In silico sampling reveals the effect of clustering and shows that the log-normal rank abundance curve is an artefact

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

The impact of clustering on rank abundance, species-individual (*S-N*) and species-area curves was investigated using a computer programme for *in silico* sampling. In a rank abundance curve the abundances of species are plotted on log-scale against species sequence. In an *S-N* curve the number of species (*S*) is plotted against the log of the total number of individuals (*N*) in the sample, in a species-area curve *S* is plotted against log-area. The results from *in silico* sampling confirm the general shape of *S-N* and species-area curves for communities with clustering, i.e., a curve that starts with a smaller slope but that later is temporarily steeper than the curve expected for Poisson-distributed species. Extrapolation of *S-N* and species-area curves could therefore be misleading. The output furthermore shows that sigmoid rank abundance curves (curves of the type of a log-normal or broken stick) can be an artefact of the standard procedure of first sorting the species in sequence of abundance in combination with clustering in the low abundant and rare species. This makes the usual explanation given to the log-normal rank abundance curve dubious. An extension of the negative-binomial rank abundance curve-fit model is discussed to make it suitable for also fitting sigmoid rank abundance curves.

Additional keywords: negative-binomial rank abundance curve, species-individual curve, species-area curve, rare species, species abundance, species-diversity index

Introduction

The species-individual curve (*S-N* curve) and species-area curve are largely determined by two factors: (1) the pattern of species abundance distribution in the community, and (2) the degree of clustering of individuals within species (He & Legendre, 2002; Neuteboom & Struik, 2005a, b). In an *S-N* curve the number of species is plotted against the total number of individuals in the sample; in a species-area curve the number of species is plotted against the area of the sample. With the total number of individuals or area plotted on log-scale and assuming that number of individuals is proportional to area, the *S-N* curve and species-area curve are identical curves with only a shifted horizontal axis (Neuteboom & Struik, 2005b).

Neuteboom & Struik (2005b) stated that clustering will result in a species accumulation curve (*S-N* curve or species-area curve) that starting with a smaller slope can temporarily be steeper compared with the curve expected for Poisson-distributed species. So extrapolation of such a curve could be misleading. The predicted shape of the curve is based on the expectation that the effect of clustering will gradually decrease with increasing sample size.

In this paper we shall verify the predicted general shape of the *S-N* curve and the species-area curve, using a computer programme for *in silico* sampling. Also, statements on the rank abundance curve made earlier (Neuteboom & Struik, 2005a, b) will be tested. The rank abundance curve is the curve in which (as usual) the number of individuals per species is plotted on log-scale against species sequence.

The computer programme for *in silico* sampling has been used before in a relatively simple form to analyse the Dry Weight Rank (DWR) method for estimating botanical composition of grassland ('T Mannetje & Haydock, 1963). The analysis showed that the DWR method works thanks to the fact that in grassland the species usually grow in clusters of individuals (Neuteboom *et al.*, 1998). Clustering within species is so common that it almost inevitably must have an effect on the *S-N* curve and the species-area curve. This is to be expected for two reasons: (1) the number of species counted in a sampling area or sampling quadrat equals the sum of the probabilities of presence of all species present in the community (Coleman, 1981), and (2) with clustering, the probability that a species will be present in a sample is expected to be lower (Greig-Smith, 1983).

Below we shall show that the predicted effects of clustering on the *S-N* curve and the species-area curve are indeed confirmed. However, the output from sampling yields another interesting effect that until now has been overlooked. The procedure of first sorting the species in sequence of abundance – standard in making a rank abundance curve – can lead to a sigmoid rank abundance curve when, due to clustering, part of the low abundant and rare species in the tail of the curve are heavily underestimated or totally missed by sampling. The resulting curve can even be a perfect log-normal rank abundance curve. The results from sampling throw serious doubts on the ecological explanation usually given to the log-normal rank abundance curve.

First, the outline of the sampling programme is discussed. The sampling includes marking species as primary and secondary patch-species to assess the effects of clustering. The principles of the computer programme for *in silico* sampling could be used in other applications as well. We also briefly summarize the negative-binomial rank abundance curve-fit model (Neuteboom & Struik, 2005a) and shall refer to the log-normal rank abundance curve-fit model of Wilson (1991) because we use that model for fitting the sigmoid rank abundance curve resulting from sampling. In the discussion section an extension of the negative-binomial rank abundance curve-fit model will be discussed to make the model also suitable for fitting sigmoid rank abundance curves.

Appendix 1 contains a glossary with some relevant terms, parameters and symbols used in the present paper. The reader is also referred to the appendices of Neuteboom & Struik (2005a, b) with additional information on the use of terms and symbols.

Computer programme for *in silico* sampling

Outline of the programme

In the computer programme for *in silico* sampling, plants of a given density are randomly distributed over the computer screen and assigned individually (Poisson-distributed species) or group-wise (patch species in randomly distributed circles) a species name. Names are randomly drawn from an array in which each species is represented with a frequency that corresponds to its relative yield as defined by De Wit (1960), or – assuming that all species have the same plant size and the same productivity per plant – to its mass proportion. The mass proportions of the species are calculated from the negative-binomial rank abundance curve-fit model (Neuteboom & Struik, 2005a) for given parameter values (κ, μ and c), for an in principle infinite number of species. However, in the practical execution of the programme, species abundance proportions are calculated up to and including a lowest proportion of 10^{-8}. The sampling on screen is repeated for consecutive samples. After each sample, the plants on the screen are replaced.

Sampling is executed with a set of eight nested sampling quadrats. Recordings on the presence of plants in the quadrats are made in sequence from the smallest to the largest quadrat. Whether a plant is present in a quadrat, is calculated from its x and y co-ordinates, its radius and the x and y co-ordinates (centre point) and radius of the quadrat. A plant is recorded as present even if it only touches the edge of the quadrat. With the consecutive quadrat sizes the effect of sample size on the rank abundance curve can be investigated and even a species-individual (S-N) curve and a species-area curve can be made.

Primary and secondary patch-species

A distinction is made between primary and secondary patch-species. Primary patch-species are used to calculate – from their total relative cover and a given patch size – the number of patches per unit area (patch density, see below). The computer programme asks the number of species that are assigned as primary patch-species. Assignment is at random or in a distinct order. In the default option of the programme, species are assigned as primary patch-species sequentially, starting with the second species ($S2$), i.e., species $S2$, $S3$, $S4$, etc. in sequence of abundance, up to and including species Sn, in which Sn has to be given. Secondary patch-species are all the remaining species added as inclusions to the patches. These remaining species can also include the first dominant species $S1$. But always 10% of the number of individuals of $S1$ remains in the space outside patches.

From the total available primary and secondary patch-species, limited numbers (groups) of continuously different species can be selected for allotting to the plants in single patches. These numbers are called the numbers of primary patch-species per patch and the numbers of secondary patch-species per patch, respectively. Drawing lots determines which of the total available primary and secondary patch-species are assigned to a single patch. The chance that a primary or secondary patch-species is allotted to a patch is proportional to its abundance.

The further allotment of species from the pre-selected groups to the single plants in a patch is at random. Primary and secondary patch-species can be allotted more than once to a pre-selected group. Since this will happen more frequently with the more abundant species – present among the primary as well as the secondary patch-species – the average cluster sizes of the species are roughly correlated with their abundance proportions.

With secondary patch-species, patch density is kept unchanged but the size of the patches is enlarged and thus re-calculated. Working with two different categories of patch species – with for both categories the option to vary the number of available species per patch – allows for simulations with two groups of species with different cluster sizes in patches. All sorts of realistic simulations with patches of mixed species composition are possible. But also simple simulations with for instance all species randomly distributed, or simulations with only primary patch-species in patches and all remaining species randomly distributed in the space outside patches. The reason to work with patches of mixed species composition is that we ultimately also want to fit for all species the distribution of the numbers of individuals over replicate samples to the negative-binomial distribution (Neuteboom & Struik, 2005b). That is possible only with a sufficient amount of variation. In simulations with only one species per patch, problems can occur with the calculation of negative-binomial k-values for single species.

Calculation of the cover by patch species and patch density

Essential is the calculation of the number of patches per unit area (patch density). Patch density (d_{patch}) is calculated from the Patch Area Index (PAI), using the relation $PAI = -\ln(1-c_{patch})$, where c_{patch} is the relative cover by patches. As stated before, c_{patch} is calculated from the relative cover of primary patch-species, which, assuming that all species have the same plant size and the same productivity per plant, equals the sum of their relative yields or the sum of their mass proportions. Given the patch radius R_{patch}, d_{patch} follows from the relation $d_{patch} = PAI/\ \pi R^2_{patch}$

Also patches are assigned x and y co-ordinates drawn by lot from an array, and a sequential number. For plants in the overlap zone of patches always the species name assigned to the highest patch number applies. So species names assigned earlier can be overruled. With secondary patch species in the simulation and d_{patch} kept unchanged (see above), the enlarged area per patch (patch size) is calculated from a c_{patch} based on the relative cover of the total of all primary and secondary patch-species.

Calculation of the numbers of plants per patch for primary and secondary patch-species

Knowing the area per patch calculated for primary patch-species (a_{patch1}) and the re-calculated area per patch (a_{patch2}) for the total of primary and secondary patch-species, the average numbers of plants assigned to primary and secondary patch-species can be calculated. Let n_{ppp} be the average number of plants per patch. Then the number of plants that get the name of a primary patch-species (n_{prim}) is n_{ppp} (a_{patch1} / a_{patch2}). The number of plants that get the name of a secondary patch-species (n_{sec}) is n_{ppp} (a_{patch2} – a_{patch1}) / a_{patch2}). Whether the system of allotments of species is right can always be checked, because in case of a sufficient number of replicates the average species abundance proportions from sampling finally must always equal the species proportions given to the sampling programme.

Two approaches to constructing species-area curves

Condit *et al.* (1996) discussed two approaches to constructing *S-N* and species-area curves. The usual approach for species-area curves (approach 1) is to count species in nested series of expanding quadrats. The other approach (approach 2) is to add species counts from one-size sampling quadrats randomly taken from across the whole sampled plot. Approach 1 results in a curve showing the increase of the number of species in a geographically expanding area, the appropriate way to compare different sites. The species accumulation curve derived from approach 2 is not comparable across sites because, as explained by Condit *et al.* (1996), that curve is dependent on the size (area) of the plot sampled. Our computer programme for artificial sampling is in line with approach 1.

Curve-fit equations and calculation procedures

Negative-binomial rank abundance curve-fit model

In the negative-binomial curve-fit model (Neuteboom & Struik, 2005a), the abundance proportions (*p*) of consecutive species are calculated in an unusual way from the expected relative frequencies of occurrence of samples with n = 0, 1, 2, 3, ..., *j* individuals according to the negative-binomial distribution. The frequency *f*(0) is used for fitting the abundance proportion p_1 of the first (dominant) species, the frequency *f*(1) for fitting the abundance proportion p_2 of the second species, etc., or (*R* is species rank):

$$p_R = f(R - 1)$$

or, adapted for the case of curve fitting,

$$p_R = f(R - 1) \;/\; c \tag{1}$$

The method for calculating the relative frequencies $f(n)$ from the negative-binomial distribution was discussed by Davies (1971). First, the expected relative frequency in the first class for $n = 0$ individuals is calculated from:

$$f(0) = 1/[1 + (m\ /\ k)]^k \tag{1.1}$$

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:

$$f(n) = f(n - 1)\ m\ (k + n - 1)\ /\ ((m + k)n) \tag{1.2}$$

Here, parameters m and k are the mean of the negative-binomial distribution and a measure of the amount of clustering, respectively. They determine the shape of the fitted rank abundance curve. With abundance proportions plotted on log-scale, parameter c corrects for a level difference between the fitted curve and the points from observation. Parameter c is calculated from the mean of the calculated relative frequencies (λ_f) from curve fit and the mean of the species abundance proportions (λ_p) from the data according to:

$$c = \lambda_f\ /\lambda_p \tag{1.3}$$

Correction for level difference is carried out at each step of the iteration before calculating the least squares of the deviancies between the data and the fitted curve. The value of c can also be found by treating it as a third iterable parameter. But that requires a lot more calculation time.

However, as stated before (Neuteboom & Struik, 2005a), in the way 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. Both are degraded in the negative-binomial rank abundance curve-fit model to pure curve-fit coefficients without any further statistical meaning and are therefore given a different name. In order to distinguish them from the original m and k in the negative-binomial distribution they will be referred to in the remaining of this paper, like before (Neuteboom & Struik, 2005a, b), as μ and κ.

Fitting 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 always be converted or re-converted into numbers of individuals by multiplying them by the total number (N) of individuals in the sample. So the numbers of individuals from curve fit (z) are (see also Neuteboom & Struik, 2005a):

$$z_R = p_R N \tag{2}$$

Calculation of E(S(1,∞)) as site discriminant and species-diversity index

From the fitted rank abundance curve the expected number of singleton species in an infinitely large sample $E(S(1,\infty))$ can be calculated. Its calculation is based on the fact that in the asymptotically ending tail of the concave rank abundance curve the calculated relative frequencies $f(n)$ from Equation 1.2 ultimately differ by a constant factor $\nu = \mu / (\kappa + \mu)$. From that point on the number of singleton species $E(S(1))$ that can be calculated from the fitted curve and the Poisson-distribution for a theoretical single sample becomes a constant. $E(S(1,\infty))$ is independent of clustering and could therefore be used as a site discriminant and species-diversity index (Neuteboom & Struik, 2005a).

$E(S(1,\infty))$ is calculated from (Neuteboom & Struik, 2005a):

$$E(S(1, \infty)) = \sum_{1}^{R' = \infty} [z_o \nu^{(R'-1)} \exp\{-z_o \nu^{(R'-1)}\}] \tag{3}$$

where z_o is a number set at 30 (see Neuteboom & Struik, 2005a) and R' a species number in a series of fictitious sequential species ranging from 1 to infinite; $\nu = \mu / (\kappa + \mu)$.

The single sample rank abundance curve

A rank abundance curve is usually made by totalizing per species the numbers of individuals in a series of 'replicate' samples after which these totalized numbers are plotted on log-scale against the species' rank in sequence from most to least abundant. As explained before (Neuteboom & Struik, 2005a), the 'total' rank abundance curve is with a level difference the same as the curve for the average numbers of individuals per species per sample (the 'average' rank abundance curve) and like the latter curve in principle independent of clustering. The number of species that is plotted against the total number of individuals in an *S-N* curve or species-area curve is the number of species in an average single sample (Neuteboom & Struik, 2005b). That number can be theoretically derived from the average rank abundance curve using a function describing the pattern of distribution of the numbers of individuals within species over replicate samples. We use as before (Neuteboom & Struik, 2005b) the Poisson-distribution, and for detecting clustering the negative-binomial distribution (distinct from the negative binomial rank abundance curve fit model) with separately fitted *k*-values per species. Both distribution functions calculate from the average numbers of individuals per species the contribution of each species to the theoretically expected numbers of species with $n = 1, 2, 3, ..., j$ individuals in an average single sample (Neuteboom & Struik, 2005b). The frequency distribution obtained can be compared with the frequency distribution of the actual average numbers of species in $\log_2$-classes of numbers of individuals per species according to the method of Preston (1948). With the χ^2-test it can be tested whether there is a statistically significant difference.

The frequency distribution 'expected' can be transformed into a rank abundance curve in the usual graphical presentation with the abundance per species plotted on log-scale against species sequence (the 'single sample' rank abundance curve). In case individuals within species actually behave as Poisson-distributed, the 'single sample'-rank abundance curves expected for Poisson- and negative-binomial-distributed individuals will both coincide with the curve for the actual mean numbers of individuals per species from sampling. In case of clustering, the curve for negative-binomial-distributed individuals will deviate (Neuteboom & Struik, 2005b).

Log-normal distribution; the general log-normal

Wilson (1991) fitted a general log-normal to data by iteratively adjusting the mean and standard deviation (σ) of the ln-transformed species abundances of what he calls an abundance-class log-normal curve. For details of the curve-fit model we refer to the original paper. We applied Wilson's log-normal curve-fit model to data from sampling with clear indications of a log-normal rank abundance curve

Output from sampling

Rank abundance curve, species-individual (*S*-*N*) curve and species-area curve

Rank abundance curve

Figure 1 shows for an arbitrary case of clustering the output from sampling. Species are assigned as primary and secondary patch-species for allotting to the plants in patches. Primary patch-species are species *S*2 up to and including *S*10. Secondary patch-species are all remaining species excluding 10% of the individuals of species *S*1. The plant density is 6 plants dm^{-2} (point-sized plants). The area per patch after re-calculation for secondary patch-species is 50.5 dm^2. The average number of plants per patch is 303. The average number of plants sampled with the largest quadrat 8 is 769. The number of primary patch-species per patch is 1. The secondary patch-species are randomly distributed over the total space occupied by patches.

Presented in Figure 1a are (1) the rank abundance curve given to the sampling programme and calculated from the negative-binomial rank abundance curve-fit model (Neuteboom & Struik, 2005a) for $\kappa = 0.5$, $\mu = 3$ and $c = 1$ (curve 1), and (2) the numbers of individuals per species from sampling (□). The numbers of individuals are the total numbers of individuals per species in 500 samples taken with the largest of the set of 8 nested sampling quadrats (128 dm^2). Figure 1b shows (1) the numbers of individuals (□) per species after sorting the species in sequence from most to least abundant, and (2) the re-fitted 'total' rank abundance curve using the negative-binomial rank abundance curve fit model (curve 2). Figure 1c shows the average numbers (□) of individuals per species per sample, the same re-fitted rank abundance curve (curve 2, now given as 'average' rank abundance curve) and the 'single-sample' rank abundance curve (curve 3). Calculation and meaning of the 'total' and 'average' rank abundance curve have been explained in Neuteboom & Struik (2005a), the 'single-sample' rank abundance curve has been explained in Neuteboom & Struik (2005b). The single-sample rank abundance curve was calculated from the 'average' rank abundance curve using the negative-binomial distribution with separately fitted negative-binomial *k* values per species (see Neuteboom & Struik, 2005b).

The conclusion from Figures 1b and 1c is that, as expected (Neuteboom & Struik, 2005a, b), clustering in the common species *S*2, *S*3, ..., *S*10 hardly affected the 'total' and 'average' rank abundance curve. Both the 'given' (curve 1) and the re-fitted rank

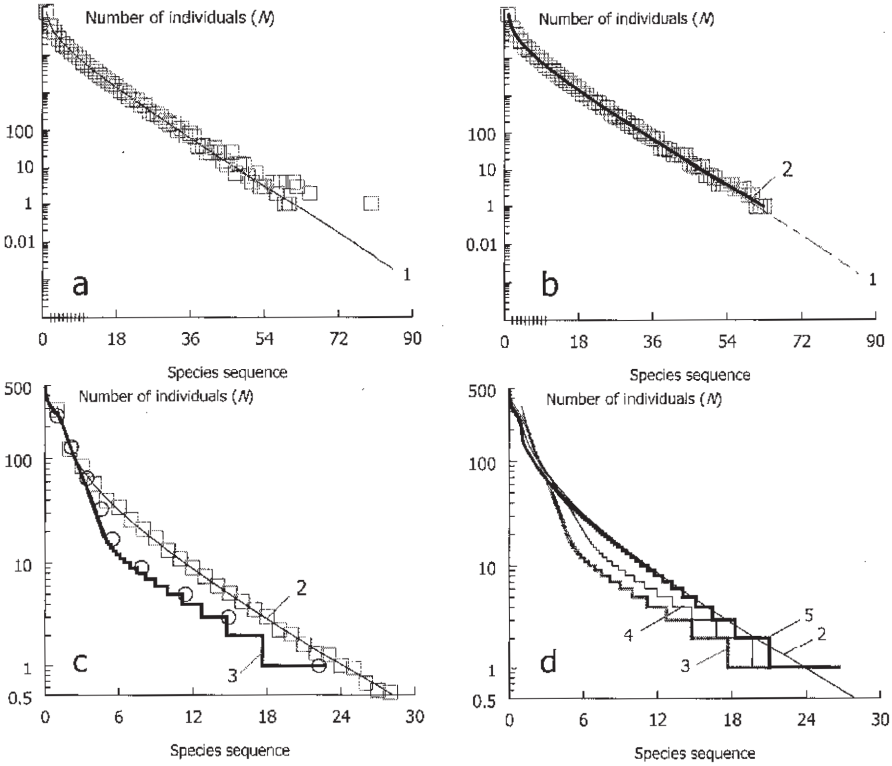

Figure 1. Species abundances from artificial sampling, sampling quadrat 8 (128 dm^2). Plant density $d = 6$ plants dm^{-2}, point-sized plants. Species in patches of 50.5 dm^2 with species $S2$, $S3$, ..., $S10$ assigned as primary patch-species and all remaining species ($S11$, $S12$,....., Sn and 90% of species $S1$) assigned as secondary patch-species. (a), (b) and (c) Patches with 1 primary patch-species and secondary patch-species randomly distributed over the total space occupied by patches. (a) and (b) show the totalized species individual numbers (□) in 500 replicate samples, unsorted (a) and after sorting the species in sequence of abundance (b). (c) shows in '□' the average numbers of individuals per species per sample. Curve 1 in (a) and (b) is the rank abundance curve from which the given species abundance proportions in the sampled community were calculated. The curve is calculated from the negative-binomial curve-fit model for $\kappa = 0.5$, $\mu = 3$ and $c = 1$. The total number of species in the sampled community (calculated from the negative-binomial curve-fit model up to and including a theoretically lowest proportion for species abundance defined at 10^{-8}) is 107. Curve 2 in (b) and (c) is the curve refitted through the data by the negative-binomial rank abundance curve-fit model. Curve 3 in (c) and (d) is the 'single sample' rank abundance curve calculated from the contributions per species to the expected numbers of species with 1, 2, 3, ..., etc. individuals in an average single sample. The contributions per species were calculated from the 'average' rank abundance curve using the negative binomial distribution with separately fitted negative-binomial k values per species. Curve 3 is confirmed by the species abundances (○) derived from the frequency distribution of the average numbers of species per sample actually found from sampling and classified in log$_2$-classes of numbers of individuals. (d) Single-sample rank abundance curves for 1 primary patch-species and 1 (curve 3) and 3 secondary patch-species (curve 4) per patch. Curve 5 in (d) is the curve in case all species would have had Poisson-distributed individuals.

abundance curve (curve 2) had approximately the same value for $E(S(1,\infty))$ (6.49 and 6.80, respectively). However, clustering strongly affected the 'single-sample' rank abundance curve (Figure 1c). The validity of that curve (curve 3) is sustained by the points (O) derived from the frequency distribution of the actually sampled average numbers of species in $\log_2$-classes of numbers of individuals.

Figure 1d shows what happens if we let the number of primary patch-species per patch increase from 1 (curve 3) to 3 (curve 4). The single sample rank abundance curve becomes less concave and approaches the shape of the curve for Poisson-distributed species (curve 5). Clustering results in a lower number of species: 23 (curve 3) vs. 27 (curve 5).

Species-individual (S-N) curve and species-area curve

Figures 2a and 2b confirm (see Neuteboom & Struik, 2005a) that for Poisson-distributed species the expected number of singleton species ($E(S(1))$) calculated from the fitted rank abundance curve (via the single-sample curve) for consecutive values of N, is the slope of the S-N curve. The points marked with a large dot in Figure 2a present the mean numbers of species plotted against the mean numbers of individuals (N) sampled in each of the 8 nested sampling quadrats (quadrats Q1, Q2,....Q8; quadrat sizes 1, 2, 4, 8, 16, 32, 64 and 128 dm^2, respectively); curve 1 shows the theoretical numbers of species calculated from the 'given' rank abundance curve for consecutive values of N. The points marked with a large dot in Figure 2b are the slope values calculated as tangent at the consecutive curve positions Q2, Q3, ..., Q7 in Figure 2a. They fully coincide with the curve for $E(S(1))$ calculated from the rank abundance curve. The calculation of the slope as tangent is explained in the caption of Figure 2. $E(S(1,\infty))$ in Figure 2b is the theoretical slope of the S-N curve for an infinitely large sample.

Figures 3a, 3c, 3b and 3d show the consequences of clustering for the shape and the slope of the species accumulation curve. The simulation conditions are the same as those in Figure 1. However, the patches are smaller. The area per patch after re-calculation for secondary patch-species is 5.61 dm^2. Like in Figure 2, the data of the complete set of 8 nested sampling quadrats are used. The species accumulation curve is presented now as a species-area curve.

The curves in Figure 3a (note that the slope values were calculated as tangent in Figure 3b) are for Poisson-distributed species (curve 1) and for species in patches with 1 (curve 2), 3 (curve 3) and 9 primary patch-species per patch (curve 4). All remaining species (except 10% of the individuals of $S1$) are randomly distributed as inclusions (secondary patch-species) in the patches. In Figure 3c (note that the slope values were calculated as tangent in Figure 3d) the curves are for Poisson-distributed species (curve 1) and for species in patches with 1 primary patch-species and 98 (curve 2), 10 (curve 3), 3 (curve 4) and 1 (curve 5) secondary patch-species per patch. Curves 2, 3, 4 and 5 in Figure 5c may be seen as approaching realistic mixtures with all species occurring in smaller (the low abundant and rare secondary patch-species) or larger clusters of individuals (primary patch-species).

Figures 3a, 3b, 3c and 3d confirm the earlier statement made (Neuteboom & Struik, 2005b) that clustering can result in a species-accumulation curve (S-N curve or species-area curve) that starting with a smaller slope can temporarily be steeper compared with

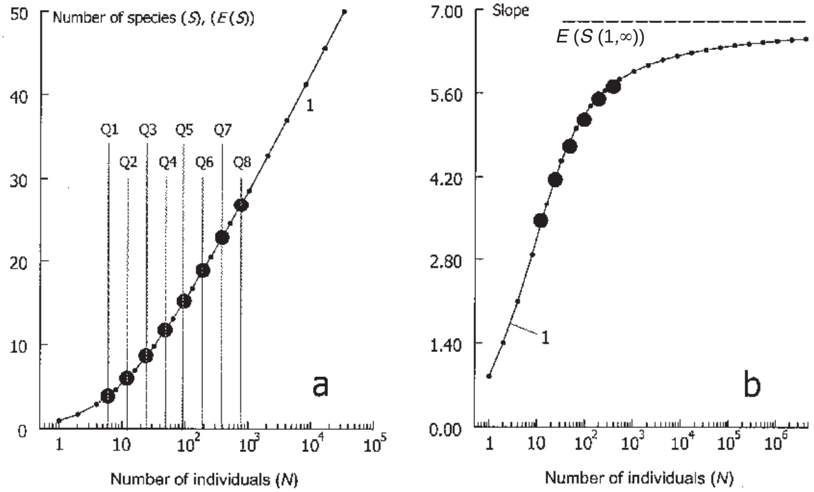

Figure 2. (a) S-N curves theoretically calculated with expected numbers of species $E(S)$ for consecutive sample sizes in terms of total numbers of individuals N (curve 1 and marker •) and, numbers of species (S) observed (●) versus total numbers of individuals N, obtained from sampling with 8 nested sampling quadrats: Q1(1 dm^2), Q2 (2 dm^2), Q3 (4 dm^2), Q4 (8 dm^2), Q5 (16 dm^2), Q6 (32 dm^2), Q7 (64 dm^2) and Q8 (128 dm^2). (b) Slope values theoretically calculated as expected numbers of singleton species $E(S(1))$ for curve 1 and slope values calculated as tangent at consecutive positions along the suggested curve through data from observation. The slope values as tangent were calculated from $(S_{Q3} - S_{Q1}) / (\ln N_{Q3} - \ln N_{Q1})$, $(S_{Q4} - S_{Q2}) / (\ln N_{Q4} - \ln N_{Q2})$, etc. $E(S(1,\infty))$ in (b) is the theoretically expected number of singleton species in an infinitely large sample, calculated from Equation 3 (see text).

the curve expected for Poisson-distributed species. So extrapolating species-area curves for calculating numbers of species for larger samples can be misleading (curve 2' in Figure 3c). Figures 3b and 3d show that with clustering, the slope of the curve can even temporarily exceed the value of $E(S(1,\infty))$, the theoretical value of the slope for an infinitely large sample. According to the negative-binomial rank abundance curve-fit model (Neuteboom & Struik, 2005a) all curves in Figures 3a and 3c will finally end with a value for the slope equal to $E(S(1,\infty))$. In case the low abundant and rare species (secondary patch-species) are approximately Poisson-distributed (Figure 3a), the curve for clustering will ultimately merge with the theoretical curve for Poisson-distributed species.

The sigmoid rank abundance curve

Figure 4 shows what happens if the number of primary patch-species and the number of secondary patch-species are both reduced to one species per patch. The conditions with regard to plant density, plant size, patch size, κ and μ values of the 'given' rank abundance curve, and the total number of available primary patch-species (9 species,

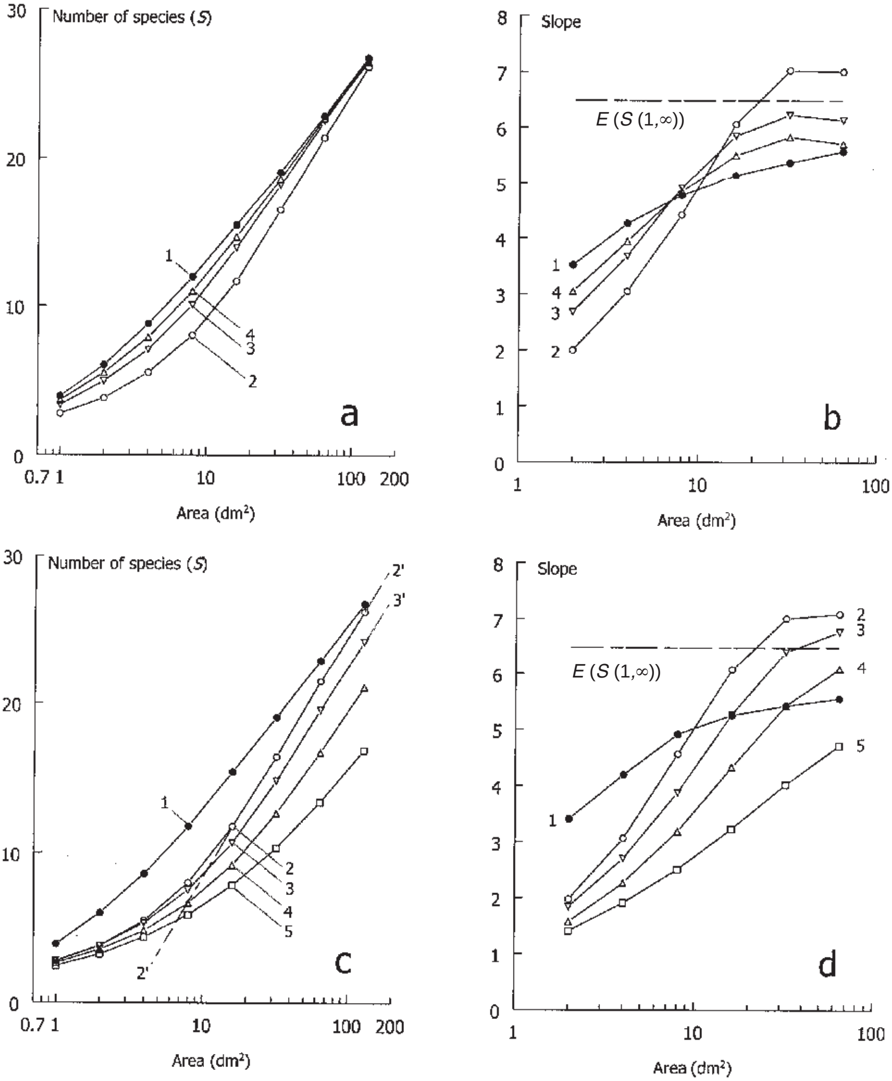

Figure 3. Effect of clustering on the species-area curve. Parameters and sampling conditions as in Figure 1, but the patches are smaller, i.e., after re-calculation for secondary patch-species, 5.61 dm^2. Like in Figure 2, the data of the complete set of 8 nested sampling quadrats are used. (a) All species with Poisson-distributed individuals (curve 1) and species in patches with 1 (curve 2), 3 (curve 3) and 9 primary patch-species per patch (curve 4). All remaining species (except 10% of the individuals of *S*1) are randomly distributed as inclusions (secondary patch-species) in the patches. (b) Slope values of the curves in (a) calculated as tangent. (c) Poisson-distributed species (all species, curve 1) and species in patches with 1 primary patch-species and 98 (curve 2), 10 (curve 3), 3 (curve 4) and 1 (curve 5) secondary patch-species per patch. (d) Slope values of the curves in (c) calculated as tangent.

species *S2*, *S3*,....., *S10*) are the same as in Figure 1. The points marked □ in Figure 4a are the species individual numbers obtained from sampling. The same points □ in Figure 4b are the numbers of individuals per species after first sorting the species in sequence of abundance, i.e., the standard procedure in making a rank abundance curve. The result is a sigmoid rank abundance curve. The enforced downward curvature in the tail of the curve is caused by the fact that due to strong clustering, part of the low abundant and rare species in the system are strongly underestimated or totally lacking in the samples. The resulting curve after sorting the species in sequence of abundance is, fitted with Wilson's log-normal curve fit model (Wilson, 1991), a perfect log-normal rank abundance curve with (given the *S* and given the abundance proportions from sampling) a standard deviation $\sigma = 2.82$.

Figure 5 presents a second example of a sigmoid rank abundance curve resulting from first sorting the species in sequence of abundance when the rare species in the tail of the curve show strong clustering. The 'given' rank abundance curve is calculated from the negative-binomial curve-fit model for $\kappa = 0.3$, $\mu = 3$ and $c = 1$. However, now all (162) species present in the system, except all individuals of species *S1*, act as primary patch-species while the number of primary patch-species per patch of 28.26 dm^2 (patch radius = 3 dm) is 1, i.e., now all species are present in large clusters. This in fact also applies to species *S1* since species force clustering upon each other.

Sampling was again executed with the largest of the set of 8 nested sampling quadrats (quadrat area = 128 dm^2). Given a plant density of 3 plants dm^{-2}, the average number of plants per quadrat was 384. Taken were 1000 samples. The total numbers of

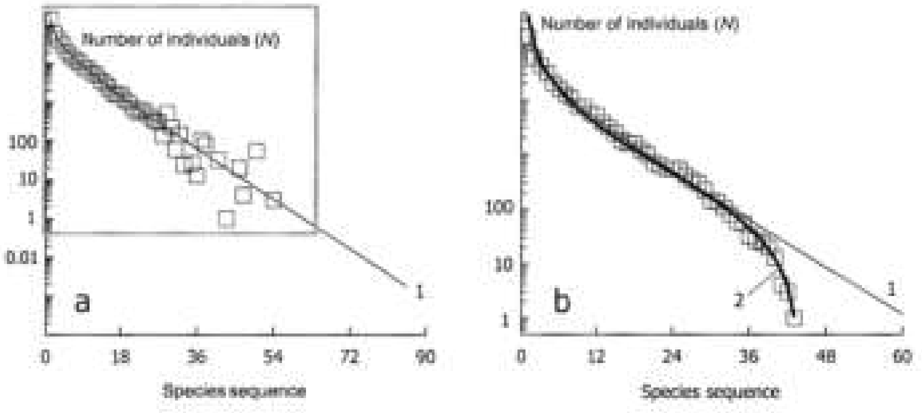

Figure 4. Output from artificial sampling, sampling quadrat 8. Sampling and system parameters as in Figure 1, but primary as well as secondary patch-species reduced to one species per patch. The points marked '□' in (a) are the species individual numbers obtained directly from sampling. The same points in (b) are the numbers of individuals per species after first sorting the species in sequence of abundance, standard procedure in making a rank abundance curve. Curve 1 in (a) and (b) is the rank abundance curve 'given' to the sampling programme. Curve 2 in (b), fitted with Wilson's log-normal rank abundance curve-fit model, is a perfect log-normal rank abundance curve (Wilson, 1991).

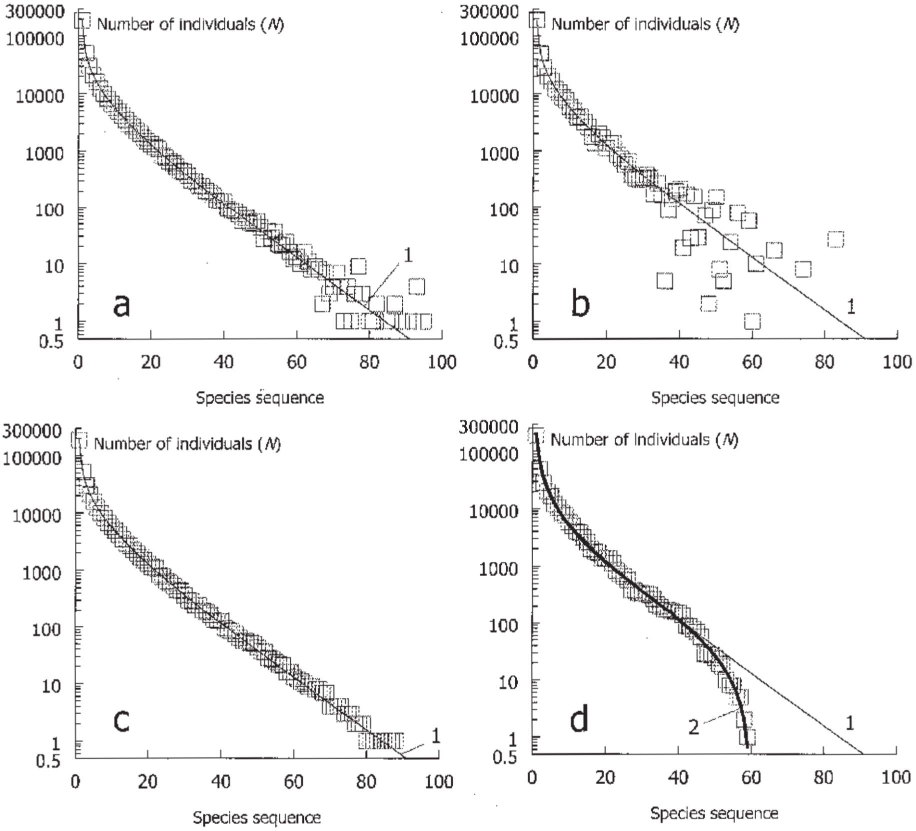

Figure 5. Output from artificial sampling (sampling quadrat 8) for point-sized plants and a plant density of 3 plants dm^{-2}. The given species abundances in the sampled community are calculated from the negative-binomial rank abundance curve-fit model using $\kappa = 0.3$, $\mu = 3$ and $c = 1$. Different from Figure 4 now all species present in the sampled community exclusive of the total species $S1$ (Poisson-distributed in the space outside patches), act as primary patch-species. The radius per patch is 3 dm (patch area = 28.27 dm^2). The simulated case is for 1 primary patch-species per patch, which means that all species (except species $S1$) occur in large patches. (a) and (c) Species with Poisson-distributed individuals; (b) and (d) Species in patches. The points marked ☐ in (a) and (b) are the species individual numbers direct from sampling. The same points in (c) and (d) are the numbers of individuals per species after first sorting the species in sequence of abundance. Curve 1 in (a), (b), (c) and (d) represents the rank abundance curve of the community 'given' to the sampling programme. Curve 2 in (d) is fitted with Wilson's log-normal rank abundance curve-fit model (Wilson, 1991), a perfect log-normal rank abundance curve.

individuals per species are plotted as usual on log-scale against species sequence. The species abundances in Figures 5a (Poisson-distributed species) and 5b (all species except $S1$ in equally sized patches) are the abundances directly from sampling, those plotted in Figures 5c (Poisson-distributed species) and 5d (species in patches) are the abundances after first sorting the species in sequence of abundance. Curve 1 in Figures 5a, 5b, 5c and 5d represents the 'given' rank abundance curve as calculated from the negative-binomial curve-fit model ($E(S(1,\infty)) = 10.6$). Curve 2 in Figure 5d, fitted with Wilson's log-normal curve-fit model (Wilson, 1991), is an almost perfect log-normal curve with (given the S and given the abundance proportions from sampling) a standard deviation of 2.65.

Figures 4b and 5d show that the strong clustering in all species especially had an effect in the tail of the rank abundance curve. The reason that clustering hardly affected the upper part of the curve can be explained from the fact that for the dominant and common species it does not matter so much whether one cluster more or one cluster less is caught in the samples. On the other hand, one cluster more or one cluster less can have a strong effect on the low abundant and rare species because this can lead to a strong under- or overestimation and even to a total absence from sampling.

Discussion

The output from *in silico* sampling confirms that the 'total' and 'average' rank abundance curves are in principle independent of clustering but that clustering has a strong effect on the 'single sample' rank abundance (Figure 1). The single sample rank abundance curve is the curve from which the number of species is calculated for plotting in an S-N curve or species-area curve, which means that also these two curves are strongly affected by clustering. The output confirms that clustering results in a species-accumulation curve (S-N curve or species-area curve) that starts with a smaller slope, but which can temporarily be steeper than the curve expected for Poisson-distributed species (Figures 3a, 3b, 3c and 3d). The temporarily steeper slope is due to an effect of catching up caused by the fact that the effect of clustering gradually decreases with increasing sample size. The sampling programme appears to confirm the conclusion (Neuteboom & Struik, 2005b) that clustering can make the slope of the S-N curve and species-area curve unsuitable for site characterization and for use as species-diversity index.

The output from sampling also clearly shows that the sigmoid rank abundance curve can be an artefact of the standard procedure of first sorting the species in sequence of abundance in combination with clustering in the low-abundant and rare species. The actual curve for clustered species without sorting (Figures 4a and 5b) is a concave rank abundance curve. The curve after first sorting the species can even be a perfect log-normal curve (Figures 4b and 5d). This is something totally different from the frequently cited view (e.g. May, 1975; Magurran, 1988) that the log-normal pattern of species abundance distribution arises as the response to the statistical properties of large numbers and as a consequence of the central limit theorem. The central limit theorem states that when a large number of factors act to determine the value of a variable, random variation in those factors will result in that variable being normally distributed. In case of a log-normal pattern of species abundance distribution, the variable and the determining factors are conceived as the number of individuals per species and all the processes that govern the community, respectively.

The log-normal distribution was first applied to abundance data by Preston (1948).

Preston also presented the canonical log-normal (Preston, 1962a, b) that later was given an ecological explanation in Sugihara's sequential breakage model (Sugihara, 1980). The canonical log-normal is seen as a special form in which the mode of the individuals curve coincides with the abundance of the most abundant species. The canonical log-normal has a standard deviation that is constrained between narrow limits. Sugihara's model is one of the resource-partitioning models used for explaining patterns of species abundance distribution. The idea behind these models is that the species abundances are a reflection of niches in the sampled community.

A log-normal rank abundance curve is often associated with communities in equilibrium with competitive species interactions (May, 1975a, b; Sugihara, 1980; Gray, 1987). The log-normal model could be used in this respect in combination with other rank abundance models for monitoring disturbance (Hill *et al.*, 1995; see for an extensive discussion on the subject Basset *et al.*, 1998). However, the deceiving effect of the standard procedure of first sorting the species in sequence of abundance seems to invalidate the log-normal as a useful descriptor of the pattern of species abundance distribution in a community. Lambshead & Platt (1985) stated that the log-normal is never found in genuine ecological samples. They argue that data sets fitting the log-normal are not true samples but collections or amalgamations of non-replicate samples. Our results from artificial sampling might support their view because samples from totally different origins put together in one large sample more or less behave as replicate samples from a community with strong clustering. Sigmoid rank abundance curves are occasionally fitted with the broken stick model of MacArthur (1957). For the same reasons as given above, that model too could be deceiving as descriptor of the species abundances in a community.

The shape of the species-area relation predicted from artificial sampling and the factors determining it clearly indicate that the most common models used to fit species-area curves, i.e., Gleason's exponential model (Gleason, 1922) and Arrhenius' power function model (Arrhenius, 1921) can only yield a rough fit. Gleason's exponential model fits the approximately linear part of the species-area curve for large samples as a straight line using $S = \varepsilon \log A + \delta$. Arrhenius' power function model fits the entire curve as a straight line, using $\log S = \rho \log A + \log \tau$. According to Preston (1960; 1962a, b) and May (1975a), the power function for the species-area relation can be mathematically derived from the log-normal pattern of species abundance distribution. But as stated before, that pattern probably does not exist. That in turn would imply that the power function model would lose its justification in the log-normal.

Figure 6 illustrates that an alternative explanation of the better fit of a log S-log area equation could be an in origin S-log area relation for Poisson-distributed species that is changed by clustering. The species-area curves presented are the same as in Figure 3c. The power function model transforms the exponential part of an S-log area curve into a linear curve. That part is stretched out over the horizontal axis by clustering. The result is that particularly curve 5 with the strongest clustering has changed into an almost linear curve (Figure 6, curve 5').

In vegetation science, species-area curves are usually constructed on the basis of presence recordings of species in sampling quadrats. The mere touching of the quadrat's edge is usually enough for a species to be recorded as present. In that case

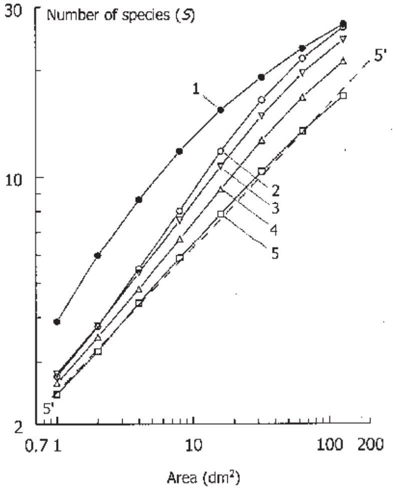

Figure 6. The same species-area curves from artificial sampling as in Figure 3c. But now with both S and *area* plotted on log-scale. Curve 5 is now approximately a straight line (curve 5'). For further explanation see text.

also plant size (plant area) is a determining factor in the shape of the curve, at least in the range of small samples. Moreover, a limited (maximum) number of species (S_{max}) present in the sampled system will cause a levelling off at the end of the curve (Neuteboom & Struik, 2005b). Plant size and S_{max} are not the issues in this paper. Nevertheless we have briefly summarized their effects with a further analysis from artificial sampling for the case of Poisson-distributed species (Figure 7). A larger plant size (plant radius) does not change the rank abundance curve (not shown in a graphical presentation). Nor does it alter the shape of the S-N curve (Figure 7a). But it changes the shape of the species-area curve by giving a higher number of species in the range of small samples (Figure 7b). In conclusion, in principle at least the following factors seem to determine the shape of the species-area curve: (1) the pattern of species abundance distribution, (2) clustering of individuals within species, (3) plant size, and (4) the number of species present in the sampled system. These factors can be included in a formal equation for the species-area curve (J.H. Neuteboom & P.C. Struik, unpublished data).

Remains the question what could be done with the information from sigmoid rank abundance curves. One possibly could make use of the suggestion from Figures 4b and 5d that the first part of the sigmoid curve with the dominant and common species is hardly affected by clustering. That suggests that in principle for that part of the curve an $E(S(1,\infty))$ can be calculated, using the negative-binomial rank abundance curve fit model (Neuteboom & Struik, 2005a).

We are investigating an extension of the negative-binomial rank abundance curve fit model that separates the basic concave curve component with the $E(S(1,\infty))$ from the

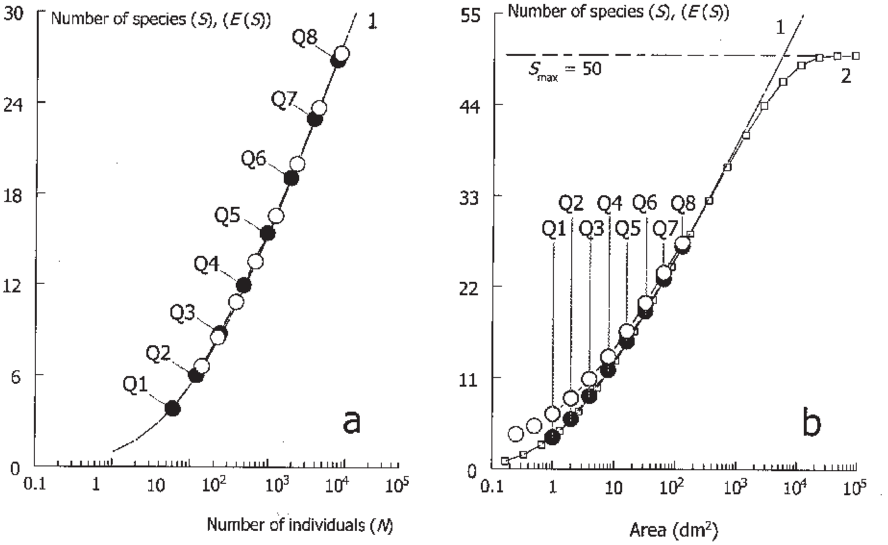

Figure 7. (a) Species-individual curves (*S-N* curves) and (b) Species-area curves with numbers of species (*S*) from a sampling with 8 nested sampling quadrats: Q1 = 1 dm², Q2 = 2 dm², Q3 = 4 dm², Q4 = 8 dm², Q5 = 16 dm², Q6 = 32 dm², Q7 = 64 dm² and Q8 = 128 dm². Poisson-distributed species with in both (a) and (b) two cases: 1 (●), all species in the sampled community with a plant radius of 0.0001 cm, and 2 (○) all species in the sampled community with a plant radius of 3 cm. Curve 1 in (a) is the theoretical *S-N* curve and curve 1 in (b) the theoretical species-area curve with *E*(*S*)-values calculated from the rank abundance curve given to the sampled community, using $\kappa = 0.5$, $\mu = 3$ and $c = 1$. For the plant radius of 3 cm in (b) two points are added for *S* sampled in smaller quadrats (0.25 and 0.5 dm²); these points were obtained by changing a scaling parameter in the programme. Curve 2 marked '☐' in (b) is the curve expected for a maximum number of species (S_{max}) present in the sampled community; $S_{max} = 50$.

sigmoid rank abundance curve. One option is to extend the curve fit model by adding a term *T* to Equation 1: $p_R = [f(R-1) - T] / c$ in which $T = R\ (1 - e^{-t})$. Application of the extended model to sampling data can work reasonably well as is demonstrated in Figure 8. Fitted are the same data as used for the log-normal curve in Figure 5d. Curve 1 of Figure 8 is the rank abundance curve 'given' to the sampling programme. Curve 2 is the complete curve fitted by the extended model; curve 3 is the curve stripped of the (fitted) effect of *t*. The functioning of *t* is further illustrated in curve 4 that shows the increase of the component *T/c* with species sequence.

Fitted values for κ, μ and *c* (Figure 8) were 0.328, 3.126 and 1.027074, respectively. Values fitted for *t* are always very low; in this case $t = 0.0000005385$. For ν of the stripped curve (curve 3) calculated from $\mu / (\kappa + \mu)$ and needed in Equation 3, a value was calculated of 0.905038. The value ultimately calculated for $E(S(1,\infty))$ (Equation 3) was 10.03, which was close to the value $E(S(1,\infty)) = 10.492$) calculated from the κ and μ given to the sampling programme.

We have further tested the extended model using species abundance data from the

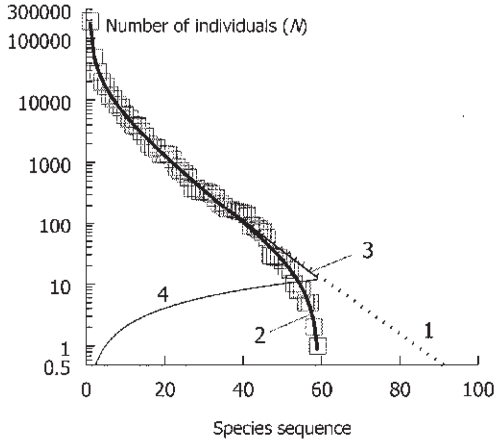

Figure 8. The same data from artificial sampling as in Figure 5d, but fitted now with an extension of the negative-binomial rank abundance curve-fit model. The fitted species abundance proportions in Equation 1 (see text) are not calculated from $p_R = [f(n)_{R-1})] / c$ but from $p_R = [f(n)_{R-1}) - T] / c$ where $T = R\,(1 - e^{-t})$. Curve 1 is the rank abundance curve 'given' to the sampling programme. Curve 2 is the complete rank abundance curve fitted by the extended negative-binomial rank abundance curve-fit model. Curve 3 is the curve stripped of the (fitted) effect of t. The functioning of t is illustrated in curve 4 that shows the increase of the component T/c with species sequence.

literature with clear indications of a sigmoid rank abundance curve. The extended model appeared in all cases to fit the data better than the log-normal or the broken stick model (not shown because it can easily be checked). However, a problem could be that in principle the effects of κ, μ and t, and thus the effects of $E(S(1,\infty))$ and t are confounded. Whether this really causes a problem with the interpretation of $E(S(1,\infty))$ as species-diversity index needs further investigation. The extended model could at least provide a way to test whether there really is a significant downward curvature in the tail of the rank abundance curve.

Overview and concluding remarks

Our series of three papers on rank abundance, species-individual and species-area relations (Neuteboom & Struik, 2005a, b and the current paper) can be briefly summarized as follows.

The negative-binomial rank abundance curve-fit model presented by Neuteboom & Struik (2005a) fits linear to deeply concave rank abundance curves and expresses species diversity in one index, the $E(S(1,\infty))$, defined as the number of singleton species in an infinitely large sample. The rank abundance curve is in our approach the usual curve made by totalizing per species the numbers of individuals in a series of replicate samples. The curve fit is on proportions. The species abundance proportions from curve fit are re-converted into numbers of individuals by multiplying them by the total number of individuals in the totalized replicate samples, or by the average number of individuals (totalized for all species) per replicate sample. The resulting curves are the 'total' rank abundance curve and the 'average' rank abundance curve, respectively. Both are curves with species individual numbers on a continuous scale and in principle independent of clustering.

The advantage of the new model over the log-series model is that it can fit a much broader range of concave rank abundance curves. Moreover, its species-diversity index [the $E(S(1,\infty)]$ can also be calculated for cases in which the abundances of the species are available only in terms of proportions. The new model links the geometric series and log-series model and is more flexible than any of the existing curve-fit models for linear or concave rank abundance curves.

New in Neuteboom & Struik (2005b) are the analysis of the variation in rank abundance replicate samples and the calculation of the 'single-sample' rank abundance curve with discrete species individual numbers as the link between the 'average' rank abundance curve and the *S-N* curve or species-area curve. The 'single-sample' rank abundance curve is calculated from the 'average rank' abundance curve, using the negative-binomial distribution for quantifying the amount of variation in species individual numbers in replicate samples. That variation can be expressed in one common negative-binomial *k* for all species. The single-sample rank abundance curve yields the number of species from curve fit for plotting in an *S-N* curve or species-area curve. All three, the single-sample rank abundance curve, the *S-N* curve and the species-area curve are strongly affected by clustering. Neuteboom & Struik (2005b) clearly showed (1) that the slope of the *S-N* and species-area curves calculated as tangent is not constant and therefore unsuitable as species-diversity index and site discriminant, and (2) that extrapolating *S-N* curves or species-area curves can be misleading.

In the current paper, the conclusions with regard to the *S-N* and species-area curves are confirmed by means of a computer programme for *in silico* sampling. This paper also shows the deceiving effect of the procedure of first sorting the species in sequence of abundance, standard in making a rank abundance curve. Until now this effect has been overlooked but it makes clear that giving ecological explanations to the shape of the rank abundance curve without information on the amount of clustering in the sampled community is premature.

The extended negative-binomial rank abundance curve-fit model reduces all types of rank abundance relations (inclusive the log-normal and broken stick) to a linear to concave curve and could therefore be seen as a first step towards a universal rank abundance curve-fit model. Such a universally applicable model could be a useful tool in the context of macro-ecology, the new discipline aiming at explaining patterns of species diversity on different scales varying from the level of a single community to the level of landscapes and higher (see Nee, 2002). Our paper demonstrates the benefit of a computer-sampling programme for the analysis of theoretical questions with regard to rank abundance and species-area curves.

References

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

Basset, Y., V. Novotny, E.M. Scott & N.D. Springate, 1998. Assessing the impact of forest disturbance on tropical invertebrates: some comments. *Journal of Applied Ecology* 35: 461–466.

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.

De Wit, C.T., 1960. On Competition. Verslagen Landbouwkundig Onderzoek No 66.8. Pudoc, Wageningen, 82 pp.

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

Gray, J.S., 1987. Species-abundance patterns. In: J.H.R. Gee & P.S. Giller (Eds), Organization of Communities, Past and Present. Blackwell Scientific, Oxford, pp. 53–67.

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

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

Hill, J.K., K.C. Hamer, L.A. Lace & W.M.T. Banham, 1995. Effects of selective logging on tropical forest butterflies on Buru, Indonesia. *Journal of Applied Ecology* 32: 754–760.

Lambshead, J. & H.M. Platt, 1985. Structural patterns of marine benthic assemblages and their relationships with empirical statistical models. In: Proceedings of the 19th European Marine Biology Symposium, Plymouth, 1984. Cambridge University Press, Cambridge, pp. 371–380.

MacArthur, R. H., 1957. On the relative abundance of bird species. *Proceedings of the National Academy of Sciences (USA)* 43: 293–295.

Magurran, A.E., 1988. Ecological Diversity and its Measurement. Croom Helm, London, 179 pp.

May, R.M., 1975a. Patterns of species abundance and diversity. In: M.L. Cody & J.M. Diamond (Eds), Ecology and Evolution of Communities. Harvard University Press, Cambridge, Massachusetts, pp. 81–120.

May, R.M., 1975b. Successional patterns and indices of diversity. *Nature* 258: 285–286.

Nee, S., 2002. Thinking big in ecology. *Nature* 417: 229–230.

Neuteboom, J.H., E.A. Lantinga & P.C. Struik, 1998. Evaluation of the dry weight rank method for botanical analysis of grassland by means of simulation. *Netherlands Journal of Agricultural Science* 46: 285–304.

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. Variation in rank abundance replicate samples and impact of clustering. *NJAS – Wageningen Journal of Life Sciences* 53: 199–221.

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

Preston, F.W., 1960. Time and space and the variation of species. *Ecology* 41: 611–627.

Preston, F.W., 1962a. The canonical distribution of commonness and rarity I. *Ecology* 43: 185–215.

Preston, F.W., 1962b. The canonical distribution of commonness and rarity II. *Ecology* 43: 410–432.

Sugihara, G., 1980. Minimal community structure: an explanation of species abundance patterns. *American Naturalist* 116: 770–787.

'T Mannetje, L. & K.P. Haydock, 1963. The dry weight rank method for the botanical analysis of pasture.

Journal of the British Grassland Society 18: 268–275.

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

Appendix 1

Glossary of terms, parameters and symbols relevant to this paper. See also appendices of Neuteboom & Struik (2005a, b)

Primary patch-species	Species used to calculate patch density (d_{patch}).
Secondary patch-species	Species allotted as inclusions to patches; on the basis of the abundance proportions of secondary patch-species the area per patch is re-calculated.

a_{patch1}	Area per patch calculated for primary patch-species.
a_{patch2}	Re-calculated area per patch for the total of primary and secondary patch-species.
c_{patch}	Relative cover by patches.
d_{patch}	Patch density.
n_{ppp}	Number of plants per patch.
n_{prim}	Number of plants of primary patch-species per patch.
n_{sec}	Number of plants of secondary patch-species per patch.
PAI	Patch Area Index.
R_{patch}	Patch radius.
Sn	Species rank when assigned as primary patch-species.
$S_{\max}$	Maximum number of species present in the sampled system.
T	Term in the extended negative-binomial rank abundance curve-fit curve model.
t	Element of T.
x and y	Patch co-ordinates in *in silico* sampling programme.
δ, ε	Coefficients in Gleason's equation for the species-area curve.
ρ	Coefficient in Arrhenius' equation for the species-area curve.
σ	Standard deviation.
τ	Coefficient in Arrhenius' equation for the species-area curve.