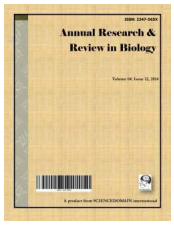
Extrapolation of the species accumulation curve associated to "Chao" estimator of the number of unrecorded species: a mathematically consistent derivation.

Jean Béguinot



Annual Research & Review in Biology
11(4): 1-19, 2016; Article no.ARRB.30522
ISSN: 2347-565X, NLM ID: 101632869

SCIENCEDOMAIN *international*
www.sciencedomain.org

Extrapolation of the Species Accumulation Curve Associated to "Chao" Estimator of the Number of Unrecorded Species: A Mathematically Consistent Derivation

Jean Béguinot[1*]

[1]Department of Biogéosciences, Université de Bourgogne, F 21000 – Dijon, France.

Author's contribution

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

Article Information

DOI: 10.9734/ARRB/2016/30522
Editor(s):
(1) Jin-Zhi Zhang, College of Horticulture and Forestry Science, Huazhong Agricultural University, China.
(2) George Perry, Dean and Professor of Biology, University of Texas at San Antonio, USA.
Reviewers:
(1) Loc Nguyen, Sunflower Soft Company, Vietnam.
(2) Surendra S Bargali, Kumaun University, India.
(3) Suntud Sirianutnapiboon, King Mongkut's University of Technology Thonburi, Thailand.
(4) Manoel Fernando Demétrio, Universidade Federal da Grande Dourados, Brazil.
Complete Peer review History: http://www.sciencedomain.org/review-history/17184

Method Article

Received 15th November 2016
Accepted 5th December 2016
Published 9th December 2016

ABSTRACT

Incomplete samplings are doomed to become common practice for many inventories of biodiversity, thereby inviting to *extrapolate* what the rate of accumulation of newly recorded species would be if sampling was to be continued any further. For this purpose, a new derivation is provided for the extrapolation of the Species Accumulation Curve associated to the "Chao" estimator of the number of unrecorded species. This new derivation *strictly complies* with the general mathematical relationship constraining the shape of any expression of the Species Accumulation Curve, while the extrapolation previously proposed by Chao & Chiu [1] does not. The mathematically relevant formulation for the extrapolation R(N) of the Species Accumulation Curve associated to "Chao" estimator is thus: $R(N) = R(N_0) + [f_1^2/(2\ f_2)](1 - \exp[-(2f_2/f_1/N_0)\ .\ (N - N_0)])$, with N_0 as the actual sample size, f_1 and f_2 as the numbers of species actually recorded once and twice and R(N) as the extrapolated number of species expected to be recorded as a function of sample size N ($N > N_0$). Accounting for the constraining relationship mentioned above is also essential in another respect: it

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

allows to extrapolate *separately* the numbers of species expected to be recorded 0-, 1-, 2-, >2-times, thereby permitting to analyse rationally the process of species accumulation during continuously growing sampling. At last, the preferred range of applicability of both "Chao" estimator and the associated extrapolation of the Species Accumulation Curve estimator is discussed, by comparison with the alternative type estimator Jackknife-2.

Keywords: Chao estimator; extrapolation; species accumulation curve; Jackknife; incomplete sampling; species richness.

1. INTRODUCTION

Incomplete inventories of biodiversity are likely doomed to become increasingly frequent, as surveys progressively address new taxonomic groups more difficult to cope with, in particular those groups giving rise to species assemblages with high number of species made of tiny individuals, such as, for example, small or micro-invertebrates. In addition, more commonly investigated taxonomic groups, as well, are likely doomed to remain more or less incompletely surveyed at the *local scale*, due to sampling efforts often being far less at these small scales than they usually are in larger areas [2,3].

Incomplete samplings raise two important questions, of high practical relevance:

- how many "missing" species might be left unrecorded by the incomplete sampling and, accordingly, what would be the estimated total species richness of the assemblage (that is the expected number of recorded species if the sampling was ideally complete) ;
- what would be the extrapolated shape of the so called "Species Accumulation Curve", beyond the currently achieved sampling-size, that is, how the rate of discovery of new species would vary with increasing sampling size, beyond the currently achieved sampling-size. A major practical interest of extrapolating the species accumulation curve being the possibility to predict quantitatively the level of additional sampling effort that would be required to obtain any desired increment of sampling completeness. In other words, extrapolation offers the possibility to gauge the ratio between the expected gain in newly recorded species and the corresponding additive sampling effort needed.

Regarding the first issue – *the estimation of the number Δ of missing (unrecorded) species* – a lot of non-parametric estimators have been proposed during the last decades (reviewed in [4,5]). All these estimators are based upon the numbers f_x of species recorded x-times during the considered incomplete sampling, especially the two first numbers, f_1 and f_2. Among the more commonly implemented non-parametric estimators are: (i) the "*Chao*" estimator ($\Delta_{Ch} = f_1^2/(2\ f_2)$) and (ii) the series of "*Jackknife*" estimators at different orders. Jackknife estimators are more commonly implemented at orders 1 and 2 (i. e. $\Delta_{J1} = f_1$ and $\Delta_{J2} = 2f_1 - f_2$) but higher orders, up to 5, should be considered however, when samplings are substantially incomplete [6,7].

Now, regarding the second issue - *the extrapolation of the species accumulation curve* - a series of parametric models are classically considered (reviewed in [8]). These models are expected to fit more or less the main common feature of species accumulation curves considered as a whole (that is: species accumulation rate monotonically decreasing with additional sampling efforts, finally slowing to zero when total species richness is reached). Yet, none of these formal models have direct relevance to the process of species accumulation itself, during progressive sampling and, accordingly, none of these models explicitly satisfy the general mathematical relationship (equation (1)) that systematically constrains any kind of species accumulation curves.

In fact, as might have been expected, it has been previously demonstrated that a *specific expression* of the extrapolation of the species accumulation curve is associated to *each type of non-parametric estimator* of the number of unrecorded species [5,7,9]. This is so because both the number of unrecorded species and the shape of the species accumulation curve are jointly dependent upon a *same* cause: the particular Distribution of Species Abundances (the so-called "S.A.D.") within the sampled assemblage of species (as has been already

suggested implicitly in [6]). More specifically, this linkage between the type of estimator of the number of unrecorded species and the expression of the extrapolated species accumulation curve is precisely ruled by the constraining mathematical relationship mentioned above (equation (1)).

2. EXTRAPOLATION OF THE SPECIES ACCUMULATION CURVE ASSOCIATED TO THE "CHAO" ESTIMATOR OF THE NUMBER OF UNRECORDED SPECIES

* a new derivation, complying with the mathematical requirements constraining the shape of any Species Accumulation Curve

The successive derivatives, $\partial^x\Delta(N)/\partial N^x$, of the number $\Delta(N)$ of species expected to remain unrecorded after a sampling of size N are respectively related to the numbers, $f_{x\ (N)}$, of species recorded x-times during this sampling of size N:

$$[\partial^x \Delta_{(N)}/\partial N^x] = (-1)^x f_{x\ (N)}/C_{N,\,x} \quad (1)$$

with $C_{N,\,x} = N!/X!/(N-x)!$. A detailed proof of this general theorem is given in Appendix.

Leaving aside the very beginning of sampling (of no practical relevance), the sampling size N rapidly exceeds widely the numbers x of practical concern, so that, in practice, the preceding equation simplifies as:

$$\partial^x \Delta_{(N)}/\partial N^x = (-1)^x (x!/N^x)\, f_{x\ (N)} \quad (2)$$

In particular,

$$\partial\Delta_{(N)}/\partial N = -f_{1\ (N)}/N \quad (3)$$

$$\partial^2 \Delta_{(N)}/\partial N^2 = 2\, f_{2\ (N)}/N^2 \quad (4)$$

These relations have *general relevance* because their derivation does not require any specific assumption relative to the particular shape of the distribution of species abundances ("S.A.D.") in the sampled assemblage of species. Accordingly, the general equation (2) and its successive forms (3), (4),… actually constrain any theoretical form of Species Accumulation Curves.

Let now focus upon the case of the "Chao" estimator of the number of missing (still unrecorded) species in a sample of size N:

$$\Delta_{(N)} = (f_{1\ (N)})^2/(2\, f_{2\ (N)}) \quad (5)$$

Applying the general relation (2) and its particular consequences (3) and (4) to the definition (5) of the "Chao" estimator yields:

$$\Delta_{(N)} = (f_{1\ (N)})^2/(2\, f_{2\ (N)}) = (\partial\Delta_{(N)}/\partial N)^2/(\partial^2 \Delta_{(N)}/\partial N^2) \quad (6)$$

The general solution of this differential equation (6) is:

$$\Delta_{(N)} = k'.\exp(k.N)$$

with k and k' as constants, independent of N.

Now, let f_1 and f_2 be the numbers of species recorded once and twice in the actually realised sampling of size N_0 (that is: $f_1 = f_{1\ (No)}$ and $f_2 = f_{2\ (No)}$). Then, according to "Chao" estimator, Δ_0 (= $\Delta_{(No)}$) = $f_1^2/(2f_2)$ is the estimated number of unrecorded species after a sampling of size N_0. Accordingly, $\Delta_{(No)} = \Delta_0$ requires $k' = \Delta_0.\exp(-kN_0)$. And satisfying equations (3) and (4) implies $k = -2f_2/(N_0.f_1)$.

For samplings of sizes greater than the size N_0 of the actually realised sample, the expression of the extrapolation of the Species Accumulation Curve specifically associated to "Chao" type estimator is thus:

$$\Delta_{(N)} = \Delta_0.\exp[-2f_2/(N_0.f_1).(N - N_0)] \quad (7)$$

or, as well, accounting for $\Delta_0 = f_1^2/(2\, f_2)$:

$$\Delta_{(N)} = [f_1^2/(2\, f_2)]\, \exp[-(2f_2/f_1).(N/N_0 - 1)] \quad (8)$$

Thus, the extrapolation of the species accumulation curve, $R(N)$ [= $R(N_0) + \Delta_{(N0)} - \Delta_{(N)}$] takes the following form, when associated to "Chao" type estimator:

$$R(N) = R(N_0) + [f_1^2/(2f_2)](1 - \exp[-(2f_2/f_1).(N/N_0 - 1)]) \quad (9)$$

* comparison with previous formulations of the extrapolation of the Species Accumulation Curves associated to "Chao" estimator

An extrapolation of the Species Accumulation Curve specifically associated to "Chao" type estimator was previously proposed by Chao & Chiu [1]. This formulation (their equation (9)), converted in our own notations, is:

$$R(N) = R(N_0) + \Delta_0.(1 - [1- f_1/(N_0.\Delta_0 + f_1)]^{(N - N0)}) \quad (10)$$

As may easily be verified, this expression is formally different from expression (9) derived above and, as such, does not satisfy – as it should do – the relationships (3) and (4) which actually constrain all kinds of Species Accumulation Curve. Indeed, following equation (10):

$$\Delta_{(N)} = \Delta_0.[1- f_1/(N_0.\Delta_0 + f_1)]^{(N - N0)}$$

and thus, the first derivative of $\Delta_{(N)}$ at $N = N_0$ is:

$$\partial\Delta_{(N)}/\partial N]_{N0} = -\Delta_0.\ln[1- f_1/(N_0.\Delta_0 + f_1)]$$

which formally differs from the required value, $-f_1/N_0$, given by equation (3).

Although this non-compliance with mathematical requirements has relatively limited quantitative consequences in practice, it does remain unsatisfactory on theoretical ground. At last, another expression for the extrapolation of the Species Accumulation Curve associated to "Chao" estimator has been formerly proposed [5,7], which, also, does note cope correctly with equations (3) and (4) and, for this reason, should thus be discarded.

3. SEPARATE EXTRAPOLATIONS OF THE NUMBERS OF SPECIES EXPECTED TO BE RECORDED ONCE, TWICE & MORE THAN TWICE, ACCORDING TO "CHAO" ESTIMATOR

The number R(N) of recorded species is, of course, nothing else that the sum of the numbers $f_{1\ (N)}$, $f_{2\ (N)}$, $f_{3\ (N)}$, …, $f_{x\ (N)}$,… of those species respectively recorded 1-, 2- ,3-,…, x-times…

Accordingly, the evolution of the number R(N) of recorded species with sample size N has a *complex determinism*, resulting from the additive contributions of all the $f_{x\ (N)}$, each of them having its own pattern of evolution with increasing sampling size N. Disentangling these respective contributions may thus shed some light on the complex mechanism underlying the evolution with N of the number R(N) of recorded species.

For this purpose, it is necessary to consider separately the extrapolations of each of the numbers $f_{x\ (N)}$. And, precisely, this is made possible thanks to considering the general mathematical relationship (1) (and the associated equations (3) & (4)).

Here, I shall consider the separate extrapolations of the numbers $f_{1\ (N)}$, $f_{2\ (N)}$, $f_{>2\ (N)}$, which, altogether, govern the evolution of R(N) with increasing sample size N:

$$R(N) = \Sigma_x\,[f_{x\ (N)}] = f_{1\ (N)} + f_{2\ (N)} + f_{>2\ (N)}$$

In the specific context of implementation of the estimator "Chao" and, thus, in accordance with equations (3) and (8):

$$f_{1(N)}/N = [f_1^2/(2f_2)].\exp[-(2f_2/f_1).(N/N_0 - 1)].(2f_2/f_1/N_0)$$

that is:

$$f_{1\ (N)} = (N/N_0).f_1.\exp[-(2f_2/f_1).(N/N_0 - 1)] \quad (11)$$

Also, according to equations (4) and (8):

$$2\ f_{2\ (N)}/N^2 = [f_1^2/(2f_2)].\exp[-(2f_2/f_1).(N/N_0 - 1)].(4f_2^2/f_1^2/N_0^2)$$

that is:

$$f_{2\ (N)} = (N/N_0)^2.f_2.\exp[-(2f_2/f_1).(N/N_0 - 1)] \quad (12)$$

At last, the number of species expected to be recorded more than twice, $f_{>2\ (N)}$, is:

$$f_{>2\ (N)} = R(N) - f_{1\ (N)} - f_{2\ (N)}$$

that is, according to equations (9), (11), (12):

$$f_{>2\ (N)} = R(N_0) + [f_1^2/(2\ f_2)](1 - \exp[-(2f_2/f_1).(N/N_0 - 1)]) - (N/N_0).f_1.\exp[-(2f_2/f_1).(N/N_0 - 1)] - (N/N_0)^2.f_2.\exp[-(2f_2/f_1).(N/N_0 - 1)]$$

that is:

$$f_{>2\ (N)} = R(N_0) + f_1^2/(2\ f_2) - [f_1^2/(2\ f_2)+(N/N_0).f_1+(N/N_0)^2.f_2].\exp[-(2f_2/f_1).(N/N_0 - 1)] \quad (13)$$

For the sake of generality, the contribution of the number of actually recorded species, $R(N_0)$, may be cancelled by considering $f_{>2\ (N)} - f_{>2\ (No)}$ instead of $f_{>2\ (N)}$. It follows:

$$f_{>2(N)} - f_{>2\,(No)} = [f_1^2/(2\,f_2) + f_1 + f_2] - [f_1^2/(2f_2) + (N/N_0).f_1 + (N/N_0)^2.f_2].\exp[-(2f_2/f_1).(N/N_0 - 1)] \quad (14)$$

Equations (8), (11), (12), (14), thus rule, respectively, the extrapolations of the numbers $f_{0\,(N)}$ (= $\Delta_{(N)}$), $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{>2\,(N)} - f_{>2\,(No)}$, of species expected to be recorded zero, once, twice and the increment of the number of species recorded more than twice, if the sampling effort was further continued beyond the actual inventory (sample size N_0).

Figs. 1 to 5 provide representative illustrations of the different patterns of variations of each of these numbers { $f_{0\,(N)}$, $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{>2\,(N)}$–$f_{>2\,(No)}$ } for different ratios f_1/f_2 of the numbers of species recorded once and twice in the actually achieved sample. Namely: f_1/f_2 = 8.3, 5.0, 2.5, 1.0, 0.5 (= 25/3, 25/5, 25/10, 25/25, 25/50, respectively). Thereby, these figures highlight the underlying complex process by which the number of recorded species, $R(N) = \Sigma_x\,[f_{x\,(N)}]$, steadily increases with sampling size N, as the result of the cumulative contributions of $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{>2\,(N)}$. A very complex process indeed, since the contribution of each of the $f_{x\,(N)}$ successively increases, then decreases, *at its own pace and out of phase*.

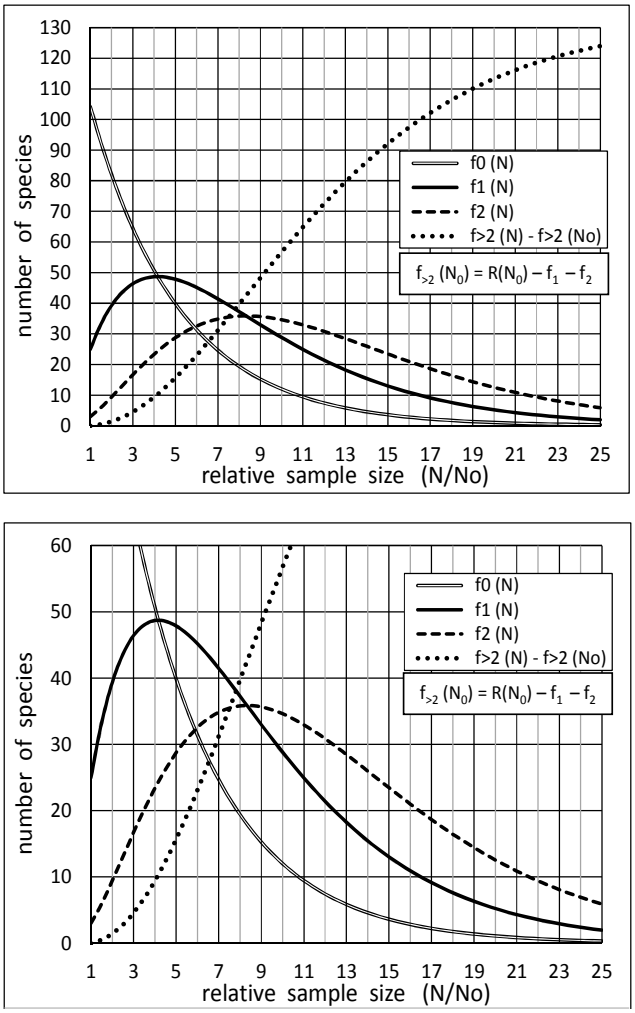
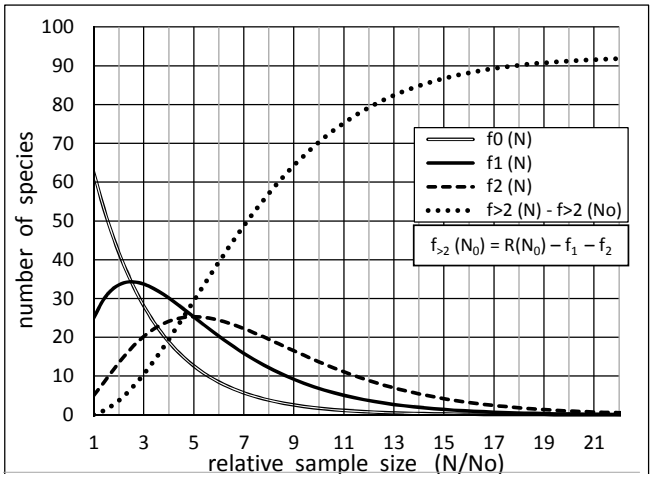


Fig. 1 and 1bis. Extrapolations of the numbers $f_{0\,(N)}$, $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{>2\,(N)} - f_{>2\,(No)}$ of species expected to be recorded zero, once, twice and more than twice, as the sampling effort is going on further beyond the actual inventory size N_0. Here, f_1/f_2 (= $f_{1\,(No)}/f_{2\,(No)}$) = 8.3 (in fact, f_1 = 25 and f_2 = 3)



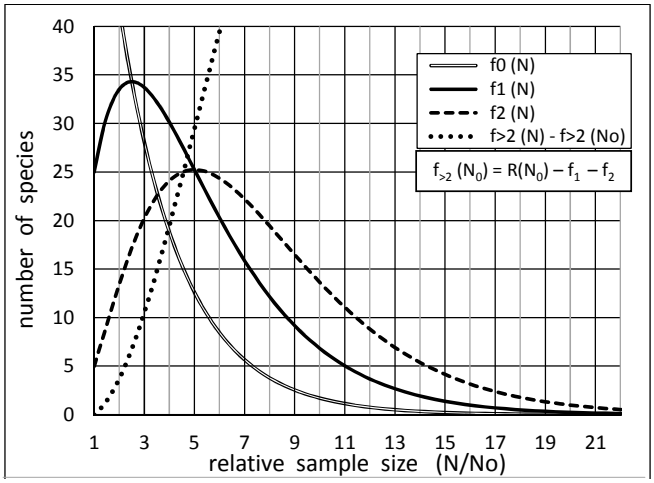
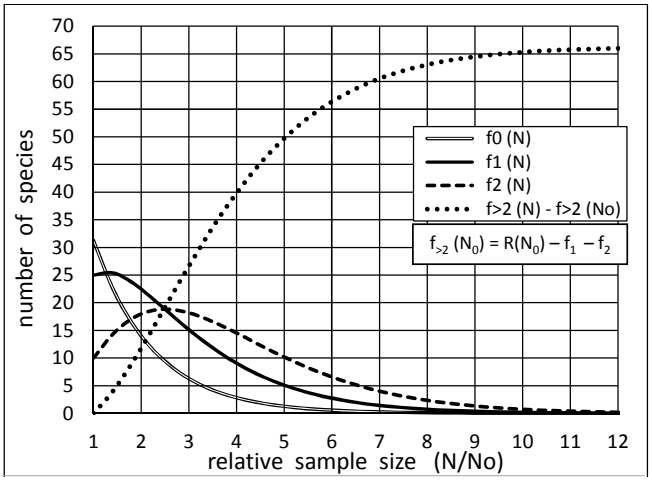


Fig. 2 and 2bis. Extrapolations of the numbers $f_{0\,(N)}$, $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{>2\,(N)} - f_{>2\,(No)}$ of species expected to be recorded zero, once, twice and more than twice, as the sampling effort is going on further beyond the actual inventory size N_0. Here, f_1/f_2 (= $f_{1\,(No)}/f_{2\,(No)}$) = 5.0 (in fact, $f_1 = 25$ and $f_2 = 5$)



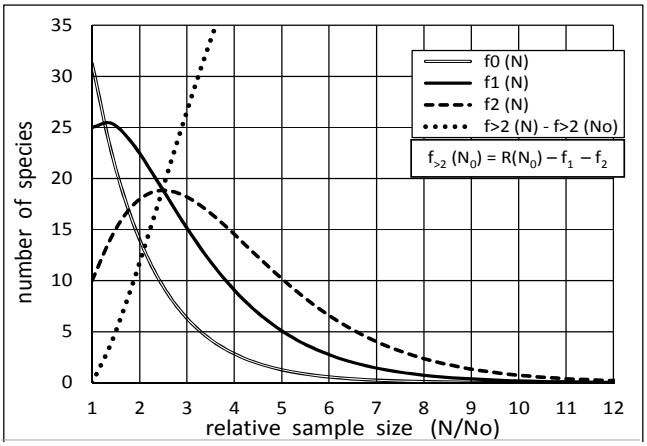
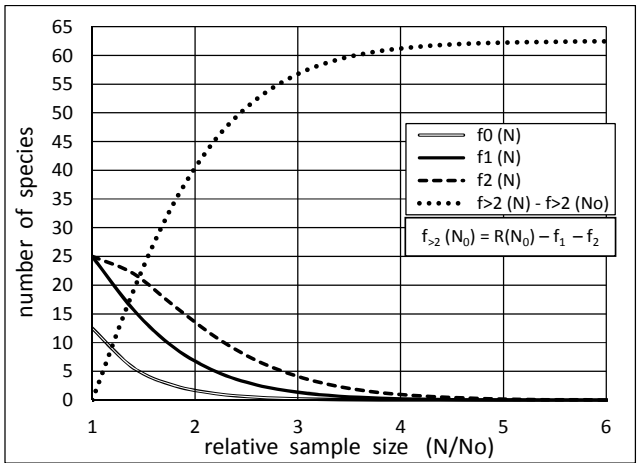


Fig. 3 and 3bis. Extrapolations of the numbers $f_{0\,(N)}$, $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{>2\,(N)} - f_{>2\,(No)}$ of species expected to be recorded zero, once, twice and more than twice, as the sampling effort is going on further beyond the actual inventory size N_0. Here, f_1/f_2 (= $f_{1\,(No)}/f_{2\,(No)}$) = 2.5 (in fact, f_1 = 25 and f_2 = 10)



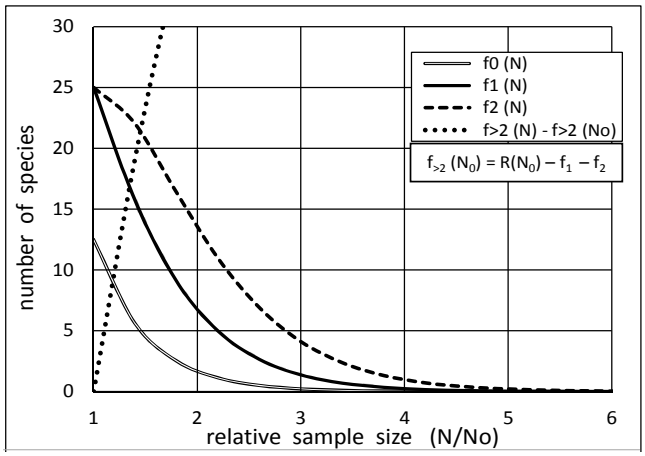


Fig. 4 and 4bis. Extrapolations of the numbers $f_{0\,(N)}$, $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{>2\,(N)} - f_{>2\,(No)}$ of species expected to be recorded zero, once, twice and more than twice, as the sampling effort is going on further beyond the actual inventory size N_0. Here, f_1/f_2 (= $f_{1\,(No)}/f_{2\,(No)}$) = 1.0 (in fact, f_1 = 25 and f_2 = 25)

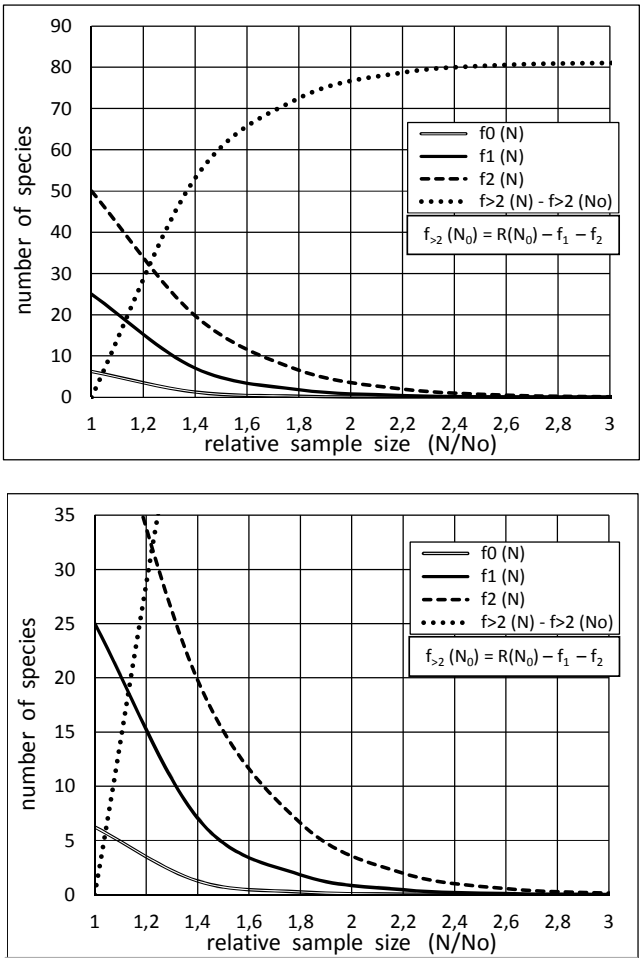


Fig. 5 and 5bis. Extrapolations of the numbers $f_{0\,(N)}$, $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{>2\,(N)} - f_{>2\,(No)}$ of species expected to be recorded zero, once, twice and more than twice, as the sampling effort is going on further beyond the actual inventory size N_0. Here, f_1/f_2 (= $f_{1\,(No)}/f_{2\,(No)}$) = 0.5 (in fact, f_1 = 25 and f_2 = 50)

High values of f_1/f_2 characterise, of course, samplings that remain substantially incomplete and, accordingly, decreasing values of f_1/f_2 are signatures of increasing degrees of sampling completeness. Keeping this in mind, the comparison between the patterns of variations of the $f_{x\,(N)}$ (in particular, here, $f_{1\,(N)}$ and $f_{2\,(N)}$) according to the ratio f_1/f_2 are very suggestive:

(i) by increasing sampling completeness, starting from a low level (say: f_1/f_2 decreasing from 8.3 to 5 and even to 2.5 [Figs. 1, 2, 3]), both $f_{1\,(N)}$ and $f_{2\,(N)}$ begin to grow, then successively pass through a maximum and finally slowly decrease asymptotically towards zero (while $f_{>2\,(N)}$ steadily increases, at a rate sufficient to more than compensate the decreases of $f_{1\,(N)}$ and $f_{2\,(N)}$, so that R(N) steadily remains monotonically increasing with N, as is expected of course).

(ii) then, by continuing to increase sampling size towards higher degrees of completeness (say from f_1/f_2 = 1.0 to 0.5 [Figs. 4, 5]), the maxima of both $f_{1\,(N)}$ and $f_{2\,(N)}$ are now let behind so that both $f_{1\,(N)}$ and $f_{2\,(N)}$ are already in process of monotonic decrease towards zero (while $f_{>2\,(N)}$ steadily increases at sufficient rate to more than compensate for these decreases).

Incidentally, the following general trends should be noticed:

(i) $f_{0\,(N)}$ intersects $f_{1\,(N)}$ precisely when $f_{1\,(N)}$ reaches its maximum and, similarly, $f_{1\,(N)}$

intersects $f_{2\,(N)}$ precisely when $f_{2\,(N)}$ reaches its maximum;

(ii) the maximum of $f_{2\,(N)}$ is reached at a sample size exactly double of the sample size when $f_{1\,(N)}$ reaches its own maximum.

Indeed, these trends are general properties for the extrapolations of all the $f_{x\,(N)}$ associated to "Chao" estimator, as demonstrated below.

From equations (8) and (11), it follows that $f_{0\,(N)}$ ($= \Delta_{(N)}$) intersects $f_{1\,(N)}$ at N such that:

$$[f_1^2/(2\,f_2)].\exp[-(2f_2/f_1).(N/N_0 - 1)] = (N/N_0).f_1.\exp[-(2f_2/f_1).(N/N_0 - 1)]$$

that is:

$$N = ½\,N_0.f_1/f_2$$

and from equation (11), the maximum of $f_{1\,(N)}$ is reached for N such that $\partial f_{1\,(N)} / \partial N = 0$:

$$\partial f_{1(N)} /\partial N = [f_1/N_0 - f_1.(N/N_0).2f_2/f_1/N_0].\exp[-(2f_2/f_1).(N/N_0 - 1)] = 0$$

which leads to the *same* value of N as just above:

$$N = ½\,N_0.f_1/f_2$$

Now, from equations (11) and (12), it follows that $f_{1\,(N)}$ intersects $f_{2\,(N)}$ at N such that:

$$(N/N_0).f_1.\exp[-(2f_2/f_1).(N/N_0 - 1)] = (N/N_0)^2.f_2.\exp[-(2f_2/f_1).(N/N_0 - 1)]$$

that is:

$$N = N_0.f_1/f_2$$

and from equation (12), the maximum of $f_{2\,(N)}$ is reached for N such that $\partial f_{2\,(N)} / \partial N = 0$:

$$\partial f_{2\,(N)} / \partial N = [2f_2.N/N_0^2 - f_2.(N/N_0)^2.2f_2/f_1/N_0].\exp[-(2f_2/f_1).(N/N_0 - 1)] = 0$$

which leads to the *same* value of N as just above:

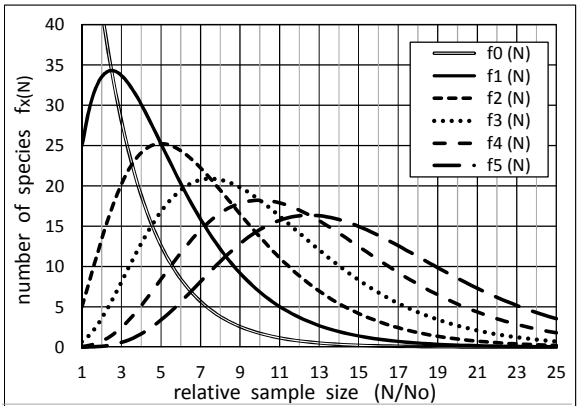
$$N = N_0.f_1/f_2$$

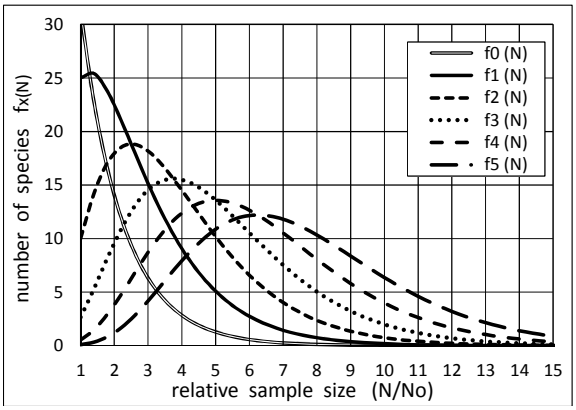
Indeed, these trends are quite consistent, remaining unchanged farther, for any f_x, as exemplified at Figs. 6 and 7: $f_{2\,(N)}$ intersect $f_{3\,(N)}$ precisely when $f_{3\,(N)}$ reaches its maximum (at $N = 1.5\,N_0.f_1/f_2$); $f_{3\,(N)}$ intersect $f_{4\,(N)}$ precisely when $f_{4\,(N)}$ reaches its maximum (at $N = 2.0\,N_0.f_1/f_2$); $f_{4\,(N)}$ intersect $f_{5\,(N)}$ precisely when $f_{5\,(N)}$ reaches its maximum (at $N = 2.5\,N_0.f_1/f_2$); and so on…

In complement to the mathematical demonstration above, it is interesting to highlight the underlying "physical" process behind this general pattern. Consider a sample of any size N (i.e. N individuals already observed) extracted from an assemblage of species having an ideally *even* distribution of species abundances – the ideal condition for "Chao" estimator being relevantly applied, as demonstrated in the next section. Under this specific condition, the next individual collected (thus making sample size growing from N to N+1) may concern *with equal probability* any species (either a species previously unrecorded, or a species already recorded once, or a species already recorded twice, …, or a species already recorded x-times, etc....). Now, the probability of drawing a species previously observed x-times is expected to be proportional to its relative abundance, reflected by the number, f_x, of those species already recorded x-times. Accordingly, the number f_x of species already recorded x-times will tend to:

- increase if the probability of drawing a species already recorded x-1 times exceeds the probability of drawing a species already recorded x times because, thus, the probability for f_x to increase by one exceeds the probability for f_x to decrease by one;
- decrease if the probability of drawing a species already recorded x-1 times is less than the probability of drawing a species already recorded x- times because, thus, the probability for f_x to decrease by one exceeds the probability for f_x to increase by one.

Therefore, f_x is expected either (i) to increase, (ii) to pass by a maximum, (iii) to decrease, depending on f_{x-1} being either (i) larger, (ii) equal, (iii) less than f_x respectively. This, indeed, is the fundamental – "mechanical" – reason which explains the general trend highlighted above. In other words, this argumentation unravels the basic *underlying process* behind *the pattern* described and mathematically demonstrated above and graphically exemplified at Figs. 6 and 7.





Figs. 6 and 7. Extrapolations of the numbers $f_{0\,(N)}$, $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{3\,(N)}$, $f_{4\,(N)}$, $f_{5\,(N)}$ of species expected to be recorded 0-, 1-, 2-, 3-, 4-, 5-times, as the sampling effort is going on further beyond the actual inventory size N_0.

above : f_1/f_2 (= $f_{1\,(No)}/f_{2\,(No)}$) = 5.0 (f_1 = 25 and f_2 = 5) ; below: f_1/f_2 (= $f_{1\,(No)}/f_{2\,(No)}$) = 2.5 (f_1 = 25 and f_2 = 10)

4. DISCUSSION

*** the Chao estimator and the associated extrapolation of the Species Accumulation Curve are especially relevant when species abundances distribution is even or, at least, close to ideal evenness**

As already pointed by Chao & co-authors [10], the "Chao" estimator provides accurate, point estimation in the specific case only when the species abundances are evenly distributed in the sampled assemblage of species. This, of course, stands also for the extrapolation of the Species Accumulation Curve associated to "Chao" estimator.

Indeed, let consider an assemblage of S species with ideally evenly distributed species abundances (all abundances the same). In this particular case, the expected number of recorded species after sampling N individuals is, classically, $R(N) = S.[1 - \exp(-k.N)]$ with k as a constant, independent of N. The number of still unrecorded species is thus: $\Delta_{(N)} = S.\exp(-k.N)$. At $N = N_0$, $\Delta(N) = \Delta_0$ $(= f_1^2/(2 f_2))$, so that:

$$\Delta_{(N)} = \Delta_0.\exp(-k.(N - N_0)) \qquad (15)$$

The formal correspondence between the preceding expression (15) and the expression of $\Delta_{(N)}$ associated to the "Chao" estimator (equation (7)), confirms that the extrapolation R(N) associated to the "Chao" estimator (equation (9)) corresponds, ideally, to the progressive sampling of a species assemblage with *evenly distributed species abundances*, as was expected.

Now, as no species is comparatively rarer than any other one when abundances are evenly distributed, the progressive sampling, in such a case, is expected to reach completeness comparatively faster than for any other assemblage of the same total species richness but having a less even distribution of species abundances.

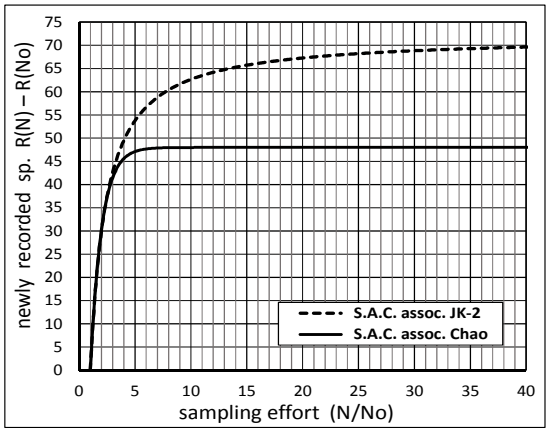


Fig. 8. Extrapolations of the Species Accumulation Curves beyond $\{N_0, R(N_0)\}$, respectively associated to "Chao" (solid line) and "Jackknife-2" (dashed line) estimators ; $f_1 = 48$, $f_2 = 24$; $\Delta_{0\ Chao} = 48$, $\Delta_{0\ JK\text{-}2} = 72$.

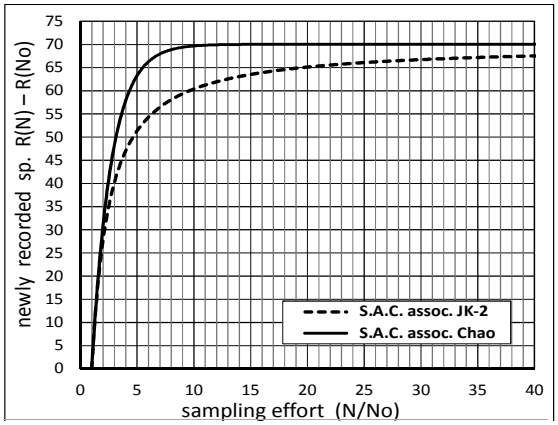


Fig. 9. Extrapolations of the Species Accumulation Curves beyond $\{N_0, R(N_0)\}$, respectively associated to "Chao" (solid line) and "Jackknife-2" (dashed line) estimators ; $f_1 = 41$, $f_2 = 12$; $\Delta_{0\ Chao} = 70$, $\Delta_{0\ JK\text{-}2} = 70$.

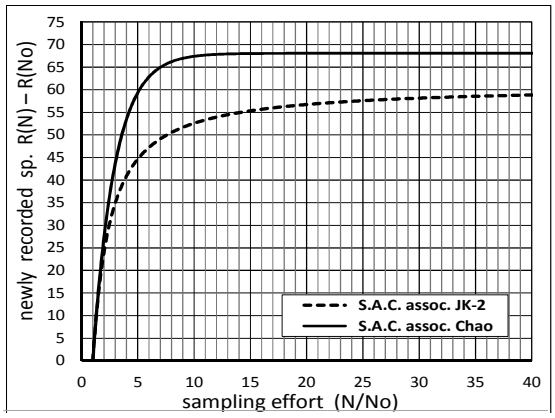


Fig. 10. Extrapolations of the Species Accumulation Curves beyond $\{N_0, R(N_0)\}$ respectively associated to "Chao" (solid line) and "Jackknife-2" (dashed line) estimators ; $f_1 = 35$, $f_2 = 9$; $\Delta_{0\ Chao} = 68$, $\Delta_{0\ JK\text{-}2} = 61$.

Indeed, *faster achievement of completeness* is a characteristic feature of the Species Accumulation Curve associated to "Chao" estimator, as compared to Species Accumulation Curves associated to any other type of estimator, in the same context.

This is highlighted considering three examples, where comparisons are made between the extrapolated Species Accumulation Curves respectively associated to "Chao" estimator and to "Jackknife-2" estimator (both estimators relying upon the numbers f_1 and f_2 of species recorded once and twice): Figs. 8, 9, 10. In these examples, the pair of values f_1 and f_2 are chosen to examine three cases: "Chao" estimate of the number of unrecorded species ($= f_1^2/(2f_2)$) being either (i) smaller, (ii) equal, (iii) larger than the corresponding "Jackknife-2" estimate ($= 2f_1 - f_2$).

In all three cases, regardless of the sign of the gap between "Chao" and "Jackknife-2" estimates (negative [Fig. 8], zero [Fig. 9] or positive [Fig. 10]), the Species Accumulation Curve associated to "Chao" estimator always reaches its asymptote far more rapidly than the Species Accumulation Curve associated to "Jackknife-2" estimator. This, once more, is in agreement with the fact that the extrapolation associated to "Chao" clearly refers to the hypothesis of an ideally homogeneous distribution of the species abundances in the sampled assemblage.

*** Comparing the extrapolations associated to "Chao" and "Jackknife" respectively: an illustrative case study**

A survey of butterfly fauna at Mount Gariwang-san, Korea [11], was conducted along years 2010 to 2015, encompassing 2037 observed individuals and 105 recorded species, with f_1 = 13.6 and f_2 = 15.2 (values obtained after prescribed regression of the crude values of the f_x, in order to reduce the consequences of stochastic dispersion [7]). Accordingly, the estimated number of unrecorded species is $\Delta_{Ch} = f_1^2/(2\ f_2)) = 6.1$ according to "Chao" estimator and $\Delta_{J2} = 2f_1 - f_2 = 12.0$ according to "Jackknife-2" estimator (with corresponding total species richness estimated to 111 or 117 species respectively).

The extrapolations of the Species Accumulation Curve, respectively associated to "Chao" and "Jackknife-2" estimators are plotted at Fig. 11.

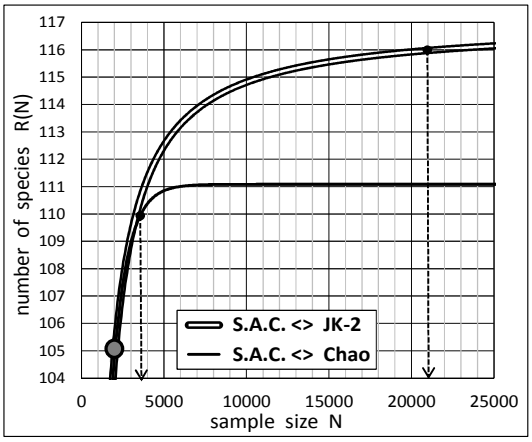


Fig. 11. Extrapolations of the Species Accumulation Curve beyond $\{N_0, R(N_0)\}$ respectively associated to "Chao" and "Jackknife-2" estimators, for a survey of Lepidoptera of Gariwang-san (field data from [11] → $\Delta_{0\ Chao} = 6$, $\Delta_{0\ JK\text{-}2} = 12$). Accordingly, the total species richness is estimated to 111 and 117 species respectively. In fact, in agreement with the procedure of selection of the less biased estimation [5, 7], it is the "Jackknife-2" estimator and its associated extrapolation which are to be adopted rather than "Chao"

These extrapolations may serve to predict the sampling effort that would be necessary to reach any given level of sampling completeness. In particular, the sampling efforts predicted to reach a quasi-exhaustive species inventory (say reaching total species richness minus one; that is 110 and 116 species respectively) are strikingly different, depending on whether "Chao" estimator or "Jackknife-2" estimator is selected. For the extrapolation associated to "Chao" the sampling effort required is N = 4600 against N = 21000 for the extrapolation associated to "Jackknife-2". This clearly highlights the importance of selecting the less-biased extrapolation [5, 7]. Here, between the extrapolations associated to "Chao" and to "Jackknife-2", it is the latter which ought to be adopted, according to the procedure of selection described in [5]. The level of sampling completeness of this inventory of the butterfly fauna at Mount Gariwang-san, 105/117 = 90% thus appears fairly good.

5. CONCLUSION

Incomplete inventories of biodiversity invite to extrapolate the species accumulation process beyond the actually reached sample size,

ultimately trying to estimate the asymptotic, total species richness of the sampled assemblage of species. In this perspective, many attempts have been made in recent decades to find appropriate expressions for the extrapolation of the Species Accumulation Curve (reviewed in [8]), each of these expressions supposed to be as close as possible to some hypothetical “characteristic feature” of the Species Accumulation Curves. In fact, all these attempts were doomed to some form of failure, being confronted with the severe difficulty of identifying a common and generalizable feature for an entity as polymorphous as the Species Accumulation Curve actually is.

Hence, the recent attempt by Chao & Chiu [1] to postpone the difficulty by limiting the scope and focusing only upon the very specific case when the species abundance distribution is ideally even or close to be so. This, indeed, considerably reduces the polymorphism of the Species Accumulation Curve, which, accordingly may be extrapolated more accurately. But, yet, not derived in a strictly satisfying manner, as has been shown above.

In fact, as suggested previously, and demonstrated here, a general feature valid for all theoretical forms of Species Accumulation Curves R(N) (i.e. independently of the type of species abundance distribution) does exist indeed, derived from equation (1) above, that is : $[\partial^x R_{(N)}/\partial N^x] = (-1)^{x-1} f_{x\,(N)}/C_{N,\,x}$.

This equation actually constrains the detailed shape of any kind of Species Accumulation Curve, by means of controlling the series of its derivatives, $\partial^x R_{(N)}/\partial N^x$.

Accordingly, satisfying this general relationship is a prerequisite to any relevant attempt to extrapolate the species accumulation process beyond actual incomplete sampling. And the general relevance of this constraining relationship allows to address, in turn, the extrapolation of the Species Accumulation Curve for any type of species abundance distribution as well.

Coming back to the specific case of an ideally even distribution of species abundances, dealt with by Chao & Chiu [1], we proposed an alternative expression for the extrapolation, which is *mathematically relevant* (that is, in accordance with equation (1)). As such, this formulation actually differs formally from the expression proposed by the preceding authors and should therefore be considered as more reliable.

ACKNOWLEDGEMENTS

"The publications by Anne CHAO and by Ulrich BROSE, (in particular those mentioned here) have stimulated my interest to tackle the quantitative aspects of the process of species accumulation during progressive sampling of species assemblages on a theoretical basis."

COMPETING INTERESTS

Author has declared that no competing interests exist.

REFERENCES

1. Chao A, Chiu CH. Nonparametric estimation and comparison of species richness. In: eLS. John Wiley & Sons, Ltd: Chichester; 2016.
DOI: 10.1002/9780470015902.a0026329
2. Kittur B, Swamy SL, Bargali SS, Jhariya MK. Wildland fires and moist deciduous forests of Chhattisgarh, India: Divergent component assessment. Journal of Forestry Research. 2014;25(4):857-866.
3. Behera SK, Sahu N, Mishra AK, Bargali SS, Behra MD, Tuli R. Aboveground biomass and carbon stock assessment in Indian tropical deciduous forest and relationship with stand structural attributes. Ecological Engineering (In Press); 2016.
4. 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. Waltham, MA: Academic Press. 2013;5:195-211.
5. Béguinot J. Extrapolation of the species accumulation curve for incomplete species samplings: A new nonparametric approach to estimate the degree of sample completeness and decide when to stop sampling. Annual Research & Review in Biology. 2015;8(5):1-9.
DOI: 10.9734/ARRB/2015/22351
6. Brose U, Martinez ND, Williams RJ. Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. Ecology. 2003;84(9): 2364-2377.

7. 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>
8. Thompson GG, Withers PC, Pianka ER & Thompson SA. Assessing biodiversity with species accumulation curves; inventories of small reptiles by pit-trapping in Western Australia. Austral Ecology. 2003;28:361–383.
9. 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. International Scholarly Research Notices – Ecology. 2014;2014:6 Article ID 847328.
DOI: 10.1155/2014/847328
10. 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
11. Lee CM, Kim SS, Kwon TS. Butterfly fauna in Mount Gariwang-san, Korea. Journal of Asia-Pacific Biodiversity. 2016; 9:198-204.
12. Lee SM, Chao A. Estimating population size via sample coverage for closed capture-recapture models. Biometrics. 1994;50(1):88-97.

APPENDIX

A.1 - Derivation of the constraining relationship between $\partial^x R_{(N)}/\partial N^x$ and $f_{x(N)}$

The shape of the theoretical Species Accumulation Curve is directly dependent upon the particular Species Abundance Distribution (the "S.A.D.") within the sampled assemblage of species. That means that beyond the common general traits shared by all Species Accumulation Curves, each particular species assemblage give rise to a specific Species Accumulation Curve with its own, unique shape, considered in detail. Now, it turns out that, in spite of this diversity of particular shapes, all the Species Accumulation Curves are, nevertheless, *constrained by a same mathematical relationship* that rules their successive derivatives (and, thereby, rules the details of the curve shape since the successive derivatives altogether define the local shape of the curve in any details). Moreover, it turns out that this general mathematical constraint relates bi-univocally each derivative at order x, [$\partial^x R_{(N)}/\partial N^x$], to the number, $f_{x(N)}$, of species recorded x-times in the considered sample of size N. And, as the series of the $f_{x(N)}$ are obviously directly dependent upon the particular Distribution of Species Abundance within the sampled assemblage of species, it follows that this mathematical relationship between $\partial^x R_{(N)}/\partial N^x$ and $f_{x(N)}$, ultimately reflects the indirect but strict dependence of the shape of the Species Accumulation Curve upon the particular Distribution of the Species Abundances (the so called S.A.D.) within the assemblage of species under consideration. In this respect, this constraining relationship is central to the process of species accumulation during progressive sampling, and is therefore at the heart of any reasoned approach to the extrapolation of any kind of Species Accumulation Curves.

This fundamental relationship may be derived as follows.

Let consider an assemblage of species containing an unknown total number 'S' of species. Let R be the number of recorded species in a partial sampling of this assemblage comprising N individuals. Let p_i be the probability of occurrence of species 'i' in the sample This probability is assimilated to the relative *abundance* of species 'i' within this assemblage or to the relative *incidence* of species 'i' (its proportion of occurrences) within a set of sampled sites. The number Δ of missed species (unrecorded in the sample) is $\Delta = S - R$.

The estimated number Δ of those species that escape recording during sampling of the assemblage is a decreasing function $\Delta_{(N)}$ of the sample of size N, which depends on the particular distribution of species abundances p_i:

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

with Σ_i as the operation summation extended to the totality of the 'S' species 'i' in the assemblage (either *recorded* or *not*)

The expected number f_x of species recorded *x times* in the sample, is then, according to the binomial distribution:

$$f_x = [N!/X!/(N-x)!] \Sigma_i [(1-p_i)^{N-x} p_i^x] = C_{N,x} \Sigma_i (1-p_i)^{N-x} p_i^x \qquad \text{(A1.2)}$$

with $C_{N,x} = N!/X!/(N-x)!$

We shall now derive the relationship between the successive derivatives of $R_{(N)}$, the theoretical Species Accumulation Curve and the expected values for the series of 'f_x'.

According to equation (A1.2):

► $f_1 = N \Sigma_i [(1-p_i)^{N-1} p_i] = N \Sigma_i [(1-p_i)^{N-1} (1- (1-p_i))] = N \Sigma_i [(1-p_i)^{N-1}] - N \Sigma_i [(1-p_i)^{N-1}(1-p_i))] = N \Sigma_i [(1-p_i)^{N-1}] - N \Sigma_i [(1-p_i)^N]$.

Then, according to equation (A1) it comes: $f_1 = N\,(\Delta_{(N-1)} - \Delta_{(N)}) = -\,N\,(\Delta_{(N)} - \Delta_{(N-1)})$
$= -\,N\,(\partial\,\Delta_{(N)}/\partial N) = -\,N\,\Delta'_{(N)}$

where $\Delta'_{(N)}$ is the first derivative of $\Delta_{(N)}$ with respect to N. Thus:

$$f_1 = -\,N\,\Delta'_{(N)} \quad (= -\,C_{N,1}\,\Delta'_{(N)}) \tag{A1.3}$$

Similarly:

► $f_2 = C_{N,2}\,\Sigma_i\,[(1-p_i)^{N-2}\,p_i^2]$ according to equation (A1.2)

$= C_{N,2}\,\Sigma_i\,[(1-p_i)^{N-2}\,(1-(1-p_i^2))] = C_{N,2}\,[\Sigma_i\,[(1-p_i)^{N-2}] - \Sigma_i\,[(1-p_i)^{N-2}(1-p_i^2)]]$
$= C_{N,2}\,[\Sigma_i\,[(1-p_i)^{N-2}] - \Sigma_i\,[(1-p_i)^{N-2}(1-p_i)(1+p_i)]] = C_{N,2}\,[\Sigma_i\,[(1-p_i)^{N-2}] - \Sigma_i\,[(1-p_i)^{N-1}(1+p_i)]]$
$= C_{N,2}\,[(\Delta_{(N-2)} - \Delta_{(N-1)}) - f_1/N]$ according to equations (A2.1) and (A1.2)
$= C_{N,2}\,[-\Delta'_{(N-1)} - f_1/N] = C_{N,2}\,[-\Delta'_{(N-1)} + \Delta'_{(N)}]$ since $f_1 = -\,N\,\Delta'_{(N)}$ (*cf.* equation (A1.3)).
$= C_{N,2}\,[(\partial\,\Delta'_{(N)}/\partial N)] = [N(N-1)/2]\,(\partial^2\,\Delta_{(N)}/\partial N^2) = [N(N-1)/2]\,\Delta''_{(N)}$

where $\Delta''_{(N)}$ is the second derivative of $\Delta_{(N)}$ with respect to N. Thus:

$$f_2 = [N(N-1)/2]\,\Delta''_{(N)} = C_{N,2}\,\Delta''_{(N)} \tag{A1.4}$$

► $f_3 = C_{N,3}\,\Sigma_i\,[(1-p_i)^{N-3}\,p_i^3]$ which, by the same process, yields:

$= C_{N,3}\,[\Sigma_i\,(1-p_i)^{N-3} - \Sigma_i\,(1-p_i)^{N-2} - \Sigma_i\,[(1-p_i)^{N-2}\,p_i] - \Sigma_i\,[(1-p_i)^{N-2}\,p_i^2)]]$
$= C_{N,3}\,[(\Delta_{(N-3)} - \Delta_{(N-2)}) - f_1^*/(N-1) - 2\,f_2/(N(N-1))]$ according to equations (A2.1) and (A1.2)

where f_1^* is the number of singletons that would be recorded in a sample of size (N - 1) instead of N.

According to equations (A1.3) & (A1.4):

$$f_1^* = -\,(N-1)\,\Delta'_{(N-1)} = -\,C_{N-1,1}\,\Delta'_{(N-1)} \quad \text{and} \quad f_2 = [N(N-1)/2]\,\Delta''_{(N)} = C_{N-1,2}\,\Delta''_{(N)} \tag{A1.5}$$

where $\Delta'_{(N-1)}$ is the first derivate of $\Delta_{(N)}$ with respect to N, at point (N-1). Then,

$f_3 = C_{N,3}\,[(\Delta_{(N-3)} - \Delta_{(N-2)}) + \Delta'_{(N-1)} - \Delta''_{(N)}] = C_{N,3}\,[-\Delta'_{(N-2)} + \Delta'_{(N-1)} - \Delta''_{(N)}]$
$= C_{N,3}\,[\Delta''_{(N-1)} - \Delta''_{(N)}] = C_{N,3}\,[-\partial\,\Delta''_{(N)}/\partial N] = C_{N,3}\,[-\partial^3\,\Delta_{(N)}/\partial N^3] = C_{N,3}\,\Delta'''_{(N)}$

where $\Delta'''_{(N)}$ is the third derivative of $\Delta_{(N)}$ with respect to N. Thus :

$$f_3 = -\,C_{N,3}\,\Delta'''_{(N)} \tag{A1.6}$$

Now, generalising for the number f_x of species recorded *x times* in the sample:

► $f_x = C_{N,x}\,\Sigma_i\,[(1-p_i)^{N-x}\,p_i^x]$ according to equation (A1.2),

$= C_{N,x}\,\Sigma_i\,[(1-p_i)^{N-x}\,(1-(1-p_i^x))] = C_{N,x}\,[\Sigma_i\,(1-p_i)^{N-x} - \Sigma_i\,[(1-p_i)^{N-x}\,(1-p_i^x)]]$
$= C_{N,x}\,[\Sigma_i\,(1-p_i)^{N-x} - \Sigma_i\,[(1-p_i)^{N-x}\,(1-p_i)(\Sigma_j\,p_i^j)]]$

with Σ_j as the summation from j = 0 to j = x-1. It comes:

$f_x = C_{N,x}\,[\Sigma_i\,(1-p_i)^{N-x} - \Sigma_i\,[(1-p_i)^{N-x+1}\,(\Sigma_j\,p_i^j)]]$
$= C_{N,x}\,[\Sigma_i\,(1-p_i)^{N-x} - \Sigma_i\,(1-p_i)^{N-x+1} - \Sigma_k\,[(\Sigma_i\,(1-p_i)^{N-x+1}\,p_i^k)]]$

with Σ_k as the summation from k = 1 to k = x-1 ; that is:

$f_x = C_{N,x}\,[(\Delta_{(N-x)} - \Delta_{(N-x+1)}) - \Sigma_k\,(f_k^*/C_{(N-x+1+k),k})]$ according to equations (A1.1) and (A1.2))

where $C_{(N-x+1+k),\,k} = (N-x+1+k)!/k!/(N-x+1)!$ and f_k^* is the expected number of species recorded k times during a sampling of size (N-x+1+k) (instead of size N).

The same demonstration, which yields previously the expression of f_1^* above (equation (A1.5)), applies for the f_k^* (with k up to x-1) and gives:

$$f_k^* = (-1)^k \, (C_{(N-x+1+k),\,k}) \, \Delta^{(k)}_{(N-x+1+k)} \tag{A1.7}$$

where $\Delta^{(k)}_{(N-x+1+k)}$ is the k^{th} derivate of $\Delta_{(N)}$ with respect to N, at point (N-x+1+k). Then,

$$f_x = C_{N,\,x} \, [(\Delta_{(N-x)} - \Delta_{(N-x+1)}) - \Sigma_k \, ((-1)^k \, \Delta^{(k)}_{(N-x+1+k)})]$$

,

which finally yields :

$f_x = C_{N,\,x} \, [(-1)^x \, (\partial \Delta^{(x-1)}_{(N)}/\partial N)] = C_{N,\,x} \, [(-1)^x \, (\partial^x \Delta_{(N)}/\partial N^x)]$. That is:

$$f_x = (-1)^x \, C_{N,\,x} \, \Delta^{(x)}_{(N)} = (-1)^x \, C_{N,\,x} \, [\partial^x \Delta_{(N)}/\partial N^x] \tag{A1.8}$$

where $[\partial^x \Delta_{(N)}/\partial N^x]$ is the x^{th} derivative of $\Delta_{(N)}$ with respect to N, at point N.

Conversely:

$$[\partial^x \Delta_{(N)}/\partial N^x] = (-1)^x \, f_x / C_{N,\,x} \tag{A1.9}$$

Note that, in practice, leaving aside the beginning of sampling, N rapidly increases much greater than x, so that the preceding equation simplifies as:

$$[\partial^x \Delta_{(N)}/\partial N^x] = (-1)^x \, (x!/N^x) \, f_{x(N)} \tag{A1.10}$$

In particular:

$$[\partial \Delta_{(N)}/\partial N] = f_{1(N)}/N \tag{A1.11}$$

$$[\partial^2 \Delta_{(N)}/\partial N^2] = 2 \, f_{2(N)}/N^2 \tag{A1.12}$$

This relation (A1.9) has general relevance since it does not involve any specific assumption relative to either (i) the particular shape of the distribution of species abundances in the sampled assemblage of species or (ii) the particular shape of the species accumulation rate. Accordingly, this relation constrains any theoretical form of species accumulation curves. As already mentioned, the shape of the species accumulation curve is entirely defined (at any value of sample size N) by the series of the successive derivatives $[\partial^x R_{(N)}/\partial N^x]$ of the predicted number R(N) of recorded species for a sample of size N:

$$[\partial^x R_{(N)}/\partial N^x] = (-1)^{(x-1)} \, f_x / C_{N,\,x} \tag{A1.13}$$

with $[\partial^x R_{(N)}/\partial N^x]$ as the x^{th} derivative of $R_{(N)}$ with respect to N, at point N and $C_{N,\,x} = N!/(N-x)!/x!$ (since the number of recorded species $R_{(N)}$ is equal to the total species richness S minus the expected number of missed species $\Delta_{(N)}$).

As above, equation (A1.13) simplifies in practice as:

$$\partial^x R_{(N)}/\partial N^x = (-1)^{(x-1)} \, (x!/N^x) \, f_{x(N)} \tag{A1.14}$$

Equation (A1.13) makes quantitatively explicit the dependence of the shape of the species accumulation curve (expressed by the series of the successive derivatives $[\partial^x R_{(N)}/\partial N^x]$ of R(N)) upon the shape of the distribution of species abundances in the sampled assemblage of species.

A2 - An alternative derivation of the relationship between $\partial^x R_{(N)}/\partial N^x$ and $f_{x(N)}$

Consider a sample of size N (N individuals collected) extracted from an assemblage of S species and let G_i be the group comprising those species collected i-times and $f_{i(N)}$ their number in G_i. The number of collected individuals in group G_i is thus $i.f_{i(N)}$, that is a proportion $i.f_{i(N)}/N$ of all individuals collected in the sample. Now, each newly collected individual will either belong to a new species (probability $1.f_1/N = f_1/N$) or to an already collected species (probability $1– f_1/N$), according to [12]. In the latter case, the proportion $i.f_{i(N)}/N$ of individuals within the group G_i accounts for the probability that the newly collected individual will contribute to increase by one the number of species that belong to the group G_i (that is will generate a transition [i-1 → i] under which the species to which it belongs leaves the group G_{i-1} to join the group G_i). Likewise, the probability that the newly collected individual will contribute to reduce by one the number of species that belong to the group G_i (that is will generate a transition [i → i+1] under which the species leaves the group G_i to join the group G_{i+1}) is $(i+1).f_{i+1(N)}/N$.

Accordingly:

$$\partial f_{i(N)}/\partial N = [i.f_{i(N)}/N – (i+1).f_{i+1(N)}/N](1 – f_1/N)$$

Leaving aside the very beginning of sampling, and thus considering values of sample size N substantially higher than f_1, it comes:

$$\partial f_{i(N)}/\partial N \approx i.f_{i(N)}/N – (i+1).f_{i+1(N)}/N \quad (A2.1)$$

Let consider now the Species Accumulation Curve R(N), that is the number R(N) of species that have been recorded in a sample of size N. The probability that a newly collected individual belongs to a still unrecorded species corresponds to the probability of the transition [0 → 1], equal to $i.f_{i(N)}/N$ with i = 1, that is: $f_{1(N)}/N$ (as already mentioned).

Accordingly, the first derivative of the Species Accumulation Curve R(N) at point N is

$$\partial R_{(N)}/\partial N = f_{1(N)}/N \quad (A2.2)$$

In turn, as $f_{1(N)} = N.\partial R_{(N)}/\partial N$ (from equation (A2.2)) it comes:

$$\partial f_{1(N)}/\partial N = \partial[N(\partial R_{(N)}/\partial N)]/\partial N = N(\partial^2 R_{(N)}/\partial N^2) + \partial R_{(N)}/\partial N$$

On the other hand, according to equation (A2.1):

$\partial f_{1(N)}/\partial N = 1.f_{1(N)}/N – 2.f_{2(N)}/N = f_{1(N)}/N – 2f_{2(N)}/N$, and therefore:

$$N(\partial^2 R_{(N)}/\partial N^2) + \partial R_{(N)}/\partial N = f_{1(N)}/N – 2f_{2(N)}/N$$

And as $\partial R_{(N)}/\partial N = f_{1(N)}/N$ according to equation (A2.2):

$$\partial^2 R_{(N)}/\partial N^2 = – 2f_{2(N)}/N^2 \quad (A2.3)$$

Likewise, as $f_{2(N)} = –N^2/2.(\partial^2 R_{(N)}/\partial N^2)$, it comes:

$$\partial f_{2(N)}/\partial N = \partial[–N^2/2.(\partial^2 R_{(N)}/\partial N^2)]/\partial N = – N(\partial^2 R_{(N)}/\partial N^2) – N^2/2.(\partial^3 R_{(N)}/\partial N^3)$$

As $\partial f_{2(N)}/\partial N = 2f_{2(N)}/N – 3f_{3(N)}/N$, according to equation (A2.1), it comes:

$$– N(\partial^2 R_{(N)}/\partial N^2) – N^2/2.(\partial^3 R_{(N)}/\partial N^3) = 2f_{2(N)}/N – 3f_{3(N)}/N$$

and as $\partial^2 R_{(N)}/\partial N^2 = -2f_{2(N)}/N^2$, according to equation (A2.3), it comes:

$$\partial^3 R_{(N)}/\partial N^3 = +6f_{3(N)}/N^3 \quad \text{(A2.4)}$$

More generally:

$$\partial^x R_{(N)}/\partial N^x = (-1)^{(x-1)}\,(x!/N^x)\,f_{x(N)} \quad \text{(A2.5)}$$

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/17184