POLISH JOURNAL OF ECOLOGY (Pol. J. Ecol.)	47	3	293–305	1999

Werner ULRICH

Nicholas Copernicus University in Toruń, Department of Animal Ecology, Gagarina 9,
87-100 Toruń; Poland, e-mail: ulrichw @ cc.uni.torun.pl

ESTIMATING SPECIES NUMBERS BY EXTRAPOLATION II: ESTIMATING THE ADEQUATE SAMPLE SIZE

ABSTRACT: On the basis of large model assemblages estimators are developed to predict the sample size necessary to sample a given fraction of the total species number. The classical method that takes the point of leveling off of the species accumulation curves proved to be less efficient than the use of the second order jackknife in determining the sample size necessary to collect exactly half of the species number ($N_{0.5}$). The present paper studies eight newly developed estimators for $N_{0.5}$ and shows that estimators based on a Michaelis-Menten formula and a negative exponential model give even better results with minimal sampling effort. The quality of all estimators was not correlated with simple measures of community structure.

KEY WORDS: Sample size, species numbers, model populations, Michaelis-Menten formula, species accumulation curve, extrapolation, community.

1. INTRODUCTION

Despite the fact that many authors have dealt with species richness estimators the question of sample size has only seldom been a matter of study (Moravec 1973, Efron and Thisted 1976, Soberon and Llorente 1993, Keating *et al.* 1998). Often, the sample size is described as or assumed to be "sufficiently large" (e.g. Heck *et al.* 1975, Menkens and Anderson 1988, Baltanas 1992, Colwell and Coddington 1994). The fact is astonishing because exhaustive sampling is often time and energy consuming and it would be valuable to take only the minimum number of samples necessary for answering a given problem (Colwell and Coddington 1994). For many studies, for instance in atlas studies, it would also be welcome to have standardized sampling efforts that allow comparisons of different habitats to be made (Elphick 1996).

An early approach to relate sampling effort and species numbers was made by Good and Toulmin (1956). They used a Bayes approach to estimate species richness and found this estimate to be related to sample size. Southwood (1978) – following the earlier work of Wald (1948), Waters (1955), Kuno (1969), and Green (1970) summarized methods of adjusting the sampling effort if the distribution of a population can be fitted to a negative binomial. Southwood (1978) admits, however, that extensive preliminary work is necessary to establish the type and the parameters of the distribution. Heck *et al.* (1975) gave an estimator of minimum sample size using rarefaction. However, for their method to work, the total number of species has already to be known, a fact that reverses the actual problem. De Caprariis *et al.* (1976) developed an estimator of optimal sample size based on species accumulation curves by using the equation y = ax/(1+bx), a model similar to the Michaelis-Menten approach of enzyme kinetics. But they acknowledged that the accuracy of this estimator heavily depends on the fit of the model and that no extrapolation beyond the sample interval can be made. This, however, is nearly always necessary in real samplings. Keating (1998) reviewed the use of the Michaelis-Menten formula for estimating species numbers.

Baltanas (1992; see also Mingoti and Meeden 1992) mentioned that sample size and reliability of an estimator of species richness are directly related. Although acknowledging that there is no direct measure of sufficient sample size he gave some features of representativeness: high density, low proportion of rare species and small numbers of species in an area that is not too large. Bunge and Fitzpatrick (1993), when reviewing several non-parametric richness estimators noticed that for some estimators there is a threshold below which an estimator does not give reliable results. For the most common used estimators (e.g. jackknife estimators or the Chao estimator) this threshold is roughly the sample size necessary to detect half of the total number of species (in this paper denoted as $N_{0.5}$).

The classical method to "measure" sufficient sampling or sampling completeness is to take the point when the species accumulation curve of a sampling program levels off (Preston 1948, Balogh 1958, Pielou 1977, Moravec 1973, Trojan 1992). However, this method has the severe drawbacks that both type and type II errors frequently occur. In heterogeneous and unevenly distributed assemblages there may be either no asymptotic behavior of the accumulation curve until all species are found or there are pseudo-asymptots (hence underestimating $N_{0.5}$, see below). In very evenly distributed communities the method frequently overestimates the real effort necessary by far (Heck *et al.* 1975, Pielou 1977, Miller and Wiegert 1989).

Recently, Keating *et al.* (1998) tested 11 estimators (mainly based on species – area relationship models) of additional species expected in further n samples (a problem closely related to that of this paper) and found the negative binomial estimator of Efron and Thisted (1976) and the Michaelis-Menten model (De Caprariis *et al.* 1976) the best ones, although negatively biased and dependent on the underlying relative abundance distribution.

The first part of this paper (Ulrich 1999) dealt with the question how to estimate the total number of species of a community from a series of samples. It was shown that the crucial point is the fraction of species already sampled. If more than 33% of the total species number (TS) is already represented in the sample one of the corrected data analytical methods will work sufficiently well, for more than 66% sampled the jackknife estimators give reasonably good estimates. In this second part model communities generated in the same way as in the first part will be used to develop estimators for appropriate sample sizes. It will mainly deal with the question how large has the sample to be to get half of total species number: $N_{0.5} = f(TS_{0.5})$. Other sample sizes (for instance to get 1/3 or 2/3 of TS) may then be inferred by interpolation.

2. METHODS

In this study the data set used in part I (78 model communities) will be used to derive estimators for $N_{0.5}$. The computation process, the program used, and the parameters of these model species assemblages are already described in part I.

The newly developed estimators will then be tested using 50 new model communities (48 to 995 species) computed with the same program. Again power functions (in 30 cases), normal distributions (10 cases), and log-normal distributions (10 cases) were used as underlying density – weight distribution. The SD-values (standard deviations of $\log_2$ densities) of the samples of these communities are given in Fig. 1. The mean aggregation (Lloyd index) of their species ranged between 0.96 (random distribution) and 6.94 (highly aggregated) with a mean of 1.92. Again, a grid of 100 × 100 cells was used taking samples from 1 to 50 cells. Maximum species density was set to 100 ind./cell (total of 1 000 000 ind.), the minimum density ranged between 0.001 and 0.0005 ind./cell (total of 5 to 10 ind.).

Testing the estimators

The estimators developed below based on the Michaelis-Menten and the negative exponential model require an initial rough estimate of TS, the true number of species in the community. Such a previous estimate is possible in many studies, because generally there will be comparisons possible with other studies or an upper boundary is known. To develop the estimators I used the true value of TS and two times TS in the models. The models were then tested with four different estimates of TS (random numbers in these

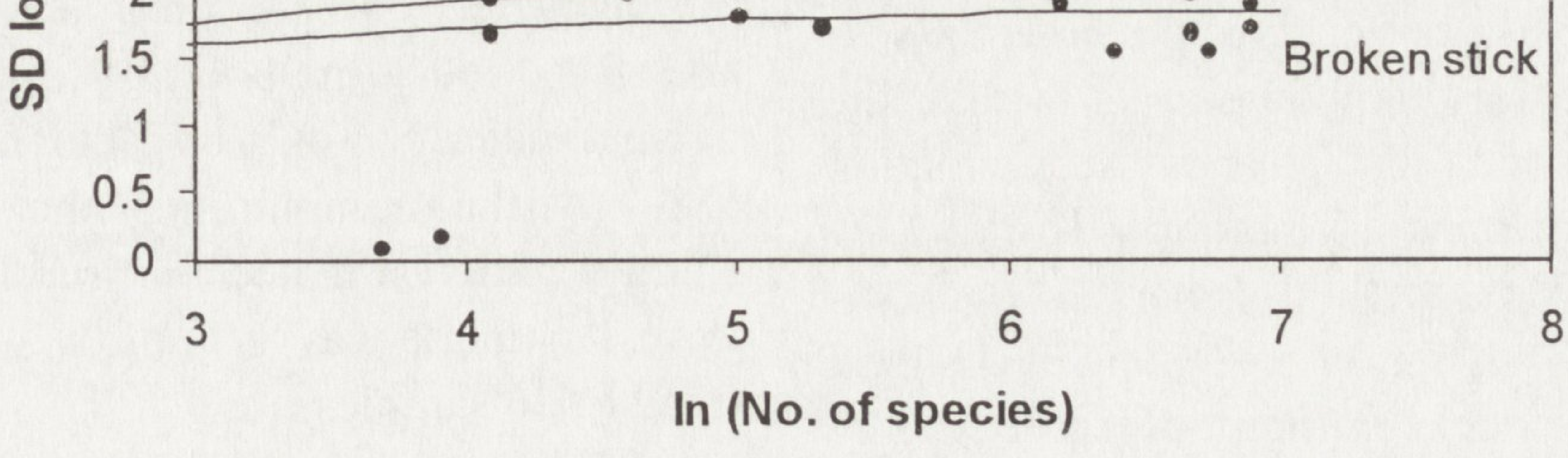

Fig. 1. Standard deviations (SD) of $\log_2$ densities of 50 model assemblages with species numbers between 48 and 995. Given are also the theoretical values of 5 species – rank order distributions. The data for the power fraction and the Sugihara model (log normal distribution) are redrawn from Tokeshi (1996).

ranges): 0.67TS to 1.5TS, 0.5TS to 2TS and 0.67(2TS) to 1.5(2TS), 0.5(2TS) to 2(2TS).

These models used therefore a wide range from 0.5TS to 4TS. The higher the initial estimate of TS, the more conservative will be the estimate of $N_{0.5}$, and the more exact the estimate, the better will be the latter estimate of $N_{0.5}$.

The upper and lower boundaries in Fig. 3 to 6 were computed using the regression method of Blackburn *et al.* (1992) with 5 logarithmic classes. For the $N_{0.5|narrow}$ estimate (see below) 10 logarithmic classes were taken.

3. RESULTS

In principal there are four possible conditions under which a minimal sample size may be inferred.

1. It may be possible to take a large number of samples and to construct a species accumulation curve.

2. It is possible only to take a limited number of samples.

3. Only one sample was taken.

4. There are not previous samples.

In all four cases there may be or may not be a preliminary estimate of the total number of species S.

Case 4 will not be dealt with because without *a priori* knowledge of the assemblage given it seems not to be possible to infer how large the sample has to be to estimate population parameters and species numbers.

Case 1

The most often used method to assess the necessary sample size is to take the point when the species accumulation curve becomes asymptotic (Heck *et al.* 1975). This method generally overestimates $N_{0.5}$ by far. When testing the method with 50 model communities (see methods) the factor of overestimation was 17, ranging from 2 times to 49 times. The factor of overestimation of the sample size to sample 2/3 of TS ($N_{0.67}$) was 7. In none of the cases did the method underestimate the true $N_{0.5}$. It is therefore a very robust estimate but results in a much too high sampling effort.

A better method to estimate $N_{0.5}$ is to use one of the estimators of species numbers (see part I of this paper: Ulrich 1999). Most of these estimators are negatively biased but become asymptotic if the sample size contains more than a certain fraction of the true species number. In this respect the second. order jackknife estimator (E_{J2}) seems most useful because of the clear asymptotic behavior and the small variance. E_{J2} becomes asymptotic if more than half of TS was found.

In a plot of sample size versus E_{J2} – estimate and using 50 model communities four types of curves were found (Fig. 2). Most often (73% of the model communities) occurred types A and C. These types are characterized by a clear leveling off and in the mean the point of leveling off overestimated $N_{0.5}$ only by factor of 1.8. In 18% of the communities the method slightly underestimated the true $N_{0.5}$ The lowest estimate was 89% of $N_{0.5}$ = 32 instead of 36 samples).

In Type B the number of samples is yet too low to reach the point of leveling off. In most of these cases one or more pseudo-levels occurred. Care has there-

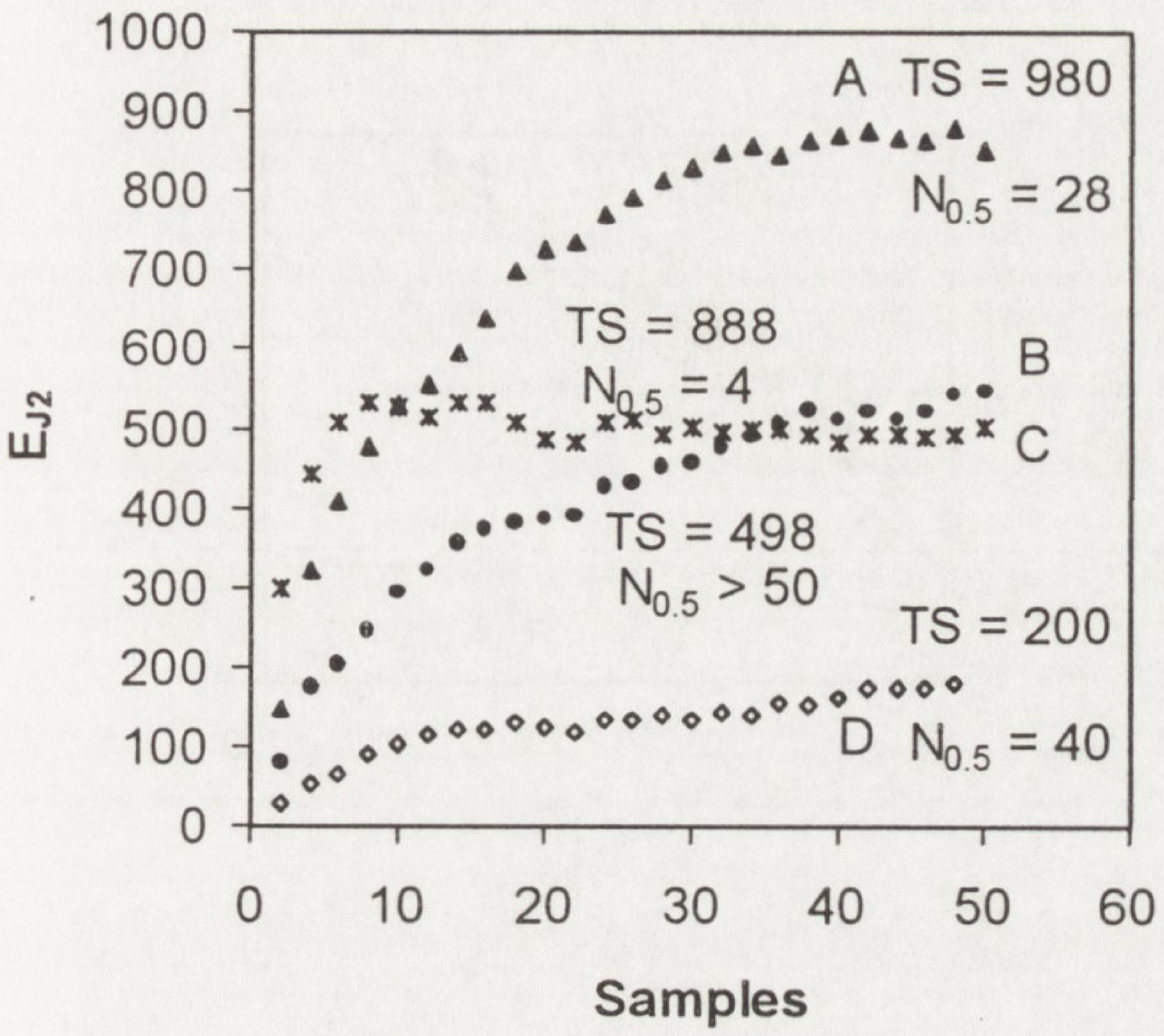

Fig. 2. Estimates (E_{J2}) of the second order jackknife estimator dependent on sample size. Given are also the true species number (TS) and the sample size to collect exactly half of TS ($N_{0.5}$). The four types (A–D) are typical examples of species accumulation curves from 50 model communities and are discussed in the text.

fore to be taken when interpreting the plots and a sufficient number of additional samples (in my communities 10 proved to be enough) have to be considered. The most difficult situation occurs in type C. The communities have a very uneven distribution with many rare species and (often) a low fraction of species per sample unit. In this case, the curve is very flat without a distinct point of leveling off even after $N_{0.5}$ is passed. In this case additional methods which are described below have to be used.

Case 2

Is it possible to estimate $N_{0.5}$ if only a limited number of samples (for instance in a preliminary study) can be taken? There are two estimators of species richness, the Michaelis-Menten and the negative binomial model, which contain TS in their equation and which allow an easy computation of $N_{0.5}$. They have the further advantage that their variation is comparably low (see part I: Ulrich 1999). Therefore, both models were used to develop estimates of $N_{0.5}$ taking only three samples (the minimum possible number) from the model assemblages.

In the case of the Michaelis-Menten approach, $N_{0.5}$ equals the constant B (in the following denoted as B_{MM}) of the equation

$FS(n) = (TS\,n)/(B + n)$. (1)

In the negative binomial model

$FS(n) = TS\,(1 - e^{-Kn})$ (2)

the estimate of $N_{0.5}$ is given by

$B_{NE} = \ln(0.5)/{-K}$ (3)

where FS(n) is the cumulative number of species after n samples. B denotes the sample size to find exactly half of the total number of species (TS). K is a constant which determine the shape of the function and which is derived from the fitting process.

Both estimators require an initial estimate of TS (see the methods section). Because the model communities cover the whole range of real assemblages the lower boundary lines of the plots in Fig. 3 may be used to construct robust estimators of $N_{0.5}$:

Michaelis-Menten, exact value of TS (MM_{TS}):

$B_{MM} = (B/0.7948)^{1/0.7466}$ (4)

Michaelis-Menten, 2TS (MM_{2TS}):

$B_{MM} = (B/2.5081)^{1/0.639}$ (5)

Negative exponential, exact value of TS (NE_{TS}):

$B_{NE} = ([\ln(0.5)/K]/0.9608)^{1/0.622}$ (6)

Negative exponential, 2TS (NE_{2TS}):

$B_{NE} = ([\ln(0.5)/K]/2.5553)^{1/0.5316}$ (7)

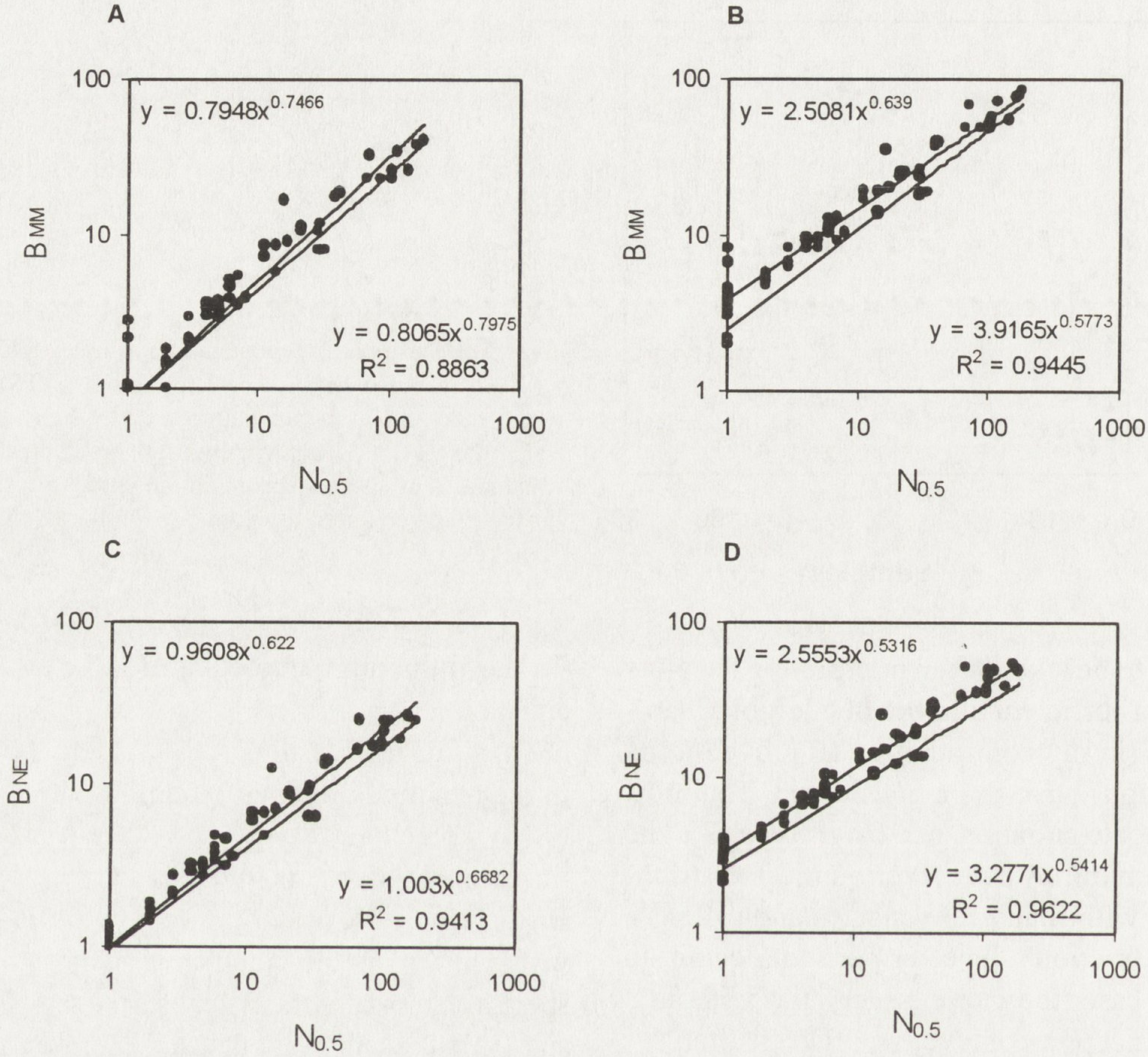

Fig. 3. Estimated $N_{0.5}$ (B_{MM} and B_{NE}) dependent on the true $N_{0.5}$. A, B: Michaelis-Menten model using the true TS (A) and two times TS (B). Given are the regression functions (below) and the function of the lower boundary lines (upper equation). C, D: negative exponential model with TS (C) and two times TS (D). Abbreviations as in Fig. 2.

These formulas were tested using 50 assemblages with estimates of TS ranging from 0.5TS to 4TS (see Methods).

First, Fig. 3 and Table 1 compare the estimates B_{MM} and B_{NE} of formulas 1 and 3 with real $N_{0.5}$. Both values are highly correlated. Table 2 shows that both methods are able to estimate $N_{0.5}$. If TS is roughly known – cases MM_{TS} (4) and NE_{TS} (6) – the methods overestimate the true value by factors of 1.8 to 3.2 and 10 to 12% of the estimates are below the true value. If TS is less known – cases MM_{2TS} (5) and NE_{2TS} (7) 6 to 10% of the estimates of $N_{0.5}$ were below the true value and the mean factor of overestimation was 2.4 to 2.8. As expected, the more exact the estimate of TS, the smaller the standard deviation of the estimate of $N_{0.5}$.

Case 3

The same type of reasoning was also used to develop estimators that are based only on the number of species per sampling unit (SM_1) or on the relation between species numbers found in one (SM_1) and in two cells (SM_2). These cases resemble situations in which only one or two samples were taken from the community under study. Figures 4 and 5 show that again there is a correlation between $N_{0.5}$ and SM_1/TS and (SM_2–SM_1)/TS.

Table 1. Eight methods to assess the number of samples to collect at least 50% of the true species number of an animal assemblage. Data from 78 model assemblages with 28 to 997 species. Given is the percentage of estimates that is below the true sample size to detect half of the species ($N_{0.5}$). The factor of overestimation of $N_{0.5}$ is the quotient of estimate and true value In the case of (SM_2 – SM_1)/TS only communities were taken in which $N_{0.5} > 1$.
Symbols are the same as used in formulae 1 to 11 in the text. TS: estimate using the true species number; 2TS: estimate sung two times TS $SM_{1,2}$: number of species found in 1, 2, respectively, units of area (cells)

Methods using initial estimates of TS

Method	Percentage estimates below $N_{0.5}$	Mean factor of overestimation $N_{0.5}$	SD of overestimation
Michaelis-Menten ($MM_{TS)}$	6.4	1.6	0.8
Michaelis-Menten MM_{2TS}	5.1	2.1	1.1
Negative exponential NE_{TS}	5.1	1.5	0.5
Negative exponential NE_{2TS}	5.1	2.0	0.7
SM_1/TS	3.8	2.5	1.1
(SM_2 – SM_1)/TS	6.4	2.9	5.3

Methods not using initial estimates of TS

Method	Percentage estimates below $N_{0.5}$	Mean factor of overestimation $N_{0.5}$	SD of overestimation
$N_{0.5|narrow}$	10.3	7.4	12.8
$M_{0.5|wide}$	3.8	11.1	18.4

Table 2. Test of eight methods to assess the nu.2,ber of samples to collect at least 50% of the true total species number. Data from 50 model assemblages with 48 to 995 species. Given is the percentage of estimates that is below the true sample size to detect half of the species. The factor of overestimation of $N_{0.5}$ is the quotient of estimate and true value. In the case of (SM_2 – SM_1)/TS only communities were taken in which $N_{0.5} \geq 1$
Symbols are the same as in Table 1 and in formulas 1 to 11 in the text.

Methods using initial estimates of TS

Method	Percentage estimates below $N_{0.5}$	Mean factor of overestimation $N_{0.5}$	SD of overestimation	Percentage estimates below $N_{0.5}$	Mean factor of overesti -mation $N_{0.5}$	SD of overesti -mation
	2/3TS < estimate of TS < 3/2TS			1/2TS < estimate of TS < 2TS		
MM_{TS}	10.	1.9	1.1	12.0	3.2	2.5
MM_{2TS}	6.0	2.4	1.2	8.0	2.7	2.0
NE_{TS}	12.0	1.8	1.2	12.0	3.1	2.3
NE_{2TS}	6.0	2.5	1.3	10.0	2.8	2.0
SM_1 / TS	8.0	2.8	1.8	8.0	4.4	3.2
(SM_2 – SM_1)/TS	8.0	2.5	1.7	4.0	4.4	5.3

Methods not using initial estimates of TS

Method	Percentage estimates below $N_{0.5}$	Mean factor of overestimation $N_{0.5}$	SD of overestimation
$N_{0.5|narrow}$	8.0%	11.5	14.8
$M_{0.5|wide}$	2.0%	16.3	20.6

The equations of the upper boundary lines may therefore serve as estimators of $N_{0.5}$.

SM_1/TS:

$$B_{SM1} = 1.0108\,(SM_1/TS)^{-1.74} \quad (8)$$

$(SM_2-SM_1)/TS$:

$$B_{SM1,2} = 149.64\,e^{-24.544\,(SM2-SM1)/TS} \quad (9)$$

Is it possible to estimate $N_{0.5}$ without any preliminary estimate of TS. Figure 6 shows that in the model communities there was indeed a correlation between SM_1 and $N_{0.5}$ which can be used to construct estimators. With the method of Blackburn *et al.* (1992) it is possible to construct two boundary lines, a wider and a more narrow one (see methods).

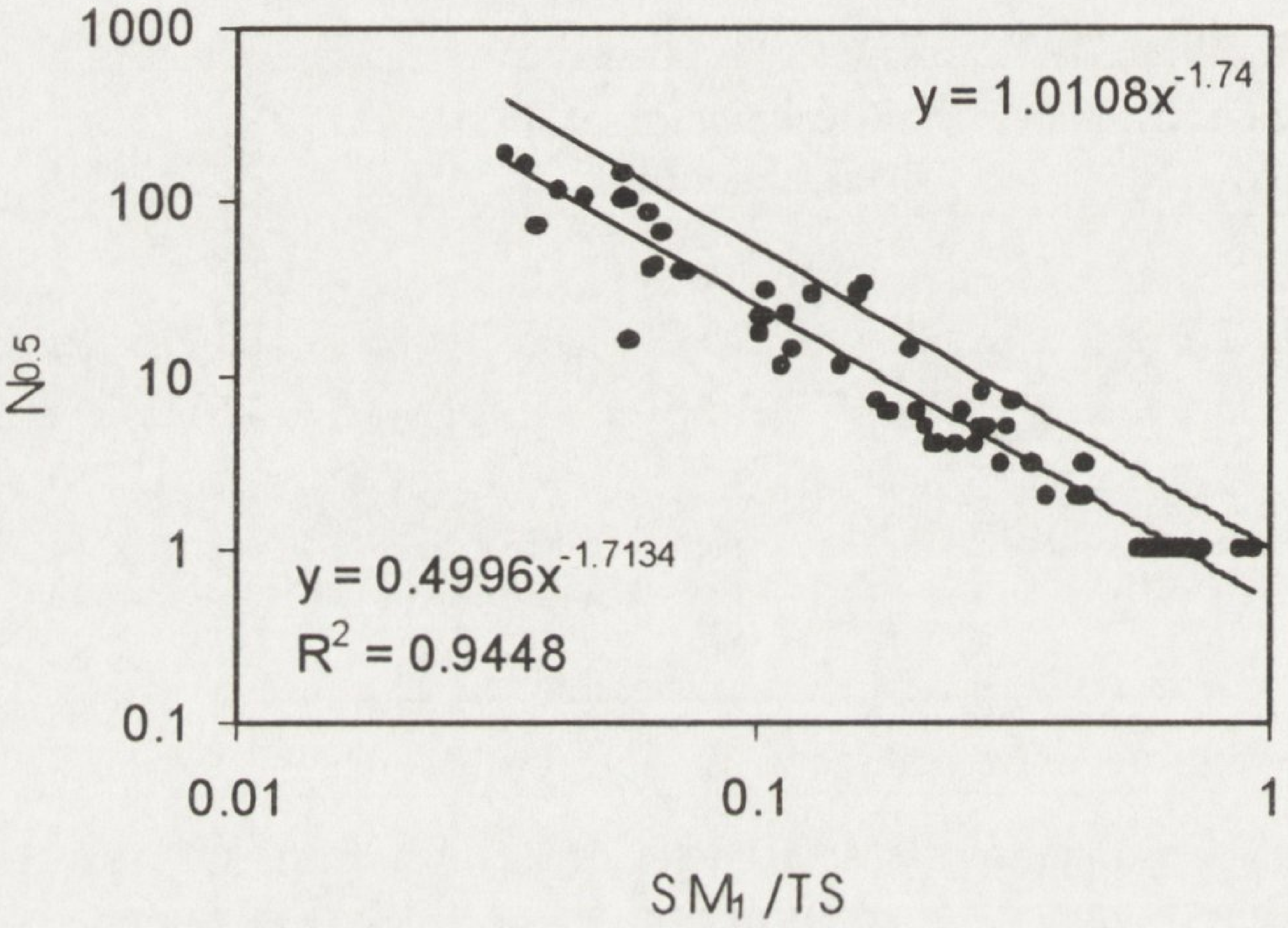

Fig. 4. $N_{0.5}$ versus the quotient of species number per sampling unit (cell) (SM_1) and TS. The upper equation gives the regression of the upper boundary line. Abbreviations as in Fig. 2.

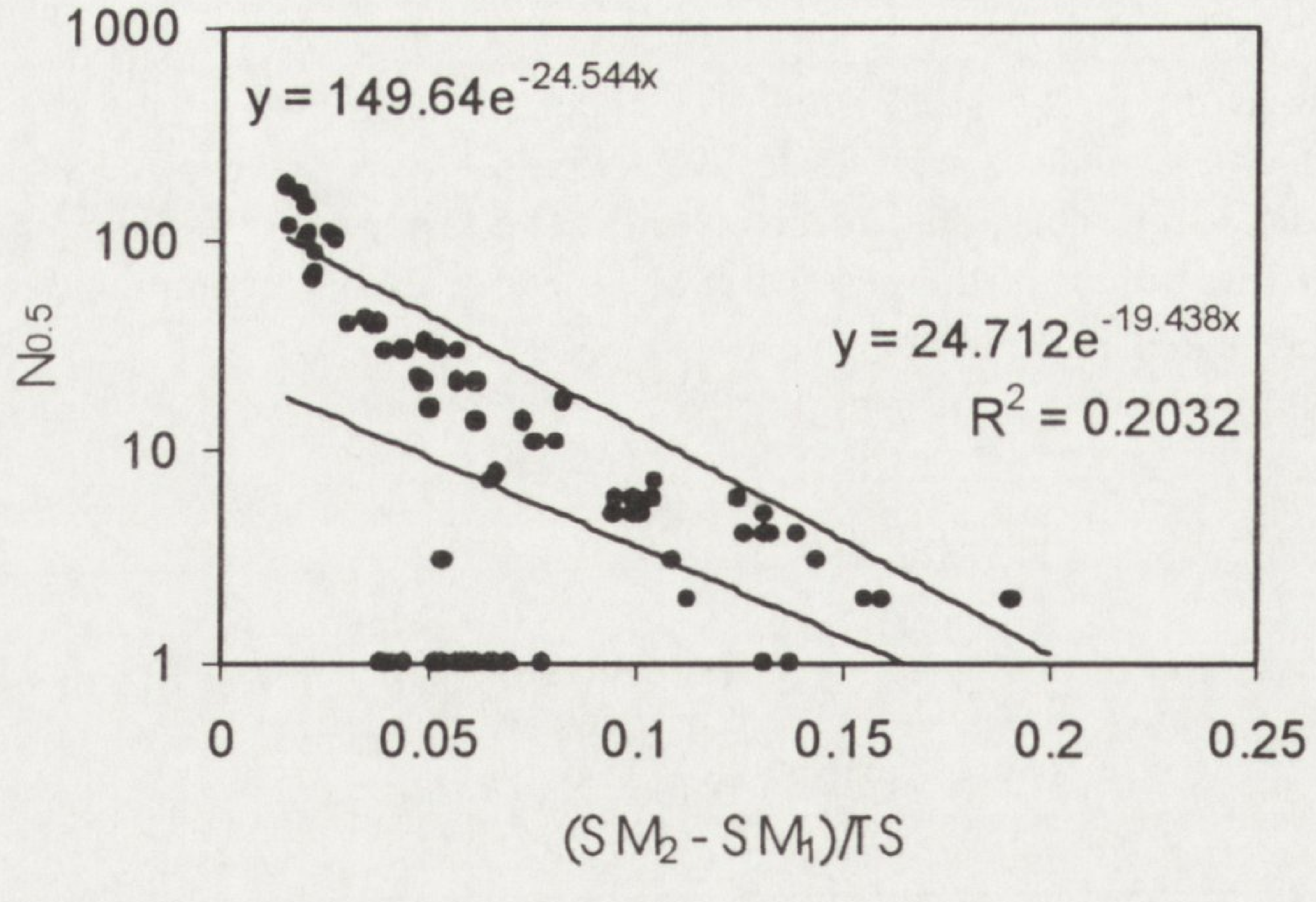

Fig. 5. $N_{0.5}$ versus the quotient of (SM_2–SM_1) and TS. The upper equation gives the regression of the upper boundary line. The regression was computed excluding the $N_{0.5}$ = 1 values. SM_1, SM_2 are the species numbers found in one and two samples, respectively.

Table 2 shows that indeed both estimators are able to predict $N_{0.5}$, but due to the more limited information used, the results are not as good as in the case of the MM– and NE-approaches. Between 4 and 8% of the estimates underestimate the true value and the mean factor of overestimation was 2.5 to 4.4. The standard deviations also indicate slightly worse results in comparison with the two above estimators.

$N_{0.5|narrow}$

$$B = 5223.4\,SM_1^{-1.4157} \quad (10)$$

$N_{0.5|wide}$

$$B = 5492.5\,SM_1^{-1.314} \quad (11)$$

As expected, both estimators are very robust and only 8 and 2%, respectively, of their estimates were below the true value

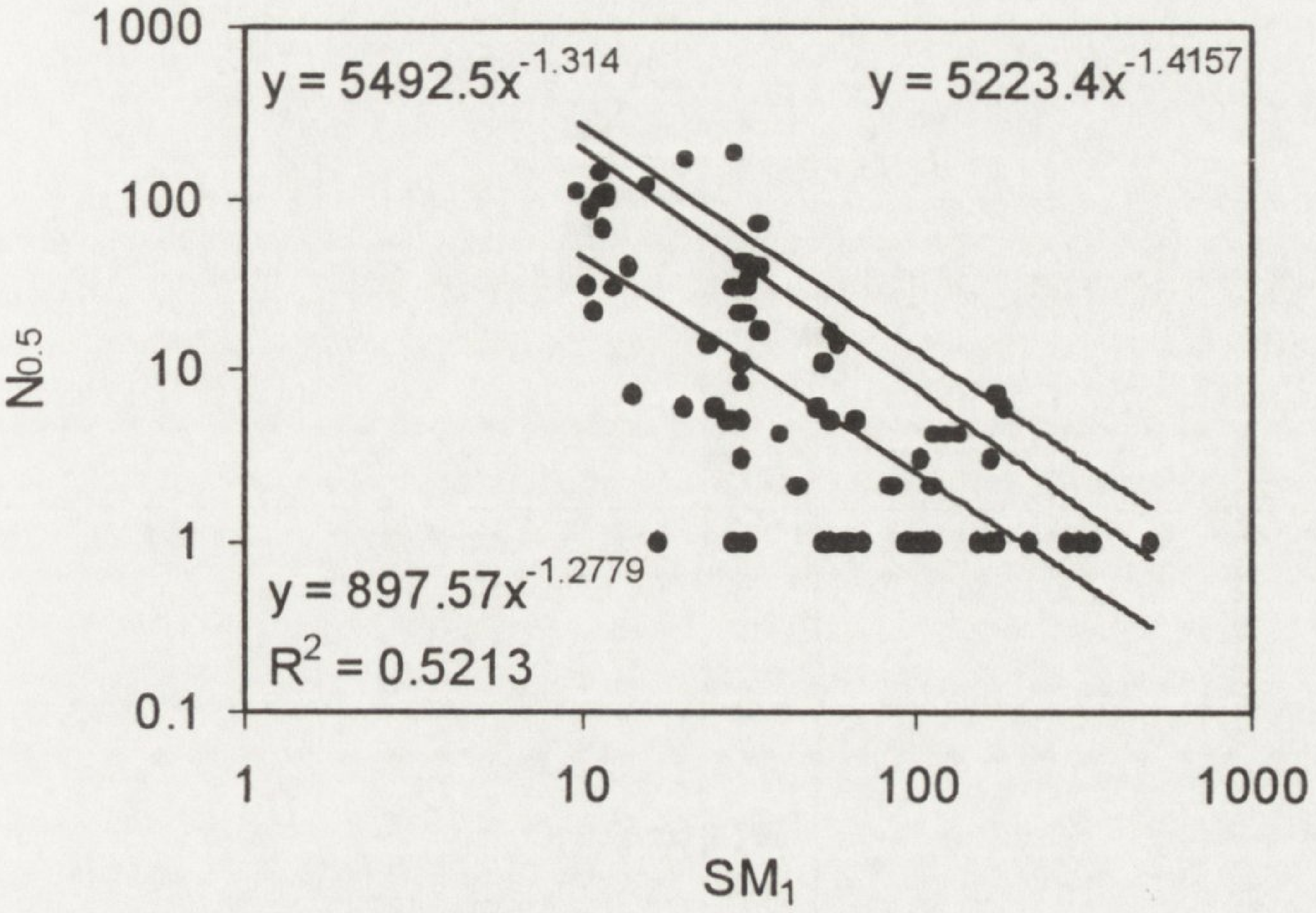

Fig. 6. $N_{0.5}$ versus the species number per sampling unit (cell). The upper equation on the left side gives the regression of the outer upper boundary line ($N_{0.5lwide}$) (formula 11 in the text), the equation on the right side is the regression of the inner upper boundary line ($N_{0.5lnarrow}$) (formula 10 in the text).

of $N_{0.5}$ of the test communities. However, in the mean they overestimated the true $N_{0.5}$ by factors of 11.5 and 16.3 combined with a relatively high variance. These results are comparable to the first leveling off method of case 1 but with much less sampling effort.

Interpolating other fractions of species

With the Michaelis-Menten and the negative exponential approach it is easy to compute $N_{0.33}$, $N_{0.67}$ or other sample numbers. The general formulas are (x denotes the proportion of species to be found):

Michaelis-Menten (MM):

$$N_x = xB/(1-x) \quad (12)$$

Negative exponential (NE):

$$N_x = N_{0.5}\ln(1-x)/\ln(0.5) \quad (13)$$

In the MM-approach $N_{0.33}$ is exactly $0.5N_{0.5}$ and $N_{0.67} = 2N_{0.5}$. The NE model gives slightly different results: $0.6N_{0.5}$ and $1.6N_{0.5}$.

Table 3 compares such interpolations (for x = 0.33, 0.67, and 0.9) with the real values and shows that both estimators are indeed able to predict the true N_x in a reasonable way. This is especially true for low fractions. In the case of $N_{0.33}$ nearly all estimates were above the true value. The results of the Michaelis-Menten model were slightly better than that of the negative exponential. At higher fractions the performance reverses. The Michaelis-Menten model overestimates the true sampling effort more than the negative exponential model. However, the negative exponential tends to underestimate the real sampling effort. Between 14 and 36% of the estimates (x = 0.67 and 0.9) ranged below the true value.

Table 3. Interpolation of estimates of $N_{0.33}$, $N_{0.67}$, and $N_{0.9}$ with the formulas given in the text. Methods, symbols and assemblages are the same as in Tables 1 and 2

Method	Percentage estimates below $N_{0.33}$	Mean factor of overestimation $N_{0.33}$	SD of overestimation	Percentage estimates below $N_{0.67}$	Mean factor of overestimation $N_{0.67}$	SD of overestimation	Percentage estimates below $N_{0.9}$	Mean factor of overestimation $N_{0.9}$	SD of overestimation
	2/3TS < estimate of TS < 3/2TS			2/3TS < estimate of TS < 3/2TS			2/3TS < estimate of TS < 3/2TS		
MM_{TS}	0.00%	1.8	0.8	17.95%	1.9	1.1	19.05%	3.6	3.9
MM_{2TS}	0.00%	2.4	1.1	17.95%	2.5	2.0	7.69%	3.4	2.5
NE_{TS}	2.00%	2.2	1.2	35.90%	1.4	0.8	23.81%	2.1	1.6
NE_{2TS}	0.00%	3.2	1.7	25.64%	2.1	1.6	23.81%	2.5	1.8
	1/2TS < estimate of TS < 2TS			1/2TS < estimate of TS < 2TS			1/2TS < estimate of TS < 2TS		
MM_{TS}	8.00%	2.9	2.0	15.38%	3.2	2.3	14.29%	5.5	4.3
MM_{2TS}	4.00%	2.6	2.1	12.82%	2.7	2.0	19.05%	4.7	3.8
NE_{TS}	4.00%	3.6	2.5	23.08%	2.6	1.9	14.29%	3.4	2.1
NE_{2TS}	0.00%	2.8	1.0	28.21%	2.8	1.0	28.57%	2.8	1.0

4. DISCUSSION

The above analysis showed that it is possible to estimate sample sizes to sample a given fraction of species of a community having only minimal knowledge about this community. In a larger series of samplings, the use of the second order jackknife estimator proved to give better results than the classical method which determines the point of leveling off of the species accumulation curve.

If only a limited number of samplings is available (3 in this paper) the use of a negative exponential model gave the best results to determine $N_{0.5}$ or larger fractions. For smaller fractions of the true species number (TS) to be found the Michaelis-Menten approach was more efficient. Both methods overestimate the real N_x (a given proportion x of TS) by factors between 1.8 and 5.5. For both methods to work an initial estimate of TS is necessary. If such an estimate is not available, the number of species per sampling unit leads to an estimate which is frequently too high (factor 11.5 in the mean) but which is still better than the leveling off method.

Which factors determine the quality of the estimate? From samplings of the model communities the SD-value, the mean aggregation of the species, the true species numbers, the number of species per sampling unit, and the underlying density – weight distribution with their parameters were known. Table 4 shows the results of a MANCOVA and a multiple regression to study which factors influence the quality of the estimate (measured by the factor of overestimation). The surprising result appears that only in a few cases do the predicators correlate significantly with the estimates. Especially the underlying distributions – measured by the SD-value and the type of density – weight dis-

Table 4. MANCOVA and multiple regression to detect the influence of underlying distributions and SD values on the performance (estimate/true value) of eight methods to estimate $N_{0.5}$. The MANCOVA tested the influence of the underlying density – weight distribution on the performance of the estimators (log-normal, normal or power function). SD was used as a covariate. The multiple regression used SD (the standard deviation of the mean aggregation (all species with more than 50 individuals in the sample, see part I: Ulrich 1999), SM_1 and TS.

Symbols are the same as in Table 1 and in formulae 1 to 11 in the text. β: β-weight of the multiple regression; *F*: *F* – statistic of the MANCOVA. ***: $P < 0.001$, ****: $P < 0.0001$.

MANCOVA			Multiple regression			
Method	All effects	Covariate	SD	Aggregation	SM_1	TS
	F	SD	β	β	β	β
MM_{TS}	0.33	–0.18	–0.18	0.07	0.01	0.25
MM_{2TS}	0.4	–0.1	–0.1	0.14	0.17	–0.05
NE_{TS}	0.37	–0.05	0.05	0.15	–0.24	0.4
NE_{2TS}	0.18	–0.06	0.04	0.19	–0.04	0.15
SM_1/TS	0.48	–0.15	0.01	0	–0.33	0.61***
$(SM_2 - SM_1)/TS$	11.1****	0.25	0.07	0.23	0.01	0.25
$N_{0.5lnarrow}$	1.05	–0.01	–0.15	–0.16	–0.08	–0.52***
$M_{0.5lwide}$	1.15	–0.01	–0.16	–0.16	–0.05	–0.55***

tribution (normal, log-normal or power) had only a minor influence on the outcome of the estimate. In the SM_1/TS-method (8) the quotient of estimate and true value is correlated with TS. Therefore, the lower the number of species, the better the estimate. The opposite is true for the $N_{0.5|narrow}$ and $N_{0.5|wide}$ estimators (10 and 11). In the case of the (SM_2-SM_1)/TS-estimator (9) a normal distribution gave worse results (factor of overestimation 5.5) than the two others (factor 2.1). This means that in the latter cases this estimator may serve as an alternative to the two parametric ones.

The estimators are developed on the basis of model assemblages. Because these assemblages span a wide range of community types, they are surely also applicable to natural communities. However, the next step has to be to test the above developed formulas using real communities.

Of course, the estimators are far from being perfect. A mean factor of overestimation of 2 mesan that instead of for instance the necessary 10 samples the estimators give 20. $N_{0.5|narrow}$ (formula 10) gives 115 – a value that is hardly satisfactory. However, under the wide range of community structures to be considered and the very limited amount of data used no better results seem to be possible. To obtain better estimates detailed knowledge of community structure and community specific sample theoretical modeling would be necessary, tasks for which the knowledge of the sufficient sample size is already required. It seems that this circle cannot be solved.

ACKNOWLEDGMENTS: I thank Prof. J. Buszko and Dr. Kartanas for critical and valuable suggestions on the manuscript. Miss H. Pearson kindly improved my English.

5. SUMMARY

The present paper developed methods to estimate the sample size to sample a given fraction to the total species number (TS) of large animal communities (Table 1, Figs 1 to 6). It is shown that a Michaelis-Menten and a negative exponential model are able to predict sample size sufficiently well if there is an initially (rough) estimate of TS (Tables 2 and 3) (B and K are the parameters of the Michaelis-Menten and the negative exponential model developed from the species accumulation curve; $B_{0.5}$ is the estimate of" TS):

Michaelis-Menten, using the exact value of TS (MM_{TS}):

$$B_{0.5} = (B/0.7948)^{1/0.7466}$$

Michaelis-Menten, using two times TS (MM_{2TS}):

$$B_{0.5} = (B/2.5081)^{1/0.639}$$

Negative exponential, using exact value of TS (NE_{TS}):

$$B_{0.5} = ([\ln(0.5)/K]/0.9608)^{1/0.622}$$

Negative exponential, using two times TS (NE_{2TS}): $B_{0.5} = ([\ln(0.5)/K]/2.5553)^{1/0.5316}$

In the mean these estimators overestimated the real value by factors of 1.8 to 3.2.

If TS is not known the number of species per sampling unit can be used:

$$B_{0.5|narrow} = 5223.4\ SM_1^{-1.4157}$$

$$B_{0.5|wide} = 5492.5\ SM_1^{-1.314}$$

Due to the reduced amount of data used the latter estimators give less exact results and overestimate the real values by factors of 11.5 and 16.3.

It was not possible to predict the quality of the estimates from the parameters of the model assemblages (Table 4).

6. REFERENCES

Balogh J. 1958 – Lebensgemeinschaften der Landtiere – Berlin (Akad. Verlag).

Baltanas A. 1992 – On the use of some methods for the estimation of species richness – Oikos, 65: 484–492.

Blackburn, T. M., Lawton J. H., Perry J. N. 1992 – A method of estimating the slope of upper bounds of plots of body size and abundance in natural animal assemblages – Oikos, 65: 107–112.

Bunge J., Fitzpatrick M. 1993 – Estimating the number of species: a review – J. Am. Stat. Assoc., 88: 364–373.

Colwell R. K., Coddington J. A. 1994 – Estimating terrestrial biodiversity through extrapolation – Phil. Trans. R. Soc. Lond. B: 345: 101–118.

De Caprariis P., Lindemann R. H., Collins C. M 1976 – A method for determining optimum sample size in species diversity studies – Math. Geol., 8: 575–581.

Elphick Ch. S. 1996 – Correcting avian richness estimates for unequal sample effort in atlas studies – Ibis, 139: 189–190.

Efron B., Thisted R. 1976 – Estimating the number of unseen species: how many words did Shakespeare know? – Biometrika, 63: 435–447.

Good I. J., Toulmin G. H. 1956 – The number of new species, and the increaese in population coverage, when a sample is increased – Biometrika, 43: 45–63.

Green R. H. – On fixed precision level sequential sampling – Res. Pop. Ecol., 12: 249–251.

Heck K. L., van Belle G., Simberloff G. 1975 – Explicit calculation of the rarefaction diversity measurement and the determination of the sufficient sample size – Ecology, 56: 1459–1461.

Keating K. A. 1998 – Estimating species richness: the Michaelis–Menten model revisited – Oikos, 81: 411–416.

Keating K. A., Quinn J. F., Ivie M. A., Ivie L. L. 1998 – Estimating the effectiveness of further sampling in species inventories – Ecol. Apll., 8: 1239–1249.

Kuno E. (1969) – A new method of sequential sampling to obtain the population estimates with a fixed level of precision – Res. Pop. Ecol., 11: 127–136.

Menkens G. E., Anderson S. H. 1988 – Estimation of small mammal population size – Ecology, 69: 1952–1959.

Miller, R. I., Wiegert R. G. 1989 – Documenting completeness, species–area relations, and the species abundance distribution of a regional flora – Ecology, 70: 16–22.

Mingoti S. A, Meeden G. 1992 – Estimating the total number of distinct species using presence and absence data – Biometrics, 48: 863–875.

Moravec J. 1973 – The determination of the minimal area of phytocenoses – Folia Geob. Phytotax., 8: 23–47.

Pielou E. C. 1977 – Mathematical Ecology – New York (Wiley & Sons), 1–385.

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

Soberon M. J., Llorente B. J. 1993 – The use of species accumulation functions for the prediction of species richness – Cons. Biol., 7: 480–488.

Southwood T. R. E. 1978 – Ecological Methods – London (Chapman & Hall), 1–523.

Tokeshi M. 1996 – Power fraction: a new explanation of relative abundance patterns in species-rich assemblages – Oikos, 75: 543–550.

Trojan P. 1992 – Analysis of faunal structure – Memorabilia Zoologica, 47: 1–120.

Ulrich W. 1999 – Estimating species numbers by extrapolation I: Comparing the performance of various estimators using large model communities – Pol. J. Ecol. 47: 271–291.

Wald A. 1948 – Sequential Sampling – New York (Wiley & Sons).

Waters W. E. 1955 – Sequential sampling in forest insect surveys – For. Sci., 1: 68–79.

(*Received after revising April 1999*)