Advances in Research

13(4): 1-24, 2018; Article no.AIR.39002
ISSN: 2348-0394, NLM ID: 101666096

How to Extrapolate Species Abundance Distributions with *Minimum Bias* When Dealing with Incomplete Species Inventories

Jean Béguinot[1*]

[1]*Biogéosciences, UMR 6282, CNRS, Université Bourgogne Franche-Comté, 6, Boulevard Gabriel, 21000 Dijon, France.*

Author's contribution

The sole author designed, analyzed, interpreted and prepared the manuscript.

Article Information

DOI: 10.9734/AIR/2018/39002
Editor(s):
(1) Farzana Khan Perveen, Founder Chairperson & Associate Professor, Department of Zoology, Shaheed Benazir Bhutto University (SBBU), Main Campus, Pakistan.
Reviewers:
(1) Manoel Fernando Demétrio, Federal University of Grande Dourados, Brazil.
(2) Richa Pandey, India.
(3) José Roberto Pereira de Sousa, Universidade Estadual do Maranhão, Brazil.
(4) Loc Nguyen, International Engineering and Technology Institute, Vietnam.
(5) Azizur Rahman, University of Toronto, Canada.
Complete Peer review History: http://www.sciencedomain.org/review-history/23066

Method Article

Received 26th November 2017
Accepted 2nd February 2018
Published 7th February 2018

ABSTRACT

The total number of co-occurring species ("true species richness") and the way their respective abundances are distributed ("species abundance distribution") are two major descriptive traits of species assemblages, in numerical terms. Moreover, beyond mere description, the species abundance distribution may help to infer how ecological factors/constraints are currently shaping the hierarchical structure of species assemblages and thereby, may contribute to shed light upon general traits regarding the functional organisation within communities of species. Unfortunately, both *total* species richness and *exhaustive* abundance distributions are not available when dealing with more or less *incomplete* species inventories, a situation which becomes increasingly frequent with the generalisation of the so-called "quick surveys" and "rapid biodiversity assessments", which are almost unavoidable when addressing very species-rich assemblages, such as, for example, invertebrate communities. Hence, the necessity of extrapolating with *minimum bias* (i) the species *accumulation* curve, thereby deriving reliable estimates of the total species richness of sampled assemblage and (ii) the species *abundance* distribution to get an exhaustive pattern including the

**Corresponding author: E-mail: jean-beguinot@orange.fr;*

full set of co-occurring species. Previous reports from the same author already dealt with the least-biased extrapolation of species accumulation and the associated derivation of total species richness. Now, an appropriate method is proposed, hereafter, to extrapolate with *minimum bias* the Species Abundance Distribution itself, when having to deal with only partial species inventories. The method shares in part some theoretical results that had already served to support the extrapolation of species accumulation process. The procedure leading to the extrapolation of the Species Abundance Distribution is first detailed in principle and then put into practice utilising a few examples. Improvements as compared to an earlier attempt at the same goal are discussed.

Keywords: Quick survey; rapid biodiversity assessment; ranked abundance distribution; least-biased extrapolation; unveiled S.A.D; least-biased estimation; broken-stick; log-normal; species community, butterfly, Bhutan.

1. INTRODUCTION

Numerical characterisation of species communities requires, first of all, the evaluation of the total (true) species richness, as accurately as possible. As important as the evaluation of total species richness may be [1,2], this remains, however, insufficient to thoroughly characterise a species assemblage numerically. Assessing also the more or less uneven distribution of species abundances within species communities is an essential complement to the estimation of total species richness [3–10]. This, in turn, can help shed light on the processes at work to shape the internal structuring of species communities. Hence the long-standing interest devoted to the specific shape of the Species Abundance Distribution (the "S.A.D.") – also designed as "Ranked Abundance Distribution" when species abundances are conveniently ranked by decreasing order of values, as will be the case all along the text below.

Thorough representations of "S.A.D.s", including the *whole set* of species that occur in the studied assemblage of species, is indeed required to get a deeper understanding of the processes involved in the hierarchical structuration of species assemblages [11]. This, however, would imply, first of all, achieving a (sub-) exhaustive inventory of the whole set of co-occurring species within the community of interest. Unfortunately, samplings may rarely be carried out exhaustively in practice, especially when dealing with highly multi-species communities, as is often the case, for example, with invertebrate faunas. Hence, the increasingly frequent recourse to the so-called "*quick surveys*" or "*rapid biodiversity assessments*", imposed as a frustrating but inevitable compromise between the desires of thorough investigations on the one hand and the limiting practical constraints on the other hand. Thus, time-limited sampling scenarios often result in sample sizes that are far too small to capture the (sub-) total species richness of the sampled communities [1,7,12].

Numerical extrapolation is a suitable surrogate to compensate for the lack of completeness of sampling, taking account of unrecorded species, at least numerically.

Various numerical extrapolation procedures have been implemented for the estimation of the number of unrecorded species and, thereby, for the evaluation of the total species richness of the studied community, using nonparametric estimators of the number of unregistered species [13,14]. Several attempts to improve the accurateness of estimations have been proposed more recently: see [15–20].

On the opposite, the numerical extrapolation of "S.A.D.s" has hardly been addressed so far. Yet, a significant attempt in this direction has recently been proposed by Chao et al. [21], but with some substantial limitations, resulting in particular from the sole recourse to "Chao-1" estimator, known to exceedingly underestimate the number of still unrecorded species in most cases [15–20].

Hereafter, I report a new procedure to *extrapolate Species Abundance Distributions* with *minimised bias*, based on the prior, least-biased extrapolation of the corresponding "Species Accumulation Curve" [19,20]. The practical application of this new procedure is subsequently illustrated using a series of examples that, moreover, highlight the interest of considering *fully extrapolated* rather than incomplete Species Abundance Distributions.

2. METHODS

A BRIEF DESCRIPTION OF THE NEW PROCEDURE OF LEAST-BIASED EXTRAPOLATION OF "SPECIES ABUNDANCE DISTRIBUTIONS"

As already underlined by Chao et al. [21], two main steps are to be considered in order to elaborate a relevant and complete representation of the Species Abundance Distribution:

1) first, it is recommended to provide relevant bias corrections to the as-recorded part of the "S.A.D.", by inferring, as accurately as possible, the *true abundances* of species on the basis of their crude *recorded frequencies*;
2) then, comes the extrapolation of the missing part of the "S.A.D." (i) by estimating the number of the still unrecorded species and then, (ii) by extrapolating their expected abundance distribution.

Conversion of the as-recorded frequencies into true species abundances

The frequencies of occurrence of recorded species within a sample of finite size provide a biased evaluation of the true abundances of species within the sampled assemblage. This is already intuitive and may be confirmed by numerical simulations. Consider, for the argumentation, a community of species all of them having ideally equal levels of abundance. Obviously, a finite sampling extracted from this community will inevitably lead to some scatter among the recorded frequencies of occurrence of these species, with the range of scatter increasing with decreasing sampling size. Crude recorded frequencies of species occurrence will thus provide biased evaluations of the real proportional abundances of these species in the community. Typically, this bias (i) will show up differences between abundances even when such differences yet do not really exist and (ii) will tend to exaggerate the magnitude of differences of abundances when differences truly exist. So that converting frequencies in *true* abundances actually requires bias corrections.

According to Appendix 1 (equation A1.14), the estimated true abundance $\tilde{a}_i$ of species 'i', having a recorded frequency $p_i = n_i/N_0$ in a sample of size N_0, is given by:

$$\tilde{a}_i = p_i.(1+1/n_i)/(1+R_0/N_0).(1-f_1/N_0) \quad (1)$$

where N_0 is the achieved sample size, R_0 (=$R(N_0)$) the number of recorded species, among which a number f_1 are singletons (species recorded only once), n_i is the number of recorded individuals of species 'i', so that $p_i = n_i/N_0$ is the recorded frequency of occurrence of species 'i', in the sample.

The crude recorded part of the "S.A.D." – expressed in terms of the series of as-recorded frequencies $p_i = n_i/N_0$ – should then be replaced by the corresponding series of expected true abundances, $\tilde{a}_i$, estimated according to equation (1).

Extrapolation of the missing (unrecorded) part of the "S.A.D."

The extrapolation of the "S.A.D." beyond the (previously bias corrected) recorded part, i.e. for ranks $i > R(N_0)$, involves two complementary aspects:

i. the least-biased estimation of the number Δ of still unrecorded species,
ii. the estimation of their abundance distribution, thereby extrapolating the "S.A.D." beyond its recorded part until reaching its full completeness.

As regards point (i), a relevant approach has now become instrumental for the extrapolation of the species accumulation curve. Applying this method allows, in particular, to select the *least-biased* type of estimator of the number Δ of still unrecorded species, among the large set of now available types of estimators [19, 20].

As regards point (ii), several options are possible:

** ***Option n° 1***: this is the simplest option which consists, following Chao et al. [21], in simply assuming a *uniformly log-linear shape* of the "S.A.D." all along its extrapolated part. Given the number Δ of unrecorded species and the fact that the cumulated abundances of unrecorded species is equal to $A_{u(N)} = f_1/N_0$ (Appendix 1, equation (A1.5)), the uniform slope of the log-linear extrapolation of the "S.A.D." is thereby entirely determined.

** ***Option n° 2***: the extrapolation of the "S.A.D." is now *"articulated" in two successive parts*; this offers better opportunities to approach a little more realistically the various possible shapes of "S.A.Ds." towards their end. Indeed, beyond their

archetypal, grossly log-linear shape [8], "S.A.D.s" most often display, in details, a large range of variations in shape. In conventional, log-transformed abundance representation, one may roughly distinguish three main categories [11]: (i) more or less symmetric sigmoidal shapes as in Preston "log-normal" or MacArthur "broken stick" distributions, (ii) shapes remaining more or less (sub-) log-linear as in geometric- or log-series and (iii) shapes consistently retaining, all along, a positive curvature (power law models). In any cases, an articulated, two-part extrapolation may comply more easily with such variations of shape towards the end of the "S.A.D.".

In practice, let Δ_1 and Δ_2 (with $\Delta_1+\Delta_2 = \Delta$) be the respective extents of the two successive stages of the extrapolation (respective ranks [R_0+1 to $R_0+\Delta_1$] and [$R_0+\Delta_1$+1 to $R_0+\Delta$]). These two successive stages may be either:

****option n° 2.1***: *both log-linear*, with different slopes s_1 and s_2, that is $a_i/a_{i+1} = s_1$ and then $a_i/a_{i+1} = s_2$; the first slope, s_1, being chosen to match the slope of the end of the recorded part of the "S.A.D.";

****option n° 2.2***: a *log-linear* first-part, with the slope s_1 (once again chosen to match the slope at the end of the recorded part of the "S.A.D.") is then followed by an *incurved* second-part along which the slope is consistently increasing in module (as in log-normal or broken-stick models) or decreasing (as in Zipf's models). For this second part, an appropriate expression may be of the type $(a_i/a_{i+1}) = s_1.(1+(i - i_0)^a)$, with i_0 as the species rank at the beginning of the second stage ($i_0 = R_0+\Delta_1+1$) and the exponent 'a' being positive, for example a = 3.

Let focus a little further upon this more flexible and adaptable option. Four parameters characterise this type of extrapolation, namely: the numbers Δ_1 and Δ_2 of species involved in each of the two successive stages, the slope s_1 of the log-linear first part and the rate of variation of the slope during the second stage (related to 'a'). In turn, four parameters are thus necessary to determine the shape of the extrapolation; these will be:

i. the least-biased estimates of Δ, which constrains the sum $\Delta_1+\Delta_2$;
ii. the recorded ratio f_1/N_0 which constrains the cumulated abundances of the Δ unrecorded species, according to TURING relationship (see equation (A1.5) in Appendix 1);
iii. the slope at the end of the recorded part of the "S.A.D.", to which the slope s_1 of the first part of the extrapolation is expected to conform in order to respect the continuity of the first derivative in theoretical "S.A.D.s";
iv. the estimated abundance a_{min} (= $a_{(Ro+\Delta)}$) of the last, rarest unrecorded species.

The estimation of this last parameter proceeds from equation (1)

$$a_{min} = (1/N_f).(1+1/1)/(1+R_0/N_f).(1-f_1/N_f)$$
$$= (2/N_f)/(1+R_0/N_f).(1-f_1/N_f)$$

with N_f as the sample size when the last species is just being recorded for the first time. N_f is obtained from the least-biased extrapolation of the species accumulation curve R(N) ([19, 20]; see also Appendix 1 for the expression of this extrapolation). In practice, as the species accumulation curve reaches the last species asymptotically, we follow the convention suggested by CHAO and coworkers [22]: N_f is the computed sample-size which allows to reach total species richness minus 1 or 0.5 (i.e. $R(N_f) = R_0 + \Delta - 1$ or 0.5).

In practice, as $N_f >> R_0$ and $N_f >> f_1$, it comes:

$$a_{min} \approx 2/N_f \quad (2)$$

** ***Option n° 3***: a third (and *quite preferred*) alternative solution to relevantly extrapolate "S.A.D.s" takes advantage of the prior, least-biased extrapolation of the species accumulation curve itself [19, 20]. This, approach features particularly relevant since the rate of species accumulation along progressive sampling is directly dependent upon the distribution of species abundances in the sampled assemblage of species [12, 21]. Indeed, consider the species 'i' of rank 'i' in the "S.A.D." and let N_i be the sample size when this species is detected *at first* during progressive sampling. At sampling size N_i, the number n_i of individuals of species 'i' is thus $n_i = 1$ and species 'i' is then assigned a frequency $p_i = 1/N_i$.

Then, according to equation (A1.14) in Appendix 1, it comes:

$$a_i = (1/N_i).(1+1/1)/(1+R(N_i)/N_i).(1-f_1(N_i)/N_i)$$

that is:

$$a_i = (2/N_i)/(1+R(N_i)/N_i).(1-f_1(N_i)/N_i) \quad (3)$$

with N_i defined by $R(N_i) = i$.

The number $f_1(N_i)$ of singletons when the sample reaches size N_i is related to the first derivative $\partial R(N)/\partial N$ of the expression of the species accumulation curve R(N) (see equation (A1.16) or [19]:

$$f_1 = N.\partial R(N)/\partial N).$$

Accordingly, here, $f_1(N_i) = N_i.|\partial R(N)/\partial N|_{Ni}$ with $|\partial R(N)/\partial N|_{Ni}$ as the first derivative of R(N) at $N = N_i$.

Thus, it comes finally:

$$a_i = (2/N_i)/(1+ R(N_i)/N_i).(1– [\partial R(N)/\partial N]_{Ni}) \quad (4)$$

which, in practice, comes down to:

$$a_i \approx (2/N_i)/(1 + R(N_i)/N_i) \quad (4 \text{ bis})$$

since $f_1(N)$ and, thus, $\partial R(N)/\partial N$ already becomes quite negligible as compared to N along the extrapolated part.

This equation provides the extrapolated distribution of the species abundances a_i (for $i > R(N_0)$) as a function of the extrapolated species accumulation curve R(N) (for $N > N_0$), with 'i' being equal to $R(N_i)$. The expression of R(N) to be selected is provided at Appendix 2.

Nota – In actual "S.A.D.s", the abundances a_i are expressed as ratios of integers, thus giving "S.A.D.s" a discontinuous, "staircase-like" shape. In particular, the lowest level of *recorded* abundance (that is when $n_i = 1$) is represented by a step often comprising not only one species but a number usually > 1, since the number f_1 of singletons may often exceeds 1. The connection *in continuity* between the end of the recorded part of the "S.A.D." and the beginning of the extrapolated part is thus located at $i = R_0 – f_1/2$ (rather than $i = R_0$, see Figs. 5 and 8); therefore, equation (4) is standardised accordingly.

At last, from a more "*heuristic*" point of view, it should be noted that equation (4) clearly highlights the *tight articulation* that exists between:

1) the "Species Abundance Distribution" [i.e. the species *abundance* a_i as a function of the species *rank* i : $a_i = a(i)$]
   and
2) the "Species Accumulation Process" [i.e. the *sampling size* N_i when the *species of rank* i is expected to be first detected : $i = R(N_i)$];

finally leading to the linkage $a_i = a(i) = a(R(N_i))$, detailed by equation (4).

What can be reasonably learned from Species Abundance Distributions

Beyond their purely descriptive contribution, it is usually expected from the "S.A.D.s" some additional insights on the procedural pathways that makes the corresponding assemblage of species hierarchically structured as it actually is, in terms of abundances distribution.

**** comparison to classical "S.A.D.s" models***

A common practice consists in trying to select, among a series of referential models, which of them looks closest to the studied "S.A.D." A large – and steadily increasing number of referential models (see [23, 24]) – is currently available. Most of these models often seem, however, more or less equally appropriate, for the bulk of empirical "S.A.D.s" [23]. For example, similarly high correlation coefficients, comprised between 0.90 and 0.94, are reported by ALROY [24] when 1055 empirical "S.A.D.s" are tested against each of four classic models, namely: geometric series, double-geometric series, log series, log-normal (although these models respectively refer to quite different causal mechanisms!). Strong disagreements may thus occur among the resulting interpretations, see for example the sharply different points of view that oppose Baldridge et al. [25] to either McGill et al. [23] or Alroy [24]. Indeed, this somewhat confusing situation has already been emphasised previously [8]. This especially holds true when having to deal with *incomplete* surveys.

An alternative or comparative avenue would consist to *compare in quantitative terms* – rather than trying to identify – the studied "S.A.D" with an appropriate model that may serve as a "null" model, having explicit simple significance. In this perspective, models referring to either strictly or statistically *even* abundances distributions would feature particularly adequate. The purpose being, here, to characterise (i) the degree of unevenness of abundance distribution as a whole and, moreover, (ii) to evaluate the respective contributions to unevenness of such or such particular species. Two types of "null" models can fairly well match these objectives:

- a basically *deterministic* model, the trivial "ideally *strictly* even abundance distribution" $\{e_i\}$, with abundance e_i of the i^{th} species (labelled 'e' for 'even'), defined, independently of rank 'i', as:

$$e_i = 1/S_t \quad (5)$$

with S_t as the total species richness of the assemblage of species (previously derived by *least-biased* extrapolation);

- a basically *stochastic* model, the MacArthur "broken-stick" distribution $\{r_i\}$, expected to provide the stochastic outcome issued from the randomisation of an ideally even distribution of abundances (that is, in practice, the statistically random apportionment of abundances values between all the co-occurring species [26]:

$$r_i = (1/S_t).\Sigma\ (1/n) \quad (6)$$

with r_i as the abundance of the i^{th} species (labelled 'r' for 'random') and the summation Σ being extended from n = i to n = S_t. The "broken-stick" distribution had already been suggested as an appropriate "null" by WILSON [27].

The degree of unevenness of any empirical "S.A.D." is, of course, always greater than the zero unevenness of the deterministic "strictly even" model, while it may be either larger, equal or inferior to the unevenness level of the corresponding "broken-stick" model, depending on whether the structuring process at work in the species assemblage has a stronger, equal or weaker influence than has the randomisation process involved in the "broken-stick" model. In this respect, the "broken-stick" model may be considered a more suggestive and interesting referential model than the deterministic "ideally even" model (but in fact both approaches are complementary).

Note that both "null" models require the previous knowledge of the total species richness St (equations (5) an (6)). This is a second strong reason to implement extrapolations of "S.A.D.s".

comparison between two or several "S.A.D.s"

Two or several "S.A.D.s" may also be compared directly, by considering in each "S.A.D." the abundances of species having a same rank 'i' (or for a same range of ranks) in each of compared "S.A.D.s". Yet, if the "S.A.D.s" to be compared come from species assemblages that substantially differ by their respective species richness, the direct comparison between abundances may become somewhat irrelevant and can suggest "misleading conclusions" [11]. This is because, in such case, a trivial contribution from species richness, of purely numerical order, inopportunely adds to the direct influence of ecological factors upon the distribution of species abundances. Indeed, there is an unavoidable trend for species dominance to decrease when total species richness increases; the dominance tending to be somewhat "diluted" by the number of co-occurring species [11, 28, 29]. This trend – and its essentially *numerical rather than biological* origin – are clearly exemplified (i) by the inversely proportional decrease of the flat level of abundances in the "ideally even abundance distribution" and (ii) by the decrease of the average steepness of the "broken-stick" distribution, when species richness S_t increases (see equations (5) and (6)). And it is precisely why, both "null" models can serve to cancel the influence of this non-biological trend, when "S.A.D.s" issued from communities having substantially different species richness are to be compared.

Therefore, to ensure relevant comparisons in practice, the respective species abundances of compared "S.A.D.s" should be rationalised by reference to one or the other of the two "null" models: the accordingly "rationalised" abundance at rank 'i' is then identified to the ratio: a_i/e_i or a_i/r_i (see below).

synthetic indices to reflect the intensity of structuration within species assemblages

The distribution of species abundances in a community may be understand either in term of:

- *pattern*: the "S.A.D." being, by itself, the complete and detailed *description* of the internal structuring of the assemblage;

- *process*: the relative abundance of each species being, then, considered as reflecting the species relative "*performance*" in the particular context of the assemblage. "Performance" being understand, here, *sensu latissimo*, that is encompassing the factors of all kinds which together contribute to increase (or decrease) the relative abundances of each species: these factors may be, for some of them, *intrinsic* to the species (its own capacities facing the ecological and syn-ecological context within the assemblage) and, for some others, *opportunistic* or *stochastic* (depending, in particular, upon the historical and environmental context which contribute also to the actual structuration): see schematic sketch in Fig. 1.

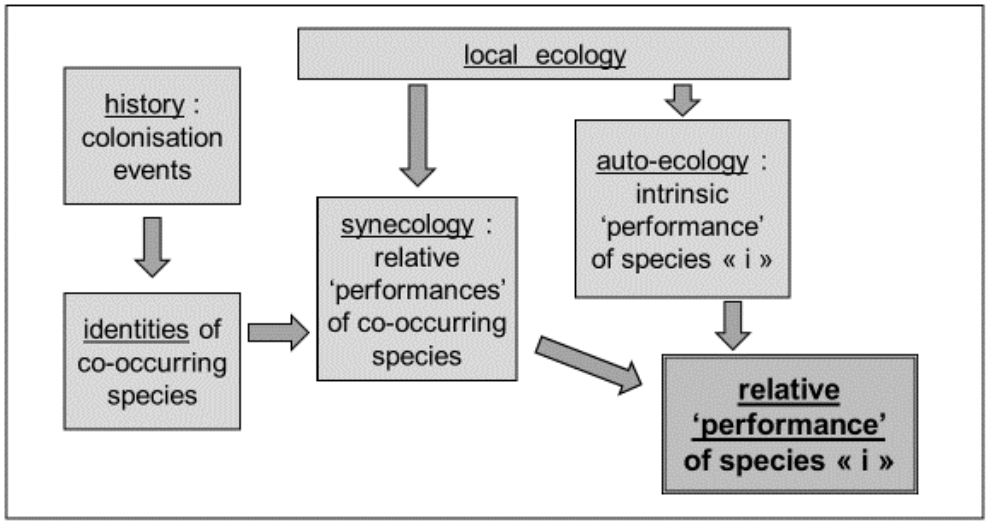

Fig. 1. Schematic sketch showing how the relative "performance" - *sensu latissimo* - of a *given* species "i", occurring in a *given* assemblage of species, depends upon both the historical and the ecological and syn-ecological contexts which are peculiar to this assemblage

Accordingly, representative indices may address either *each species in particular* or the *assemblage as a whole*.

> *Indexation per species: quantifying the relative "performance" of a species in particular*

The degree of "performance" of each species makes full sense when compared to either the "ideally even distribution" $\{e_i\}$ or the randomly apportioned abundance distribution ("broken-stick") $\{r_i\}$, giving rise to the two following indices, respectively:

$$IPe = a_i/e_i \quad (7)$$

$$IPr = a_i/r_i \quad (8)$$

Testing the statistical significance of the index IPe, comes down to test the significance of the gap between the true abundance, $a_i = p_i.(1+1/n_i)/(1+R_0/N_0).(1–f_1/N_0)$ (see equation (1)) and the reference value $e_i = 1/S_t$ (equation (5)). Which finally amounts to compare the recorded frequency of occurrence $p_i = (n_i/N_0)$ to the threshold $(1/S_t)/[(1+1/n_i)/(1+R_0/N_0).(1–f_1/N_0)]$.

Similarly, testing the statistical significance of the index IPr, comes down to test the significance of the gap between true abundance, $a_i = p_i.(1+1/n_i)/(1+R_0/N_0).(1–f_1/N_0)$ and the reference value $r_i = (1/S_t).\Sigma(1/n)$ (equation (6)). Which finally amounts to compare the recorded frequency of occurrence $p_i = (n_i/N_0)$ to the threshold $[(1/S_t).\Sigma(1/n)]/[(1+1/n_i)/(1+R_0/N_0).(1–f_1/N_0)]$.

Nota: "composite" indices, "IPc", may also be considered, each of them referring to *both* "null" models:

$$IPc1 = (a_i/e_i)/(r_i/e_i) = a_i/r_i \quad (9)$$

or:

$$IPc2 = (a_i–e_i)/(r_i–e_i) = (a_i/e_i – 1)/(r_i/e_i – 1) \quad (10)$$

These two kinds of composite indices are either > 1 or < 1, depending on whether the species abundance a_i is larger or smaller than the corresponding abundance r_i in the "broken-stick" model.

Moreover, the second composite index, IPC2, equals zero if the species abundance a_i is equal to the ideally even abundance e_i $(=1/S_t)$, which makes this index being *scaled*, profiting *both* by a "zero" threshold level and by the definition of a "unit", equal to the difference $(r_i – e_i)$ between the randomly apportioned abundance distribution and the ideally even abundance distribution.

> *Indexation relative to the whole assemblage: quantifying its relative degree of structuration*

Once again, this indexation makes full sense when compared to either the "ideally even distribution" $\{e_i\}$ or the "broken-stick distribution" $\{r_i\}$.

Two complementary factors may be considered:

i. the number of species, s_e or s_r, whose abundances exceed the corresponding abundance in either the "ideally even" or the "broken-stick" models respectively,
ii. the average values, IPe* or IPr*, of the indices of performance IPe or IPr (defined above), for each of these s_e or s_r species respectively.

A composite index may be derived accordingly:

$$Ie = s_e.IPe^* \quad (11)$$

$$Ir = s_r.IPr^* \quad (12)$$

The *steepness* of the decreasing abundance slope, along all or part of the "S.A.D.", also offers a complementary synthetic characterisation. Comparisons may be simply carried out between slopes as such. But, as already emphasised, slopes may advantageously be previously standardised to the slope of the corresponding "broken-stick" (or "ideally even") distribution, so as to cancel the direct contribution of total species richness and, thereby, highlight separately the influence of all the other parameters determining the species assemblage structuration.

Testing the statistical significance of differences between "S.A.D.s" or between "S.A.D." and a referential model

The statistical significance of differences between recorded abundances, as well as the statistical significance of the indices derived above, can be tested using conventional statistical methods. Yet, the *Bayesian inference* approach (see equation A1.7 in Appendix 1) now offers an improved way to conduct accurate tests.

PRACTICAL IMPLEMENTATION OF THE NUMERICAL EXTRAPOLATION OF THE SPECIES ABUNDANCE DISTRIBUTION

In accordance with the mainly methodological objective of this contribution, the following examples are, first of all, intended at illustrating the practical implementation of the procedure of extrapolation of "S.A.D.s" described above. That means that the resulting ecological and biological implications pertaining to each of the two examples below, as interesting as they are, will be treated elsewhere.

Among the different options of treatments provided above, the option n° 3 is duly selected in these examples for its better accurateness. Indeed, option n° 3 takes advantage of the *least-biased extrapolation* of the corresponding species accumulation curve, serving as the steering guide to the extrapolation of the "S.A.D.s" itself.

Example 1: partial inventory of butterfly fauna at "Manas Range Park" (Bhutan)

The first example relates to a subtropical butterfly community at "Royal Manas Range National Park" (Bhutan), partially surveyed by Nidup et al. [30]. Based on the reported field data (a list of R_0 = 91 recorded species including their respective abundances issued from sampling of N_0 = 1319 individuals), an extrapolation of the Species Accumulation Curve was computed, after selection of the least-biased type of estimator of the number of still unrecorded species: in this case, the 'Jackknife-5' estimator, leading to an estimated number of 28 unrecorded species. The total species richness of butterfly fauna in the sampled ecosystem at "Manas Range" is thus evaluated at 119 species: 91 recorded + 28 unrecorded (resulting completeness level of the inventory: 76%). The results above, as well as the extrapolated Species Accumulation Curve were derived in [31].

Based on this prior extrapolation of the Species Accumulation Curve, the extrapolation of the "S.A.D." is subsequently obtained, by applying equation (4).

The completed "S.A.D.", including the least-biased extrapolation (ranks i = 92 to 119), is provided in Figs. 2 to 5. While Fig. 2 is according to classical representation, using log-transformed abundances, the following figures comply with the convention of representation originally adopted by MACARTHUR [26], involving untransformed (rather than log-transformed) species abundances, a representation which provides a more straightforward visual appreciation of the relative abundances.

Note that, restricted to its as-recorded part, the shape of the "S.A.D." would likely comply with a "log-series" distribution (Fig. 2) leading to assume that only one (or, at most, very few) major factor(s) are expected being at work to shape the abundance structuration of the butterfly assemblage. Now, considering the whole "S.A.D." – fully completed thanks to

extrapolation – the pattern of distribution and the associated conclusion will strongly differ – hence the importance of implementing reliable numerical extrapolation. The completed distribution (Fig. 2) shows a sigmoidal shape, slightly dissymmetric (as is also, for example the "broken-stick") which now looks closer to a "log-normal" distribution, but slightly dissymmetrically skewed. This is more in favour of a multiplicity of mutually independent factors involved together in the process of hierarchical structuration of the community [3].

Interestingly, Ulrich et al. [11] also emphasised the fundamental importance of distinguishing between *fully censused* and *incompletely sampled* communities, when trying to provide a relevant interpretations of "S.A.D.s". Also, they reported that "S.A.D.s" issued from completely censused animal communities often tend to follow the "log-normal" model.

the practical computation procedure briefly reviewed step by step

The prior least-biased extrapolation of the Species Accumulation Curve for the butterfly assemblage at Manas Range Park, reported in detail in [31], provides all numerical data necessary [N_0 = 1319, $R_0 = R(N_0)$ = 91, f_1 = 17 and the least-biased expressions of the extrapolated species accumulation $R_5(N)$ and its first derivative $\partial R_5(N)/\partial N$)] to proceed, in turn, with the corresponding least-biased extrapolation of the Species Abundance Distribution. This data is subsequently introduced:

- in equation (1), which provides the *bias-corrected* estimates of abundances for the *already recorded part* of the "S.A.D." ;
- in equation (4), which provides the least-bias *extrapolation* of the abundance distribution of the *still unrecorded species*.

Figs. 2 to 5 provide the graphical expressions of the results derived from equations (1) and (4).

Example 2: partial inventory of butterfly fauna at "Sankosh River catchment" (Bhutan)

This second example relates to a tropical butterfly inventory at "Sankosh River catchment", partially surveyed by SINGH [32]. Based on the reported field data (a list of R_0 = 213 recorded species including their respective abundances issued from sampling of N_0 = 1731 individuals),

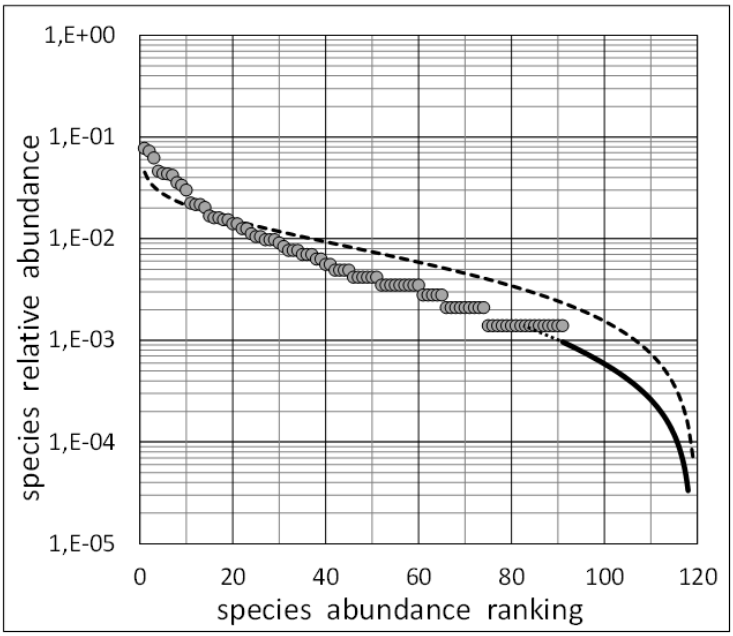

Fig. 2. The *completed* Species Abundance Distribution derived from the partial inventory of butterfly fauna at "Royal Manas Range National Park" (Bhutan). Recorded data: 91 species (ranks i = 1 to 91). Least-biased extrapolation: 28 unrecorded species (ranks i = 92 to 119). The dashed line accounts for the corresponding MacArthur "broken-stick" model. Note that species abundances are presented on a *log-transformed scale*.

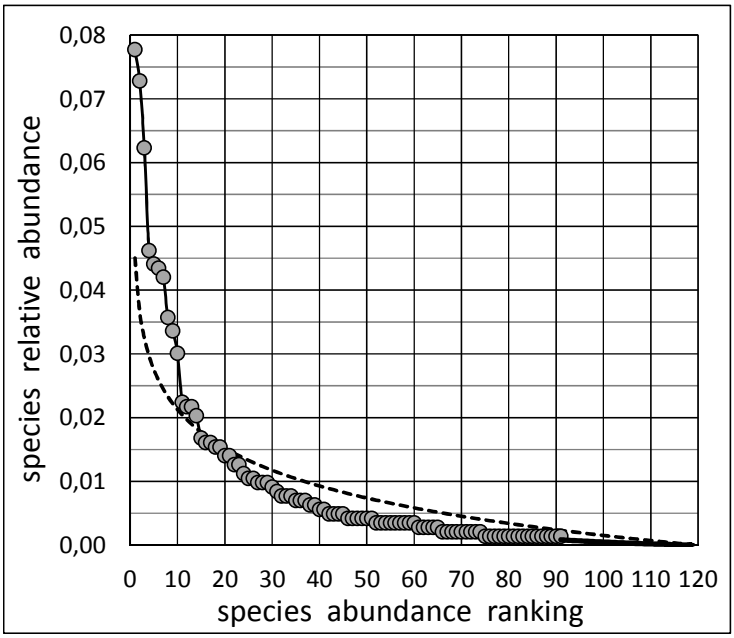

Fig. 3. The *completed* Species Abundance Distribution of Fig. 2, with species abundances presented on an untransformed scale, in compliance with the convention of representation of MACARTHUR [26]. The dashed line accounts for the corresponding "broken-stick" model

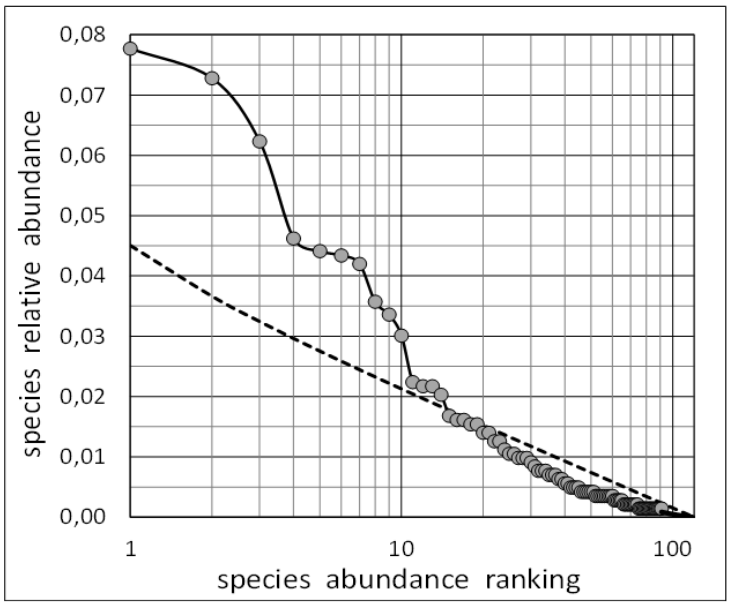

Fig. 4. Same as Fig. 3, with logarithmic scale for species ranking, which provides a more easy reading of the beginning of the "S.A.D.". The dashed line accounts for the corresponding MacArthur "broken-stick" model.

an extrapolation of the Species Accumulation Curve was implemented after selection of the least-biased type of estimator of the number of still unrecorded species: in this case, the 'Jackknife-2' estimator, leading to an estimated 68 unrecorded species. The total species richness of butterfly fauna at "Sankosh river" is thus evaluated at 281 species: 213 recorded +

68 unrecorded (resulting completeness level of the inventory: 76%) and the extrapolated Species Accumulation Curve is given in [33]. The completed "S.A.D." including the derived least-biased extrapolation (ranks 214 to 281) is provided in Figs. 6 to 8.

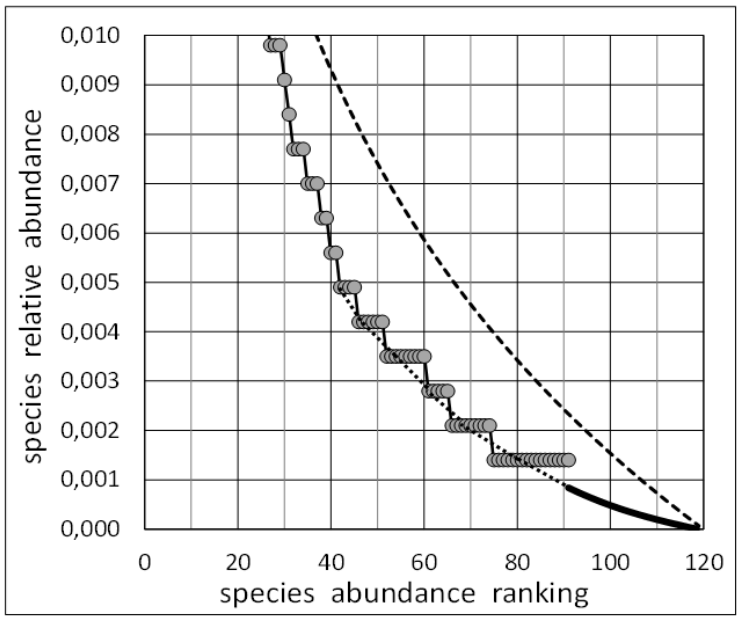

Fig . 5. Same as Fig . 3, with expanded scale for abundances, providing a more easy reading of the following part of the "S.A.D." Note that the extrapolated part (ranks 92 to 119) definitely supports the expectation that the abundance distribution of the unrecorded species still stays lower than the "broken-stick" model (dashed line), as already initiated as soon as rank i ≈ 15.

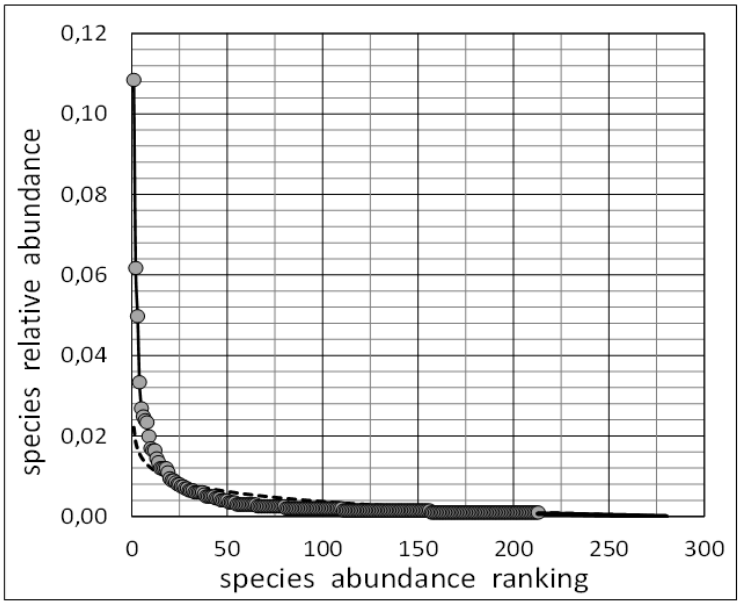

Fig. 6. The *completed* Species Abundance Distribution derived from the partial inventory of butterfly fauna at "Sankosh River catchment" (Bhutan). Recorded data: 213 species (ranks i = 1 to 213). Least-biased extrapolation: 68 unrecorded species (ranks i = 214 to 281). The dashed line accounts for the corresponding MacArthur "broken-stick" model.

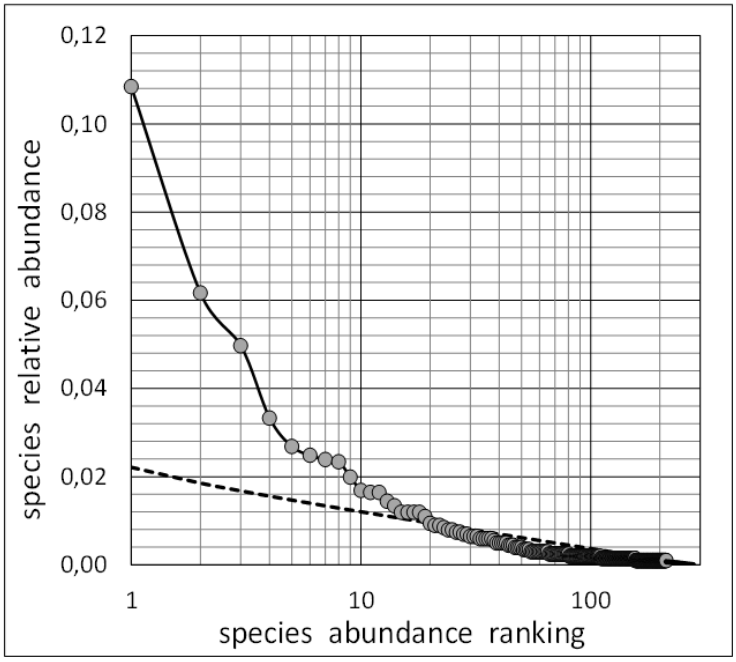

Fig. 7. Same as Fig. 6, with logarithmic scale for ranks providing a more easy reading of the beginning of the "S.A.D."

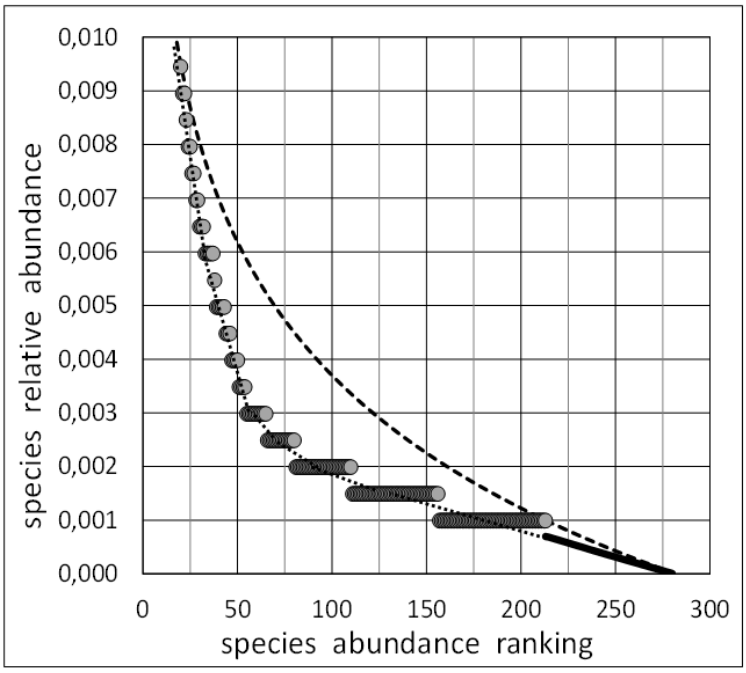

Fig. 8. Same as Fig. 6, with expanded scale for abundances, providing a more easy reading of the following part of the "S.A.D." Note that the extrapolated part (ranks 214 to 281) definitely supports the expectation that the unrecorded species abundances distribution actually continue staying slightly lower than the "broken-stick" model, as initiated from rank ≈ 20.

Making relevant comparisons between species abundance distributions, issued from different species assemblages: comparing "Manas Range" and "Sankosh River"

The butterfly assemblages considered above, at "Manas Range" and "Sankosh River", markedly differ as regards their total species richness, with respectively S_t = 119 and S_t = 281. The (recorded) abundances of, say, the ten most abundant species (i = 1 to 10) in each assemblage are plotted in Fig. 9. On average, apart from some differences rank by rank, the abundances in both assemblages are relatively similar. But, in fact, the comparison is appreciably biased by the substantial difference of true species richness S_t existing between the two assemblages, as already underlined above. Thus, in the comparison, the abundances at "Sankosh River" (twice as species-rich as "Manas Range") might be considered as "disfavoured" due to the highest number of co-coccurring species at "Sankosh". Accordingly, to put aside the effect of difference in total species richness, abundances can be rationalised by reference to an appropriate "null" model, the latter opportunely taking account of the specific contribution of total species richness alone. The rationalisation of abundances thus makes it possible to identify and evaluate separately the respective contributions to the hierarchical structuring of species abundances of (i) the total species richness and (ii) the bulk of other ecological factors, which, they, are of specific interest.

The rationalised abundances values Ipr (see equation (8)), computed by reference to the "broken-stick" model are plotted in Fig.10: $Ipr = a_i/r_i = a_i/[(1/S_t).\Sigma\ (1/n)]$, with S_t = 281 for Sankosh and S_t = 119 for Manas. This rationalisation of abundances reveals that the processes at work in both assemblages, leading to the hierarchical structuring of abundances, are considerably stronger (≈ 3 times) at "Sankosh" than at "Manas", once deduced the specific effect of the difference of species richness. This, indeed, could not have been suspected from looking at the crude data from Fig. 9 alone. Hence the interest of this rationalisation with respect to "null" models.

The rationalised abundances values Ipe (see equation (7)), by reference to the other "null" model, the "ideally even abundance distribution" ($Ipe = a_i/e_i = a_i/(1/S_t)$), are plotted in Fig. 11. The highlighted trend remains quite similar to those derived from rationalisation to "broken-stick" model.

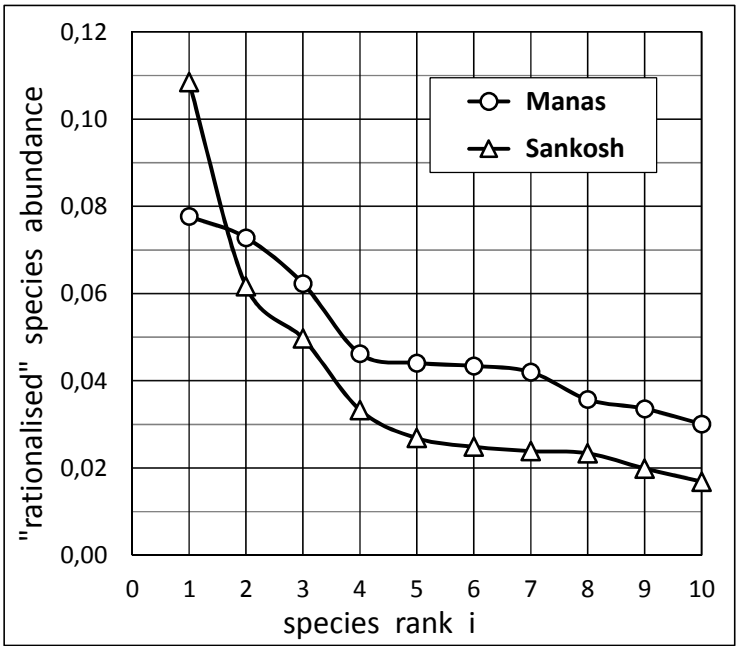

Fig. 9. The abundances of the ten most frequent species in the butterfly assemblages at "Manas Range" and at "Sankosh River" (Bhutan)

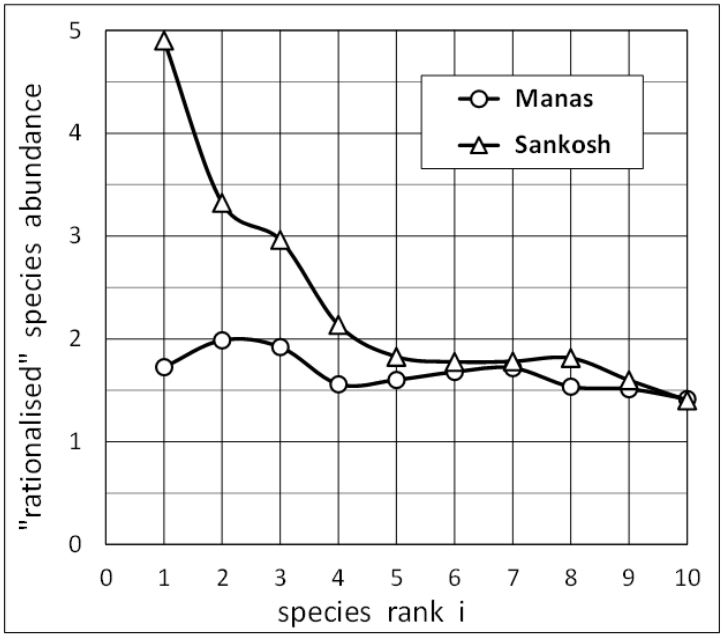

Fig. 10. The normalised abundances of the ten first (most frequent) species in the butterfly assemblages at "Manas Range" and at "Sankosh River" (Bhutan), after rationalisation of the corresponding abundances by reference to the "broken-stick" model. The rationalisation relevantly cancels the specific contribution of the difference of species richness (twice larger at "Sankosh River")

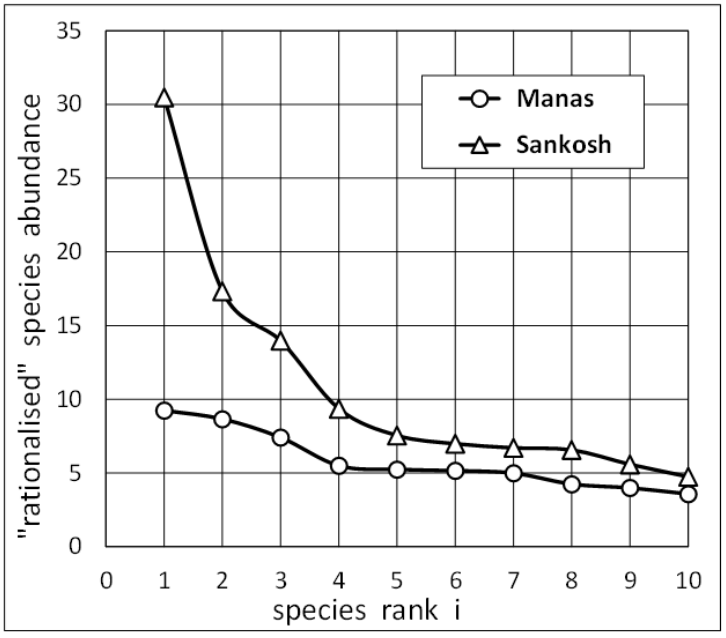

Fig. 11. The normalised abundances of the ten first (most frequent) species in the butterfly assemblages at "Manas Range" and at "Sankosh River" (Bhutan), after rationalisation of the corresponding abundances by reference to the "ideally even abundance distribution" model. The rationalisation relevantly cancels the specific contribution of the difference of species richness (twice larger at "Sankosh River")

3. DISCUSSION

Species richness is often considered as the major numerical parameter dedicated to characterize a community of species [34-37]. Yet, an ecological community is not simply a collection of species, the number of which would suffice to summarize all that can be said about this community. The particular pattern of the *species abundance distribution* within the community admittedly conveys a great deal of *additional information* about its internal structure and functionality [3-5,8-10,38].

This is why incomplete "S.A.D.s", resulting from only partial inventories, would remain deprived of a significant part of valuable data, unless they are *properly extrapolated*. And, first of all, a reliable evaluation of the number of still unrecorded species and, thereby, a reliable estimation of the total species richness of the focused assemblage of species is needed.

A relevant procedure of estimation of the number of unrecorded species is a prerequisite to the appropriate extrapolation of "S.A.D.s"

As mentioned in Introduction, Chao et al. [21] have already derived a method to extrapolate Species Abundance Distributions. The procedure advocated by these authors relies upon two main assumptions:

1) that Chao 1 estimator of the number of still unrecorded species provides relevant estimates of the true value, which, indeed, supposes either: (i) unusually even distributions of species abundances [18,39], which, unfortunately, almost never occurs in practice and / or (ii) species inventories having already reached a level of completeness close to exhaustivity [15–17]. But, here also, this second requirement is quite difficult to satisfy in practice, at least with highly species-rich assemblages which, most often, are the subject of only "rapid surveys", thus remaining substantially incomplete. Table 1 shows more precisely that "Chao" estimator becomes appropriate only when sampling completeness reaches 95% at least, a scarcely reached level in common practice. Brose et al. [15,16] even went so far as to discard the "Chao" estimator whatever the levels of sampling completeness, substituting Jackknife-1 estimator at highest completeness levels. Improper selection of the type of estimator generally results in considerable bias in the estimation of the number of unrecorded species, as shown, for example, in Fig. 12.

As the reliable estimation of the number of unrecorded species is crucial for the relevant computation of the extrapolation of "S.A.D.s", the arguments above should draw attention to what could well be a severe limitation of the range of application of the method proposed by Chao et al. [21], as, indeed, suggested by the authors themselves, in a somewhat ambiguous manner: only "when sample size is *large enough*, this lower bound approaches the true number of undetected species" [21, p. 1195].

2) that the extrapolated part of the "S.A.D.s" follows a log-linear trend which, although qualified of "natural", is far from being adequate in most circumstances [3, 23]. Indeed, models commonly taken as reference, such as log-normal, broken-stick, double-geometric series, are all ending by a more or less pronounced downwards curvature in the log-transformed representation.

Table 1. Estimated levels of sampling completeness of 37 partial inventories of species communities (34 butterfly and 3 marine gastropod inventories) and the types of nonparametric estimators of the number of unrecorded species which were selected as being the *least biased*, among a set of six nonparametric estimators: "Chao" and the five first Jackknifes, JK-1 to JK-5 (BÉGUINOT, *unpublished data*)

Sampling completeness	Chao	JK-1	JK-2	JK-3	JK-4	JK-5
> 95%	3	0	0	0	0	0
85% – 94%	0	3	1	0	0	0
75% – 84%	0	3	2	2	0	2
65% – 74%	0	0	1	4	3	3
55% – 64%	0	0	0	2	1	2
40% – 54%	0	0	0	0	0	5

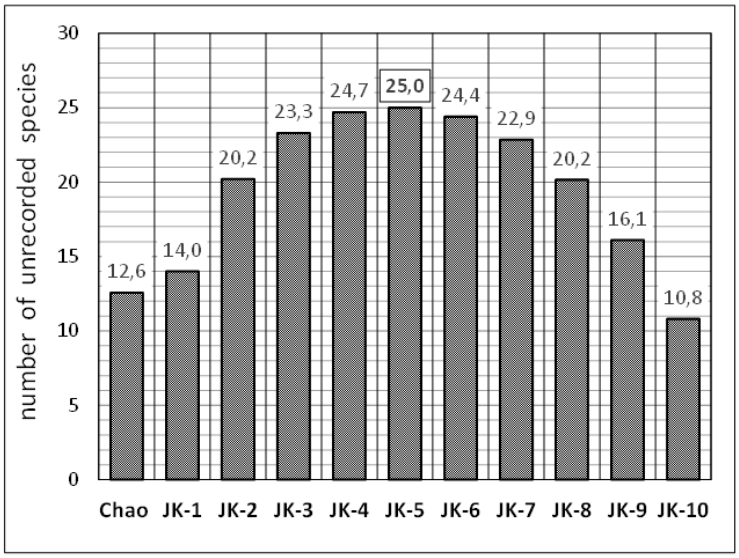

Fig. 12. The estimated number of still unrecorded species in the partial inventory of butterfly fauna at "Gwangneung Forest Biosphere Reserve", according to different types of nonparametric estimators: Chao and the ten first Jackknifes JK-1 to JK-10 (BÉGUINOT, *unpublished*). The estimates vary from simple to double according to the types of estimators, thus highlighting the interest of selecting among them which type of estimator provides the *least-biased* evaluation (in practice, the estimator providing the highest estimation, as demonstrated previously [40].

4. CONCLUSION

When dealing with partial inventories – as it becomes most often the case in practice – it is highly desirable trying to extrapolate the Species Abundance Distribution beyond its uniquely recorded part. As discussed and illustrated above, completing Species Abundance Distributions by proper extrapolation has, indeed, major implications in both descriptive and functional perspectives (pattern and process).

A first attempt in this direction was achieved by Chao et al. [21], but yet remains likely reserved to the too scarce species inventories already enjoying "large enough sampling size", that is, in fact, with completeness level 95% at least.

Otherwise, when sampling completeness is less than ≈ 95% – which, indeed, encompasses the great majority of cases in practice – another alternative approach should be considered, implying:

i. to select first of all, in each case, the *least-biased* type among classically available types of nonparametric estimators of the number of unrecorded species;
ii. then, to compute the least-biased estimation of the number of still unrecorded species and derive, accordingly, the corresponding least-biased extrapolation of the Species Accumulation Curve, associated to the selected estimator;
iii. at last, to derive, from the latter, the related extrapolation of the Species Abundance Distribution, which will thereby benefit from a minimised level of bias.

Basing the extrapolation of the Species Abundance Distribution on the previously derived extrapolation of the Species Accumulation Curve [19,20,41,42] avoids having to *arbitrarily assume* an hypothetical shape for the extrapolated section of the Species Abundance Distribution.

Accordingly, a new method has been proposed, here, to reliably extrapolate the Species Abundance Distribution beyond the already recorded part, in order to get a *complete* and *thorough* description of this Distribution. This new method benefits by a strongly enlarged domain of applicability, encompassing most of the usual range of incompleteness of partial inventories of biodiversity, includingy when dealing with very

species-rich communities. Besides, being able to derive exhaustive Species Abundance Distributions, by means of reliable extrapolations is not a final achievement in itself, as interesting as it is. Completed Species Abundance Distributions moreover open up interesting prospects for going further in the overall understanding of the processes likely instrumental in the hierarchical structuring of relative species abundances within biological communities.

ACKNOWLEDGEMENTS

Five anonymous Reviewers are gratefully acknowledged for their comments and/or suggestions.

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Cam E, Nichols JD, Sauer JR, Hines JE. On the estimation of species richness based on the accumulation of previously unrecorded species. Ecography. 2002;25: 102-108.
2. Colwell RK, Coddington JA. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society London B. 1994;345 :101-118.
3. May RM. Patterns of species abundance and diversity. In Cody ML & Diamond JM. Ecology and Evolution of Communities. The Belknap Press of Harvard University. 1975;81-120.
4. May RM. The search for patterns in the balance of nature: advances and retreats. Ecology. 1986;67(5):1115-1126.
5. Sugihara G. Minimal community structure: an explanationof species abundance patterns. The American Naturalist. 1980; 116(6):770-787.
6. Magurran AE. Ecological Diversity and its Measurement. Princeton University Press; 1988.
7. DeVries PJ, Walla TR, Greeney HF. Species diversity in spatial and temporal dimensions of fruit-feeding butterflies from two Ecuadorian rainforests. Biological Journal of the Linnean Society. 1999;68: 333-353.
8. Stevens MHH, Petchey OL, Smouse PE. Stochastic relations between species richness and the variability of species composition. Oikos. 2003;103:479-488.
9. Locey KJ, White EP. How species richness and total abundance constrain the distribution of abundance. Ecology Letters. 2013;16:1177-1185.
10. Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proceedings of the National Academy of Sciences USA. 2016;113(21):5970-5975.
11. Ulrich W, Ollik M, Ugland KI. A meta-analysis of species-abundance distributions. Oikos. 2010;119:1149-1155.
12. Rajakaruna H, Drake DAR, Chan FT, Bailey SA. Optimizing performance of nonparametric species richness estimators under constrained sampling. Ecology and Evolution. 2016;6:7311-7322.
13. Gotelli NJ, Colwell RK. Estimating species richness. Biological Diversity: Frontiers in measurement and assessment. AE Magurran and BJ McGill (eds.). Oxford University Press, Oxford. 2010;345:39-54.
14. Gotelli NJ, Chao A. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: Levin S.A. (ed.) Encyclopedia of Biodiversity, second edition. 2013;5:195-211. Waltham, MA: Academic Press.
15. Brose U, Martinez ND, Williams RJ. Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. Ecology. 2003;84(9): 2364-2377.
16. Brose U, Martinez ND. Estimating the richness of species with variable mobility. Oikos. 2004;105:292-300.
17. Reese GC, Wilson KR, Flather CH. Performance of species richness estimators across assemblage types and survey parameters. Global Ecology and Biogeography. 2014;23:585-594.
18. Chiu CH, Wang YT, Walther BA & Chao A. An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula. Biometrics. 2014;70(3).
DOI : 10.1111/biom.12200
19. Béguinot J. Theoretical derivation of a bias-reduced expression for the extrapolation of the Species Accumulation Curve and the associated estimation of total species richness. Advances in Research. 2016;7(3):1-16.

DOI:10.9734/AIR/2016/26387;<hal-01367803>

20. Béguinot J. Extrapolation of the Species Accumulation Curve associated to "Chao" estimator of the number of unrecorded species: A mathematically consistent derivation. Annual Research & Review in Biology. 2016;11(4):1-19.
DOI:10.9734/ARRB/2016/30522;<hal 01477263 >
21. Chao A, Hsieh TC, Chazdon RL, Colwell RK, Gotelli NJ. Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. Ecology. 2015;96(5):1189-1201.
22. Chao A, Colwell RK, Lin CW, Gotelli NJ. Sufficient sampling for asymptotic minimum species richness estimators. Ecology. 2009;90(4):1125-1133.
23. McGill BJ, Etienne RS, Gray JS et al. Species abundance distributions: Moving beyond single prediction theories to integration within an ecological framework. Ecology Letters. 2007;10:995-1015.
24. Alroy J. The shape of terrestrial abundance distributions. Science Advances. 2015;1:e1500083:2-8.
25. Baldridge E, Harris DJ, Xiao X, White EP. An extensive comparison of species-abundance distribution models. PeerJ. 2016;4:e2823.
26. MacArthur RH. On the relative abundance of bird species. Proceedings of the National Academy of Sciences U.S.A. 1957;43:293-295.
27. Wilson JB. Would we recognise a Broken-Stick community if we found one? Oikos. 1993;67(1):181-183
28. Komonen A, Elo M. Ecological response hides behind the species abundance distribution: community response to low-intensity disturbance in managed grasslands. Ecology and Evolution. 2017; 7:8558-8566.
29. MacDonald ZG, Nielsen SE, Acorn JH. Negative relationships between species richness and evenness render common diversity indices inadequate for assessing long-term trends in butterfly diversity. Biodiversity Conservation. 2017;26:617-629.
30. Nidup T, Dorji T, Tshering U. Taxon diversity of butterflies in different habitat types in Royal Manas National Park. Journal of Entomology and Zoology Studies. 2014;2(6):292-298.
31. Béguinot J, Nidup T. Least-biased extrapolation of a partial inventory of butterfly fauna in Manas Range (Royal Manas National Park, Bhutan). Asian Journal of Environment & Ecology. 2017; 2(2):1-14.
DOI: 10.9734/AJEE/2017/32701
32. Singh AP. Lowland forest butterflies of the Sankosh River catchment, Bhutan. Journal of Threatened Taxa. 2012;4(12):3085-3102.
33. Béguinot J. "Pollard-walk" butterfly survey does not warrant the equity of sampling-completeness among butterfly families: a case study with tropical butterfly fauna in Bhutan. Journal of Applied Life Science International. 2017;12(2):1-12.
DOI: 10.9734/JALSI/2017/34058
34. Williams PH, Gaston KJ. Measuring more of biodiversity: Can higher-taxon richness predict wholesale species richness? Biological Conservation. 1994;67(3):211-217.
35. Pogue MG. Preliminary estimates of Lepidoptera diversity from specific sites in the Neotropics using complementarity and species richness estimators. Journal of the Lepidopterists Society. 1999;53(2):55-71.
36. Stirling G, Wilsey B. Empirical relationships between species richness, evenness and proportional diversity. The American Naturalist. 2001;158(3):286-299.
37. Garcia D, Martinez D. Species richness matters for the quality of ecosystem services: A test using seed dispersal by frugivorous birds. Proceedings of the Royal Society B. 2012;279:3106–3113.
38. Locey KJ, Lennon JT. Powerful predictions of biodiversity from ecological and scaling models. Proceedings of the National Academy of Sciences USA. 2016; 113(35):5097.
39. Chao A, Colwell RK. Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. Sort. 2017; 41(1):3-54.
40. Béguinot J. Extrapolation of total species richness from incomplete inventories: application to the Gastropod fauna associated to coral reefs in 'Mannar Gulf Biosphere Reserve', India. Asian Journal of Environment and Ecology. 2017;4(3):1-14.
DOI: 109734/AJEE/2017/36831

41. Béguinot J. Basic theoretical arguments advocating Jackknife-2 as usually being the most appropriate nonparametric estimator of total species richness. Annual Research & Review in Biology. 2016; 10(1):1-12.
DOI:10.9734/ARRB/2016/25104>;<hal-01300828>
42. Béguinot J. On general mathematical constraints applying to the kinetics of species discovery during progressive sampling: consequences on the theoretical expression of the Species Accumulation Curve. Advances in Research. 2016;8(5): 1-17.
DOI:10.9734/AIR/2016/31791.<hal-01516141>.
43. Good IJ. The population frequencies of species and the estimation of population parameters. Biometrika. 1953;40: 237-264.
44. Béguinot J. An algebraic derivation of Chao's estimator of the number of species in a community highlights the condition allowing Chao to deliver centered estimates. ISRN Ecology. 2014;6. Article ID: 847328.
DOI:10.1155/2014/847328 ;<hal-01101415>
45. Béguinot J. When reasonably stop sampling? How to estimate the gain in newly recorded species according to the degree of supplementary sampling effort. Annual Research & Review in Biology. 2015;7(5):300-308.
DOI:10.9734/ARRB/2015/18809;<hal-01228695>

APPENDIX 1

Preliminary: the sum of the abundances of the unrecorded species

According to the two first equations of Appendix 1 in [19, 20, 42]:

- the expected number $\Delta_{(N)}$ of still unrecorded species in a sample of size N is:

$$\Delta_{(N)} = \Sigma_i (1-p_i)^N \quad \text{(A1.1)}$$

where p_i is the proportional abundance (identified to the probability of drawing during sampling) of species 'i' and Σ_i is the summation extended to the totality of the 'S_t' species 'i' present in the sampled assemblage;

-the expected number f_x of species recorded *x times* in a sample of size N, is:

$$f_x = [N!/x!/(N-x)!] \Sigma_i [(1-p_i)^{N-x} p_i^x]$$

and accordingly, for x = 1, the expected number f_1 of singletons is:

$$f_1 = N. \Sigma_i [(1-p_i)^{N-1} p_i] \quad \text{(A1.2)}$$

Now, the expected value $A_{u(N)}$ of the cumulated abundances of the $\Delta_{(N)}$ *still unrecorded* species in a sample of size N is:

$$A_{u(N)} = \Sigma_i [p_i.(1-p_i)^N] \quad \text{(A1.3)}$$

Accordingly,

$$f_1 = N.A_{u(N-1)} \quad \text{(A1.4)}$$

As, in practice, samplings of interest have sizes N (N recorded individuals) considerably larger than the species richness of the sampled assemblage, the difference between $\Delta_{(N)}$ and $\Delta_{(N-1)}$ is quite negligible, as also negligible is the difference between $A_{u(N)}$ and $A_{u(N-1)}$, so that, with a very good approximation:

$$A_{u(N)} = f_1/N \quad \text{(A1.5)}$$

Accordingly, the cumulated abundances, $A_{r(N)}$, of the *already recorded* species is the complement to 1 of $A_{u(N)}$, that is:

$$A_{r(N)} = 1 - f_1/N \quad \text{(A1.6)}$$

This general relationship was originally derived by A. TURING [43].

Correction to be applied when estimating the true abundance of species, based on the corresponding recorded frequency of occurrence, in a sample of finite size

Consider a sample of size N (N recorded individuals) with R(N) recorded species among which a number f_1 of them are singletons (species recorded only once). Let $p_i = n_i/N$ be the frequency of occurrence of species 'i', and let 'a_i' be the true proportional abundance of this species in the sampled community. A way to evaluate the bias of the recorded frequency p_i relative to the corresponding true abundance a_i is to consider a Bayesian inference based on the binomial distribution. Accordingly, the probability '$\partial\pi_i$' that the abundance of species 'i' is comprised between 'a' and 'a + ∂a' is:

$$\partial\pi_i(a)/\partial a = (N+1).C(N, n_i).a^{ni}.(1–a)^{(N–ni)} \quad \text{(A1.7)}$$

with: $C(N, n_i) = N!/(N–n_i)!/n_i!$

The probability $\pi_i(a)$ reaches its maximum (modal) value for $a = p_i$ (= n_i/N), as is easily demonstrated. And the average value $\tilde{a}_i$ of a_i, which will be considered as providing the least-biased evaluation of the true abundance a_i,, is computed as follows.

In a first step, and accounting for equation (A1.7), $\tilde{a}_i$ is identified to:

$$\int a.\partial\pi_i(a) = \int a.(\partial\pi_i(a)/\partial a)\, da = (N+1).C(N, n_i).\int a^{(ni+1)}.(1–a)^{(N–ni)})\, da \quad (A1.8)$$

with the integral $\int$ extended from a = 0 to a = 1.

The integral may be resolved through successive iterations, accounting for the general classical formula:

$$\int x^m.(1–x)^n\, dx = – x^m.(1–x)^{(n+1)}/(n+m+1) + m/(n+m+1).\int x^{(m–1)}.(1–x)^n\, dx$$

with, for m = 1: $\int x.(1–x)^n\, dx = – (1–x)^{(n+1)}.(x.(n+1)+1)/(n+1)/(n+2)$ (A1.9)

which finally yields:

$$\int a^2.(1–a)^{(N–1)}\, da = 2/(N+1)/(N+2)^2$$

$$\int a^3.(1–a)^{(N–2)}\, da = 6/(N+1)/(N+2)^3$$

$$\int a^4.(1–a)^{(N–3)}\, da = 24/(N+1)/(N+2)^4$$

…….

$$\int a^{(ni+1)}.(1–a)^{(N–ni)})\, da = (n_i+1)!\,/(N+1)/(N+2)^{(ni+1)} \quad (A1.10)$$

As, at this first step of computation, $\tilde{a}_i$ is identified to $(N+1).C(N, n_i).\int a^{(ni+1)}.(1–a)^{(N–ni)})\, da$, it comes :

$$\tilde{a}_i = (N+1).N!/(N-n_i)!/n_i!.[(n_i+1)!/(N+1)/(N+2)^{(ni+1)}] \quad (A1.11)$$

$$\tilde{a}_i = (n_i+1).N!/(N-n_i)!/(N+2)^{(ni+1)}$$

$$\tilde{a}_i \approx (n_i+1)/N \quad (A1.12)$$

since sampling size N of interest are always far larger than n_i.

Now, according to equation (A1.6), the sum of abundances of the recorded species is $(1 – f_1/N)$, with f_1 as the recorded number of singletons (species recorded only once):

$$\Sigma\, \tilde{a}_i \approx \Sigma\, (n_i+1)/N = (1 – f_1/N)$$

Now, $\Sigma\, (n_i+1)/N = [\Sigma\, (n_i) + \Sigma\, (1)] = [N + R]/N = (1 + R/N)$, with R as the number of recorded species. It then follows that a standardisation coefficient $(1 – f_1/N)/(1 + R/N)$ is to be applied to the preceding first step evaluation of $\tilde{a}_i$. In the frame of Bayesian approach, this standardisation coefficient corresponds to the initial setting of the so called "probabilities *a priori*").

Finally, the expression of $\tilde{a}_i$ is thus:

$$\tilde{a}_i = (n_i/N+1/N).(1–f_1/N)/(1+R/N) \quad (A1.13)$$

and, accounting for $p_i = n_i/N$, it comes:

$$\tilde{a}_i = p_i.(1+1/n_i).(1–f_1/N)/(1+R/N) \quad (A1.14)$$

The bias correction applied to p_i, to obtain the true abundance estimates $\tilde{a}_i$, thus includes: (i) the correction $(1+1/n_i)/(1+R/N)$ for the bias resulting from the finite size N of the sample, a bias which cancels, as expected, when N (and thus also $n_i = N.p_i$) tend to infinity; (ii) the correction $(1–f_1/N)$ resulting from the existence of the set of still unrecorded species, which cancels, as expected, when sampling reaches exhaustivity, that is when f_1 is falling down to zero.

Note that the estimated true abundances are less scattered than are the recorded frequencies. Indeed, considering the ratio of estimated abundances, $\tilde{a}_i/\tilde{a}_j$, between two species 'i' and 'j', it comes:

$$\tilde{a}_i/\tilde{a}_j = (p_i/p_j).(1+1/n_i)/(1+1/n_j) \qquad \text{(A1.15)}$$

which confirms that, if $p_i/p_j > 1$ then $\tilde{a}_i/\tilde{a}_j < (p_i/p_j)$ and, reciprocally, if $p_i/p_j < 1$ then $\tilde{a}_i/\tilde{a}_j > (p_i/p_j)$.

APPENDIX 2

Bias-reduced extrapolation of the Species Accumulation Curve and the associated bias-reduced estimation of the number of missing species, based on the recorded numbers of species occurring 1 to 5 times

Consider the survey of an assemblage of species of size N_0 (with sampling effort N_0 typically identified either to the number of recorded individuals or to the number of sampled sites, according to the inventory being in terms of either species abundances or species incidences), including $R(N_0)$ species among which f_1, f_2, f_3, f_4, f_5, of them are recorded 1, 2, 3, 4, 5 times respectively. The following procedure, designed to select the less-biased solution, results from a general mathematical relationship that constrains the theoretical expression of *any* theoretical Species Accumulation Curves R(N) (see [19, 44,45]):

$$\partial^x R_{(N)}/\partial N^x \;=\; (-1)^{(x-1)}\, f_{x(N)}\,/C_{N,\,x} \;\approx\; (-1)^{(x-1)}\,(x!/N^x)\, f_{x(N)} \quad (\approx \text{as } N >> x) \qquad (A2.1)$$

Compliance with the mathematical constraint (equation (A2.1)) warrants a *reduced-bias* expression for the extrapolation of the Species Accumulation Curves R(N) (i.e. for $N > N_0$). Below are provided, accordingly, the polynomial solutions $R_{\mathbf{x}}(N)$ that respectively satisfy this mathematical constraint, considering increasing orders x of derivation $\partial^x R_{(N)}/\partial N^x$. Each solution $R_{\mathbf{x}}(N)$ is appropriate for a given range of values of f_1 compared to the other numbers f_x (according to reference [19]):

* *for* f_1 *up to* f_2 → $R_{\mathbf{1}}(N) = (R(N_0) + f_1) - f_1.N_0/N$

* *for* f_1 *up to* $2f_2 - f_3$ → $R_{\mathbf{2}}(N) = (R(N_0) + 2f_1 - f_2) - (3f_1 - 2f_2).N_0/N - (f_2 - f_1).N_0^2/N^2$

* *for* f_1 *up to* $3f_2 - 3f_3 + f_4$ → $R_{\mathbf{3}}(N) = (R(N_0) + 3f_1 - 3f_2 + f_3) - (6f_1 - 8f_2 + 3f_3).N_0/N - (-4f_1 + 7f_2 - 3f_3).N_0^2/N^2 - (f_1 - 2f_2 + f_3).N_0^3/N^3$

* *for* f_1 *up to* $4f_2 - 6f_3 + 4f_4 - f_5$ → $R_{\mathbf{4}}(N) = (R(N_0) + 4f_1 - 6f_2 + 4f_3 - f_4) - (10f_1 - 20f_2 + 15f_3 - 4f_4).N_0/N - (-10f_1 + 25f_2 - 21f_3 + 6f_4).N_0^2/N^2 - (5f_1 - 14f_2 + 13f_3 - 4f_4).N_0^3/N^3 - (-f_1 + 3f_2 - 3f_3 + f_4).N_0^4/N^4$

* *for* f_1 *larger than* $4f_2 - 6f_3 + 4f_4 - f_5$ → $R_{\mathbf{5}}(N) = (R(N_0) + 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5) - (15f_1 - 40f_2 + 45f_3 - 24f_4 + 5f_5).N_0/N - (-20f_1 + 65f_2 - 81f_3 + 46f_4 - 10f_5).N_0^2/N^2 - (15f_1 - 54f_2 + 73f_3 - 44f_4 + 10f_5).N_0^3/N^3 - (-6f_1 + 23f_2 - 33f_3 + 21f_4 - 5f_5).N_0^4/N^4 - (f_1 - 4f_2 + 6f_3 - 4f_4 + f_5).N_0^5/N^5$

The associated non-parametric estimators of the number Δ_J of missing species in the sample [with $\Delta_J = R(N = \infty) - R(N_0)$] are derived immediately:

* $0.6\, f_2 < \mathbf{f_1} \leq f_2$ → $\Delta_{J1} = f_1$; $R_{\mathbf{1}}(N)$

* $f_2 < \mathbf{f_1} \leq 2f_2 - f_3$ → $\Delta_{J2} = 2f_1 - f_2$; $R_{\mathbf{2}}(N)$

* $2f_2 - f_3 < \mathbf{f_1} \leq 3f_2 - 3f_3 + f_4$ → $\Delta_{J3} = 3f_1 - 3f_2 + f_3$; $R_{\mathbf{3}}(N)$

* $3f_2 - 3f_3 + f_4 < \mathbf{f_1} \leq 4f_2 - 6f_3 + 4f_4 - f_5$ → $\Delta_{J4} = 4f_1 - 6f_2 + 4f_3 - f_4$; $R_{\mathbf{4}}(N)$

* $\mathbf{f_1} > 4f_2 - 6f_3 + 4f_4 - f_5$ → $\Delta_{J5} = 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5$; $R_{\mathbf{5}}(N)$

N.B. 1: As indicated above (and demonstrated in details in [19], this series of inequalities define the ranges that are best appropriate, respectively, to the use of each of the five Jackknife estimators, JK-1 to JK-5. That is the respective ranges within which each estimator will benefit of minimal bias for the predicted number of missing species.

Besides, it is easy to verify that another consequence of these preferred ranges is that the selected estimator will *always* provide the *highest* estimate, as compared to the other estimators. Interestingly, this mathematical consequence, of general relevance, is in line with the already admitted opinion that all non-parametric estimators provide *under*-estimates of the true number of missing species [13,14], so that the least-biased estimator is expected to be the one providing the highest estimate. Also, this shows that the approach initially proposed by Brose et al. [15] – which has regrettably suffered from its somewhat difficult implementation in practice – might be advantageously reconsidered, now, in light of the very simple selection key above, of *far much easier practical use*.

N.B. 2: In order to reduce the influence of drawing stochasticity on the values of the f_x, the as-recorded distribution of the f_x should preferably be smoothened: this may be obtained either by rarefaction processing or by regression of the as-recorded distribution of the f_x versus x.

N.B. 3: For f_1 falling beneath 0.6 x f_2 (that is when sampling completeness closely approaches exhaustivity), then Chao estimator may be selected: see reference [20].

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sciencedomain.org/review-history/23066