POLISH JOURNAL OF ECOLOGY (Pol. J. Ecol.)	49	3	299–305	2001

short contribution

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: A CAUTIONARY NOTE

ABSTRACT: This paper evaluates the accuracy any estimator of species numbers may achieve if only a limited fraction (up to 3/4) of the species number in the community has been sampled. From the impossibility to infer the relative abundance distribution (RAD) the rare and not sampled species follow it is shown that it is only possible to give a lower and an upper boundary of the species number. The lower boundary may be inferred either from a fit of a log-normal type RAD or by a graphical method. In the latter case, the lower boundary is $S_{min} = (\ln (d_{min}) - 2\ icpt) / slope$ with d_{min} being the minimal possible relative density in the community and *icpt* and slope being the intercept and the slope of the geometric series fitted through the linear part of the log-normal distribution. The upper boundary is found through an extrapolation of this geometric series up to d_{min} $[S_{max} = (ln\ (d_{min}) - icpt) / slope]$. For any estimator to work d_{min} has to be known.

KEY WORDS: species numbers, extrapolation, species accumulation curves, relative abundance distribution, power fraction, jackknife estimators

The estimation of the number of species in a habitat from a series of samples is an important problem in community ecology and conservation. It takes therefore no wonder that a lot of different estimation methods have been developed and that they find a wide range of applications in studies of community ecology and in biodiversity assessment and conservation. These methods are based on four different kinds of reasoning.

The most often used approach is to extrapolate from species accumulation curves and to fit asymptotic or non-asymptotic curves to them. A variety of such curves has been used (Stout and Vandermeer 1975, de Caprariis *et al.* 1976, Lauga and Joachim 1987, Soberon and Llorente 1993, Edwards 1997, Winklehner *et al.* 1997, Keating 1998, Ulrich 1999a) but the reviews of Palmer (1990), Colwell and Coddington (1994) and Ulrich (1999a) showed that most of them have large and unknown error terms. If less than two thirds of the species had been sampled none of them gives reliable results. Ulrich (1999a) found the asymptotic linear model to be the best of these estimators.

A second approach uses plots of the number of new species found versus sample size. Examples may be found Hilpert (1989) and Ulrich (1999b). Ulrich (1999a) showed that this technique requires even larger sample sizes than the previous one to give fairly reliable results.

A third kind of methods is based on capture-recapture models and uses numbers of species found only once or several times in the samples (Bunge and Fitzpatrick 1993). Most often used are the estimators developed by Chao and coworkers (Chao 1984, 1987, Chao and Lee 1992, Chao *et al.* 1992, Lee and Chao 1994) and the jackknife estimators of Burnham and Overton (1978, 1979). Again, reliable estimates are only obtained with large sample sizes.

At least, the relative abundance distribution (RAD) of species has been used to infer species numbers. This is the least reliable technique because beside density estimates the underlying relative abundance distribution has to be known which only is very seldom the case (Miller and Wiegert 1989). For instance, in the case of the log-normal distribution Slocomb *et al.* (1977) and Slocomb and Dickson (1978) reported 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) found it not to be better than the simple measure of the number of species detected.

All of the above methods are best applicable in small communities where most of the species can be sampled with moderate sampling effort. However, in biodiversity studies it is very often necessary to estimate species numbers of whole habitats or even regions. In this cases we have to deal with hundreds or even several thousands of species. Sampling two thirds of them would require extraordinary sampling efforts and will most often be impossible. Therefore, the question arises whether it is possible to estimate species numbers even from a more limited sample size and if the available estimators may be used for such a task.

The present paper is intended as a cautionary note. It directs attention to the fact that even from fairly complete samples we may not infer the type of underlying relative abundance distribution and whether this distribution is homogeneous. All available estimators, even the non-parametric ones, but rely on the quiet assumption of homogeneity and all published tests of them used such distributions (Palmer 1990, 1991, Colwell and Coddington 1994, Mingoti and Meeden 1992, Tackaberry *et al.* 1997, Keating 1998, Walther and Morand 1998, Ulrich 1999a). These tests may therefore underestimate the true estimation error.

Whether natural communities follow homogeneous distributions (that is whether they may be described by a single model) or whether they have to be called composite (Tokeshi 1993) is even after 40 years of research in relative abundance distributions and the publication of more than 20 different models still unknown. The large compilation of Hughes (1986), even if it contains mostly incomplete samplings, does not point to this direction. Hughes probably correctly inferred that most natural communities follow in their upper part a log-normal type but in their lower part a geometric (or log-series) type distribution. If this is true existing estimators may give a wrong impression about the species number to be estimated. However, it will be shown that it is possible to infer upper and lower boundaries of species numbers even from incomplete samplings. Because species numbers in real habitats are not fixed but change in time (sometimes considerably) such boundaries may often be more useful than simple estimates of species numbers.

If we sample a natural community we will in general find the more abundant species. To estimate species numbers we have to extrapolate into the realm of the more rare species. This is frequently done by species accumulation curves. More instructive is to use a plot of the relative densities versus the species rank order, the relative abundance distribution (RAD). This is shown in Fig. 1. This figure shows a hypothetical community of which 50 species have been sampled. In theory, there are now three types of extrapolation possible until we reach a minimum possible relative density (d_{min}). This minimum relative density depends on two values, the maximum density one of the species can achieve and, more important, the minimum possible density. The latter is determined either by the type of habitat, the area under consideration, or by the ecology of the species that does not allow for lower densities because at lower densities the species will go extinct. The concept of d_{min} is important because even for communities following a sigmoid log-normal type distribution d_{min} has to be known. This because we do not know the exact shape of the distribution of the very rare species and (more important) the distribution may be truncated (in Fig. 1 indicated by the broken line). The latter case will occur in communities in which the lowest densities predicted by the RAD are impossible because below a certain threshold species go locally extinct.

If we do not know d_{min} or the RAD the rare species follow any parametric estimator that relies on species accumulation curves does not work. But even non-parametric estimators that do not rely on a certain type of relative abundance distribution assume a homogeneous RAD (Bunge and Fitzpatrick 1993). If this prerequisite is violated an overproportional number of very rare

species or species not as rare as predicted will introduce a systematic error into the estimates.

In principle, species rank order plots may follow three different kinds of relative abundance distributions (or combinations of them) (Ulrich 2001): a Zipf-Mandelbrot type, a geometric type, or a log-normal type distribution (Fig. 1). If the community follows a Zipf-Mandelbrot distribution (D in Fig. 1) it may have a very large number of species and extraordinary high sample sizes would be necessary to estimate the number within a reasonable error range. Because a Zipf-Mandelbrot distribution has the general form (with T and k being the shape parameters and d_S the relative density of S^{th} species):

$$d_S = (S+T)^{-k} \quad (1)$$

the species number S_{max} at d_{min} is:

$$S_{max} = d_{min}^{-k} - T \quad (2)$$

Fortunately, such distributions seem to be very rare in nature. Frontier (1985) fitted them to marine species assemblages. But this was only possible when leaving out the rare species, a method of questionable value. Moulliot *et al.* (2000) fitted the model (under the name fractal model) to some species poor and incomplete samples of forest Diptera where again the very rare species were missing. In a previous study (Ulrich 2000) I could show that the community of parasitoids of necrophagous parasitoids of a beech forest on limestone follows a Zipf-Mandelbrot distribution. But this community consists of only 30 species. Existing compilations and reviews of relative abundance distributions (Hughes 1986, Tokeshi 1993) do not point to this type especially for species rich communities.

Very often the community follows in their lower part a geometric series (or log-series). In this case the species number is given by the solution of the geometric series $d_S = e^{icpt + slope\ S}$ (the straight line in Fig. 1):

$$S_{max} = \frac{\ln(d_{min}) - icpt}{slope} \quad (3)$$

Here, *icpt* is the intercept of the regression line fitted to the geometric series and slope their slope. Of course, both, the intercept and the slope have error terms that have to be measured from the sample. This may be done by a bootstrap technique. Fortunately, for large communities well above 100 species

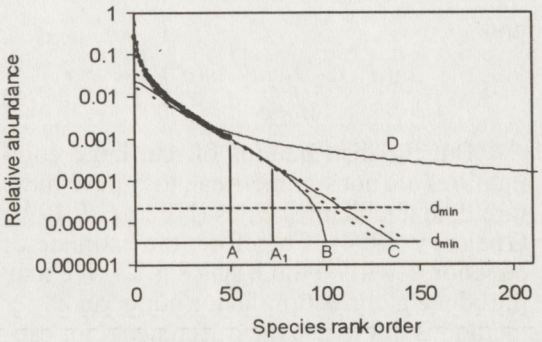

Fig. 1. Estimation of upper and lower limits of species numbers from a sample (A) of 50 species. B: species number if the community follows a Sugihara fraction relative abundance distribution, C: for a geometric series, and D: for a Zipf-Mandelbrot distribution. A_1 denotes the point of inflection of the Sugihara fraction distribution. d_{min} is the minimum possible relative abundance

these error terms will be small which will be shown below.

If the relative abundance distribution of the community has a lower curvature and follows a log-normal or Sugihara type distribution (Sugihara 1980, Ulrich 2001) it is more difficult to estimate S_{max} because we do not know the exact type of distribution. We may try to fit existing models to the sample. Such a procedure however would require the testing a whole range of species numbers (necessary for computing a distribution) and several models until a best fit is reached. A more simple graphical method is shown in Fig. 2. If we assume that the curve is symmetrical (the case of a log-normal distribution) the difference d between *icpt* and the relative abundance of the most abundant species d_1 is equal to the difference between a value X and d_{min}. Therefore

$$X = ln(d_{min}) - d = icpt + slope\ S_{max} \quad (4)$$

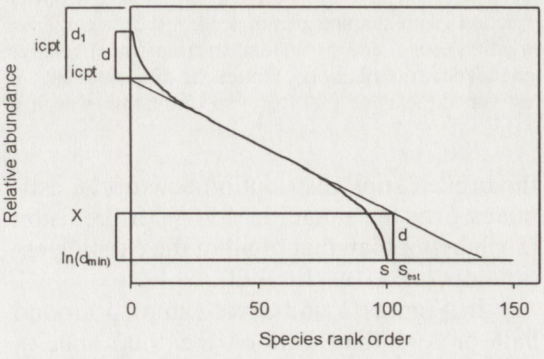

Fig. 2. Graphical estimation of species numbers of a community following a log-normal type relative abundance distribution. Further explanations in the text

and

$$S_{max} = \frac{\ln(d_{min}) - 2icpt - \ln(d_1)}{slope} \quad (5)$$

But the distributions of real large communities are not symmetrical, they have more rare than abundant species (Nee *et al.* 1991, Gregory 1994). Therefore, the estimate of equation 5 will be too high (Fig. 2). We may introduce a correction factor butter an easier method is not to use the difference d for estimating X but the intercept. The estimate simplifies then to:

$$S_{max} = \frac{\ln(d_{min}) - 2icpt}{slope} \quad (6)$$

In Fig. 3 I tested this estimate using four different relative abundance distributions, a random fraction, a Sugihara fraction, a power fraction, and a broken stick (see Tokeshi 1993, 1996 and Ulrich 2001 for definitions of these distributions and computing algorithms). Plotted are the products of estimated and real species numbers against the real species number. Figure 3 shows that at this sample size (50% of the true species number) for sigmoid log-normal type RADs except for

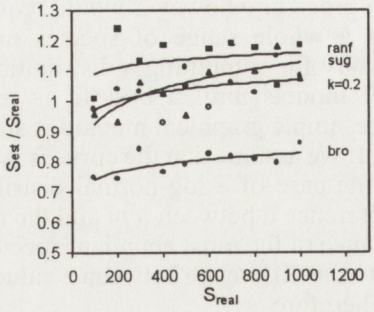

Fig. 3. Test of the goodness of estimate of equation 6. The test uses 10 communities each (with 100 to 1000 species) of each a broken stick (bro), a power fraction (with shaping parameter k = 0.2), a Sugihara fraction (sug), and a random fraction (ranf) relative abundance distribution. Plotted is the quotient of estimated (S_{est}) and real (S_{real}) species number against S_{real}

the broken stick distribution equation 6 estimates S in the range of ± 20%, a precision much better than that of all of the existing estimators (cf. Ulrich 1999).

In Figures 1 and 3 we sampled around half or less than half of the total species number. In this case the equations 3 and 6 mark the upper and lower range of possible estimates (except in some rare cases of very unusual or highly truncated relative abundance distributions). For the lower boundary we safely may take the estimate of equation 6 – 20%. For the upper limit the estimate has to include the error term (estimate + two standard deviations). These error terms have to be estimated from the sample.

In principle no estimator can do better than giving estimates inside these limits and for every estimate we have to know d_{min}. This result even holds if we enlarge the sample size up to point A_1 in Fig. 1, the point of inflection (if the rare species follow a log-normal type distribution), where around 3/4 of the species are sampled.

For large communities sampling up to point A_1 requires extraordinary high sample sizes and will often be impossible. For instance, for a community of 100 species following a log-normal distribution sampling up to A_1 requires a sample size of at least 3 000 individuals. However, in the more interesting case of a community of 500 species even more than 500 000 individuals have to be sampled. At all lower sample sizes no better estimates of species numbers will be possible than that defined above through equations 3 and 6.

From equations 3 and 6 we see that the quotient of upper and lower limit is exactly

$$\frac{S_{max}}{S_{min}} = \frac{\ln(d_{min}) - icpt}{\ln(d_{min}) - 2icpt} \quad (7)$$

The quotient is therefore independent of the slope. How large is this range defined by equation 7. To answer this question I computed 10 assemblage of 100 to 1000 species each following a broken stick, a power fraction, a Sugihara fraction, and a random fraction relative abundance distribution. Figure 4 shows the estimation ranges computed from random samples containing in every case 2/3 of the total species number. We see that except for the broken stick distribution the upper boundary not reaches above 2.5 times the lower one.

Figure 5 shows the sampling of parasitic Hymenoptera of a dry meadow on limestone (Ulrich 1999b). In 1986 254 species were collected with emergence traps and the plot of species densities against species rank gives the relative abundance distribution. An estimation of species numbers using a log-linear model (Ulrich 1999a) resulted in 629 species, the extrapolation from a plot of new species versus sample even gives more than 1100 species (Ulrich 1999b). The second order jackknife predicts 338.

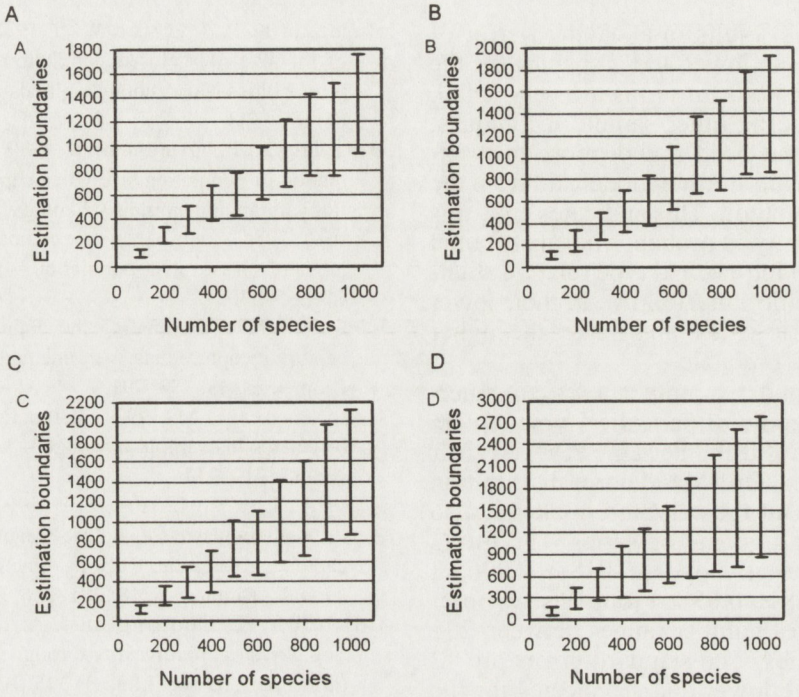

Fig. 4. Boundaries of species numbers computed with equations 3 and 6 for the communities of Fig. 3. For the computation of slope and *icpt* of the equations the middle ranking 50% of species were used. A: Random fraction, B: Sugihara fraction, C: Power fraction, D: Broken stick

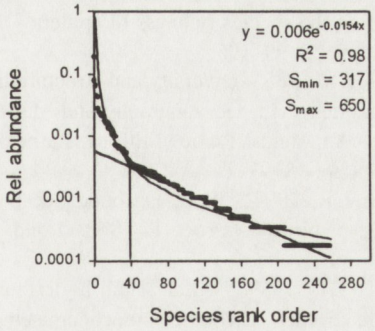

Fig. 5. Relative abundance distribution of the Hymenoptera of a dry meadow on limestone sampled in 1986 (data from Ulrich 1999b). Given are the equation of the geometric series fitted, the variance explanation and the estimates of upper and lower species numbers. The vertical line marks the species number from which on the geometric series was fitted. The black curve shows a fit of a Zipf-Mandelbrot model

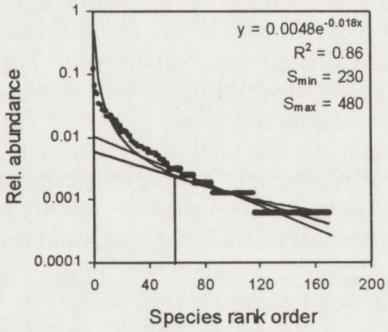

Fig. 6. Relative abundance distribution of the Hymenoptera of a beech forest on limestone sampled in 1986 (data from Ulrich 1998). Given are the equation of the geometric series fitted (the regression lines mark their 95% confidence limits), the variance explanation and the estimates of upper and lower species numbers. The vertical line marks the species number from which on the geometric series was fitted. The black curve shows a fit of a Zipf-Mandelbrot model

The first two estimates mean that less than half of the species had been sampled with the traps and the variance in estimation is very high. The species-rank order distribution is not fitted by a Zipf-Mandelbrot model, which is shown by the upper regression line. Computing the upper and lower boundaries with equations 3 and 6 results in a species number between 317 and 650 species, a range far more reliable than the three estimates. From the regression module of the STATISTICA software we also get an estimator of the error terms of slope and intercept (slope: –0.0155 ± 0.00015; intercept: –5.127

± 0.019). This results in a standard error of 3 species for the lower and 7 species for the upper limit. Both error terms are very small. In total, including other sample techniques, 343 species had been found, more than expected from a log-normal model and from the jackknife estimator. This indicates that this large community is probably not distributed according to a log-normal type relative abundance distribution but follows in their lower part a geometric series. This makes the higher estimate more probable.

The second example shows an emergence trap sample of parasitic Hymenoptera in a beech forest from the same year (Ulrich 1998). Again a Zipf-Mandelbrot distribution does not fit. But the curvature makes it also difficult to fit a geometric series. The fitting process results in: slope: –0.018 ± 0.00053; intercept: –4.8 ± 0.058. From this we infer that the species number ranges between 231 and 480 species. The standard errors are 12 species for the lower and 17 species for the upper limit. Again, the precision of the estimates is very good. In reality, with all sampling methods combined 313 species had been found (Ulrich 1998) and estimates using the second order jackknife and a log-linear estimator (Ulrich 1999c) resulting in 374 and 509 species, respectively. This again indicates a geometric series distribution of the least abundant species making the higher estimate more probable. In both cases non-parametric estimators give much too low estimates and are not applicable, mainly because less than half of the total species number had been sampled. We also see that the log-linear estimator of Ulrich (1999a) gives in both cases estimates near the upper limits. This estimator may therefore be applicable especially in cases were relative abundance distributions of large communities follow in their lower part a geometric series.

ACKNOWLEDGMENTS: I thank Prof. J. Buszko and Dr. Kartanas for critical and valuable suggestions on the manuscript. This work was in part supported by a grant from the Deutsche Forschungsgemeinschaft. The author received a scholarship from the Friedrich-Ebert-Foundation.

3. REFERENCES

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.

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. 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.

Frontier S. 1985 – Diversity and structure in aquatic ecosystems (In: Oceanography and Marine Biology – An Annual Review, Ed. M. Barnes) – Aberdeen, pp. 253–312.

Gregory R. 1994 – Species abundance patterns of British birds – Proc. R. Soc. Lond. B 257: 299–301.

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

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

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

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.

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

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.

Moulliot D., Lepretre A., Andrei-Ruiz M.-C., Viale D. 2000 – The fractal model: an new model to describe the species accumulation process and relative abundance distribution (RAD) – Oikos, 90: 333–342.

Nee S., Harvey P. H., May R. M. 1991 – Lifting the veil on abundance patterns – Proc. R. Soc. Lond. B 243: 161–163.

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.

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.

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

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 – Polish J. Ecol. 46: 261–289.

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

Ulrich W. 1999b – The Hymenoptera of a dry meadow on limestone: species composition, abundance and biomass – Polish J. Ecol. 46: 29–47.

Ulrich W. 1999c – Temporal stability of community structure of the parasitic Hymenoptera in a beech forest on limestone – Pol. J. Ecol. 47: 257–270.

Ulrich W. 2000 – Niche segregation and coexistence of parasitic Hymenoptera of the *Aspilota* genus group (Hymenoptera, Braconidae) in a beech forest on limestone – Pol. J. Ecol. 48: 225–238.

Ulrich W. 2001: Models of relative abundance distributions I: Model fitting by stochastic models – Pol. J. Ecol.: 49: 145–157.

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 December 2000)