POLISH JOURNAL OF ECOLOGY (Pol. J. Ecol.)	47	3	271–291	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 I: COMPARING THE PERFORMANCE OF VARIOUS ESTIMATORS USING LARGE MODEL COMMUNITIES

ABSTRACT: A computer program was constructed that simulates large species assemblages (28 to 997 species) with various species – rank order distributions and degrees of aggregation of the species. From these model assemblages random samples were taken to study the performance of 14 estimators of species diversity. For 6 of the estimators correction factors are developed. In sufficiently large samples (more than 2/3 of the true species number (TS) sampled) a corrected second order jackknife estimator gave the best results. 18% of the estimates ranged outside TS ± 10%. If fewer species are represented in the sample (but more than 1/3 TS) two newly developed data analytical estimators performed better. Between 23 and 24%, respectively, of their estimates ranged outside TS ± 20%. Crucial to the performance of all of the estimators is the sample size. The minimum sample size for an estimator to work has to contain at least 1/3 of the total species number.

KEY WORDS: model species assemblages, species diversity, jackknife-estimator, bootstrap-estimator, distribution, estimation, sampling, parametric models.

1. INTRODUCTION

Estimating species numbers from a series of samples is one of the more important but also one of the most difficult tasks in ecological research. Despite the fact that the problem has recently gained growing attention (Miller and Wiegert 1989, Palmer, 1990, 1991, Baltanas 1992, Chao *et al.* 1992, Mingoti and Meeden 1992, Bunge and Fitzpatrick, 1993, Hodkinson and Hodkinson 1993, Soberon and Llorente 1993, Colwell and Coddington 1994, Lee and Chao 1994, Solow 1994, Coddington *et al.* 1996, Norris and Pollock 1996, Edwards 1997, Longino and Colwell 1997, Tackaberry *et al.* 1997, Boulinier *et al.* 1998, Keating 1998, Walther and Morand 1998) no systematic performance study and comparison of the various methods proposed in the literature has up to now been undertaken.

Palmer (1990, 1991) tested eight extrapolation methods using medium large plant assemblages. He found that of the non-parametric methods the first- and

second-order jackknife estimators (Smith and van Belle 1984) performed best. Of the parametric models a log-linear estimator was least biased but resulted in too high estimates. Bunge and Fitzpatrick (1993) reviewed several estimators from a sample theoretical point of view, but dealt only briefly with data analytical methods. They did not compare estimates and real values. The review of Colwell and Coddington (1994) contains such comparisons for some non-parametric estimators and a small community. They largely confirmed the results of Palmer (1990). Keating (1998) tested the Michaelis-Menten approach and found it to be unsatisfactory in heterogeneous communities. The test of Tackaberry *et al.* (1997) again showed that uncorrected data analytical models (log-log or log linear) result in too high estimates, thus supporting the findings of Palmer (1990). Winklehner *et al.* (1997) applied the first order jackknife estimator to epigeic Collembola and confirmed that it yields reliable results if a high fraction of species is already found. Walther and Morand (1998) tested 9 estimators with real data and computer simulations and found the first order jackknife and the CHAO 2 estimator to perform best. Their study however dealt only with a small assemblage (40 species) and high fractions of species sampled (95%).

In all of these reviews small to medium sized assemblages (<150 species) were studied. Other studies used model communities constructed assuming certain fixed distributions (e.g. Heltshe and Forrester 1983, Smith and van Belle 1984, Baltanas 1992, Chao and Lee 1992, Keating 1998). However, such an assumption is not met in reality. Good estimators have to deal with various kinds of communities and the performance has to be studied under a wide range of conditions. Another aspect that up to now has only gained little attention is the performance of the estimators under the condition of small sample size, that means the degree of bias and possible threshold values.

In this and in the second part of this paper (Ulrich 1999a) I will therefore take another approach to test species diversity estimators with model assemblages. In testing 14 different estimators a newly developed program will be used that – starting from a few basic features of natural assemblages – generates a wide range of large but realistic model communities. Random samples from these communities will then be used to study the performance of each estimator.

This first part deals largely with the dependence of the quality of the estimates on the fraction of species represented in the sample. The second part will then study the question how to adjust the sample size to get a sufficient fraction for the estimators to work.

2. METHODS

Real populations are characterized by more or less distinct density – weight distributions (Lawton 1990, Currie 1993, Ulrich 1999b, c). These distributions also roughly define the ranges in which the densities of the species fluctuate (Ulrich 1999b). A second feature of most of the species of a community is that their spatial distribution is non-random but aggregated. Thirdly, species have no indefinitely low densities; however, there is a species specific lower density limit. Start-

ing with these three basic features of animal communities a FORTRAN-program was developed to generate model animal assemblages (Fig. 1). As a random generator the commercial Dran1 module of 'Numerical Recipes Software' was used, giving uniformly distributed random numbers. The program places individuals at random into the cells of a large grid (for this study a 100 × 100 grid was used), takes random samples of various size from this grid and analyses them. To mimic real assemblages the maximum density per species and per cell was set to 100, the minimum density to 0.001. One can compare this procedure with the insects in one ha area; the species then have densities between 0.001 and 100 ind. m^{-2}. These densities and the density-range (5 magnitudes of order) resemble natural ones (Schaefer 1991, 1996, Ulrich 1998).

The individuals of up to 1000 species were placed at random inside the cells of the grid. The densities of the species were random variables inside the range given by log-normal, normal, linear, power or random density-weight distributions and the minimum density. In reality density – weight distributions most often follow power functions (Currie 1993, Ulrich 1999c), and the upper boundary limits, defined in most assemblages (Gaston 1993, Scharf *et al.* 1998, Ulrich 1999c), have exponents between 0.5 and more than 2. The triangular form of many of the distributions may also be described by log-normal or normal functions. Linear and random distributions gave in all of the runs too even distributions with nearly all species present in the first sample. Therefore, in the present study 53 power functions, 15 log-normal and 10 normal distributions were tested.

To simulate aggregation 1 to 50 individuals (the number chosen at random) were placed together resulting in values of

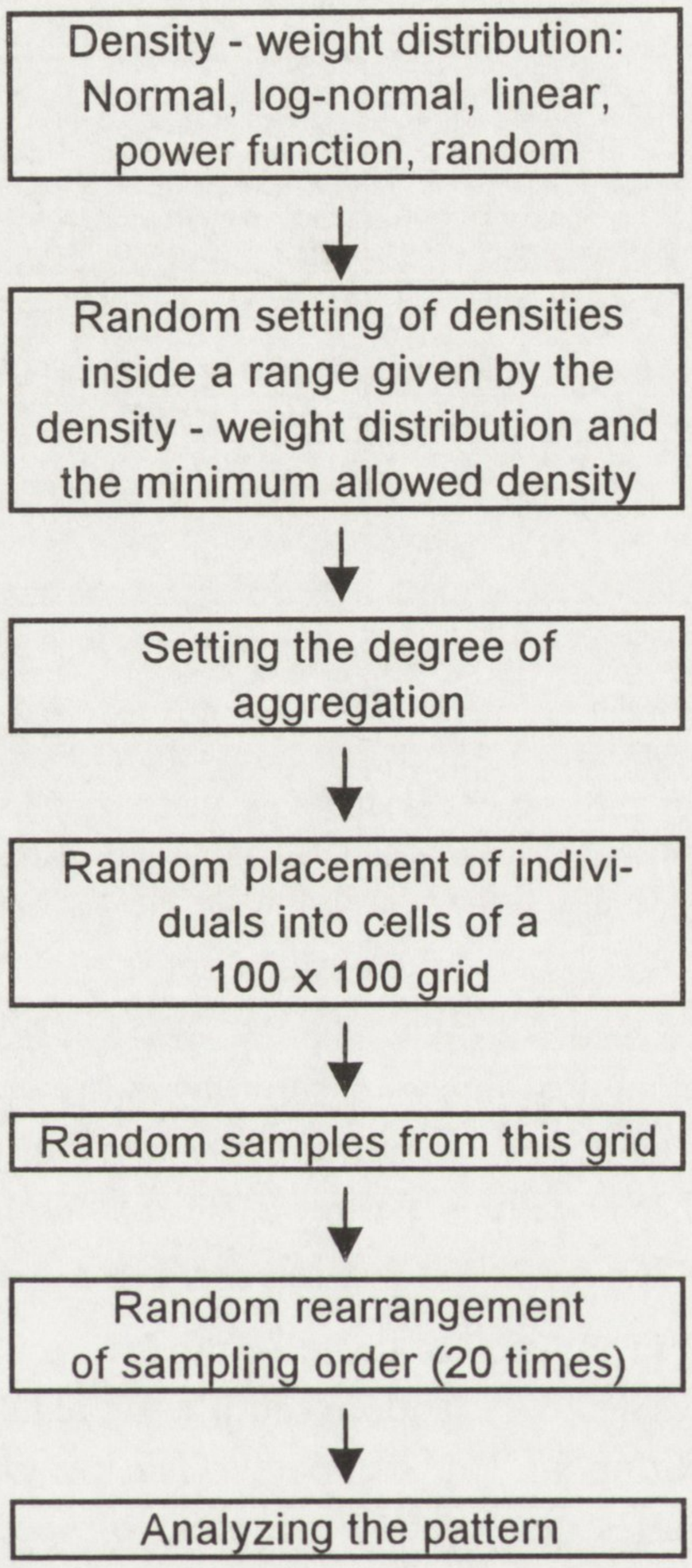


Fig. 1. Flow diagram of the FORTRAN-program for constructing and analyzing the model communities used in the present study. Further explanations in the text.

the Lloyd index between 1 and 10 which also equals natural values. With this procedure I computed 78 different model assemblages having 28 to 997 species. Fig. 2 shows that these assemblages had distributions similar to natural ones. Following Sugihara (1980) and Tokeshi (1996) the standard deviations of $\log_2$ species densities (SD) of the assemblages are plotted against species numbers. All model communities range inside an area given by the most even rank order distribution (the broken stick) and uneven distributions (Power fraction models). Real large het-

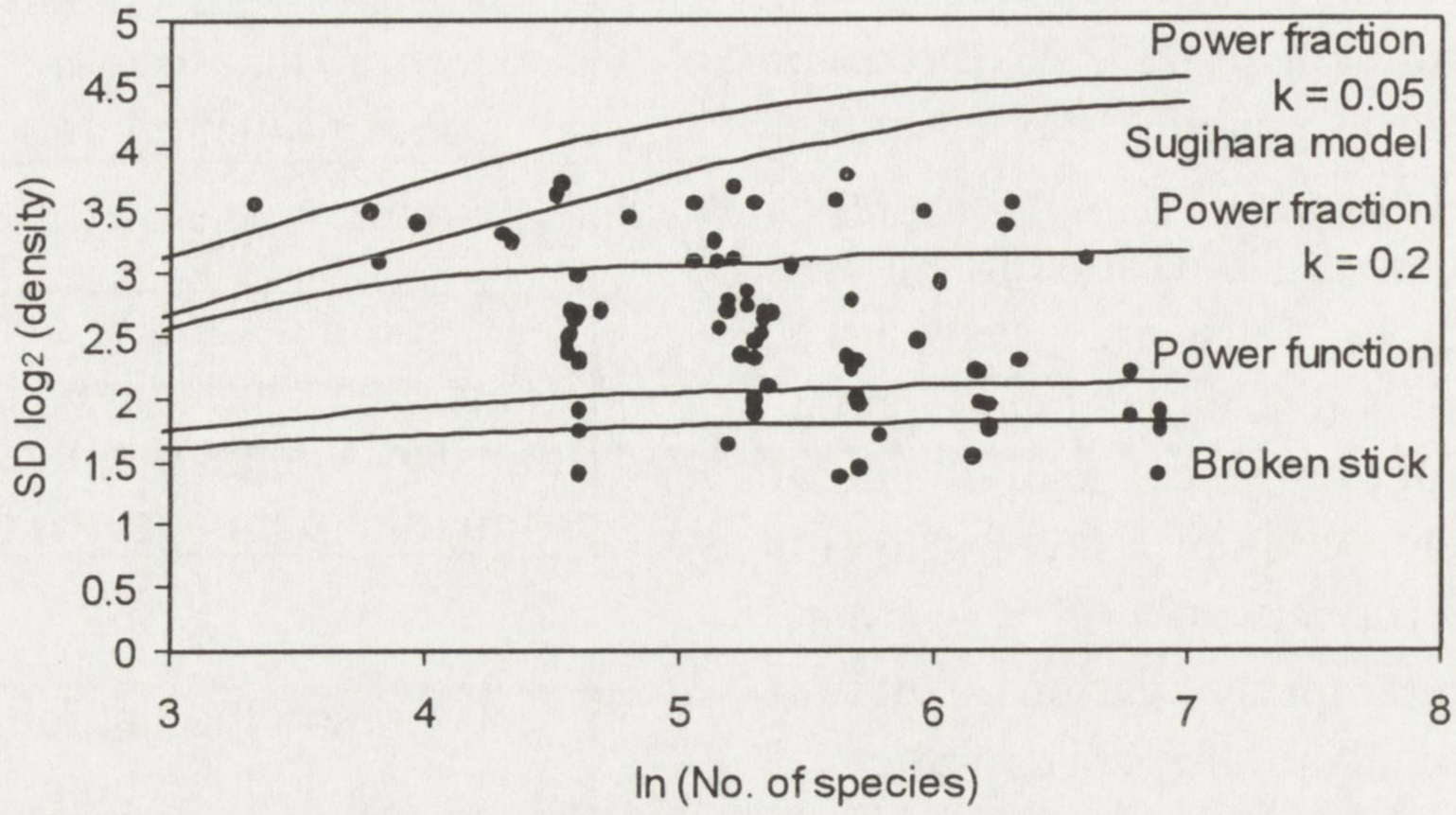


Fig. 2. Standard deviations (SD) of $\log_2$ densities of the samples of 78 model assemblages with species numbers between 28 and 997. Given are also the theoretical values of 5 species – rank order distributions. The data for the power fraction and the Sugihara model (log normal distribution) are redrawn from Tokeshi (1996).

erogeneous animal assemblages are often characterized by rather even distributions with SD values below 2 (Ulrich unpublished).

From each of these 78 model communities 8 and 48 random samples were taken resulting in 156 estimates. The program assumes quantitative sampling of species and homogeneity of the whole grid.

Because the sampling order has a large influence on the resulting species accumulation curves the sampling order was randomized 20 times and the mean values taken (following the procedure of Colwell and Coddington 1994). Of each model community the SD-value and the mean aggregation (measured by the index of Lloyd, Pielou 1977) were computed. Because the Lloyd-index – as other indices using the variance/mean ratio – is biased at low densities (McArdle *et al.* 1990), I computed the index only for those species which had more than 50 individuals in the sample.

Preliminary analyses showed that not the number of samples but the percentage of the true species number represented in the sample is the main factor determining the performance of an estimator of species diversity. Therefore, the following Figs 4 to 8 do not plot the estimate (E) versus the true species number TS (the normal procedure in the literature) but plot the quotients E/TS versus number of species found in the total sample (FS) / true number (FS/TS). Such a plot shows immediately the performance under various sample sizes.

To assess the quality of the estimators the data points of Figs 4 to 8 were fitted to a normal distribution and the skewness of the distributions was calculated. A good estimator should have a mean around 1 (E/TS), a low variance and a low skewness. A lower mean and/or marked skewness indicate a negative bias and vice versa.

The question what is a good estimate is largely a matter of choice. In small stable communities a 5% error term may be desired. In large assemblages with a fairly degree of annual species turnover even a 20% error term seems to be acceptable. The analysis showed that none of the estimators is able to catch the 5% goal (TS ±

5%). Therefore, Tables 2 and 3 show how many of the estimates ranged inside TS ± 10% and TS ± 20%. The Tables do not give the variance of the estimates. Although for all non-parametric estimators such variance terms are available (Burnham and Overton 1978, Heltshe and Forrester 1983, Smith and van Belle 1984, Colwell and Coddington 1994) and the variance of the parametric estimators can easily be inferred from the estimation process, such values are misleading, because most estimators are biased and their distributions skewed (Tables 1 and 2). However, as an empirically derived estimate of the true variance the variance of the normal fits in Tables 1 and 2 can be taken.

3. RESULTS AND DISCUSSION

Methods for estimating species numbers can be classified into two groups: non-parametric estimators derived from sample theoretic reasoning, and data analytical parametric methods. In the latter case three different kinds of species-area relationships can be used to estimate species numbers (Fig. 3): asymptotic functions (Type 1) or non-asymptotic (Type 2) functions both using species accumulation curves, and functions using a plot of the newly found species versus sample size (Type 3).

A third main group of methods uses the parameters of the log-normal distribution (Preston 1962) to infer the total number of species (e.g. Miller and Wiegert 1989, Baltanas 1992, Kobayashi and Kimura 1994). However, because real populations seldom follow exactly theoretical distributions, and the mode has to be known (Hughes 1986), which is often not the case in real samples, such computations do not result in sufficiently exact estimates and will not be analyzed in this paper. Slocomb *et al.* (1977) and Slocomb and Dickson (1978) found that such an estimator requires at least 1000 individuals in the sample and more than 80% of the true species number has to be represented. Palmer (1990) studied the method and found it not to be better than the simple measure of number of species detected.

Pielou (1977) gave a method to infer the number of species if the community fits a negative binomial distribution with $k > 0$. This distribution is seldom used in the ecological literature (see Tokeshi 1993 for a review) and probably not often applicable to real communities. The method has the further disadvantage that the necessary estimation parameters themselves have to be estimated by the empirical data set, thus enhancing the variance. I am not aware of any study using this method.

A third species – rank order distribution that easily allows species numbers to be calculated is the geometric series (Pielou 1977). This is the most uneven distribution and may be found in some small communities structured by a few severe ecological factors (Tokeshi 1993). This type of distribution is easy to recognize and needs no further methods for estimating species richness.

Uncorrected non-parametric estimators

Bunge and Fitzpatrick (1993 and literature therein) discussed various non-parametric estimators developed assuming different underlying distributions. In the present study the four most often used

Table 1. Performance of estimators of species numbers. Fit of normal distribution, skewness and % estimates outside 10 or 20% error range of true species number (TS ± 10% or 20%) The values are given for more than 1/3, more than 2/3 TS, and between 1/3 and 2/3 of TS represented in the sample. Model communities are the same as in the Figures. Normally distributed estimators (judged by the Kolmogorov-Smirnov test) are marked in bold print

Method	More than 1/3 TS found					More than 2/3 TS found					Between 1/3 and 2/3 TS found	
	Normal distribution		Skewness	outside ± 10%	outside ± 20%	Normal distribution		Skewness	outside ± 10%	outside ±20%	outside ± 10%	outside ± 20%
	Mean	Variance				Mean	Variance					
Chao (E_{CHAO})	0.86	0.02	–0.10	52%	27%	**0.92**	**0.00**	–0.33	34%	4%	93%	79%
1. order jackknife (E_{J1})	0.84	0.02	–0.93	52%	30%	**0.93**	**0.00**	–0.57	30%	3%	100%	88%
2. order jackknife (E_{J2})	0.89	0.02	–0.90	40%	21%	**0.96**	**0.00**	–0.77	19%	1%	83%	67%
Bootstrap (E_{BOOT})	**0.79**	**0.03**	–0.79	67%	45%	0.89	0.01	–0.10	51%	19%	100%	100%
LOGLOG (E_P)	3.03	4.01	1.17	100%	100%	2.74	2.11	1.45	100%	100%	100%	100%
LOGLIN 1 (E_{L1})	1.27	0.11	1.19	66%	48%	**1.25**	**0.09**	1.33	62%	43%	74%	57%
LOGLIN 2 (E_{L2})	1.05	0.04	0.30	60%	37%	1.10	0.03	0.66	50%	32%	81%	48%
Michaelis-Menten (E_{MM})	**0.83**	**0.03**	1.08	70%	42%	**0.91**	**0.02**	3.84	58%	20%	95%	88%
Negative exponential (E_{NE})	**0.72**	**0.03**	–0.55	86%	61%	**0.82**	**0.01**	0.08	80%	42%	100%	100%
Hyperbola (E_H)	1.22	0.04	0.36	94%	78%	1.52	0.23	0.34	92%	73%	98%	88%
Asymptotic power (E_{AP})	0.84	0.17	1.76	77%	58%	0.92	0.09	2.37	70%	46%	91%	86%
Asymptotic linear (E_{AL})	**0.88**	**0.02**	–0.37	50%	24%	**0.94**	**0.01**	0.09	33%	6%	88%	64%
New species LOGLIN (E_{Pnew})	**0.90**	**0.11**	0.15	74%	64%	**1.07**	**0.06**	0.35	64%	50%	95%	95%
New species LOGLOG ($_{ELnew}$)	0.98	0.30	2.09	87%	78%	0.95	0.19	2.36	83%	70%	95%	93%

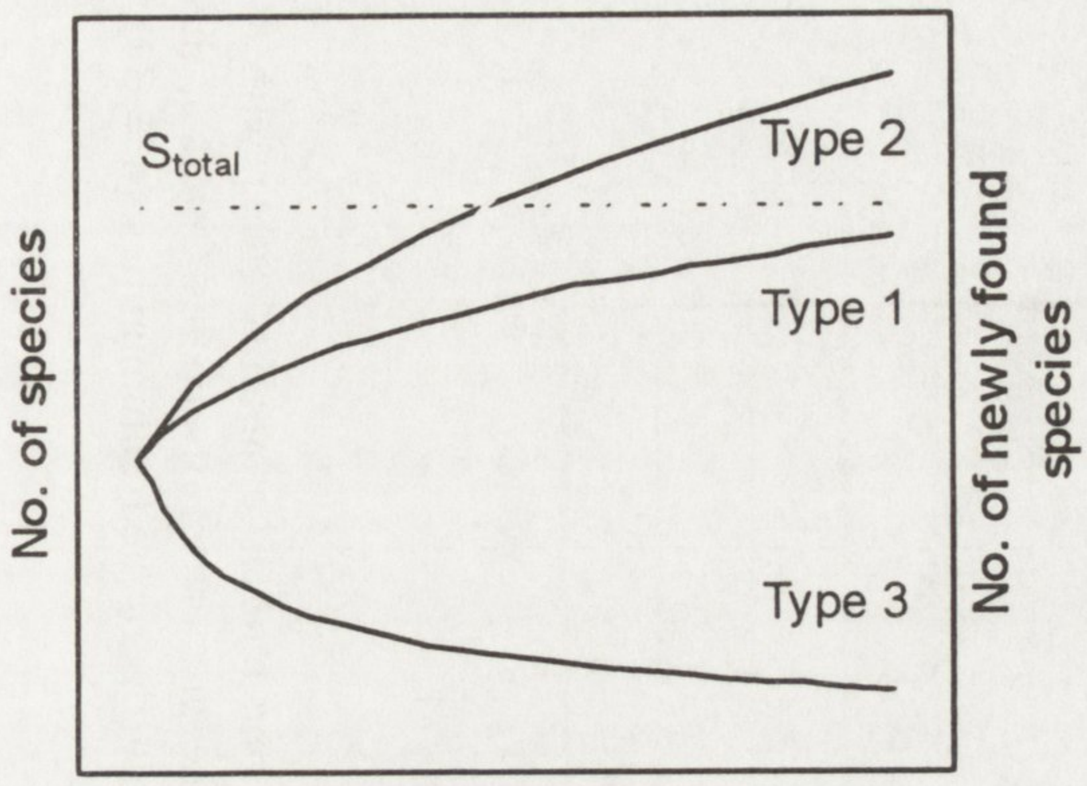


Fig. 3. Three types of species – area models that can be used to compute species numbers. Asymptotic species accumulation curve (Type 1), infinite species accumulation curve (Type 2), and a plot of the newly found species versus area (Type 3).

estimators in the ecological literature were tested: the Chao estimator (Chao 1984), the first and the second order jackknife (Burnham and Overton 1978, 1979; Smith and van Belle 1984) and a bootstrap estimator (Smith and van Belle 1984). For other non-parametric estimators see Chao (1987), Chao and Lee (1992), Lee and Chao (1994) and Norris and Pollock (1996). These latter estimators depend on closed capture-recapture models and the tests given in the two latter works indicate that they do not perform better than the four estimators tested in this paper [but see Walther and Morand (1998) for the case of high fractions of species sampled].

Chao:

$$E_{Chao} = FS + (S_1^2 / 2S_2) \quad (1)$$

First order jackknife:

$$E_{J1} = FS + S_1 (n-1) / n \quad (2)$$

Second order jackknife:

$$E_{J2} = FS + [S_1 (2n-3) / n] - [S_2 (n-2)^2 / n (n-1)] \quad (3)$$

Bootstrap:

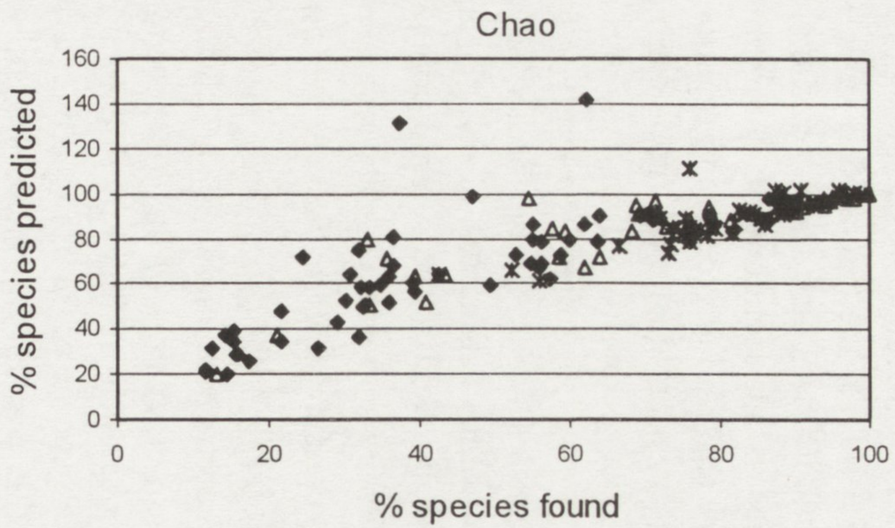
$$E_{Boot} = FS + \sum_{i=1}^{FE} (1 - p_i)^n \quad (4)$$

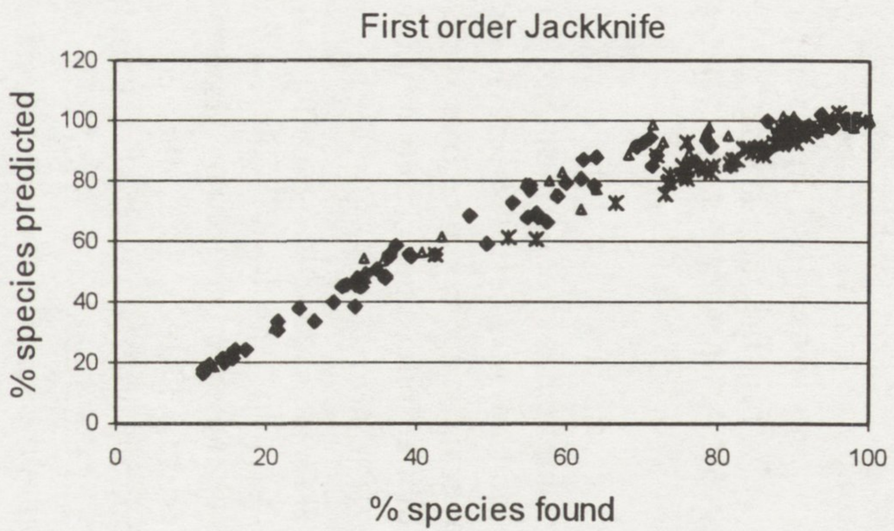
where FS is the number of species found in the sample. S_1, S_2 are numbers of species that occur in exactly 1 or 2 samples, respectively, n is the sample size, and p_i: the proportion of cells containing each species *i*.

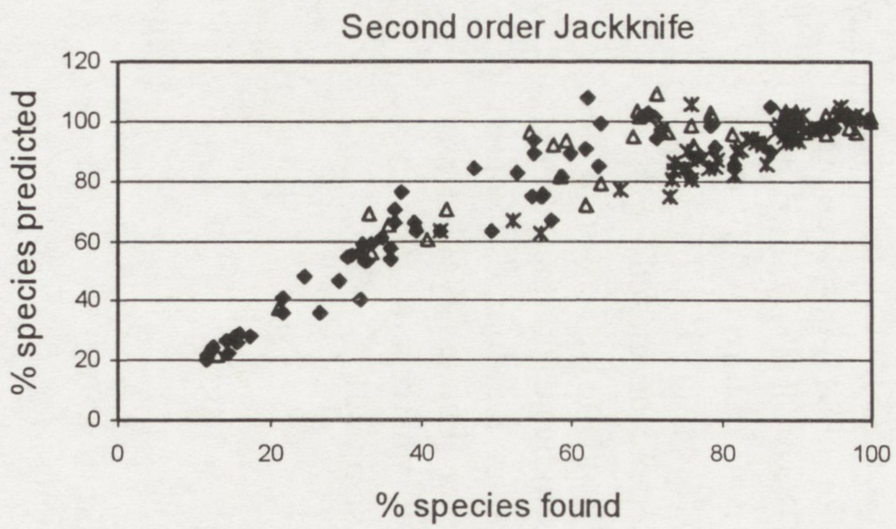
These estimators have the drawback that the sample has to contain at least 50% of TS (the total number of species). They also have a strong negative bias. Fig. 4 shows that despite some claims in the literature (Palmer 1990, Colwell and Coddington 1994) the bootstrap is not suited to predict species numbers, a fact that has already been noticed by Mingoti and Meeden (1992). The other estimators work reasonably well if already more than 2/3 of the true species number is represented in the sample with the second order jackknife giving the best results. Around 20% of the E_{J2} estimates ranged outside ± 10% of the true value (Table 1). Fig. 4 and Table 3 also show that the performance is independent of community structure (SD and degree of aggregation), a necessary prerequisite of a good estimator. If fewer species were found the bias of all four methods resulted frequently in too low estimates. Again E_{J2} performed better than the other estimators (Table 1) and may still be used if the sample contains more than 50% of TS (Fig. 4).

Uncorrected asymptotic parametric estimators (Type 1)

Theoretically there is an infinite number of estimators of this kind because every constantly rising function can be forced to become asymptotic (see Brainerd 1972 for several such functions). However, only a few such functions have been applied to estimate species numbers. This paper examines five such functions, the Michaelis-Menten formula (de Ca-







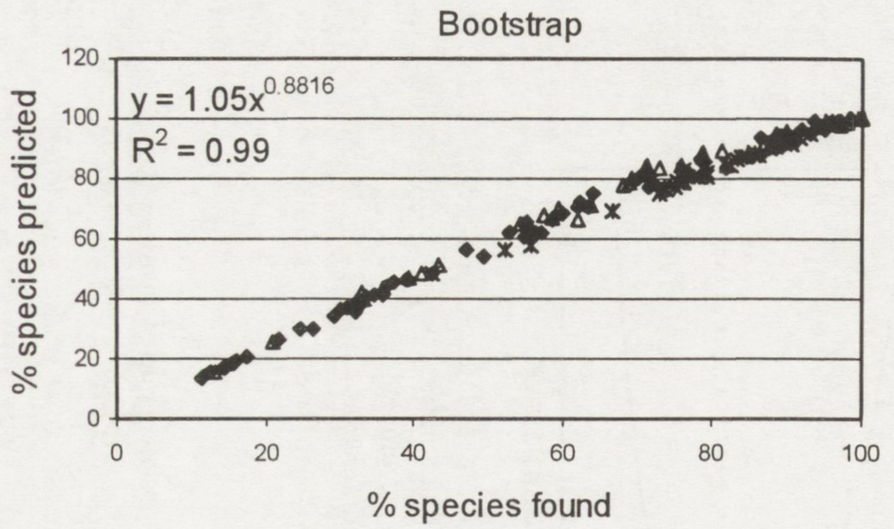


Fig. 4. Performance of non-parametric estimators for estimating species numbers. Each data point represents the estimate for one model community either with 8 or with 48 samples taken, together 156 estimates. Plotted is the proportion of estimate / true species number (E/TS) against the proportion of species found in the sample against true species number (FS/TS). *: Unevenly distributed communities (n = 23) with SD > 3, ♦: moderately evenly distributed communities (n = 34). with SD between 2 and 3, △: evenly distributed communities (n = 21) with SD < 2.

Table 2. Performance of corrected estimators of species numbers. Fit of normal distribution, skewness and % estimates outside 10 or 20% error range of true species number (TS ± 10% or 20%) The values are given for more than 1/3, more than 2/3 TS, and between 1/3 and 2/3 of TS represented in the sample. Model communities are the same as in the Figures. Normally distributed estimators (judged by the Kolmogorov-Smirnov test) are marked in bold print. Symbols as in Table 1, E_{GM}: geometric mean of all corrected parametric estimators, E_{ML}: mean of corrected E_{L1} and E_{L2}.

Corrected method	More than 1/3 TS found					More than 2/3 TS found					Between 1/3 and 2/3 TS found	
	Normal distribution		Skewness	outside ± 10%	outside ± 20%	Normal distribution		Skewness	outside ± 10%	outside ± 20%	outside ± 10%	Outside ± 20%
	Mean	Variance				Mean	Variance					
E_{L1}	**1.14**	**0.05**	0.55	68%	46%	**1.14**	**0.04**	–0.27	73%	46%	60%	55%
E_{L2}	**0.95**	**0.03**	–0.29	50%	24%	**1.01**	**0.02**	–0.52	36%	10%	80%	52%
E_{ML}	**1.05**	**0.03**	0.13	62%	24%	**1.08**	**0.02**	–0.45	60%	23%	67%	26%
E_{Pnew}	1.01	0.20	2.10	79%	52%	0.99	0.12	2.99	72%	34%	91%	80%
E_{NE}	**0.97**	**0.06**	0.06	72%	39%	**1.04**	**0.04**	0.46	67%	27%	83%	64%
E_{AL}	**0.94**	**0.03**	–0.46	50%	27%	**1.01**	**0.02**	–1.26	38%	11%	81%	62%
E_{GM}	0.99	0.03	–0.06	49%	23%	**1.03**	**0.01**	–0.46	34%	10%	76%	55%
E_{J1}	0.86	0.02	–1.02	46%	25%	**0.95**	**0.00**	–0.34	21%	0%	92%	76%
E_{J2}	0.93	0.02	–0.73	36%	19%	**1.00**	**0.01**	0.82	18%	1%	71%	57%
E_{Boot}	1.05	0.07	0.67	64%	40%	**0.99**	**0.06**	0.88	57%	30%	76%	52%

Table 3. Multiple regression of performance (% of real species numbers predicted) as dependent and standard deviation of $\log_2$ densities (SD), degree of aggregation (mean value of Lloyd-index) and % of the real species numbers found in the sample as independent variables. Given are the β-weights of the multiple regression functions. Significant β-weights ($P < 0.05$) are marked in bold print.

Method	SD	Degree of aggregation	% species detected in the sample
Chao (E_{CHAO})	-0.07	-0.02	**0.95**
1. order jackknife (E_{J1})	0.07	-0.05	**0.99**
2. order jackknife (E_{J2})	0.08	-0.06	**0.94**
Bootstrap (E_{BOOT})	0.03	-0.03	**1**
LOGLOG (E_P)	**-0.43**	0	0.21
LOGLIN 1 (E_{L1})	**0.22**	**-0.37**	**0.29**
LOGLIN 2 (E_{L2})	0.14	-0.22	**0.61**
Michaelis-Menten (E_{MM})	**0.2**	**-0.2**	**0.89**
Negative exponential (E_{NE})	0.02	**-0.09**	**1**
Hyperbola (E_H)	**-0.09**	0.1	**0.85**
Asymptotic power (E_{AP})	0.05	0.1	0.17
Asymptotic linear (E_{AL})	-0.11	**-0.19**	**0.88**
New species LOGLIN (E_{Pnew})	-0.07	0.02	**0.88**
New species LOGLOG ($_{ELnew}$)	**0.28**	**-0.44**	**0.43**

prariis *et al.* 1976), which was originally developed to describe the kinetics of enzymes (Morris 1976) but – because of its asymptotic shape – has also been applied to estimate species numbers (Keating 1998), the negative exponential (Soberon and Llorente 1993), an asymptotic power function (Stout and Vandermeer 1975), and a simple hyperbola (Lauga and Joachim 1987). Additionally, an asymptotic linear model was analyzed.

Michaelis-Menten (E_{MM}):

$$FS(n) = (TSn)/(B + n) \quad (5)$$

Negative exponential (E_{NE}):

$$FS(n) = TS(1 - e^{-Kn}) \quad (6)$$

Simple Hyperbola (E_H):

$$FS(n) = TS/n \quad (7)$$

Asymptotic power function (E_{AP}):

$$FS(n) = an^z(1 - FS(n)/TS) \quad (8)$$

this results after simple rearrangement in

$$FS(n) = a/[n^z + (a/TS)] \quad (9)$$

Asymptotic linear function (E_{AL}):

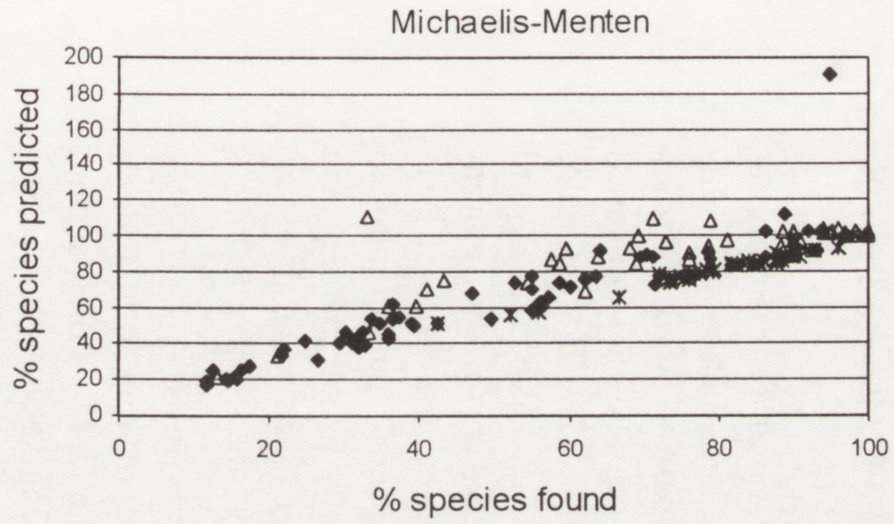
$$FS(n) = (a + bn)(1–FS(n)/TS) \quad (10)$$

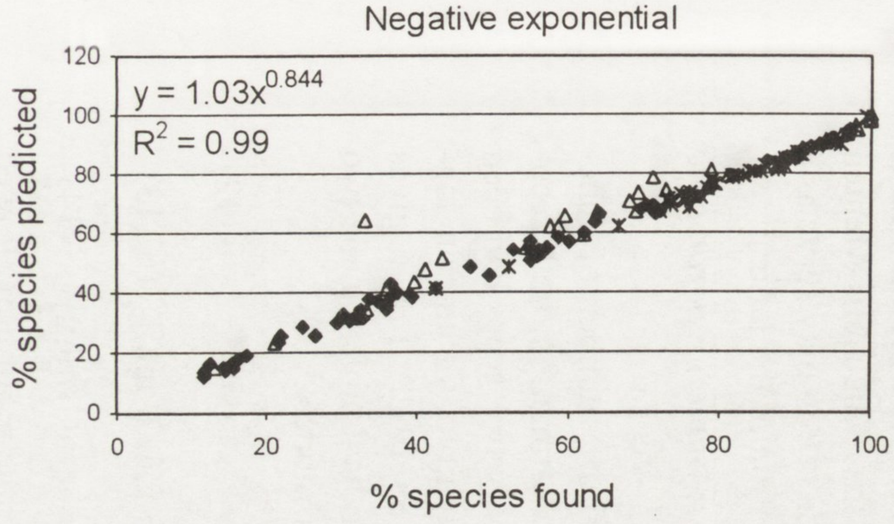
which results in

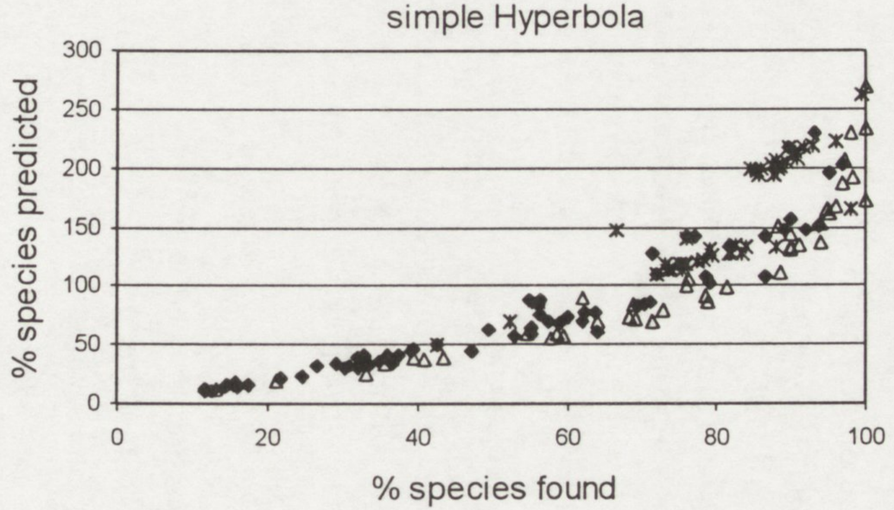
$$FS(n) = (an+b)/(1+(an+b)/TS)$$

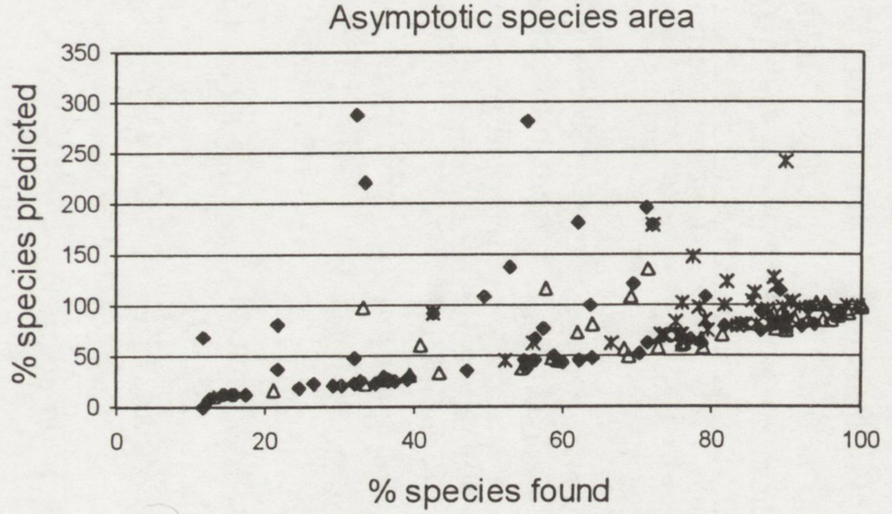
where FS(n) is the cumulative number of species after n samples. B denotes the sample size to find exactly half of the total number of species (TS). a, b, z, and K are constants which determine the shape of the functions and which are derived from the fitting process. Other symbols as in Formulas 1 to 4.

The parameters of these models were directly estimated with the nonlinear estimation procedure of the STATISTICA packet (Quasi Newton method). Care has to be taken on the initial data settings in the estimation procedure because "wrong" settings may result in highly inaccurate estimates. The best way proved to be to take the estimates of the second order jackknife as the initial value of TS.









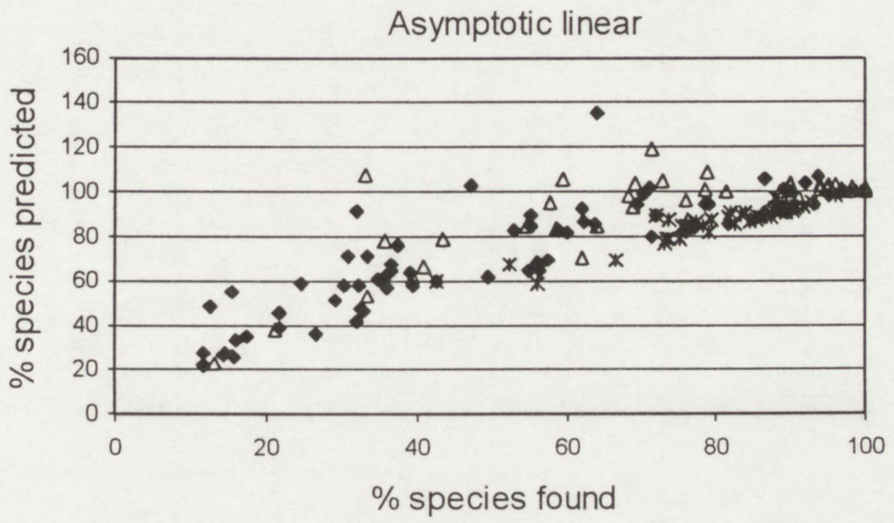


Fig. 5. Performance of asymptotic parametric estimators for estimating species numbers. Data points are the same as in Fig. 4.

Table 3 shows that (with the exception of the asymptotic power) these estimators depend on aggregation and SD. Especially the Michaelis-Menten formula (E_{MM}) is sensitive to both parameters, a fact that had already been noticed by Keating (1998). If more than 2/3 of TS were represented in the sample the first four methods gave worse results than the jackknifes or the Chao estimator (E_{Chao}) (Figs 4 and 5, Table 1). However, the asymptotic linear function (E_{AL}) performed nearly as well as E_{Chao}, E_{J1} and E_{J2} (the jackknife estimators), and this estimator has the advantage of being less biased. If fewer species are found (but more than 1/3 of TS) the estimates of the asymptotic linear were even slightly better than that of E_{J2} and than the other non-parametric estimators. The negative bias equals E_{J2}. However, the last column of Table 1 shows that all estimators even failed the TS ± 20% criterion.

In a recent paper Edwards (1997) developed an estimator based on an equilibrium of immigrations and extinctions. With immigration rates of $dS_i/dt = k_i (TS–S_l) \times$ area and extinction rates of $–dS_e/dt = k_e S_l$ he got after simple rearrangement the

Reciprocal linear:

$$1/ S_l = [k_e/(k_i TS)] \quad 1/area + 1/TS \quad (11)$$

with S_i, S_e: number of species which immigrate and get extinct, S_l: number of species in a given area, k_i, k_e: immigration and extinction constants.

The same function (under the name reciprocal linear) was independently used by Winklehner *et al.* (1997) to estimate species numbers of Collembola. However, – leaving the theoretical justification of an equilibrium of immigrations and extinctions aside – this formula appears to be nothing more than the well known Lineweaver-Burke plot of the Michaelis-Menten model (Morris 1976, Palmer 1990, Keating 1998) and it suffers the same drawback, the high variance of the intercept. The accuracy of the method is not better than the Michaelis-Menten method (but the derivation may serve as a theoretical justification of the latter).

Uncorrected non-asymptotic parametric estimators (Type 2)

Again, an infinite number of functions may serve for estimation. In the ecological literature only two types had been used, a double log function and an log-linear one (Palmer 1990). Fig. 6 gives the performance of the double log and two forms of log-linear models:

$$\text{LOGLOG:} \quad E_P = a(1/min)^z \quad (12)$$

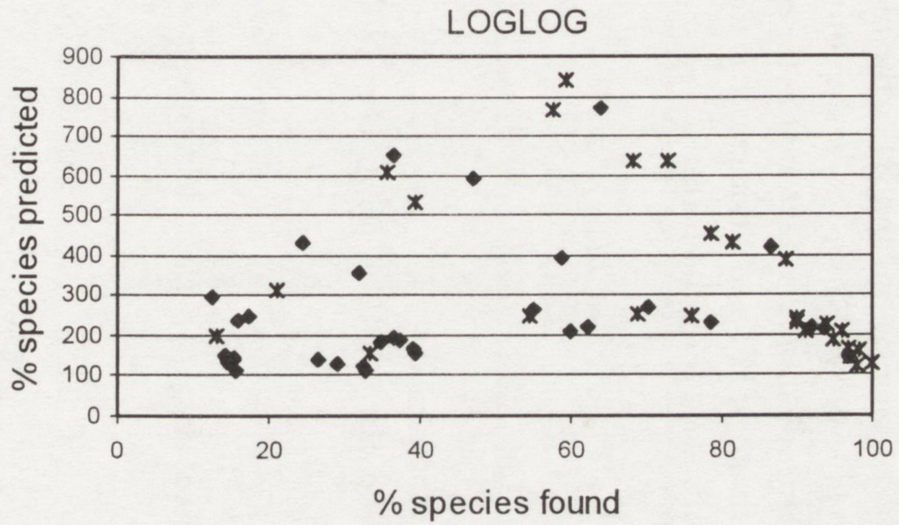
$$\text{LOGLIN1:} \quad E_{L1} = a[\ln(1/min)]^z + b \quad (13)$$

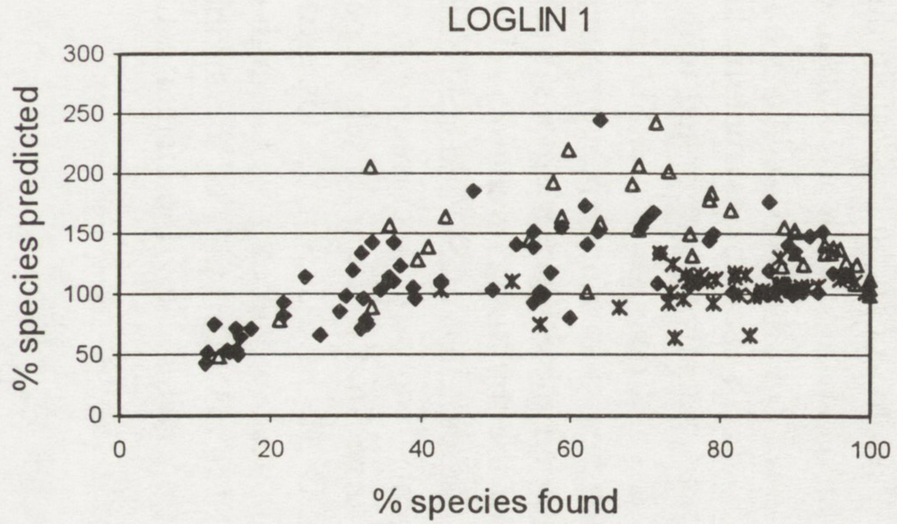
$$\text{LOGLIN2:} \quad E_{L2} = a[\ln(1/min)]^z \quad (14)$$

where min denotes the minimum allowed mean density, in our case 0.001 ind./cell. a, b and z are constants derived from the fitting process.

The first two estimators are highly dependent on SD and aggregation, their distributions are strongly skewed and they have a high variance. They are not suited to serve as estimators. The E_{L2} estimator performed better, but not as well as E_{J2} and the asymptotic linear. This finding contradicts to a certain extent the claim of Palmer (1990) who found that estimator to perform only slightly worse than E_{J1} and E_{J2}. The difference may stem from the smaller community size used in Palmer's study.

A drawback of this and some following estimators is the fact that they depend on the minimum population density (using area instead of min always results in too high estimates and is a rather artificial measure). The minimum density has at least roughly to be known. On the other





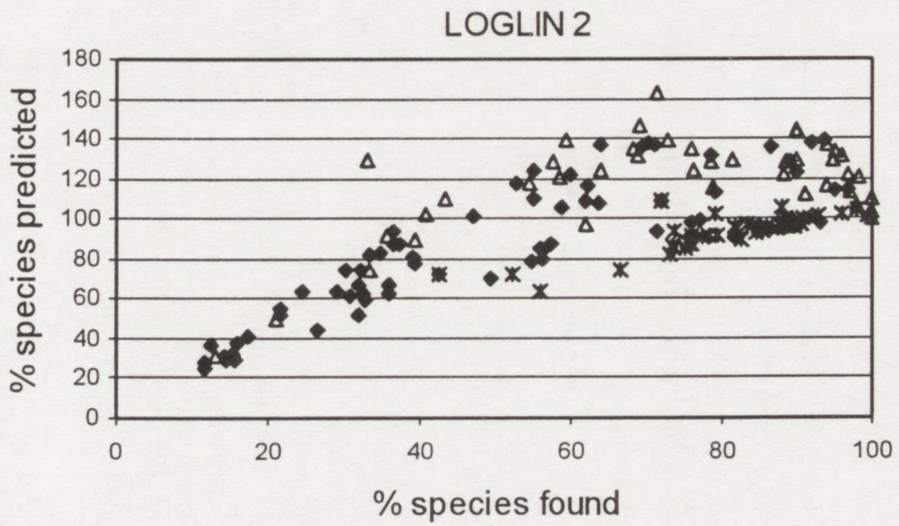


Fig. 6. Performance of infinite parametric estimators for estimating species numbers. Data points are the same as in Fig. 4. Note that for the power function not all data points are shown.

hand, it allows one to select which species have to be included in the computation. However, the behavior of this estimator under varying minimum densities gives room for further study.

Uncorrected asymptotic parametric estimators of type 3

This estimation technique is seldom used (e.g. Hilpert 1989) and requires more computational effort. I tested a double log and a log-linear function for predicting:

$$\text{LOGLOG:} \quad S(n) = an^z \qquad (15)$$

$$\text{LOGLIN:} \quad S(n) = a\ln(n) + b \qquad (16)$$

where S(n) is the number of new species found in sample number and n, a, b, and z are constants derived from the fitting process.

Because species do not have infinitely low densities it is impossible to take the integral from 0.5 to ∞ as an estimate (as did Hilpert 1989). However, one has to compute the values numerically. For S(n) – values less than 1 one also has to multiply the values with the probability of finding a species (it is not possible to find 0.3 species). This probability p is exactly S(n). For S(n) > 1 p is 1. Therefore, the estimator takes the form:

$$E_{Pnew} = \sum_{n=1}^{1/\min} S(n)p \qquad (17)$$

$$E_{Lnew} = \sum_{n=1}^{1/\min} S(n)p \qquad (18)$$

where *min* again denotes the minimal allowed density.

Fig. 7 and Table 1 show that both estimators did not give satisfactory results. More than 50% of all estimates ranged outside TS ± 20% and E_{Pnew} was strongly dependent on SD and aggregation.

Corrected estimation methods

The dependence of some of the methods on the community structure (Table 3) gives the opportunity to introduce structural parameters as correctors into the estimator. E_{L1}, E_{L2} (formulas 13 and 14) and E_{Pnew} (18) and the asymptotic linear (10) were negatively correlated with aggregation and positively with SD. Because the aggregation cannot be computed exactly for all species (see Methods) I used results of the multiple correlation between E/TS (dependent) and SD and FS/TS (independent) to construct a correction factor. E/TS turned out to be inversely dependent on SD. High values of SD lowered the estimate, low values gave too high estimates (Table 3). This fact led to a simple correction factor using the median value of SD (nearly 2):

$$\text{Corr. } E_{AL} = E_{AL} + [E_{AL}\ (\log_2(SD) - 1)/SD] \qquad (19)$$

For E_{L1}, E_{L2} and E_{Pnew} the median SD value proved to be the best quotient. After introducing the factor 0.8 to adjust the mean of E/TS to values near 1, the new estimators took the form:

$$\text{Corr. } E_{L1} = 0.8\ [E_{L1} + E_{L1}\ (\log_2(SD) - 1)/2] \qquad (20)$$

$$\text{Corr. } E_{L2} = 0.8\ [E_{L2} + E_{L2}\ (\log_2(SD) - 1)/2] \qquad (21)$$

$$\text{Corr. } E_{Pnew} = 0.8\ [E_{Pnew} + E_{Pnew}\ (\log_2(SD) - 1)/2] \qquad (22)$$

The low variance of the Bootstrap (4) and the negative exponential (6) also allowed correctors to be constructed. The plots in Figs 4 and 5 can be described by power functions of the form $y = ax^z$. Therefore

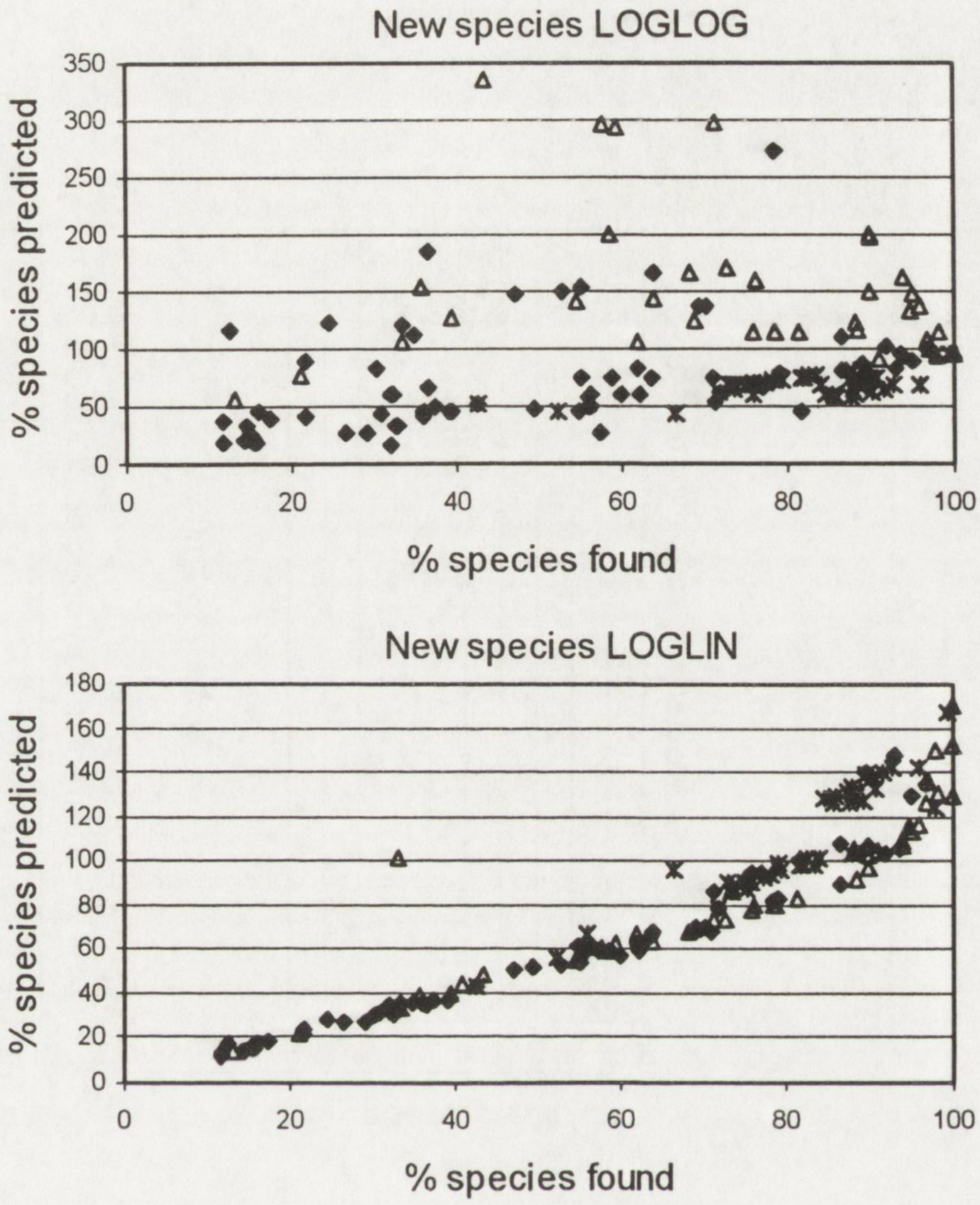


Fig. 7. Performance of type 3 parametric estimators for estimating species numbers. Data points are the same as in Fig. 4.

$$E/TS = a(FS/TS)^z. \quad (23)$$

This leads after simple rearrangement to the corrected form:

$$TS = (E/a)^{1/(1-z)} FS^{(1-z)/z} \quad (24)$$

where a and z are constants derived from the fitting process, E denotes the estimate of the Bootstrap and the negative exponential method. Because both new estimators are dependent on SD the introduction of the above corrector gives the new estimators

$$\text{Corr. } E_{Boot} = (E_{Boot}/1.05)^{8.45}\ FS^{0.134}\ (\log_2(SD) - 1)/2 \quad (25)$$

$$\text{Corr. } E_{NE} = (E_{NE}/1.03)^{6.39}\ FS^{0.185}\ (\log_2(SD) - 1)/2 \quad (26)$$

The constants are taken from Figs 4 and 5.

Both, E_{Boot} and E_{NE} were also tested by fitting a second order polynomial instead of a power function. The results, however, were worse than the above corrections and the data are therefore not shown.

For the non-parametric estimators the small sample bias correction factor of Hurvich and Tsai (1989) was used:

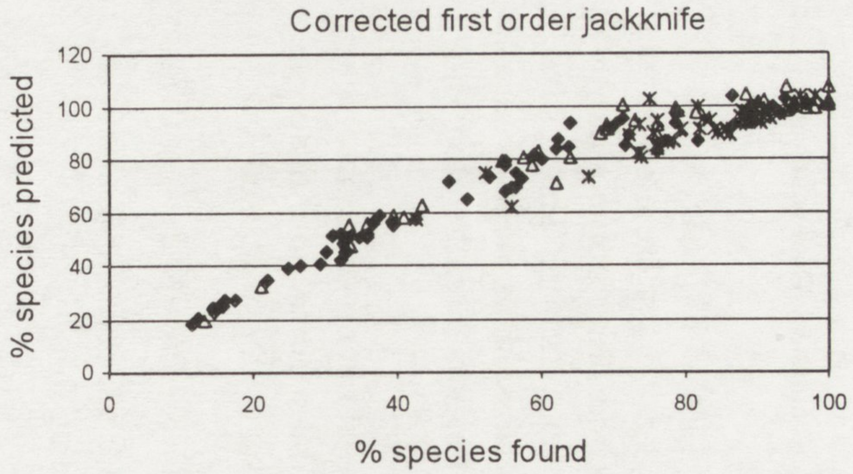
$$\text{Corr.} = 2(K+1)(K+2)/(n-K-2) \quad (27)$$

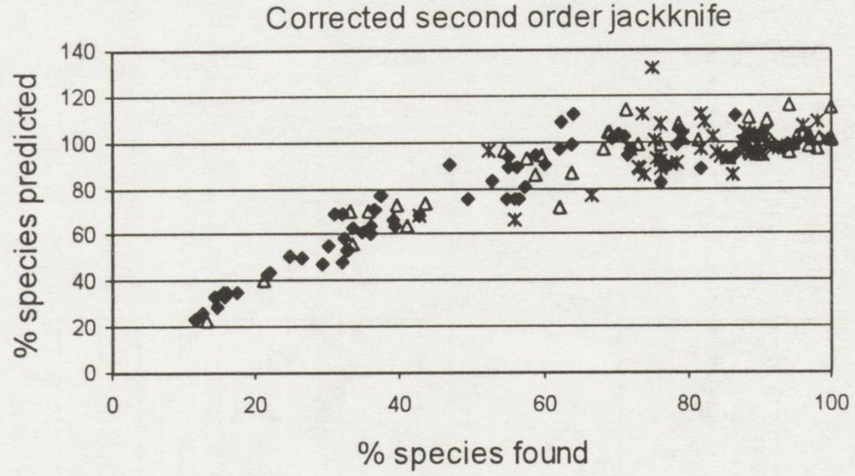
for E_{J1} and E_{J2} K becomes 2 and 3, respectively, leading to

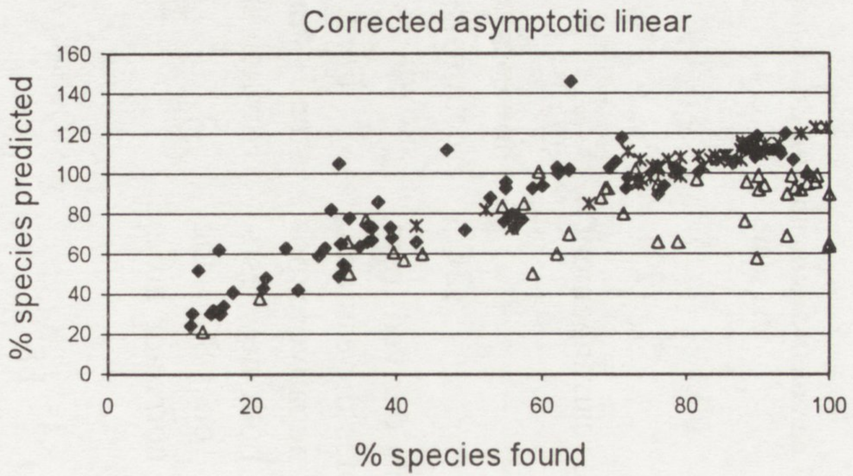
$$\text{Corr. } E_{J1} = E_{J1} + 24/(n-4) \text{ and} \quad (28)$$

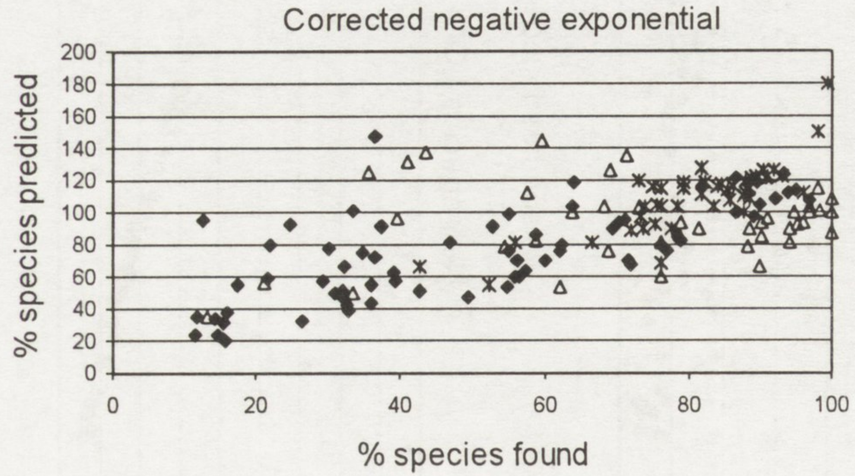
$$\text{Corr. } E_{J2} = E_{J2} + 40/(n-5). \quad (29)$$

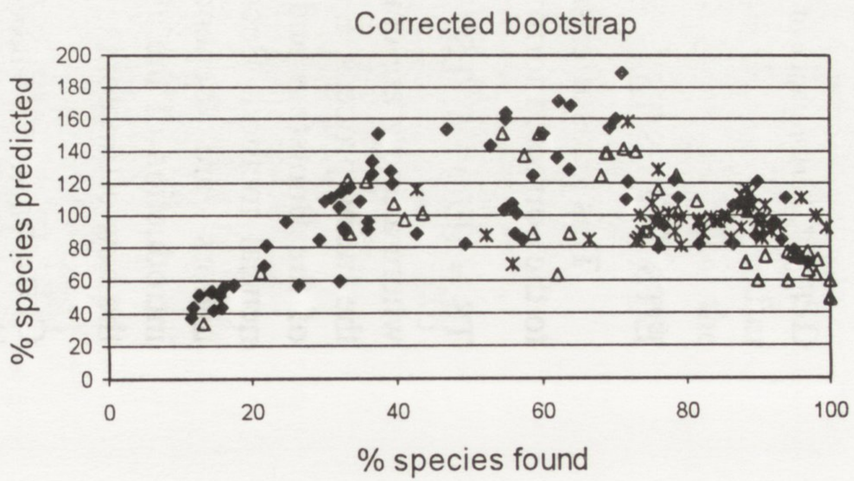
where N is the number of samples taken. Fig. 8 and Table 2 show the performance of these corrected estimators. Given are

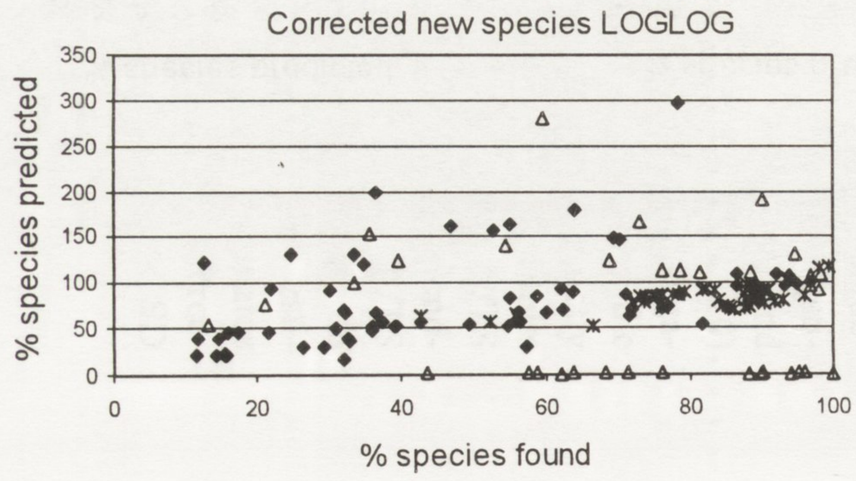

Corrected first order jackknife
% species predicted
% species found
0
20
40
60
80
100
120

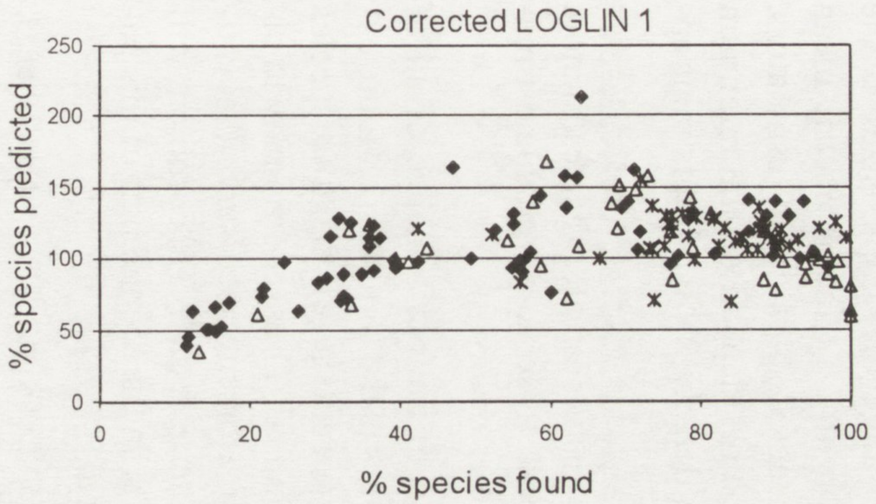

Corrected second order jackknife
% species predicted
% species found
0
20
40
60
80
100
120
140

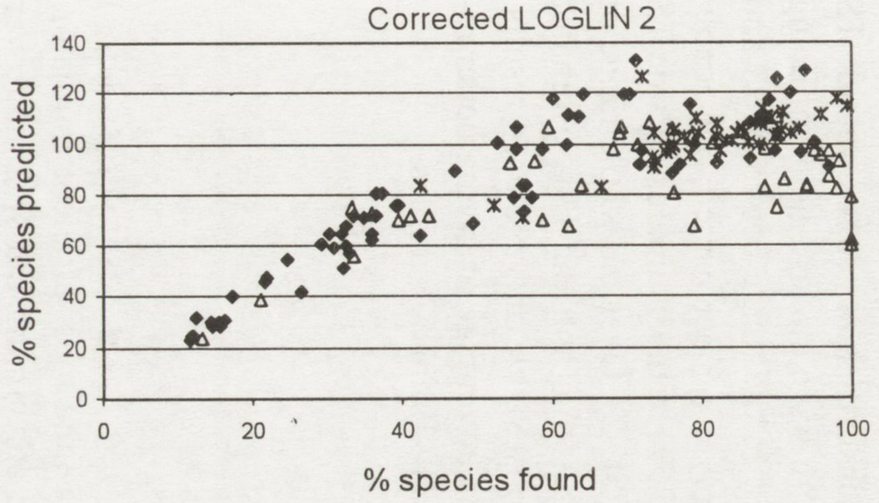

Corrected asymptotic linear
% species predicted
% species found
0
20
40
60
80
100
120
140
160

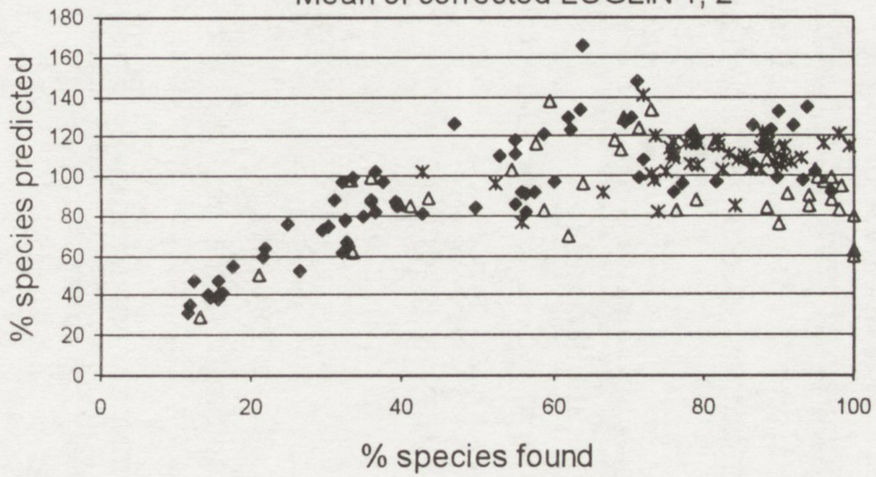

Corrected negative exponential
% species predicted
% species found
0
20
40
60
80
100
120
140
160
180
200


Corrected bootstrap
% species predicted
% species found
0
20
40
60
80
100
120
140
160
180
200


Corrected new species LOGLOG
% species predicted
% species found
0
50
100
150
200
250
300
350







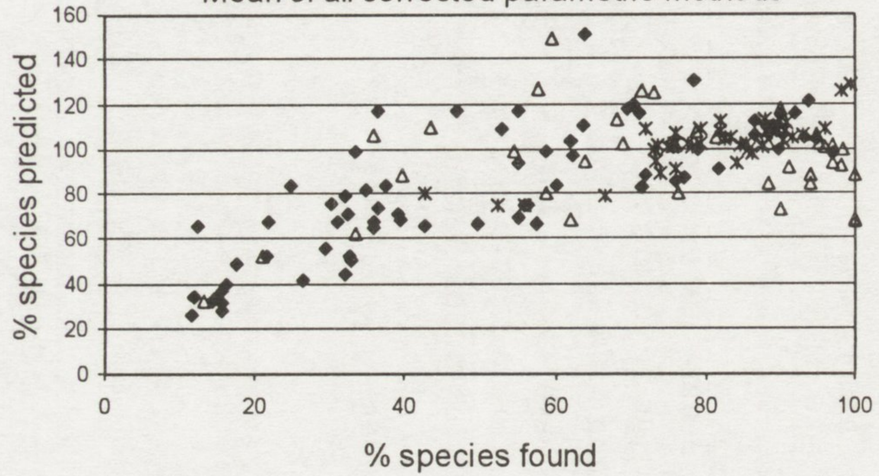


Fig. 8. Performance of corrected estimators for estimating species numbers. Data points are the same as in Fig. 4. Correction factors are given in the text.

also the mean of Corr. E_{L1} and Corr. E_{L2} (denoted as E_{ML}) and the geometric mean of all corrected parametric estimators (denoted as E_{GM}). E_{ML} is introduced because Corr. E_{L1} and Corr. E_{L2} have opposite biases. The mean levels this out (Table 2).

Fig. 8 and Table 2 show that all corrections enhanced the quality of the estimators. Corr. E_{Boot} and Corr. E_{L1} gave roughly correct estimates even if only 20% of TS was found. However, both tend to overestimate TS in the middle range. The corrected second order jackknife is still the best estimator if already 2/3 TS is represented in the sample. Only 21% of the estimates ranged outside TS ± 10%. But otherwise E_{ML} and E_{GM} performed as well under the TS ± 20% criterion. Around 25% of their estimates ranged outside TS ± 20%. Both have the advantage of being less biased. In the range between 1/3 and 2/3 TS only E_{ML} performed reasonably well with 26% of the estimates outside TS ± 20% (Table 2). Over the whole range of sample sizes this and the E_{GM} estimator gave the most stable and unbiased results. E_{GM} has the advantage of giving better results in more complete samples.

4. CONCLUSIONS

There is no simple solution to the problem of estimating total species numbers from a series of samples and, as already mentioned, it is a matter of choice or of philosophy, what degree of performance is good enough. However, from the present study some general remarks can be made.

– For any estimator to work at least 1/3 of the total species number has to be represented in the sample.

– If more than 1/5 of the species have been found the corrected Bootstrap or the corrected E_{L1} may serve as a first guess, but not more.

– In a fairly complete sample the corrected second order jackknife is the best estimator, although negatively biased. If such a bias is not desired the uncorrected asymptotic linear model (E_{AL}) may serve as an alternative.

– If the sample is not so complete (30 to 70% TS sampled) the mean of the two corrected log-linear estimators (E_{ML}) proved to be the most efficient estimator. More time consuming but spanning over the whole range of sample sizes is the computation of E_{GM}, the geometric mean of all corrected parametric estimators. Both estimators depend on the minimum population density, which limits their application.

– The results of all other estimators were too poor to serve as reliable estimators of species diversity.

Crucial to the performance of all of the estimators is the sample size, more exactly the percentage of the true species number sampled. Thus, additionally to diversity estimators are necessary other estimators for sample size. This problem will be dealt with in the second part of this paper (Ulrich 1999a).

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

A computer program was constructed which simulates large species assemblages with various species rank order distributions and degrees of aggregation of the species (Figs. 1, 2 and 3). From these model populations quantitative samples were taken to study the performance of 14 estimators of species diversity (Figs 4, 5, 6, 7 and 8). Most estimators proved to be sensitive to species distribution (Table 3). For 6 of the estimators correction factors are developed.

In sufficiently large samples (more than 2/3 of the true species number sampled) a corrected second order jackknife estimator gave the best results (Tables 1 and 2). If fewer species are represented in the sample two newly developed data analytical estimators performed better (Fig. 7).

Crucial to the performance of all of the estimators is the sample size. The minimum sample size has to contain at least 1/3 of the total species number.

6. REFERENCES

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

Boulinier Th., Nichols J. D., Sauer J. R., Hines J. E., Pollock K. H. 1998 – Estimating species richness: the importance of heterogeneity in species detectability – Ecology, 79: 1018–1028.

Brainerd B. 1972 – On the relation between types and tokens in literary text – J. Appl. Prob. 9: 507–518.

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

Burnham K. P., Overton W. S. 1978 – Estimation of the size of a closed population when capture probabilities vary among animals – Biometrika, 65: 623–633.

Burnham K. P., Overton W. S. 1979 – Robust estimation of population size when capture probabilities vary among animals – Ecology, 60: 927–936.

Chao A. 1984 – Non-parametric estimation of the number of classes in a population – Scand. J. Stat. 11: 265–270.

Chao L. 1987 – Estimating the population size for capture – recapture data with unequal catchability – Biometrics, 43: 783–791.

Chao A., Lee S. M. 1992 – Estimating the number of classes via sample coverage – J. Am. Sta. Assoc. 87: 210–217.

Chao A., Lee S. M., Jeng S. L. 1992 – Estimation of population size for capture-recapture data when capture probabilities vary by time and individual animal – Biometrics, 48: 201–216.

Coddington J. A., Young L. H., Coyle F. A. 1996 – Estimating spider species richness in a Southern Appalachian cove hardwood forest – J. Arachnology, 24: 111–128.

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

Currie D. J. 1993 – What shape is the relationship between body mass and population density – Oikos, 66: 353.

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

Edwards L. E. 1997 – A useful procedure for estimating the species richness of spiders – J. Arachnology, 25: 99–105.

Gaston K. J. 1993 – Comparing animals and automobiles: a vehicle for understanding body size and abundance relationships in species assemblages? – Oikos, 66: 172–179.

Heltshe J., Forrester N. E. 1983 – Estimating species richness using the jackknife procedure – Biometrics 39: 1–11.

Hilpert H. 1989 – Zur Hautflüglerfauna eines südbadischen Eichen-Hainbuchenmischwaldes – Spixiana, 12: 57–90.

Hodkinson I. D., Hodkinson E. 1993 – Pondering the imponderable: a probability based approach to estimating insect diversity from repeat faunal samples – Ecol. Entomol. 18: 91–92.

Hughes R. G. 1986 – Theories and models of species abundance – Am. Nat. 128: 879–899.

Hurvich C. M., Tsai C. 1989 – Regression and time series model selection in small samples – Biometrika, 76: 297–307.

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

Kobayashi S., Kimura K. 1994 – The number of species occurring in a sample of a biotic community and its connections with species--abundance relationships and spatial distribution – Ecol. Res. 9: 281–294.

Lauga J., Joachim J. 1987 – L'échantillonnage des populations d'oiseaux par la méthode des E.F.P.: interét d'une étude mathémathique de la courbe de richesse cumulée – Acta Oecol. Oecol. Gen. 8: 117–124.

Lawton J. H. 1990 – Species richness and population dynamics of animal assemblages. Patterns in body-size: abundance space – Phil. Tans. R. Soc. Lond. B 330: 283–291.

Lee S. M., Chao A. 1994 – Estimating population size via sample coverage for closed capture-recapture models – Biometrics 50: 88–97.

Longino J. T., Colwell R. K. 1997 – Biodiversity assessment using structured inventory: capturing the ant fauna of a tropical rain forest – Ecol. Appl. 7: 1263–1277.

McArdle B. H., Gaston K. J., Lawton J. H. 1990 – Variation in the size of animal populations: patterns, problems and artefacts – J. Anim. Ecol. 59: 439–454.

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.

Morris J. G. 1976 – Physikalische Chemie für Biologen – Weinheim, New York, 1–345.

Norris J. L., Pollock K. H. 1996 – Nonparametric MLE under two closed capture-recapture models with heterogeneity – Biometrics, 52: 639–649.

Palmer M. W. 1990 – The estimation of species richness by extrapolation – Ecology, 71: 1195–1198.

Palmer M. W. 1991 – The estimation of species richness: the second-order jackknife reconsidered – Ecology, 72: 1512–1513.

Pielou E. C. 1977 – Mathematical Ecology – New York, 1–385.

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

Schaefer M. 1991 – Fauna of the European temperate deciduous forest (In: Temperate deciduous forests, Eds. E. Röhrig, B. Ulrich) – Ecosystems of the world 7, Amsterdam, pp. 503–525.

Schaefer M. 1996 – Die Bodenfauna von Wäldern: Biodiversität in einem ökologischen System – Abh. Math.-Naturw. Klasse 1996,2, Stuttgart.

Scharf S. F., Juanes F., Sutherland M. 1998 – Inferring ecological relationships from the edges of scatter diagrams: comparisons of regression techniques – Ecology, 79: 448–460.

Slocomb J., Dickson K. L. 1978 – Estimating the total number of species in a biological community (In: Biological data in water pollution assessment: quantitative and statistical analyses, Eds.: K.L. Dickson, J. Cairns Jr., R.J. Livingston) – Philadelphia, pp. 38–52.

Slocomb J., Stauffer B., Dickson K. L. 1977 – On fitting the truncated lognormal distribution to species abundance data using maximum likelihood estimation – Ecology, 58: 693–696.

Smith E. P., van Belle G. 1984 – Nonparametric estimation of species richness – Biometrics, 40: 119–129.

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

Solow A. R. 1994 – On the Bayesian estimation of the number of species in a community – Ecology 75: 2139–2142.

Stout J., Vandermeer J. 1975 – Comparison of species richness for stream-inhabiting insects in tropical and midlatitude streams – Am. Nat. 109: 263–280.

Sugihara G. 1980 – Minimal community structure: an explanation of species abundance patterns – Am. Nat. 116: 770–787.

Tackaberry R., Brokaw N., Kellman M., Mallory E. 1997 – Estimating species richness in tropical forest: the missing species extrapolation technique – J. Trop. Ecol. 13: 449–458.

Tokeshi M. 1993 – Species, abundance patterns and community structure – Adv. Ecol. Res. 24: 111–186.

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

Ulrich W. 1998 – The parasitic Hymenoptera in a beech forest on limestone I: Species composition, species turnover, abundance and biomass – Pol. J. Ecol. 46: 261–289.

Ulrich W. 1999a – Estimating species numbers by extrapolation II: determining the minimum sample size to obtain a certain fraction of species in a community – Pol. J. Ecol. 47: 293–305.

Ulrich W. 1999b – The density – size and the biomass – weight distribution is generated by the species – size distribution together with

Density Fluctuations: Evidence from Model Species Distributions in the Hymenoptera – Pol. J. Ecol. 47: 87–101.

Ulrich W. 1999c – Abundance, biomass and density boundaries in the Hymenoptera: analysis of the abundance – body size relationship and differences between forest and open landscape habitats – Pol. J. Ecol. 47: 73–86.

Walther B. A., Morand S. 1998 – Comparative performance of species richness estimation methods – Parasitology, 116: 395–405.

Winklehner R., Winkler H., Kampichler C. 1997– Estimating local species richness of epigeic Collembola in temperate dry grassland – Pedobiologia, 41: 154–158.

(*Received after revising April 1999*)