



ELSEVIER

GfÖ

GfÖ Ecological Society of Germany,
Austria and Switzerland

Basic and Applied Ecology xxx (2012) xxx–xxx

Basic and
Applied Ecology

www.elsevier.com/locate/baae

Rarity in large data sets: Singletons, modal values and the location of the species abundance distribution

Gerben Straatsma^{a,*}, Simon Egli^b^aWageningen University & Research Center, Aquatic Ecology and Water Quality Management group, Wageningen, The Netherlands^bSwiss Federal Research Institute WSL, Forest Dynamics, Birmensdorf, Switzerland

Received 9 March 2011; accepted 12 March 2012

Abstract

Species abundance data in 12 large data sets, holding 10×10^3 to 125×10^6 individuals in 350 to 10×10^3 samples, were studied. Samples and subsets, for instance the summarized data of samples over years, and whole sets were analysed. Two methods of the binning of data, assigning abundance values to classes for histograms, have been applied in the past: bins of equal size and bins of exponentially increasing size ('octaves'). A hump in a histogram with exponential bins does not represent a mode of primary, non-transformed abundance values, but of log transformed abundance values. A proper interpretation of the hump is given. Moreover, the extrapolation to the left of a histogram with exponential bins, below an abundance of unity, lifting a 'veil', hiding species present in the community but absent from the sample, is rejected. The literature is confusing at these points and, as a result, prevents a proper view on the species abundance distribution. Applying bins of equal size, modal values equalled or approached unity. The number of singletons increased with sample size in some data sets but decreased in others. However, singletons remain present in large samples, subsets or sets, in agreement with the results on modal values. The relatively high number of singletons in small samples is no artefact of undersampling. The mode at unity, that is at the left end of the species abundance distribution, independent of scale (sample, subset or set), is an important statistical property of the species abundance distribution. Our results may have implications for theory development in community ecology: the selection and/or development of an accurate species abundance model, and, connected to this, the formulation of improved assembly rules, and the selection and/or development of more precise species richness estimators.

Zusammenfassung

Wir untersuchten die Abundanz von Arten in 12 großen Datensätzen, die 10×10^3 bis 125×10^6 Individuen in 350 bis 10×10^3 Proben enthielten. Die Einzelproben und Teildatensätze, z.B. über Jahre hinweg aufsummierte Einzelproben, und ganze Datensätze wurden analysiert. Zwei Methoden der Klassifizierung von Daten wurden bislang angewendet, um Abundanzwerte für Histogramme zu gruppieren: Klassen mit konstantem Umfang und Klassen mit exponentiell zunehmender Größe ("octaves"). Der Buckel in einem Histogramm von exponentiellen Klassen entspricht nicht dem Modalwert der primären, untransformierten Abundanzwerte, sondern der log-transformierten Abundanzwerte. Eine korrekte Interpretation des Buckels wird gegeben. Darüber hinaus wird die Extrapolation auf der linken Seite eines Histogramms mit exponentiellen Klassen auf Abundanzwerte kleiner als eins, womit der "Schleier", der Arten, die in der Gemeinschaft, nicht aber in der Probe enthalten sind, gelüftet würde, abgelehnt. Die Literatur ist an diesen Punkten nicht einheitlich und verhindert dadurch eine angemessene Betrachtung von Arten-Abundanz-Verteilungen. Wenn Klassen von gleichem Umfang gebildet wurden, betrug die Modalwerte eins oder nahezu eins. Die Anzahl der mit einem Individuum vertretenen Arten stieg mit der Probengröße bei manchen

*Corresponding author. Tel.: +31 77 398 5786.

E-mail address: gerben.straatsma@wur.nl (G. Straatsma).

Datensätzen, und fiel bei anderen. Dennoch bleiben Ein-Individuum-Arten in großen Proben präsent, was mit den Ergebnissen zu den Modalwerten übereinstimmt. Die relativ hohe Zahl von Ein-Individuum-Arten in kleinen Proben ist kein Artefakt unzureichender Beprobung. Der Modalwert von eins, d.h. am linken Ende der Arten-Abundanz-Verteilung, ist -unabhängig von der Skala (Einzelprobe, Teildatensatz, Datensatz) eine wichtige statistische Eigenschaft der Arten-Abundanz-Verteilung. Unsere Ergebnisse könnten Konsequenzen für die Theoriebildung in der Gemeinschaftsökologie haben, für die Auswahl und/oder Entwicklung eines genauen Arten-Abundanz-Modells und, damit verbunden, die Formulierung verbesserter Regeln zur Gemeinschaftsbildung sowie die Auswahl und/oder Entwicklung von genaueren Schätzverfahren für den Artenreichtum. © 2012 Gesellschaft für Ökologie. Published by Elsevier GmbH. All rights reserved.

Keywords: Artefact; Binning; Biodiversity; Community ecology; Mode; Species richness

Introduction

Rarity intrigues people, making them want to cherish and protect the rare organisms. In public opinion, proper ecosystem functioning depends on there being the full range of species, that 'stability' and species richness are 'good' and that rare species need protection. However, it is plausible that rare species take part in ecosystem functioning only relative to their rarity, and actually play a limited role, i.e. are 'redundant' (Gaston 1994).

Most communities have many rare species and only a few common ones (Magurran 2004; Gaston 1994; Gaston & Blackburn 2000). Reviewing species abundance distributions, McGill et al. (2007) wrote: When plotted as a histogram of number or percent of species on the y-axis vs. abundance on an arithmetic x-axis, the classic hyperbolic, 'lazy J-curve' or 'hollow curve' is produced, indicating a few very abundant species and many rare species." A histogram showing a hollow curve indicates that the lowest possible abundance, unity, is the most frequent one. The data show no central tendency and neither the (geometric) mean nor the median are suitable to estimate the location of the distribution.

A non-arithmetic x-axis is often used for species abundance distributions, where bin-sizes for abundance values increase exponentially with a factor of 2 ('octaves', Preston 1948), or 10. Such binning can result in hump-shaped histograms, where the hump indicates a modal value at some intermediate abundance value, and, additionally, may suggest a lognormal distribution (McGill et al. 2007). Although the abundance values are formally not transformed, any direct conclusions on the location and the shape of the distribution refer to log transformed values. Fitting lognormals to the mentioned hump-shaped distributions makes it tempting to extrapolate the number of species to abundance values smaller than unity. This possibility has been taken as an indication for the existence of species present in the community, but absent from the sample, as if they were 'veiled off' due to undersampling (Preston 1948, also discussed by McGill et al. 2007). Aspects of exponential bins, log transformation and the lognormal distribution for species abundance data, that can lead to misinterpretation and/or confusion, have been treated before (Dennis & Patil 1988; Blackwood 1992; Williamson & Gaston, 2005; Nekola, Sizling, Boyer, & Storch, 2008). The notion of 'missing species' and that

rarity is (partly) due to undersampling, survives (Magurran & Henderson 2003; McGill 2003; Coddington, Agnarsson, Miller, Kuntner, & Hormiga 2009).

The study of singletons (= species represented by a single individual only) is interesting if modal values equal unity. Then, the number of singleton species equals the number of species at the mode. Novotny and Basset (2000) studied quite a large data set and noted: "The number of . . . singletons in a sample . . . was increasing with the expansion of sampling . . . Increase in the number of singletons was slower than that in the total number of species . . . As a consequence, the proportion of singletons was decreasing . . . with increasing sample size. . .". They maintained that the presence of rare species is not a sampling artefact. Williamson and Gaston (2005) commented on the 'dominance of singletons' and found this generally to be true, but "that (it) is not true if the collections (data sets) are large enough". They suggest that "the dominance of singletons will be lost . . . between 100 000 and 1 000 000 individuals" (we quote their appealing terminology but add that they mean the dominance of singletons in histograms, not in communities).

The definition of a community is problematic. A community is inhomogeneous and cannot be delimited in units of space and/or time (Gleason 1926; Barkman 1989; Ricklefs 2008). Obviously, this problem affects the concept of a sample of a community. Our understanding is that a sample does not refer to randomly distributed individuals of different species but rather to the community, sampled at a particular place and time. In statistical terms, species are not 'missing', and small samples should not be ignored merely because of their size. In this study, we addressed 'rarity' by analysing modal values and numbers of singletons in several large data sets, one, managed by the second author, on mushrooms in a Swiss forest plot, and other data sets that were easily available, like some mentioned in Magurran et al. (2010, see also the supplementary material), and on a website managed by White (2011). Data sets were selected on availability and size, that they held multiple samples, but without prejudice to the outcome of any analysis. All sets provide opportunities to distinguish subsets of samples, for instance samples for specific years. Subsets or complete sets can be obtained by summing up the abundances of species present in the respective constituent samples. In the rest of this paper we will use the terms subset and set for short, for what actually are

Table 1. Characteristics of the data sets: total numbers of individuals, *n*, number of species, *S*, and numbers of samples and subsets.

Data set	<i>n</i>	<i>S</i>	Samples	Subsets
Mushrooms	108,014	408	3731	28
Fish	143,420	83	357	31
Crustaceans	1,045,302	16	348	30
Trees	206,513	295	512	32
Seedlings	289,010	254	915	24
Rodents	32,638	30	6261	26
Winter annuals	415,749	56	3781	13
Winter perennials	10,289	45	2848	14
Summer annuals	365,059	53	4565	14
Summer perennials	24,318	55	3751	14
Ants	30,360	42	8822	12
Birds	125,031,124	217	–	–

summarized subsets and summarized sets. Characteristics of the data sets are given in Table 1. Modal values and singletons were analysed at different scales, in samples, subsets and sets.

Very similar metrical patterns appear to exist in very different communities, like a common species abundance distribution (Magurran 2004; McGill et al. 2007), but until now it has not been clear whether this includes mushroom communities. We started with an interest in the species abundance distribution of mushrooms. Because of the confusion related to log transformation, binning rules and modes, we also considered modal values, number of singletons and rarity. The ultimate purpose of our study is to show the impact of rare species on the location of the species abundance distribution, at the left of the abundance axis (at unity), independent of the scale of the data considered.

Materials and methods

Twelve data sets were studied: a data set on mushrooms [1], the property of the Swiss Federal Research Institute WSL and managed by the second author. It contains data of a study in a Swiss forest. Five plots, each divided in three subplots were studied for 28 years on a weekly basis (Egli, Ayer, & Chatelain 1997; Straatsma, Ayer, & Egli 2001; Egli, Peter, Buser, Stahel, & Ayer 2006). Subsets (see below) consisted of samples for specific years. A data set on fish [2] and crustaceans [3], the property of Pisces Conservation Ltd and managed by Henderson (2011) (see also Henderson & Bird 2010). It contains ongoing catches since 1981 until present from monthly samplings at high tide at noon of sea water in the Bristol Channel, UK. As indicated, we split the set in two, one for fish and one for crustaceans. Subsets were made for years. A data set on tropical rain forest trees [4], from the Smithsonian Tropical Research Institute, Center for Tropical Forest Science, managed by Hubbell, Condit, and Foster (2005); (see also Condit et al., 1996). The data set contains information on all individuals with dbh > 10 mm (diameter at breast height), spatial coordinates included, of a

500 m × 1000 m plot on Barro Colorado Island in the Panama Canal. Using the spatial coordinates, the set can be split. We applied a grid of 512 squares of 31.25 m × 31.25 m and a grid of 32 squares of 125 m × 125 m in such a way that 16 small squares nested into one larger square, these larger squares to be used as subsets. A data set on weed seedlings [5], not protected by intellectual property rights, established during 'Farm Scale Evaluations' of four arable, conventional or genetically modified herbicide-tolerant crops at 250 sites scattered around the UK, managed by the Centre for Ecology & Hydrology (2004, see also Firbank et al., 2003). Subsets were made for 24 combinations of crop, treatment and time. Data sets on six different desert communities of rodents [6], winter annuals [7], winter perennials [8], summer annuals [9], summer perennials [10] and ant colonies [11] in the Chihuahuan desert, near Portal, Arizona, U.S.A. Twenty-four experimental plots were studied, each subdivided in 49 'stakes' for rodents and ant colonies, and in 16 'quadrats' for annual and perennial plants. Rodents and ant colonies were studied on a monthly basis for 26 and 12 years, respectively. Plants were studied in the spring and in the fall for 14 years (Ernest, Valone, & Brown 2009, see also Brown, Whitham, Ernest, & Gehring 2001). The primary data on rodents represent catches by traps. In only 64 out of 32,571 cases more than a single animal was caught (2 animals in 62 cases, 3 in one case and 4 in one case). Apparently the traps were meant to catch single individuals only and the primary samples are 'dominated' by singleton species almost by definition. We avoided to work with self-evident data and integrated the data of all the 'stakes' of a plot at a specific time as the samples for our analysis. Subsets were made for years. A data set on British breeding birds [12], treated by Williamson and Gaston (2005), is a special case. It contains a very high number of individuals (approximately 125×10^6 , representing 217 species (Stone et al., 1997, as given in Appendix 3 of Gaston & Blackburn 2000). However, in the form presented it does not consist of samples and/or subsets and it cannot be studied for scale effects. We give some comments to the bird data set in Appendix G to illustrate that definition and/or delimitation(s), as well as the accuracy of identification and counting,

are decisive for the quality of a data set. The comments were easy to conceive for the bird data and similar comments may apply to other data sets. Not all conceivable temporal, spatial and experimental combinations resulted in actual data in sets 1–11. Reasons were technical or administrative mistakes, a lack of resources, or, for ephemeral communities, the absence of organisms of interest. For data handling and analysis we primarily used ‘Excel’.

Record, sample, subset, set

We call the smallest unit of a sample a record, consisting of a number of individuals tagged for species and sample, and, eventually, samples tagged for location, time and treatment. A simple spreadsheet table of records consists of three columns for sample, species and abundance. Record tables serve as input for a cross-table showing the number of individuals of each species in each sample: a species \times sample table. Record tables can also be used to pool data for samples that belong to a specific subset, for instance, the samples of a specific year. The abundances of species over the samples of the subset are summed and a second order record table is obtained, a table with records of subsets. Similar to samples, this table serves as input for a species \times subset table.

Frequency analysis

For demonstrative purposes, we counted the number of species falling into abundance bins of exponentially increasing size. Like Preston (1948), we use exponents to the base 2 for bin boundaries. If integers of \log_2 transformed abundances were the basis for binning, values of 0, 1, 2 and so on, were to be used. Re-transformed, one gets 2^0 , 2^1 , 2^2 and so on, that equal 1, 2, 4 and so on. We take the latter values as upper bin boundaries, whereas Preston was not fully consequent for the first two bins and further assigned species with abundance values equalling bin boundaries half to the lower bin and half to the higher bin. This method of assignment is very labourious and Excel’s ‘FREQUENCY’ function does not provide for this. Preston used the term ‘octaves’ for bins and this term became common in ecological work. Because of the issues with binning, and applying values to bins, we do not use the term ‘octave’ but ‘exponential bin’ instead. For histograms with a reasonable range of bins, with reasonable numbers of species fitting the bins, quite species rich sets are required. Only the data for whole sets fulfil this requirement.

Frequencies were also analysed at the scale of samples and subsets. Of particular interest is the question if modes are higher at the larger scale of subsets than at the scale of samples. We focus directly on any difference between the groups of samples and of subsets. Each record for samples occurs exactly once in a species \times sample table and the same holds for records for subsets in a species \times subset table; records share the scale of samples or of subsets. This makes records useful for an overview of abundance frequencies of samples

and of subsets, applying bins of equal size, and for the identification of their modes. In individual samples and subsets, mathematical constraints affect the presence and the values of modes, in particular in small samples. We give examples of constraints and present results on modal values of individual samples and subsets in Appendix E.

Singletons

Singletons were counted with Excel’s ‘COUNTIF’ function. First the numbers of samples and subsets with singletons were counted and the median numbers of singletons over all samples and over all subsets were determined. It is of interest to question if numbers of singletons are related to sample and/or subset size. There is a mathematical constraint: if the total abundance of a sample equals n , there can be n species, all represented by one individual and, hence, there can be n singletons. However, $n - 1$ singletons is impossible. The possible difference between the number of singletons in samples and in subsets was analysed with the Mann–Whitney-U test.

Results

Binning

The effect of different binning methods for the mushroom data set is illustrated in Fig. 1. The frequency pattern, applying bins of equal size, is of the classical ‘hollow’ type. When exponential bins are used, a more-or-less bell-shaped pattern is obtained. These patterns are well known and occur in some of the other data sets as well (see Appendices A and B). A bell-shaped pattern can perhaps be taken as an indication of a lognormal distribution of abundance values, with a mode, representing the location of the distribution, somewhere in between of the extremes. However, we take the bell-shaped pattern at most as an indication for a normal distribution of \log_2 transformed abundances. Its mode indicates the location of the distribution of \log_2 transformed abundances, not of

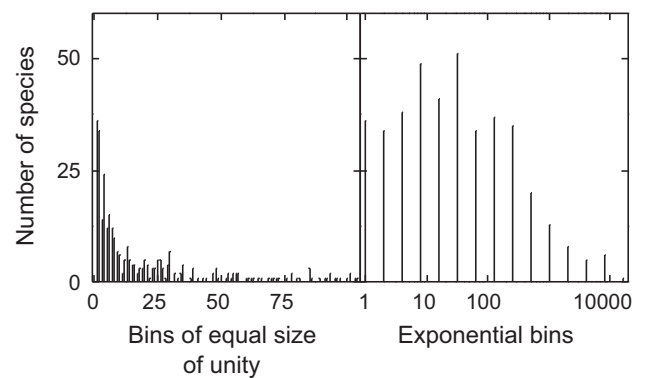


Fig. 1. Histograms of species abundances of the mushroom set. Left panel with bins of equal size of unity up to the abundance value of 105 and right panel with exponential bins to the base 2 for the whole range of abundance values.

Table 2. The effect of exponential binning on bin range, the number of specific abundance values hit and the number of involved species of the mushroom set.

Bin	Upper abundance value	Number of values hit	Number of species
0	1	1	36
1	2	1	34
2	4	2	38
3	8	4	49
4	16	8	41
5	32	15	51
6	64	21	34
7	128	31	37
8	256	32	35
9	512	20	20
10	1024	13	13
11	2048	8	8
12	4096	5	5
13	8192	6	6
14	16,384	1	1

primary abundances (the mode of the primary variable equals $\exp(\mu - \sigma^2)$, where μ and σ are the mean and standard deviation of the variable's natural logarithm; Shimizu & Crow 1988).

Why do the different binning methods lead to frequency distributions that, intuitively, show conflicting modes? Using exponential bins, the range of available abundance values increases with bin number. Not each value is actually met by a species, not all abundance values are 'hit'. We analysed available and hit values and present results for the mushroom set in Table 2 and for the other sets in Appendix C. How to read Table 2? In bin number 1, just a single abundance value, unity, is present. This single value is hit by 36 species. In bin number 10 (range 257–512), a total of 256 values are available. Of these values, 20 are hit, and, since an equal number of species fall into bin 10, each hit abundance value is hit by a single species only. Even if abundance values are hit by a single species only, the total number of hit values can result in a relatively high number of species in a bin if this bin holds a relatively large range of abundance values. A relatively high number of species in a bin does not necessarily indicate that a specific abundance value is frequently hit. In the data sets of mushrooms, fish, trees and seedlings, the ratio of number of species to abundance values actually hit is higher than unity at low bin sizes and decreases to unity at higher sizes. In the other data sets such a decreasing pattern is almost absent. The ratios of species to values are almost unity at low bin sizes already; almost all species in these sets have a different abundance.

Mode

When frequencies of species abundances are determined for bins of equal size, at a bin size of unity, modal values show

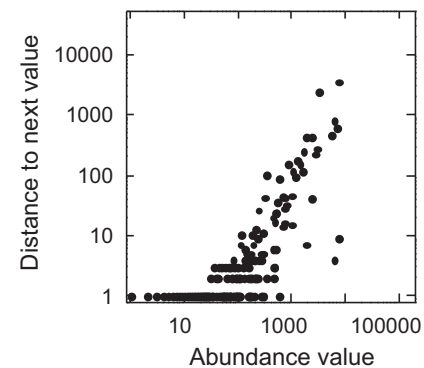


Fig. 2. Distance between abundance values in the mushroom data set.

variation among sets (Table 3A). Some sets show multiple modes. These modes occurred at the abundances of 1 and 2 for winter annuals, of 1 and 3 for winter perennials, of 2 and 39 for summer annuals and of 3, 6 and 39 for ants. Multiple modes and modes at higher abundance values tended to have a smaller size. When frequencies are determined for bins of equal size of 10, the modes occur at the first bin for all data sets (Table 3B). Plots of the 'distance' between subsequently hit abundance values against abundance value (Fig. 2 and Appendix D) indicate that 'distance' gradually increases with abundance. This pattern implies that modes will appear at the left bin at an appropriate equal bin size even if all species differ in abundance, that all hit abundance values are hit just once.

At the scale of samples and subsets with bins of equal size, species represented by one individual form the mode. Frequency declines gradually with bins. Frequencies are presented for abundance values one to five in Table 4. None of the abundance values above five shows a frequency higher than that at five, except for the subsets for rodents where a frequency of 19 occurs for the abundance value of seven.

Singletons

In the mushroom and seedling data sets, the numbers of singletons are higher in subsets than in samples and higher in the set than in the subsets (Table 5). In the fish data set, numbers of singletons are higher in subsets than in samples; the number in the set hardly surpasses the numbers in the subsets. In the tree data set, the samples show the highest numbers of singletons; they are not surpassed by the numbers in subsets or in the set. All data sets with less than 60 species show one or just a few singletons in either samples, subsets or sets. These findings are illustrated in Fig. 3, showing plots of the numbers of singletons against the numbers of species.

Table 3. (A) Modal values of species abundances in a frequency analysis with bins of equal size of unity (mult: multiple modes). (B) Species abundances in an analysis with bins of size 10. Only results for the first 15 bins are given.

	Mushrooms	Fish	Crustaceans	Trees	Seedlings	Rodents	Winter annuals	Winter perennials	Summer annuals	Summer perennials	Ants	Birds
A: Bins of equal size of unity												
Mode	1	1	2	1	1	43	mult	mult	mult	3	mult	4
S	36	9	2	17	29	2	3	4	2	3	2	4
B: Bins of equal size of 10; S values given												
Upper bin value												
10	170	35	3	65	106	6	10	14	7	13	7	13
20	41	4	0	18	24	1	3	5	2	6	3	3
30	36	5	0	15	10	0	4	1	2	6	0	3
40	13	0	1	7	12	3	0	2	4	1	2	3
50	9	3	1	9	9	3	1	0	1	2	0	3
60	11	3	2	12	5	0	0	1	1	0	0	0
70	7	2	0	10	2	1	2	3	1	2	1	0
80	7	1	0	4	2	1	1	0	1	1	2	2
90	6	2	0	4	1	0	1	0	0	0	1	1
100	8	1	0	11	2	0	0	0	2	0	1	1
110	8	1	0	6	2	0	0	1	2	0	0	1
120	3	2	1	5	0	0	0	1	0	0	0	1
130	1	0	0	5	2	0	0	2	0	1	1	1
140	3	0	0	5	0	0	0	1	1	1	1	1
150	2	0	0	5	3	1	0	0	0	0	0	2
Rest	83	24	8	114	74	14	34	14	29	22	23	182

Table 4. Frequencies of abundances of records forming the basis for samples and for subset; *n*, abundance value. Only results for the first 5 bins are given.

<i>n</i>	Mushrooms	Fish	Crustaceans	Trees	Seedlings	Rodents	Winter annuals	Winter perennials	Summer annuals	Summer perennials	Ants
Number of records (samples)											
1	6115	1650	359	15,627	3484	9022	6545	2716	5754	4227	19,635
2	2787	682	162	7017	1842	3760	3008	771	2329	1134	2786
3	1486	358	124	3907	1018	1753	1759	320	1335	443	847
4	1061	280	83	2426	740	828	1234	173	880	257	294
5	727	232	74	1733	536	553	909	94	693	165	115
Rest	3785	2039	1377	7897	4692	653	7432	288	4932	666	112
Number of records (subsets)											
1	556	257	25	949	302	45	23	53	38	87	16
2	355	132	10	547	178	21	12	24	22	41	13
3	199	81	5	445	122	19	11	18	27	25	13
4	197	51	11	304	90	4	14	11	16	21	7
5	141	47	9	257	82	6	10	13	15	18	15
Rest	1565	613	240	3281	1304	332	282	198	267	254	301

Table 5. Occurrence of singletons. Given are: the number of samples and subsets with one or more singletons and their median numbers, the results of Mann–Whitney U tests for comparison of numbers of singletons of samples and subsets (+: the number of singletons in samples has a higher average rank than that in subsets; -: reversed situation), and, finally, the numbers of singletons of the sets.

Samples	Mushrooms	Fish	Crustaceans	Trees	Seedlings	Rodents	Winter annuals	Winter perennials	Summer annuals	Summer perennials	Ants
Number of singletons	2695	347	222	512	832	4969	3003	2080	3205	2880	8203
Median number of singletons	1	4	1	30	4	1	1	1	1	1	2
Subsets											
Number of singletons	28	31	20	32	24	21	11	14	14	14	9
Median number of singletons	19	8	1	29	12	2	1	4	3	6	2
Comparison of numbers of singletons of samples and of subsets in Mann–Whitney U test											
<i>P</i>	<0.001	<0.001	0.36	–	<0.001	+	–	+	<0.001	+	<0.001
Set	36	9	1	17	29	1	3	4	1	2	1

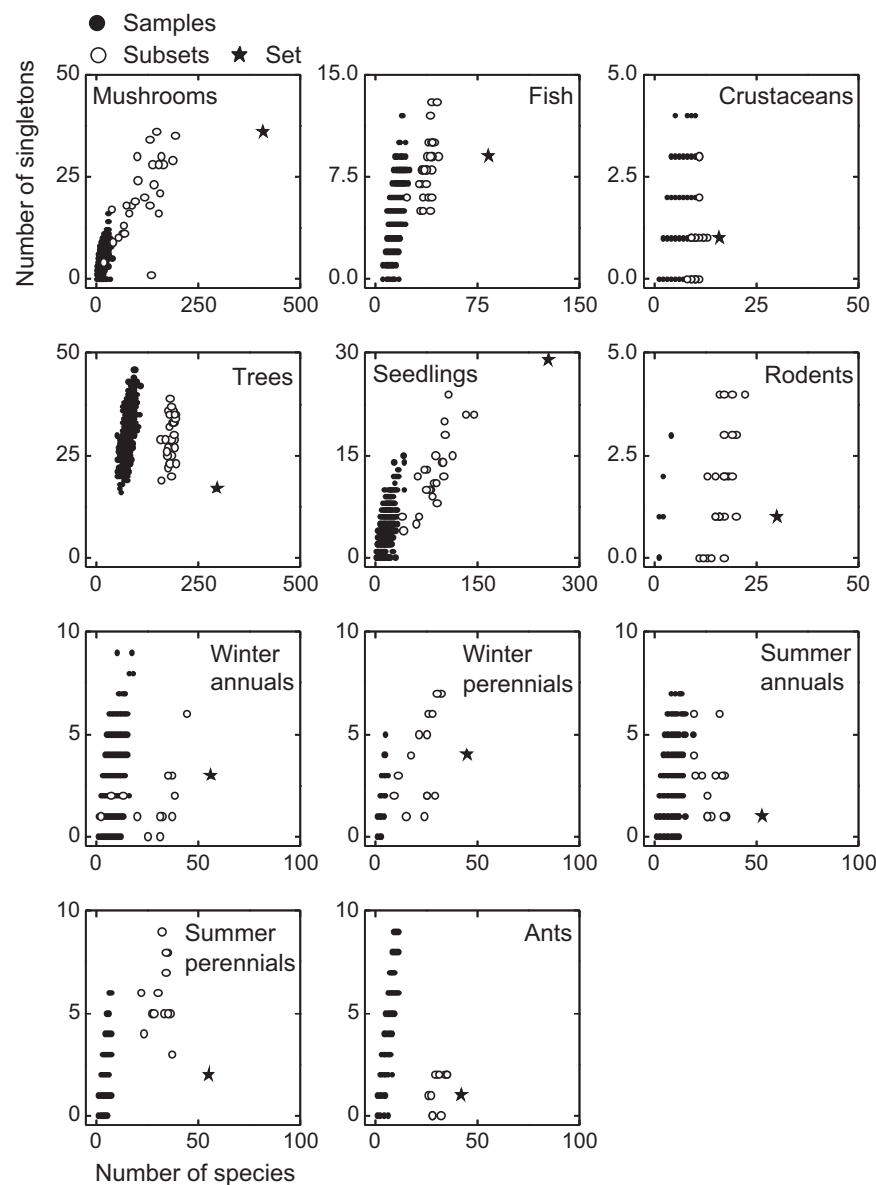


Fig. 3. Plots of the number of singleton species against sample, subset and set size. Arithmetic scales for y- and x-axis are used rather than exponential scales, to include the number of zero singletons. A unified scale ratio of 1:10 is used for ease of comparison among sets.

Discussion

Binning

More or less bell-shaped patterns occur in histograms of species abundances using exponential bins for mushrooms (Fig. 1, right panel), trees and birds (Appendix B). This is the consequence of underlying patterns for the ranges of abundance values in bins, abundance values that are hit by species and the number of species that hit specific abundances (Table 2 and Appendix C). The position of the peak of a bell-shape does not represent the mode of the species abundance distribution. Exponential binning implies log transformation of abundance values when the distribution of the binned

values is considered. Preston (1948) mentioned ‘plotting on a logarithmic base’ and applied letter codes for his ‘octaves’ but avoided the use of log transformed values. This may have been appropriate at a time when easy log transformation by computer was impossible, but it lacked consequence. Preston’s use of ‘octaves’ threw, apparently, dust in the eyes of followers who, probably, felt comfortable with non-transformation, avoiding its consequences for interpretation. We hope that our treatment illustrates the confusion caused by the usage of ‘octaves’, in fact by log transformation. We also hope that our treatment, as an addition to the work of Nekola et al. (2008), contributes to the termination of exponential binning in the analysis of species abundance distributions.

Mode

Modes at unity are the rule at the scale of samples and subsets (Table 4) but this rule is less clear at the scale of sets (Table 3A). Nevertheless modes of sets occur clearly at the first bin if larger equal bin sizes are applied (Table 3B). Except for the sets of mushrooms, fish, trees and seedling, the sets hold species that almost all differ in abundance (given the data similar to Table 2 in Appendix C on ‘number of hit values’ and ‘number of species’). All sets show a gradually increasing ‘distance’ between hit abundance values with abundance (Fig. 2 and Appendix C). This pattern implies that modes will appear at the left bin at an appropriate equal bin size.

A mode does not necessarily indicate the position of the distribution, for instance in case of a mixed, bimodal distribution with a second, minor mode. The results of the frequency analysis of abundances of records for samples and subsets (Table 4) do not indicate the presence of minor modes.

In the literature, a mode, even multiple modes, are reported when exponential binning is applied but we consider them artefacts.

Singletons

Since modal values tend to equal unity, there tend to be more singletons than species in any other abundance class. However, the number of singletons may be more or less distinct from the numbers in other classes; singletons may be more or less ‘dominating’ (Williamson & Gaston 2005). The relation between the number of singletons and the numbers of species in samples, subsets and sets, varies (Fig. 3). The data sets on mushrooms, fish and seedlings show an increase in the number of singletons with increasing total species number. The data sets on crustaceans, trees, winter annuals and ants show a decrease. A decrease is also strongly suggested by the results presented by Novotny and Basset (2000: Fig. 6) and Coddington et al. (2009: Fig. 5). We suggest that increase or decrease are related to the increase of species richness with the number of individuals over samples, subsets and the set in species-individual plots (Appendix F). The plots show differences in curvature that may indicate differences in species saturation. When the number of individuals increases beyond some point of saturation the chance for an additional individual to be a new species, a singleton, becomes low. Moreover, if it is an individual of a rare species it may belong to a yet singleton species that now loses its status.

Conclusions and implications

Our main conclusion is that the modal values of samples, subsets and sets equal or approach unity and that unity represents the location of the species abundance distribution on the abundance axis, at all scales. This formalizes the generally

shared notion of the ‘hollow curve’ for species abundance distributions (McGill et al. 2007). Our second conclusion is in agreement with Novotny and Basset (2000) who wrote: “rare species cannot be excluded from community studies as an artefact. . . they should be targeted as an interesting biological phenomenon, albeit one difficult to study”.

Our conclusions have implications for community ecology. The mode at unity implies that the distribution has no true central tendency. Neither arithmetic or geometric means nor the median are proper estimates for the location of the species abundance distribution. We suggest that an ordinary measure of variation like the variance will have little meaning as its calculation requires a mean value. This would imply that it is not useful for a preliminary characterization of species-abundance distributions to take the variance of log transformed abundance values, as has been done in the literature. We hope that our results and their implications contribute to the selection or development of an accurate species abundance model and assembly rule. We suppose that an accurate species richness estimator can be derived from a proper species abundance model. Anyway, our results indicate that rare species are not a sampling artefact which implies that richness estimators should properly deal with data on rare species (see also Uglund & Gray 2004; Coddington et al. 2009; Reichert et al., 2010). The existence of relatively many singletons and other relatively rare species in large data sets or communities is perhaps counterintuitive. Such species may be very susceptible to local extinction. However, dispersal from the ‘meta-community’ and re-establishment may compensate for local extinction.

Acknowledgements

We thank Leo van Griensven, Jac Thissen, Rene Smulders, Rampal Etienne, Martin Scheffer and Edwin Peeters for support and/or critical discussions, Karl Inne Uglund for his comments as a reviewer on an early draft, as well as anonymous reviewers and the managing editor for comments, and Silvia Dingwall for English corrections.

Appendices A–G. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <http://dx.doi.org/10.1016/j.baae.2012.03.011>.

References

- Barkman, J. J. (1989). A critical evaluation of minimum area concepts. *Vegetatio*, 85, 89–104.
- Blackwood, L. G. (1992). The lognormal-distribution, environmental data, and radiological monitoring. *Environmental Monitoring and Assessment*, 2, 193–210.

- Brown, J. H., Whitham, T. G., Ernest, S. K. M., & Gehring, C. A. (2001). Complex species interactions and the dynamics of ecological systems: long-term experiments. *Science*, *293*, 643–650.
- Centre for Ecology & Hydrology (2004). Information gateway. farm scale evaluations [internet; cited 28 May 2004 and 15 March 2005]. Available after registration from: <https://gateway.ceh.ac.uk/>.
- Coddington, J. A., Agnarsson, I., Miller, J. A., Kuntner, M., & Hormiga, G. (2009). Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. *Journal of Animal Ecology*, *78*, 573–584.
- Condit, R., Hubbell, S. P., Lafrankie, J. V., Sukumar, R., Manokaran, N., Foster, R. B., et al. (1996). Species-area and species-individual relationships for tropical trees: a comparison of three 50-ha plots. *Journal of Ecology*, *84*, 549–562.
- Dennis, B., & Patil, G. P. (1988). Applications in ecology. In E. L. Crow, & K. Shimizu (Eds.), *Lognormal distributions: theory and applications* (pp. 303–330). New York: Marcel Dekker.
- Egli, S., Ayer, F., & Chatelain, F. (1997). Die Beschreibung der Diversität von Makromyzeten. Erfahrungen aus pilzökologischen Langzeitstudien im Pilzreservat la Chaneaz. *Mycologia Helvetica*, *9*, 19–32.
- Egli, S., Peter, M., Buser, C., Stahel, W., & Ayer, F. (2006). Mushroom picking does not impair future harvests – results of a long-term study in Switzerland. *Biological Conservation*, *129*, 271–276.
- Ernest, S.K.M., Valone, T.J., & Brown, J.H. 2009. Long-term monitoring and experimental manipulation of a Chihuahuan Desert ecosystem near Portal, Arizona, USA, *Ecology*, *90*, 1708. Direct data access: Ecological Society of America, ESA, Ecological Archives [Internet; cited 19 July 2011]. Available from: <http://www.esapubs.org/archive/ecol/E090/118/>.
- Firbank, L. G., Heard, M. S., Woiwod, I. P., Hawes, C., Haughton, A. J., Champion, G. T., et al. (2003). An introduction to the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. *Journal of Applied Biology*, *40*, 2–16.
- Gaston, K. J. (1994). *Rarity*. London: Chapman & Hall., p. 159
- Gaston, K. J., & Blackburn, T. M. (2000). *Pattern and process in macroecology*. Oxford: Blackwell.
- Gleason, H. A. (1926). The individualistic concept of the plant association. *Bulletin of the Torrey Botanical Club*, *53*, 7–26.
- Henderson, P. A., & Bird, D. J. (2010). Fish and macro-crustacean communities and their dynamics in the Severn Estuary. *Marine Pollution Bulletin*, *61*, 100–114.
- Henderson, P.A. Power plant ecology/Estuarine monitoring Bristol Channel. Pisces Conservation Ltd [Internet; cited 3 August 2011]. Available on request from: <http://www.irchouse.demon.co.uk/index.html?latestreports>.
- Hubbell, S.P., Condit, R. & Foster, R.B. (2005). Barro Colorado forest census plot data. Smithsonian Tropical Research Institute, Center for Tropical Forest Science. [internet; cited 18 November 2008, downloading the 1995 files]. Available on request from: <http://www.ctfs.si.edu/group/Resources/Data>.
- Magurran, A. E., & Henderson, P. A. (2003). Explaining the excess of rare species in natural species abundance distributions. *Nature*, *422*, 714–716.
- Magurran, A. E. (2004). *Measuring biological diversity*. Oxford: Blackwell.
- Magurran, A. E., Baillie, S. R., Buckland, S. T., Dick, J. P., Elston, D. A., Scott, E. M., et al. (2010). Long-term datasets in biodiversity research and monitoring: assessing change in ecological communities through time. *Trends in Ecology and Evolution*, *25*, 574–582.
- McGill, B. J. (2003). Does mother nature really prefer rare species or are log-left-skewed SADs a sampling artefact? *Ecology Letters*, *6*, 766–773.
- McGill, B. J., Etienne, R. S., Gray, J. S., Alonso, D., Anderson, M. J., Bence, H. K., et al. (2007). Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. *Ecology Letters*, *10*, 995–1015.
- Nekola, J. C., Sizing, A. L., Boyer, A. G., & Storch, D. (2008). Artifacts in the log-transformation of species abundance distributions. *Folia Geobotanica*, *43*, 259–268.
- Novotny, V., & Basset, Y. (2000). Rare species in communities of tropical insect herbivores: pondering the mystery of singletons. *Oikos*, *89*, 564–572.
- Preston, F. W. (1948). The commonness and rarity of species. *Ecology*, *29*, 254–283.
- Reichert, K., Ugland, K. I., Bartsch, I., Hortal, J., Bremner, J., & Kraberg, A. (2010). Species richness estimation: estimator performance and the influence of rare species. *Limnology and Oceanography Methods*, *8*, 294–303.
- Ricklefs, R. E. (2008). Disintegration of the ecological community. *The American Naturalist*, *172*, 741–750.
- Shimizu, K., & Crow, E. L. (1988). History, genesis, and properties. In E. L. Crow, & K. Shimizu (Eds.), *Lognormal distributions: theory and applications*. New York: Marcel Dekker, pp. 1–25
- Stone, B. H., Sears, J., Cranswick, P. A., Gregory, R. D., Gibbons, D. W., Rehfish, M. M., et al. (1997). Population estimates of birds in Britain and in the United Kingdom. *British Birds*, *90*, 1–22.
- Straatsma, G., Ayer, F., & Egli, S. (2001). Species richness, abundance, and phenology of fungal fruit bodies over 21 years in a Swiss forest plot. *Mycological Research*, *105*, 515–523.
- Ugland, K. I., & Gray, J. S. (2004). Estimation of species richness: analysis of the methods developed by Chao and Karakassis. *Marine Ecology Progress Series*, *264*, 1–8.
- White EP. Ecological Data Wiki. [Internet]. Utah State University Department of Biology [cited 3 August 2011]. Available from: <http://ecologicaldata.org/>.
- Williamson, M., & Gaston, K. J. (2005). The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution. *Journal of Animal Ecology*, *74*, 409–422.

Available online at www.sciencedirect.com

SciVerse ScienceDirect

1 **Additional material**

2

3 **A.** Histograms of species abundances using bins of equal size of unity (expansion of the left panel of Fig
4 2 in the main text).

5

6 **B.** Histograms of species abundances using exponential bins (expansion of the right panel of Fig 2 in the
7 main text).

8

9 **C.** The effect of exponential bins on bin range, the number of specific abundance values hit and the
10 number of species (expansion of Table 2 in the main text).

11

12 **D.** Distance between abundance values (expansion of Fig 2 in the main text).

13

14 **E.** Modes of samples and subsets.

15

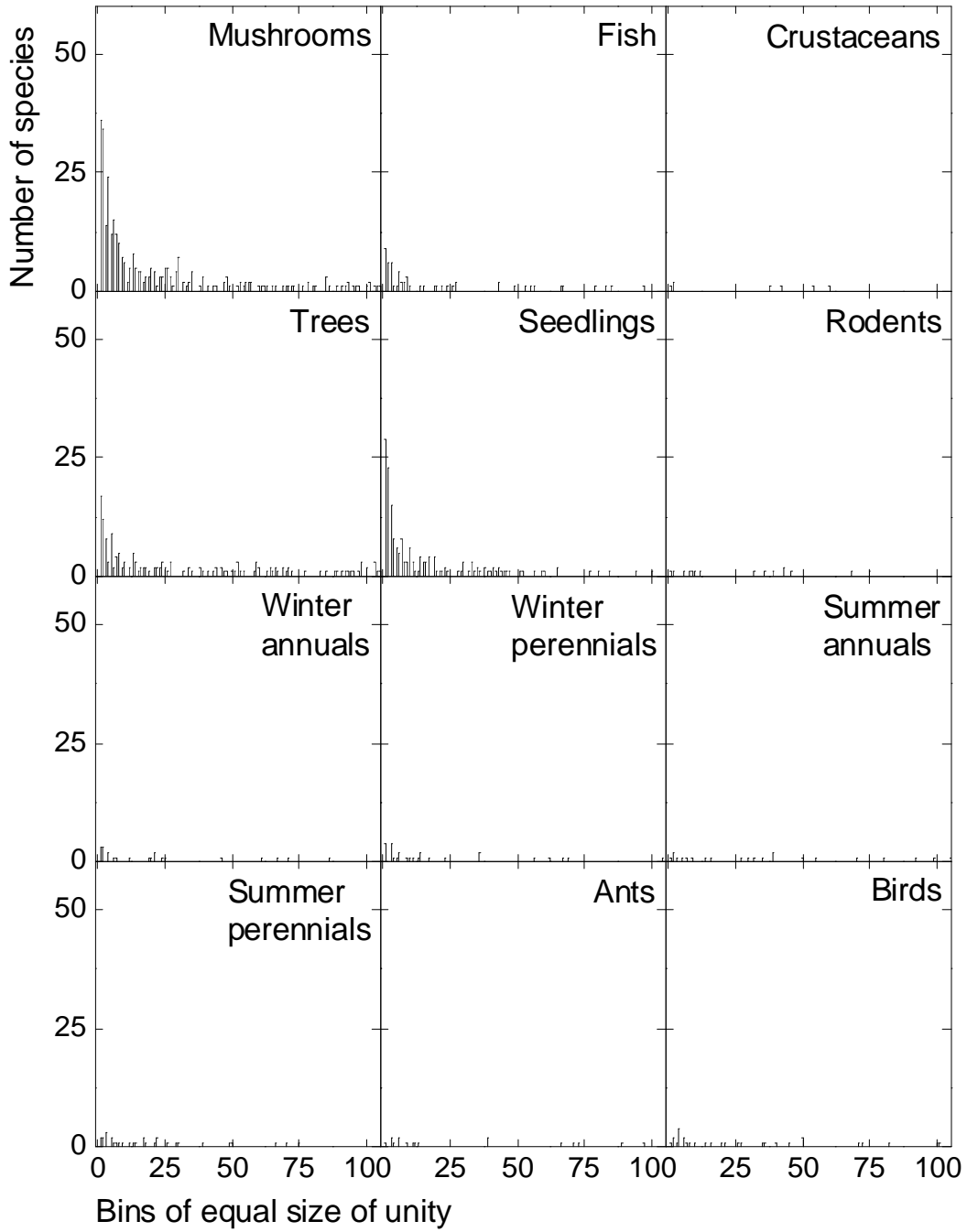
16 **F.** Species-individual plots of samples, subsets and set.

17

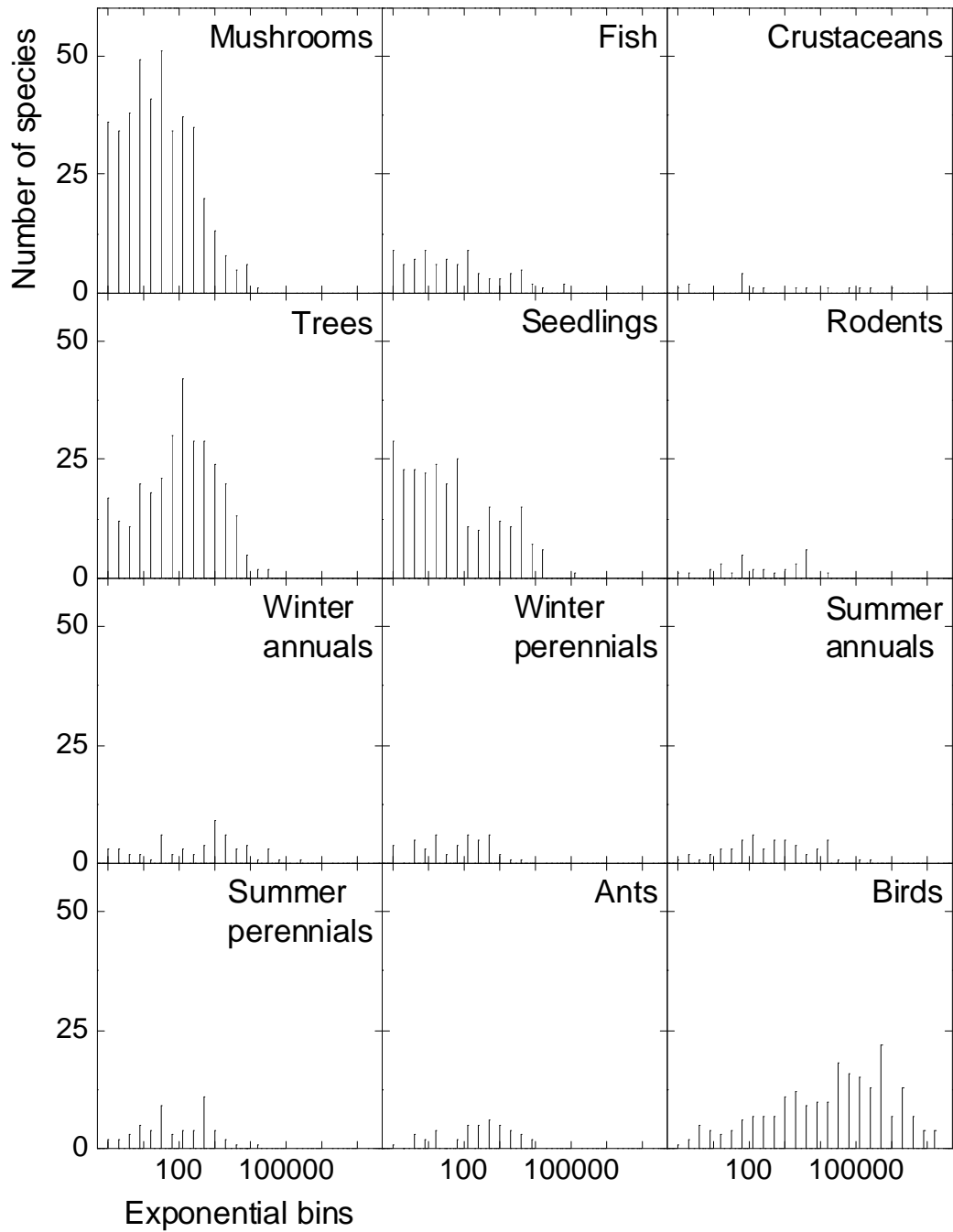
18 **G.** Bird data set.

19

20 A. Histograms of species abundances using bins of equal size of unity (expansion of the left panel of Fig
21 2 in the main text).



23 **B.** Histograms of species abundances using exponential bins (expansion of the right panel of Fig 2 in the
24 main text).



26 C. The effect of exponential bins on bin range, the number of specific abundance values hit (a) and the number of species involved (b) (expansion
 27 of Table 2 in the main text).

28

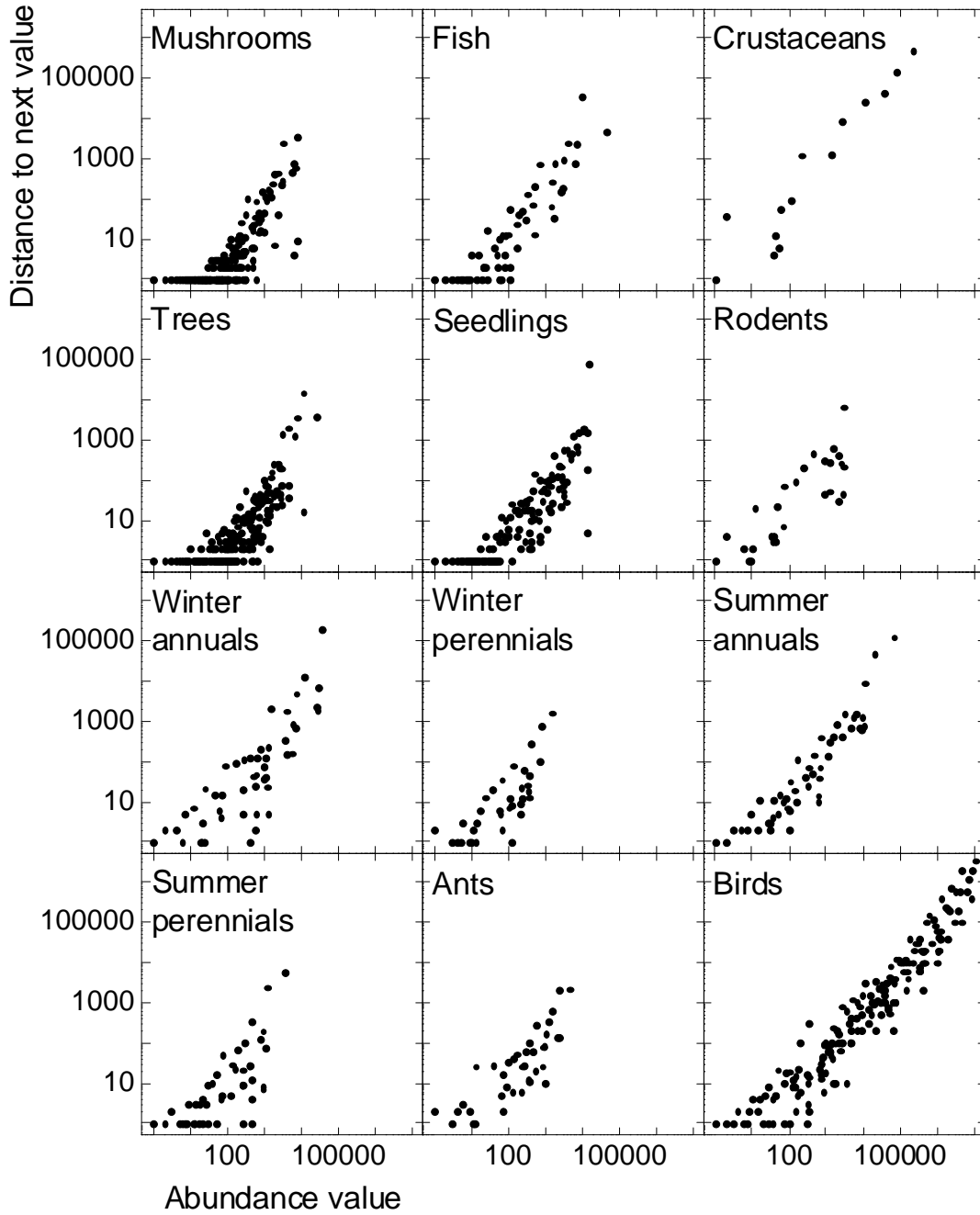
Bin	Upper value	Mushrooms		Fish		Crustaceans		Trees		Seedlings		Rodents		Winter annuals		Winter perennials		Summer annuals		Summer perennials		Ants		Birds	
		a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b	a	b		
0	1	1	36	1	9	1	1	1	17	1	29	1	1	1	3	1	4	1	1	1	2	1	1	1	1
1	2	1	34	1	6	1	2	1	12	1	23	1	1	1	3	0	0	1	2	1	2	0	0	1	2
2	4	2	38	2	7	0	0	2	11	2	23	0	0	1	2	2	5	1	1	1	3	2	3	2	5
3	8	4	49	4	9	0	0	4	20	4	22	2	2	2	2	2	3	2	2	4	5	1	2	3	4
4	16	8	41	4	6	0	0	7	18	8	24	3	3	1	1	5	6	3	3	4	4	4	4	3	3
5	32	15	51	6	7	0	0	11	21	11	20	1	1	5	6	2	2	3	3	7	9	0	0	4	4
6	64	21	34	5	6	4	4	20	30	19	25	4	5	2	2	3	4	4	5	3	3	1	2	5	6
7	128	31	37	9	9	1	1	33	42	10	11	2	2	3	3	6	6	6	6	4	4	5	5	7	7
8	256	32	35	4	4	1	1	25	29	8	10	2	2	2	2	5	5	3	3	4	4	5	5	7	7
9	512	20	20	3	3	0	0	29	29	15	15	1	1	4	4	6	6	5	5	10	11	6	6	7	7
10	1,024	13	13	3	3	0	0	24	24	12	12	2	2	9	9	2	2	5	5	4	4	5	5	11	11
11	2,048	8	8	4	4	1	1	20	20	11	11	3	3	6	6	1	1	4	4	2	2	4	4	12	12
12	4,096	5	5	5	5	1	1	13	13	15	15	6	6	3	3	1	1	2	2	1	1	3	3	9	9
13	8,192	6	6	2	2	0	0	5	5	7	7	0	0	4	4			3	3	0	0	2	2	10	10
14	16,384	1	1	1	1	1	1	2	2	6	6	1	1	1	1			5	5	1	1			7	10
15	32,768			0	0	0	0	2	2	0	0			3	3			1	1					14	18
16	65,536			2	2	1	1			0	0			1	1			0	0					15	16
17	131,072					1	1			1	1			0	0			1	1					9	15
18	262,144					1	1							1	1			1	1					11	13
19	524,288					0	0											1	1					15	22
20	1,048,576					1	1																	6	7
21	2,097,152																							13	13
22	4,194,304																							6	7
23	8,388,608																							4	4
24	16,777,216																							4	4

29

30

31 D. Distance between abundance values (expansion of Fig 2 in the main text).

32



33

34

35 E. Modes of samples and subsets.

36 Mathematical constraints

37 1) Some samples of some data sets were very small, with just a single species present. We did not take
38 their abundance values as modes. Multiple modes occur when at least two different abundance values are
39 hit by an equal, maximal, number of species of a sample. We did not accept these abundance values as
40 modes either. Since Excel's 'MODE' function does not give a warning when there are multiple modes,
41 and SPSS does, modal values were determined using SPSS.

42 2) A high modal value is impossible at a low total number of individuals in the sample. To be more
43 precise: a modal abundance value occurs at least twice, thus the maximum modal value possible equals
44 half the total number of individuals. There are other restraints, related to the combinations of abundances
45 possible at a given total abundance.

46 Mode and sample or subset size

47 Of particular interest is the question if modal values are related with sample and/or subset size. Two
48 parameters qualify for size, the total number of individuals and the number of species. Unfortunately,
49 these parameters and the mode are interdependent and correlation analyses will be difficult to interpret.
50 Alternatively, any difference between modal values of samples and of subsets, having a larger size than
51 samples by definition, can be determined. This was done with the nonparametric Mann-Whitney-U test,
52 in SPSS.

53 Modal size

54 A mode can be very distinct or it can be an abundance value with a frequency just a little higher than
55 others. Like the mode, the size of the mode is a characteristic of the species abundance distribution. A
56 distinct mode may indicate that there are relatively few different abundance values and a mode with a low
57 size may indicate the opposite and that the actual modal value may be a chance event rather than exactly
58 representing the position of the species abundance distribution. A possible relation between modal value
59 and modal size over samples was investigated with Spearman's parameter free rank correlation analysis,
60 in SPSS.

61

62 The numbers of samples and subsets with modal values were counted, thus excluding samples with a
63 single species only and samples with multiple modes. Median modal values were determined. To provide
64 for some information on the variation of modal values, frequency analyses were performed.

65

66 The mostly negative correlation of mode and mode size (Table E1) indicates that higher modes are less
67 distinct than lower modes. The modal values above unity in subsets and sets are not very distinct and can
68 be considered 'outliers'.

69 Table E1. Modes in samples, subsets and sets. Given are: the number of unimodal samples and subsets,
70 and their median modal values, the results of Spearman's rank correlation analysis of modal value and
71 modal size over samples, the results of Mann-Whitney U tests for comparison of modal values of samples
72 and subsets (+ : modal values in samples have a higher average rank than in subsets; - : reversed
73 situation), and, finally, the modes and their sizes of the sets.

74

	Mushrooms	Fish	Crustaceans	Trees	Seedlings	Rodents	Winter annuals	Winter perennials	Summer annuals	Summer perennials	Ants
Samples											
n	3,731	357	348	512	915	6,261	3,781	2,848	4,565	3,751	8,822
Single mode present, n	1,555	317	149	512	714	2,990	2,038	562	1,752	1,119	6,545
Median mode	1	1	1	1	1	1	1	1	1	1	1
Spearman's rank correlation between mode and size of mode											
r_s	-0.13	-0.32	-0.23	-0.06	-0.29	-	-0.25	-0.13	-0.24	-0.14	-0.19
p	<0.001	<0.001	0.005	0.184	<0.001	-	<0.001	0.001	<0.001	<0.001	<0.001
Subsets											
n	28	31	30	32	24	26	13	14	14	14	12
Single mode present, n	28	30	7	30	22	16	7	12	8	13	6
Median mode	1	1	5	1	1	1.5	3	1	1	1	9.5
Comparison of modes of samples and of subsets in Mann-Whitney U test											
	+	+	-	-	+	-	-	-	-	+	-
p	0.06	0.07	0.08	0.03	0.28	0.001	0.02	0.12	0.25	0.65	<0.001
Set											
Mode	1	1	2	1	1	43	mult	mult	mult	3	mult
S involved	36	9	2	17	29	2	3	4	2	3	2

75

76

77 Table E2. Frequencies of modal values in samples and subsets in the different data sets.

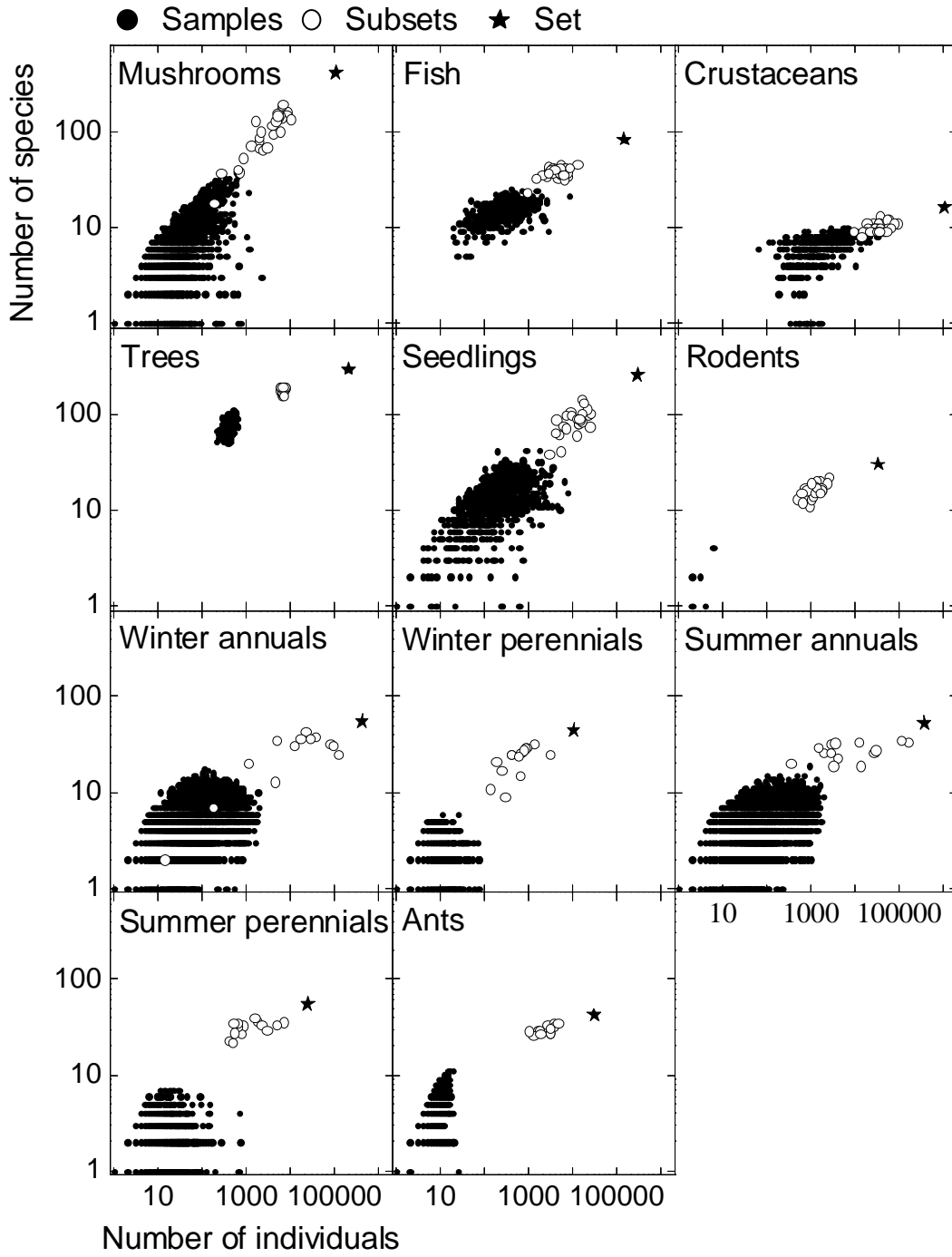
78

Modal value	Mushrooms	Fish	Crustaceans	Trees	Seedlings	Rodents	Winter annuals	Winter perennials	Summer annuals	Summer perennials	Ants
Number of samples											
1	1,224	269	99	510	543	2,378	1,479	501	1,322	989	6,032
2	219	37	17	2	121	480	300	45	251	102	444
3	55	5	14	0	33	104	108	13	87	17	55
4	23	1	5	0	7	17	51	2	34	5	8
5	12	2	2	0	1	7	31	0	26	3	5
Rest	22	3	12	0	9	4	69	1	32	3	1
Number of subsets											
1	26	29	3	29	19	8	3	9	5	12	0
2	2	1	0	0	2	3	0	2	0	1	1
3	0	0	0	1	0	2	1	1	2	0	0
4	0	0	0	0	0	0	0	0	0	0	0
5	0	0	3	0	1	0	1	0	1	0	1
Rest	0	0	1	0	0	3	2	0	0	0	4

79

80 What all sets share is the strong positive skew of the distribution of modal values, with a ‘mode of modes’
 81 equal to unity (Table E2, showing the distributions at the sample scale). The modal values of subsets were
 82 significantly higher than of samples for trees, rodents, winter annuals and ants in Mann-Whitney U tests
 83 (Table E1). This result is not necessarily in contradiction with the result that modal values of records for
 84 samples and for subsets equal unity (Table 4) and that the ‘mode of modes’ equals, or tends to, unity in
 85 subsets too. The differences may be the simple result of the mathematical constraint, that higher modal
 86 values can only occur in larger samples and/or subsets (see above).

87



91 **G. Bird data set.**

92

93 According to Gaston & Blackburn (2000), and quite inevitably, the data are population estimates rather
94 than exact counts. This is reflected in the data: maximally four digits, accompanied with some power of
95 10. For the 217 species, 186 different abundance values are given, in the range of 1 to $14.2 \cdot 10^6$, implying
96 that some species share specific abundance values. For instance, the abundance values 110 000, 320 000
97 and 400 000 are each shared by three species. It is very unlikely that such abundance values are exactly
98 shared. These details are relevant because, depending on the binning method applied, such abundance
99 values can result in (small) spikes in histograms that are artefactual rather than real. Another issue is the
100 completeness of the data set. Gaston & Blackburn (2000) provide information about the criteria that
101 species had to meet to be selected for the data set. Apparently only 217 species passed this selection. The
102 website of BOU, the British Ornithologist' Union, provides a British list with 573 species in category A,
103 9 in B and 10 in C, totalling 592 species (Category A: Species recorded in an apparently natural state at
104 least once since 1 January 1950. Category B: Species recorded in an apparently natural state at least once
105 between 1 January 1800 and 31 December 1949, but have not been recorded subsequently. Category C:
106 Species that, although introduced, now derive from the resulting self-sustaining populations). Not
107 included in the list are 15 species in category D, 235 species exclusively belonging to E and a yet
108 unknown number of species in F (Category D: Species that would otherwise appear in Category A except
109 that there is reasonable doubt that they have ever occurred in a natural state. Category E: Species recorded
110 as introductions, human-assisted transportees or escapees from captivity, and whose breeding populations
111 (if any) are thought not to be self-sustaining. Category F (list being compiled): Records of bird species
112 recorded before 1800). These details are relevant because they can be considered an indication for
113 arbitrary in- or exclusion of specific species, and for bias in the list with 'only' 217 species included of
114 possibly more than 800 considered. We assume that rare species in particular fail to meet the selection
115 criteria and consequently are underrepresented in the data set [for quite another data set on British birds
116 see Baker, Stroud, Aebischer, Cranswick, Gregory et al. 2006].

117

118 Baker, H., Stroud, D.A., Aebischer, N.J., Cranswick, P.A., Gregory, R.D., McSorley, C.A., Noble, D.G.
119 & Rehfisch, M.M. 2006. Population estimates of birds in Great Britain and the United Kingdom.
120 *British Birds*, 99, 25–44.

121 BOU, British Ornithologist' Union [Internet] . Available from: <http://bou-online.blogspot.com/>