present with n individuals in an average single sample. For n = 1, 2, 3, ... etc., $E(S(n))$ is written as $E(S(1))$, $E(S(2))$, $E(S(3))$, ... etc. Note that $E(S(1))$ is

	the expected number of singleton species in an average single sample.
$f(n)$	In the negative-binomial distribution the expected relative frequency of sampling units containing n individuals.
h	Expected number of individuals of the most abundant (dominant) species; see Equations 19 and 20.
k	Dispersion parameter of the negative-binomial distribution expressing the amount of clustering.
m	Parameter of the negative binomial distribution for the mean of a series of data.
Poiss.	Following the Poisson-distribution.
p_R	Fitted abundance proportion of the Rd species from curve fit with the negative binomial rank abundance curve fit model.
$p_{R,geom}$	Fitted abundance proportion of the Rd species from curve fit with the geometric series rank abundance curve fit model.
q	Abundance proportion of the first (dominant) species in the geometric-series model.
R	Species number in a series of species from sampling after sorting in rank of abundance.
R'	Species number in a series of fictitious sequential species ranging from 1 to infinite, used for calculating $E(S(1,\infty))$.
N	Total number of individuals of all species in a single sample or in the accumulated total of a series of replicates samples.
S	Number of species in a single sample or in the accumulated total of a series of replicate samples.
S_R	Species with rank number R in a series of species ranked from most to least abundant.
$S(n)$	Number of species present with n individuals in a single sample or in a series of replicate samples. For n = 1, 2, 3, ... etc., S(n) is written as $S(1)$, $S(2)$, $S(3)$, ... etc. Note that $S(1)$ is the number of singleton species from sampling.
v	Factor used for calculating $E(S(1,\infty))$; $v = \mu / (\kappa + \mu)$.
w	The so-called 'shifting' factor calculated for the S-log$N$ curve in the negative-binomial rank abundance curve-fit model. For S-log$N$ curves derived from negative-binomial rank abundance curves with $\kappa = 1$ (geometric series rank-abundance curves), $w = e^v = 1.781073$.
x	Constant in Fisher's log-series model approaching unity with increasing sample size.
z_R	Number of individuals of species R from curve fit with the negative-binomial rankabundance curve-fit model.
z_o	Start value used in the calculation of $E(S(1,\infty))$, and set at 30.

α	Species-diversity index in Fisher's log-series model.
γ	Euler-Mascheroni constant (0.577216).
κ	Parameter of the negative-binomial rank abundance curve-fit model.
λ_f	Mean of the calculated relative species frequencies from curve fit.
λ_p	Mean of the species abundance proportions from sampling.
μ	Parameter of the negative-binomial rank abundance curve-fit model.

Note: *j, n, o* are numerators