Theoretical derivation of a bias-reduced expression for the extrapolation of the species accumulation curve and the associated estimation of total species richness.

Jean Béguinot

Advances in Research
7(3): 1-16, 2016, Article no.AIR.26387
ISSN: 2348-0394, NLM ID: 101666096

SCIENCEDOMAIN *international*
www.sciencedomain.org

Theoretical Derivation of a Bias-reduced Expression for the Extrapolation of the Species Accumulation Curve and the Associated Estimation of Total Species Richness

Jean Béguinot[1*]

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

Author's contribution

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

Article Information

DOI: 10.9734/AIR/2016/26387
Editor(s):
(1) Martin Kröger, Computational Polymer Physics, Swiss Federal Institute of Technology (ETH Zürich), Switzerland.
Reviewers:
(1) Vinicius Costa Cysneiros, Federal University of Paraná, Brazil.
(2) Manoel Fernando Demétrio, Universidade Federal da Grande Dourado, Brazil.
(3) S. Ghollasimood, Universiti Putra Malaysia, Malaysia.
(4) Robert B. Modest, Sokoine University of Agriculture, Tanzania.
Complete Peer review History: http://sciencedomain.org/review-history/14875

Method Article

Received 14th April 2016
Accepted 24th May 2016
Published 2nd June 2016

ABSTRACT

Under-sampling becomes the current situation for an increasing part of biodiversity surveys, as more and more speciose assemblages and increasingly complex taxonomic groups are progressively addressed. Accordingly, (i) extrapolating the Species Accumulation Curve and (ii) estimating the total species richness of partially-sampled species assemblages (or taxonomic-groups) both become major issues for many naturalists nowadays. Numerous different solutions have been proposed to address these issues. Yet, no general consensus has been reached regarding which particular solution among them should be preferred according to each case. This unsatisfactory situation follows from the empirical nature of traditional approaches, especially regarding the extrapolation of the Species Accumulation Curve.
Fortunately, reconsidering the problem on decidedly more theoretical basis, including the consideration of general mathematical relationships universally constraining the expression of any theoretical (or rarefied) Species Accumulation Curves, allows a more relevant modeling for the extrapolation of species accumulation. In turn, this theoretical approach provides a rational key to

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

select the more appropriate, less biased type of species-richness estimator and the associated, less biased expression for the extrapolation of the Species Accumulation Curve, according to the context of sampling. In particular, the *wide relevance* of the series of 'Jackknife-type' estimators is highlighted (as had been already argued for specific cases, on semi-empirical basis). In practice, selecting the less biased extrapolation of the Species Accumulation Curve allows to forecast the supplementary sampling effort necessary to reach a given increase of sampling completeness more accurately than the usual procedures, involving arbitrarily chosen empirical models.

Keywords: Extrapolation; species accumulation curve; estimator; Chao; Jackknife; minimum bias; mathematical constraint; under sampling; incomplete sample.

1. INTRODUCTION

As biodiversity inventories are continuously expanding, they progressively address less surveyed taxonomical groups which usually give rise to assemblages of numerous species, with often small sized-individuals, more or less hard to detect in the field (such as, for example, assemblages of small- or micro-invertebrates). Thus, more or less *incomplete* species samplings are deemed to become increasingly frequent [1].

Under-samplings immediately raise two kinds of important questions in practice:

- How to properly estimate the *total* species richness of a partially sampled assemblage, that is the expected number of recorded species that would be recorded if the sampling was ideally complete;
- How to properly estimate the expected kinetic of discovery of new species, as sampling would be carried on further, beyond its present size, that is how to estimate adequately the shape (the governing equation) of the *extrapolated* "Species Accumulation Curve" ('SAC'), beyond the already reached sampling-size [2].

Both issues are of major importance and thus prompted researchers to propose a series of expressions for both the nonparametric estimation of total species richness (review in [3, 4]) and the extrapolation of the Species Accumulation Curve (review in [5]). Accordingly, the issue, now, is rather to select among the varied reported types of estimators, since, unfortunately, each of these types of estimators provide a substantially different results in practice [6]. Indeed, a considerable amount of work has been devoted to test comparatively these different types of estimators, mainly on an *empirical* basis [7-13]. But, as might be expected, no consensus emerged from these studies. This is because, in fact, each type of estimator may provide a centered, unbiased prediction *in a very specific case only*: for a particular shape of the Species Accumulation Curve, resulting in turn from a particular shape of the species abundance distribution within the sampled assemblage of species [3,13,14-16]. Thus, a specific shape of 'SAC' is *associated to each* type of estimator of species richness, insuring the accuracy of the resulting estimations. Accordingly, the empirical approaches above hardly help to disclose any information of general value and this is the reason why, given these ambiguities, a common but rather unsatisfactory advice consists in considering all, or at least several types of estimators concurrently, without choosing between them [4].

Now, apart from these rather unsuccessful, purely empirical approaches, a few semi-empirical studies have more recently offered suggestive insights, trying to opportunely choose among the different types of estimators provided in the literature [17-19]. These studies yet remain semi-empirical, as they are based on simulations specifically computed for particular kinds of species abundance distributions: 'broken-stick', 'random fraction', 'random assortment' [17]; 'log-normal', 'log-series' [18]; 'geometric series' [19].

Although these studies lead to broadly similar trends, in terms of selected types of estimators, they significantly differ in details, due to the different kinds of species abundance distributions involved in each study. Moreover, being based on *simulations* (computed from particular types of abundance distributions), all these studies lack a definite theoretical basis that could provide added soundness to the analysis and ensure more general applicability to the results.

Accordingly, a more appropriate approach might consist in addressing the issue from a less empirical, *more theoretical* point of view, aiming

at more general applicability for the extrapolation of the 'SAC' and for the estimation of the total species richness of the sampled assemblage.

In this perspective, a *theoretical based* guide of choice is needed for:

(i) Defining which, among the most commonly used types of species richness estimators, is able to provide the less biased estimations and, thus, should be selected according to the degree of unevenness of species abundances and the level of sampling completeness;
(ii) Defining accordingly which mathematical expression is to be selected for the associated extrapolation of the 'SAC', in accordance with the selected estimator above.

A preliminary step in this respect, addressing the problem this way, was achieved by considering at first the estimators involving only the numbers f_1 and f_2 of species already recorded once and twice respectively: Chao (= $f_1^2/(2f_2)$); Jackknife-1 (= f_1); Jackknife-2 (= $2f_1 - f_2$) [20].

Restricting to f_1 and f_2 only, yet reveals progressively insufficient, as the level of sampling completeness decreases (or as the degree of unevenness of species abundances increases), as already suggested by the semi-empirical studies cited above.

Thus, including the supplementary data provided by the values of the numbers f_3, f_4, f_5,… of those species recorded 3, 4, 5,… times respectively, is a further step necessary to cope with more incomplete samples. The corresponding issue is addressed in detail hereafter, starting from the same theoretical basis as for the previous developments.

2. GENERAL MATHEMATICAL RELATIONSHIPS CONSTRAINING THE EXPRESSION OF ANY SPECIES ACCUMULATION CURVE

The 'SAC' reflects the increasing number R(N) of recorded species with growing sampling effort N (for example, the number of collected individuals). Fundamentally, the expression of the 'SAC' is a continuously differentiable function (being understood that, in practice, the 'SAC's may not be so, due to sampling stochasticity; yet, the classical rarefaction procedure restores the theoretical shape and the genuine continuous differentiability of the 'SAC').

Two mathematical relationships of general applicability actually constrain the theoretical expression of any 'SAC':

- A general relationship between (i) the series of the successive derivatives $\partial^x R_{(N)}/\partial N^x$ of the 'SAC' and (ii) the series of the numbers $f_{x(N)}$ of species recorded x times:

$$\partial^x R_{(N)}/\partial N^x = (-1)^{(x-1)}\ f_{x(N)}\ /C_{N,\ x} \approx (-\ 1)^{(x-1)}\ (x!/N^x)\ f_{x(N)} \quad (1)$$

  with $C_{N,\ x}$ designing the number of combinations of x objects among N (the derivation of this general relationship is provided at Appendix A.1 & A.2).

- The rule of 'additivity' [21], which stipulates that if the whole set of species, within a sampled assemblage, may be distributed among several *mutually exclusive* subsets (for example, taxonomic subsets such as orders, or families, or genus), then the 'SAC' for the whole assemblage must be the sum of the 'SAC's relative to each of the constitutive subsets:

$$R(N) = \Sigma_i\ R_i(N) \quad (2)$$

The classical expressions that may be considered to extrapolate the 'SAC' beyond the size N_0 of an (incomplete) sample (see [5] for a review) are unsatisfactory in these respects: they do not satisfy the rule (2) of additivity and they satisfy the constraining relationship (1) on derivatives, at best, for the first derivative only, as may easily be verified. For example, the Clench model, R(N) = S.N/(b+N), with S standing for the total species richness, complies with the additivity rule only if the parameter 'b' takes the same value for the whole set and for each of the constitutive subsets, which is almost never the case in practice. Also, the Clench model does not satisfy the relationship on derivatives beyond the first order derivative (except very specific cases).

3. DERIVING THE EXPRESSIONS OF THE EXTRAPOLATED SPECIES ACCUMULATION CURVE AND OF THE ESTIMATED TOTAL SPECIES RICHNESS

Satisfying the rule of additivity suggests to model the extrapolation R(N) of the 'SAC', using a polynomial function in N. More precisely, the

monotonic, continuously decelerating increase of R(N) with N, and the ultimate asymptotic approach of the total species richness S, lead to the following general polynomial form:

$$R(N) = S - A.N^{-1} - B.N^{-2} - C.N^{-3} - \ldots \qquad (3)$$

with the coefficients S, A, B, C,… of this polynomial expression in N being dependent on the recorded data issued from the already achieved sampling, that is: the sample size N_0, the number of recorded species R_0 (= $R(N_0)$) and the numbers f_1, f_2, f_3,… of species already recorded 1, 2, 3,… times. In addition, to comply with the prescribed rule of additivity, the coefficient S, A, B, C,… should depend linearly upon N_0, R_0, f_1, f_2, f_3,… ; as is actually the case, as shown later. Thus, in these conditions, the addition of several 'SAC's having a polynomial form as specified by equation (3) still remains of this same polynomial form (3), as prescribed by the rule of additivity when an assemblage of species may be considered as the union of several, mutually exclusive subsets [21]. On the contrary, classical models, do not satisfy the rule of additivity in general, as already underlined above. Of course, this polynomial expression (3) is specifically designed for the *extrapolation* of 'SAC's, that is beyond the actual sample size N_0 and should not be considered for the intrapolation (when $N < N_0$), especially when N approaches low values.

Satisfying equation (1), up to the x^{th} derivative leads to the following system of x+1 equations, labelled (4):

$$\left\{\begin{array}{l} R(N_0) = R_0 \\ [\partial R_{(N)}/\partial N]_{N0} = f_1/N_0 \\ [\partial^2 R_{(N)}/\partial N^2]_{N0} = -2 f_2/N_0^2 \\ [\partial^3 R_{(N)}/\partial N^3]_{N0} = 6 f_3/N_0^3 \\ \ldots\ldots\ldots\ldots \end{array}\right\} \qquad (4)$$

with f_1, f_2, f_3,… designing the values of $f_{1(N)}$, $f_{2(N)}$, $f_{3(N)}$, for $N = N_0$.

Now, replacing $R(N_0)$ and the x first derivatives, $[\partial R_{(N)}/\partial N]_{N0}$ to $[\partial^x R_{(N)}/\partial N^x]_{N0}$ by their corresponding expressions, according to the polynomial definition (3) of R(N), converts the system (4) above in a linear system of x+1 equations, in terms of the x+1 unknown coefficients S, A, B, C, …, . This linear system, labelled (5), is, according to (3):

$$\left\{\begin{array}{l} S - A/N_0 - B/N_0^2 - C/N_0^3 = R_0 \\ A/N_0^2 + 2B/N_0^3 + 3C/N_0^4 = f_1/N_0 \\ -2A/N_0^3 - 6B/N_0^4 - 12C/N_0^5 = -2f_2/N_0^2 \\ 6A/N_0^4 + 24B/N_0^5 + 60C/N_0^6 = 6f_3/N_0^3 \\ \ldots\ldots\ldots\ldots\ldots\ldots \end{array}\right\} \qquad (5)$$

Resolving this system (5) yields the expressions of the coefficients S, A, B, C, …, each of them being, as expected, a linear function of the recorded terms R_0, f_1, f_2, f_3,… . The details of the resolution are given in Appendix A.3.

In the system above, the constraining relationship (1) is taken in consideration for the x first derivatives $\partial^x R_{(N)}/\partial N^x$ only. This, however, accounts for the essential since it is the series of the first derivatives that actually play the major role in precisely defining the shape of the (extrapolated) 'SAC'.

In practice, as will be shown later, the less complete is the achieved sampling, the larger is the number of terms in the polynomial expression (3) of R(N) which have to be considered (and accordingly, the larger the number of values of f_1, f_2, f_3,… which have to be implemented to derive the expression of the extrapolated 'SAC').

Finally, the resolution of the linear system of equations (5), considering successively a growing number (x+1) of equations, yields the following expressions for:

(i) The extrapolation of the 'SAC': Table 1;
(ii) The estimated number of missing species Δ (= $S - R_0$): Table 2;

Table 1. Successive expressions - at increasing orders m = 1 to 5 - of the *extrapolated Species Accumulation Curve* R(N), respectively associated to the "Jackknife" type estimators Δ_m (with $S_m = R_0 + \Delta_m$) defined in Table 2

* $R_1(N) = (R_0 + f_1) - f_1.N_0/N$
* $R_2(N) = (R_0 + 2f_1 - f_2) - (3f_1 - 2f_2).N_0/N - (f_2 - f_1).N_0^2/N^2$
* $R_3(N) = (R_0 + 3f_1 - 3f_2 + f_3) - (6f_1 - 8f_2 + 3f_3).N_0/N - (-4f_1 + 7f_2 - 3f_3).N_0^2/N^2 - (f_1 - 2f_2 + f_3).N_0^3/N^3$
* $R_4(N) = (R_0 + 4f_1 - 6f_2 + 4f_3 - f_4) - (10f_1 - 20f_2 + 15f_3 - 4f_4).N_0/N - (-10f_1 + 25f_2 - 21f_3 + 6f_4).N_0^2/N^2 - (5f_1 - 14f_2 + 13f_3 - 4f_4).N_0^3/N^3 - (-f_1 + 3f_2 - 3f_3 + f_4).N_0^4/N^4$
* $R_5(N) = (R_0 + 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5) - (15f_1 - 40f_2 + 45f_3 - 24f_4 + 5f_5).N_0/N - (-20f_1 + 65f_2 - 81f_3 + 46f_4 - 10f_5).N_0^2/N^2 - (15f_1 - 54f_2 + 73f_3 - 44f_4 + 10f_5).N_0^3/N^3 - (-6f_1 + 23f_2 - 33f_3 + 21f_4 - 5f_5).N_0^4/N^4 - (f_1 - 4f_2 + 6f_3 - 4f_4 + f_5).N_0^5/N^5$
* $R_m(N)$ = …… (*see appendix A.3*)

Table 2. Successive expressions - at increasing orders m = 1 to 5 - of the "Jackknife" type estimators of the estimated number Δ_m of *missing species* and of the estimated *total species richness* S_m

* $\Delta_1 = f_1$	* $S_1 = R_0 + f_1$
* $\Delta_2 = 2f_1 - f_2$	* $S_2 = R_0 + 2f_1 - f_2$
* $\Delta_3 = 3f_1 - 3f_2 + f_3$	* $S_3 = R_0 + 3f_1 - 3f_2 + f_3$
* $\Delta_4 = 4f_1 - 6f_2 + 4f_3 - f_4$	* $S_4 = R_0 + 4f_1 - 6f_2 + 4f_3 - f_4$
* $\Delta_5 = 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5$	* $S_5 = R_0 + 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5$
* Δ_m = …… (*see appendix A.3*)	* S_m = …… (*see appendix A.3*)

The general solution, derived at order m (that is for the system (5) with m+1 equations), provides the estimation of the number Δ_m of missing species and the corresponding expression R_m (N) of the associated extrapolation of the 'SAC'. The derivation of the general solution is addressed in more details in Appendix A.3.

As suggested above, the less complete is the sampling under consideration, the larger is the required number of terms to be accounted for in the polynomial expression (3) of the extrapolation R(N) ; that is, the larger is the order m to be selected in Tables 1 and 2. In other words, each order m (i.e. each expression $R_m(N)$, Δ_m) has a specific domain of appropriateness (essentially related to the degree of completeness of sampling and, also, to the degree of unevenness of the distribution of species abundances). Now, in practice, the degree of completeness and the degree of abundances unevenness are unknown *a priori*, but the choice for the appropriate order m is imposed, however, by the required compliance with the rule of continuity. This rule specifies that the continuity of value of R(N) (and of Δ as well) is to be satisfied at the boundary between the domains of appropriateness of successive orders, m and m+1. That is, $R_{m+1}(N) = R_m(N)$ and $\Delta_{m+1} = \Delta_m$, at the boundary between orders m and m+1. This, in turn, unambiguously defines the positions of the boundaries delimiting the specific domain of appropriateness of each order m and the corresponding expressions $R_m(N)$ and Δ_m to be selected in Tables 1 and 2. Thus, satisfying the continuity at the boundary between order 1 and order 2, requires that, at this boundary, $R_1(N) = R_2(N)$, that is: $f_1 = f_2$ (see Table 1). Similarly, the rule of continuity requires that (i) at the boundary between orders 2 and 3, $R_2(N) = R_3(N)$, that is: $f_1 = 2f_2 - f_3$; (ii) at the boundary between orders 3 and 4, $R_3(N) = R_4(N)$, that is, $f_1 = 3f_2 - 3f_3 + f_4$; etc… (see Table 1).

Further proceeding the same way, the following *guide of choice* is finally derived, highlighting the relevant order m to be used for the estimation Δ_m of the number of missing species and for the associated expression $R_m(N)$ of the extrapolation of the 'SAC': Table 3.

***Nota*:** Apart from the Jackknife-type series derived above, another commonly referred estimator of species richness is the *Chao*-type, which, yet, does not comply in general with the additivity rule and does not satisfy the constraining rule (1) beyond the first derivative. These are the reasons why *Chao* estimator is discarded in general (the same applying to the more recently derived [22] *i-Chao* estimator (see [21]). Yet, this lack of compliance does not hold true in the very specific (and rather unusual) circumstance when the distribution of species abundances is (sub-) even or, as well, when the sampling comes close to completeness [20,21].

Table 3. Key to select the more appropriate (less biased) choice for (i) the expression of the *extrapolated Species Accumulation Curve* R(N) and (ii) the estimates of the *number of missing species* Δ (and the resulting estimation of the *total species richness* S). The key is based upon the recorded data issued from the actually achieved sampling, i.e. the values taken by the series of numbers f_x of species respectively recorded x-times (N.B.: to minimize the consequence of sampling stochasticity, the distribution of the recorded values of f_x is preferably smoothened, as indicated in section 4 -Practical implementation)

$f_1 \leq f_2$: prefer $R_1(N)$ & $\Delta_1 = f_1$
f_1 between f_2 & $2f_2 - f_3$: prefer $R_2(N)$ & $\Delta_2 = 2f_1 - f_2$
f_1 between $2f_2 - f_3$ & $3f_2 - 3f_3 + f_4$: prefer $R_3(N)$ & $\Delta_3 = 3f_1 - 3f_2 + f_3$
f_1 between $3f_2 - 3f_3 + f_4$ & $4f_2 - 6f_3 + 4f_4 - f_5$: prefer $R_4(N)$ & $\Delta_4 = 4f_1 - 6f_2 + 4f_3 - f_4$
$f_1 > 4f_2 - 6f_3 + 4f_4 - f_5$: prefer $R_5(N)$ & $\Delta_5 = 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5$

In these particular cases, Chao estimator even provides a strictly accurate estimates and may then arguably be preferred. It is thus possible, and even advisable, to replace the first order Jackknife estimator $\Delta_1 = f_1$ by the Chao estimator $\Delta_C = f_1^2/(2f_2)$ and replace the extrapolation $R_1(N)$ associated to Jackknife-1 by the expression $R_C(N)$ associated to Chao estimator, defined by:

$$R_C(N) = (R_0 + f_1^2/(2f_2)).(1 - \exp([\mathrm{Ln}((f_1^2/(2f_2))/(R_0 + f_1^2/(2f_2)))](N/N_0)))$$

according to [20].

The two first lines of the guide of choice above (Table 3) are then modified accordingly: Table 4.

The keys provided at Tables 3 or 4 thus allow to select (i) the more appropriate (i.e. less biased) type of estimator of the total species richness of the sampled assemblage and (ii) the associated expression of the extrapolation of the Species Accumulation Curve.

4. PRACTICAL IMPLEMENTATION OF THE PROCEDURE

Consider an incomplete sampling of an assemblage of species: sample size N_0, including R_0 recorded species, among which f_1, f_2, f_3,…, f_x, of them are recorded 1, 2, 3, .., x-times respectively. In practice, to reduce the consequence of unavoidable sampling stochasticity it is generally advisable to smoothen the distribution of values of the f_x [3]. This is more specifically the case for the f_x with $x > 2$, since the $f_{x>2}$ often have comparatively lower values, as such especially sensitive to stochastic dispersion. This may be obtained by classically applying a rarefaction procedure to the recorded data [3,4], thanks to what a smoothened distribution, f_{1*}, f_{2*}, f_{3*},…, f_{x*} is derived and substituted to the originally recorded distribution f_1, f_2, f_3,…, f_x. Alternatively, the smoothened distribution f_{1*}, f_{2*}, f_{3*},…, f_{x*} may also be derived by a regression applied to the originally recorded distribution f_1, f_2, f_3,…, f_x.

The guide of choice (Tables 3 or 4) and the selected expressions for the extrapolation of the 'SAC' and the associated estimate of the number of missing species (Tables 1 and 2) should thus be considered preferably on the basis of the series of values f_{1*}, f_{2*}, f_{3*},…, f_{x*}.

- An illustrative example: A partial survey of mining moths in Burgundy involving 2605 recorded individuals encompasses 168 recorded species with $f_{1*} = 30$, $f_{2*} = 19$, $f_{3*} = 13$, $f_{4*} = 9$, $f_{5*} = 6$. The estimated number Δ of missing species and the resulting total species richness $S = R_0 + \Delta$ are computed, for the different types of estimators, according to Table 2. The results are provided at Table 5. The corresponding extrapolations of the 'SAC', computed according to each type of estimator, are plotted at Fig. 1. In turn, the evolution of sampling completeness R(N)/S with growing sampling size N is plotted at Fig. 2, for each type of estimator. The rather strong differences between the results obtained according to the different types of estimators highlight the importance of selecting appropriately the less biased among these different types of estimators. Referring to the guide of choice (Tables 3 & 4), it follows that, here, the order m = 5 is to be selected, since $f_1^* > 4f_2^* - 6f_3^* + 4f_4^* - f_5^*$. Thus, the best extrapolation of the Species Accumulation Curve is given by equation $R_5(N)$ (see Table 1) and the best estimated number of missing species is $\Delta_5 = 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5$ (Table 2).

5. DISCUSSION

As already mentioned, the common practice with the various types of estimators of species richness (and the various solutions for extrapolation of the Species Accumulation Curve ['SAC']) has long been – and still is – to consider this various types of estimators altogether, without choosing among them [4], in spite of their divergent results. Yet, this obviously unsatisfactory situation has been subsequently challenged by several authors [17-19], testing semi empirically the respective relevance of several types of estimators (Chao, Jackknife of different orders), considering different classical models of species abundance distributions ('broken-stick', 'random fraction', 'random assortment' [17], 'log-normal' and 'log-series' [18], 'geometric series' [19]). One common point, highlighted by all these studies, is that the more appropriate type of estimator is dependent upon both (i) the kind of species abundance distribution in the sampled assemblage of species and (ii) the degree of sampling completeness. This being due to the fact that the shape of the extrapolated 'SAC' – which governs the estimation of the total species richness – is dependent, itself, upon the species abundance distribution and the degree of completeness of sampling. Then, assuming (according to [17-19]):

(i) The kind of species abundance distribution, among the six cited above

(which may possibly be tested *a priori* on the basis of the already recorded data) and

(ii) The degree of sampling completeness (which may be tested *a posteriori* for each type of estimator),

The particular type of estimator to be selected is the one which provides an estimate in accordance with the resulting degree of completeness computed with this estimator [17-19]. However, it remains that having to cope and handle with the assumptions above, which require verifications *a priori* and *a posteriori*, makes those approaches not so easy to implement in practice.

Indeed, it is in this respect that the resolutely *theoretical* approach described above, has proven being able to bring substantial further progress.

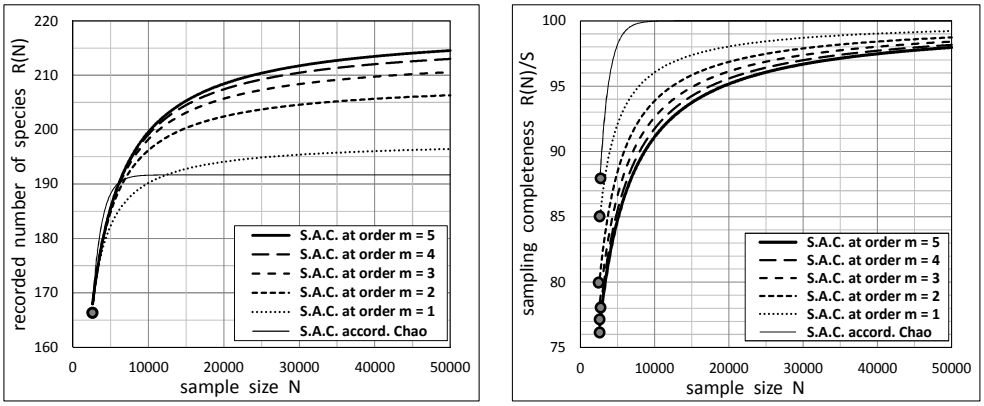
Instead of considering the recorded data (R_0 and the series of f_x) only as a mean to express the estimation of species richness (as in [17-19]), we highlighted, here, that this recorded data *carried also essential information regarding the shape of the extrapolated part of the 'SAC'*, thanks to the general relationship (1) governing the shape of any 'SAC'.

Table 4. Alternative version of the key provided at Table 3, with *"Chao" type estimator* replacing the Jackknife-1 type estimator (the modification therefore concerns only the particular case of samples already approaching full completeness)

$\mathbf{f_1} \leq \underline{0.6}\, f_2$: prefer $R_C(N)$ above & $\Delta_C = f_1^2/(2f_2)$ [i.e. $S_C = R_0 + f_1^2/(2f_2)$]
$\mathbf{f_1}$ between $\underline{0.6}\, f_2$ & $2f_2 - f_3$: prefer $R_2(N)$ & $\Delta_2 = 2f_1 - f_2$ [i.e. $S_2 = R_0 + 2f_1 - f_2$]
$\mathbf{f_1}$ between $2f_2 - f_3$ & $3f_2 - 3f_3 + f_4$: prefer $R_3(N)$ & $\Delta_3 = 3f_1 - 3f_2 + f_3$
$\mathbf{f_1}$ between $3f_2 - 3f_3 + f_4$ & $4f_2 - 6f_3 + 4f_4 - f_5$: prefer $R_4(N)$ & $\Delta_4 = 4f_1 - 6f_2 + 4f_3 - f_4$
$\mathbf{f_1} > 4f_2 - 6f_3 + 4f_4 - f_5$: prefer $R_5(N)$ & $\Delta_5 = 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5$

Table 5. Mining moths survey in Burgundy. Recorded values: sample size N_0, number of recorded species R_0, numbers f_{1*}, f_{2*}, f_{3*},… of species recorded 1, 2, 3,… times. The corresponding estimates – number Δ of missing species and total species richness S – are derived at orders m = 1 to m = 5 and for Chao. As $f_1^* > 4f_2^* - 6f_3^* + 4f_4^* - f_5^*$, the order m = 5 is selected with the extrapolation of the Species Accumulation Curve according to $R_5(N)$ and the associated estimation of the number of missing species in the sample according to $\Delta_5 = 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5$ (see Tables 1 to 4)

No	Ro	f_1^*	f_2^*	f_3^*	f_4^*	f_5^*	Δ & S m = 1	Δ & S m = 2	Δ & S m = 3	Δ & S m = 4	Δ & S m = 5	Δ & S Chao
2605	168	30	19	13	9	6	30 198	41 209	46 214	49 217	**51** **219**	24 192



Figs. 1 and 2. Mining moths survey in Burgundy. *Left*: extrapolations of the Species Accumulation Curve R(N), with thick solid line as the selected extrapolation $R_5(N)$. *Right*: corresponding extrapolations of the degree of sampling completeness R(N)/S (%), with thick solid line as the selected extrapolation

This relationship, linking the series of f_x to the shape of the 'SAC' (ruled by the successive derivatives $\partial^x R_{(N)}/\partial N^x$ of the 'SAC') is, indeed, the key tool that allows to derive an estimation of the shape of the extrapolated part of the 'SAC' from the already recorded data. This supplementary data, extracted from the series of f_x, thanks to equation (1), thus avoids having to make any assumptions *a priori*, regarding both the kind of species abundance distribution involved and the degree of sampling completeness (as is necessary according to procedures derived in [17-19]). This, indeed, is the very significant contribution of the theoretical approach of the question.

Interestingly, this theoretically based, independently derived approach comforts a most important general trend already highlighted by the semi-empirical approaches [17-19]: the relevance of selecting increasing orders of Jackknife series as the degree of sample completeness decreases (and, in particular, restricting the use of Jackknife-1 or Chao estimators to sample completeness approaching exhaustivity, i.e. when $f_1 \leq f_2$ or $f_1 \leq 0.6\ f_2$, respectively). This convergence, indeed, results from the fact that the classical kinds of species abundance distributions considered in [17-19], although exhibiting significant differences, yet satisfy, for all of them, a common trend towards a negative log-linear dependence between abundances and abundances ranking [23].

Also, defining the boundaries separating the respective domains of use of the different types of estimators on the basis of the recorded values of the f_x not only avoids having to make *a priori* assumptions on the degree of sample completeness and on the type of species abundance distribution (as just mentioned) but also satisfy the desirable continuity of the estimates at the boundaries separating the successive orders m of the estimators Δ_m and $R_m(N)$ (Tables 3 and 4). A requirement that semi-empirical approaches procedures [17-19] cannot satisfy, as may easily be verified.

At last, it should also be noted that the selection of the appropriate type among available estimators, according to the procedure described above (that is Jackknife types of various orders in the general case and Chao restricted to the vicinity of full completeness), satisfies the prescribed rule of additivity (section 2, equation (2)) [21], a required property also shared by the three semi-empirical approaches [17-19], but neglected by the traditional approach [4].

6. CONCLUSION

In short, the derivation of the procedure described here aims at taking a maximum advantage from a preliminary theoretical analysis of the process of the species accumulation during the progressive sampling of species (or of any other kind of objects), thereby resulting in a general relationship between the already recorded data (the series of f_x) and the mathematical parameters (the series of successive derivatives) defining the shape of the extrapolated Species Accumulation Curve: equation (1).

In turn, this preliminary theoretical approach allows building a *general* procedure:

- To define and select the more appropriate (less biased) type of *estimator of total species richness*, only based on the data directly issued from the incomplete sample under consideration (the f_x), without requirement of any prior assumption;
- And also to define and select the more relevant (less biased) expression for the *extrapolation of the Species Accumulation Curve*.

Yet, it should not be forgotten that "less biased" does not signify "unbiased" and it should also be kept in mind that the implemented data (the series of f_x) is submitted to stochastic dispersion, that smoothing by rarefaction or by regression may substantially reduce but does not entirely suppress.

ACKNOWLEDGEMENTS

Four anonymous reviewers are gratefully acknowledged for useful comments on the original manuscript.

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. Van Rooijen J. Estimating the snake species richness of the Santubong Peninsula (Borneo) in two different ways. Contributions to Zoology. 2009;78(4):141-147.
3. Gotelli NJ, Chao A. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: Levin SA (ed.) Encyclopedia of Biodiversity, Second Edition. Waltham, MA: Academic Press. 2013;5:195-211.
4. Colwell RK, Coddington JA. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society London B. 1994;345: 101-118.
5. 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.
6. Assessment of richness estimation methods on macroinvertebrate communities of mountain ponds in Castilla y Leon (Spain). Ann. Limnol. - Int. J. Lim. 2010;46:101–110.
7. Poulin R. Comparison of three estimators of species richness in parasite component communities. Journal of Parasitology. 1998;84(3):485-490.
8. Herzog SK, Kessler M, Cahill TM. Estimating species richness of tropical bird communities from rapid assessment data. The Auk. 2002;119(3):749-769.
9. Chiarucci A, Enright NJ, Perry GLW, Miller BP, Lamont BB. Performance of nonparametric species richness estimators in a high diversity plant community. Diversity and Distributions. 2003;9:283-295.
10. Fogo A, Attrill MJ, Frost MT, Rowden AA. Estimating marine species richness: An evaluation of six extrapolative techniques. Marine Ecology Progress Series. 2003; 248:15-26.
11. Hortal J, Borges PAV, Gaspar C. Evaluating the performance of species richness estimators: Sensitivity to sample grain size. Journal of Animal Ecology. 2006;75:274-287.
12. Unterseher M, Schnittler M, Dormann C, Sickert A. Application of species richness estimators for the assessment of fungal diversity. FEMS Microbiology Letters. 2008;282:205-213.
13. Basualdo CV. Choosing the best non-parametric richness estimator for benthic macroinvertebrates databases. Revista Sociedad Entomologica Argentina. 2011; 70(1-2):27-38.
14. Soberon MJ, Llorente BJ. The use of species accumulation functions for the prediction of species richness. Conservation Biology. 1993;7:480–488.
15. Walter BA, Morand S. Comparative performance of species richness estimation methods. Parasitology. 1998; 116:395-405.
16. Hellmann JJ, Fowler GW. Bias, precision and accuracy of four measures of species richness. Ecological Applications. 1999; 9(3):824-834.
17. Brose U, Martinez ND, Williams RJ. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. Ecology. 2003;84(9): 2364-2377.
18. 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.
19. Béguinot J. Estimer au mieux le degré de complétude d'un inventaire d'espèces: rectificatif au guide de choix. Bull. Soc. Histoire Naturelle Autun. 2016;209:9-10.
20. 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
21. Béguinot J. Basic theoretical arguments advocating Jackknife-2 as usually being the most appropriate nonparametric estimator of total species richness. Annual Research and Review in Biology. 2016; 10(1):1-8.

    DOI: 10.9734/ARRB/2016/25104
22. 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

23. Stevens MHH, Petchey OL, Smouse PE. Stochastic relations between species richness and the variability of species composition. Oikos. 2003;103:479-488.
24. 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
25. Lee SM, Chao A. Estimating population size via sample coverage for closed capture-recapture models. Biometrics. 1994;50(1):88-97.

APPENDICES

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

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:

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

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:

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

We shall now derive the relationship between the successive derivatives of $R_{(N)}$, the theoretical Species Accumulation Curve ('SAC') 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:

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

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:

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

► $f_3 = C_{N,3} \Sigma_i [(1-p_i)^{N-3} p_i^3]$ 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):

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

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:

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

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:

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

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:

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

where $[\partial^x\,\Delta_{(N)}/\partial N^x]$ is the x^{th} derivative of $\Delta_{(N)}$ with respect to N, at point N.
Now, the number of recorded species $R_{(N)}$ is equal to the total species richness S minus the expected number of missed species $\Delta_{(N)}$. Then it comes:

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

with $[\partial^x R_{(N)}/\partial N^x]$ as the x^{th} derivative of $R_{(N)}$ with respect to N, at point N and $C_{N,x} = N!/(N-x)!/x!$

Note that for sufficiently high values of N (N >> x), that is leaving aside the beginning of sampling, the preceding equation simplifies as:

$$\partial^x R_{(N)}/\partial N^x = (-1)^{(x-1)}\,(x!/N^x)\,f_{x(N)} \qquad \text{(A1.10)}$$

N.B.: Applying a Taylor development to the function R(N) and considering the relationship (A1.9, A1.10) offers a *strictly unbiased* expression for the extrapolation of the 'SAC' [24], but only for a limited range of sampling size N (from $N = N_0$ to $N \approx 1.6\,N_0$):

$$R(N) = R(N_0) + \Sigma_x\,[\partial^x R_{(N)}/\partial N^x]\,(N-N_0)^x/x! = R(N_0) + \Sigma_x\,(-1)^{(x-1)}\,f_{x(N0)}\,(N-N_0)^x/x!/C_{No,x}$$

A.2 - 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 f_1/N) or to an already collected species (probability $1–f_1/N$), according to [25]. 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 'SAC' 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 'SAC' 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):

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

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 $f_{2(N)} = –N^2/2.(\partial^2 R_{(N)}/\partial N^2)$ (from equation (A2.3)), 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:

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

More generally:

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

A.3 – Derivation of the expression of the extrapolation of the 'SAC' **in compliance with the two prescribed mathematical constraints**

According to equation (3):

$$R(N) = S - A/N - B/N^2 - C/N^3 + D/N^4 - E/N^5 + \ldots.$$

- *Derivation 1st order*

System of two equations, linear in S, A. At $N = N_0$:

$R(N_0) = S - A/N_0 = R_0$

$\partial R_{(N)}/\partial N = A/N_0^2 = f_1/N_0$

Solving the system above yields $A = f_1$;

$S = R_0 + f_1$

Accordingly:

$R_1(N) = (R_0 + f_1) - f_1.N_0/N$

- *Derivation 2nd order*

System of three equations, linear in S, A, B. At $N = N_0$:

$R(N_0) = S - A/N_0 - B/N_0^2 = R_0$

$\partial R_{(N)}/\partial N = A/N_0^2 + 2B/N_0^3 = f_1/N_0$

$\partial^2 R_{(N)}/\partial N^2 = -2A/N_0^3 - 6B/N_0^4 = -2f_2/N_0^2$

Solving the system above yields $A = (3f_1 - 2f_2).N_0$; $B = (f_2 - f_1).N_0^2$;

$S = R_0 + 2f_1 - f_2$

Accordingly:

$R_2(N) = (R_0 + 2f_1 - f_2) - (3f_1 - 2f_2).N_0/N - (f_2 - f_1).N_0^2/N^2$

- *Derivation 3nd order*

System of four equations, linear in S, A, B, C. At N = No:

$R(N_0) = S - A/N_0 - B/N_0^2 - C/N_0^3 = R_0$

$\partial R_{(N)}/\partial N = A/N_0^2 + 2B/N_0^3 + 3C/N_0^4 = f_1/N_0$

$\partial^2 R_{(N)}/\partial N^2 = -2A/N_0^3 - 6B/N_0^4 - 12C/N_0^5 = -2f_2/N_0^2$

$\partial^3 R_{(N)}/\partial N^3 = 6A/N_0^4 + 24B/N_0^5 + 60C/N_0^6 = 6f_3/N_0^3$

Solving the system above yields $A = (6f_1 - 8f_2 + 3f_3).N_0$; $B = (-4f_1 + 7f_2 - 3f_3).N_0^2$;

$C = (f_1 - 2f_2 + f_3).N_0^3$;

$S = R_0 + 3f_1 - 3f_2 + f_3$

Accordingly:

$R_3(N) = (R_0 + 3f_1 - 3f_2 + f_3) - (6f_1 - 8f_2 + 3f_3).N_0/N - (-4f_1 + 7f_2 - 3f_3).N_0^2/N^2 - (f_1 - 2f_2 + f_3).N_0^3/N^3$

- *Derivation 4nd order*

System of five equations, linear in S, A, B, C, D. At N = No:

$R(N_0) = S - A/N_0 - B/N_0^2 - C/N_0^3 - D/N_0^4 = R_0$

$\partial R_{(N)}/\partial N = A/N_0^2 + 2B/N_0^3 + 3C/N_0^4 + 4D/N_0^5 = f_1/N_0$

$\partial^2 R_{(N)}/\partial N^2 = -2A/N_0^3 - 6B/N_0^4 - 12C/N_0^5 - 20D/N_0^6 = -2f_2/N_0^2$

$\partial^3 R_{(N)}/\partial N^3 = 6A/N_0^4 + 24B/N_0^5 + 60C/N_0^6 + 120D/N_0^7 = 6f_3/N_0^3$

$\partial^4 R_{(N)}/\partial N^4 = -24A/N_0^5 - 120B/N_0^6 - 360C/N_0^7 - 840D/N_0^8 = -24f_4/N_0^4$

Solving the system above yields $A = (10f_1 - 20f_2 + 15f_3 - 4f_4).N_0$; $B = (-10f_1 + 25f_2 - 21f_3 + 6f_4).N_0^2$; $C = (5f_1 - 14f_2 + 13f_3 - 4f_4).N_0^3$; $D = (-f_1 + 3f_2 - 3f_3 + f_4).N_0^4$;

$S = R_0 + 4f_1 - 6f_2 + 4f_3 - f_4$

Accordingly:

$R_4(N) = (R_0 + 4f_1 - 6f_2 + 4f_3 - f_4) - (10f_1 - 20f_2 + 15f_3 - 4f_4).N_0/N - (-10f_1 + 25f_2 - 21f_3 + 6f_4).N_0^2/N^2 - (5f_1 - 14f_2 + 13f_3 - 4f_4).N_0^3/N^3 - (-f_1 + 3f_2 - 3f_3 + f_4).N_0^4/N^4$

- *Derivation 5nd order*

System of six equations, linear in S, A, B, C, D, E. At N = No:

$R(N_0) = S - A/N_0 - B/N_0^2 - C/N_0^3 - D/N_0^4 - E/N_0^5 = R_0$

$\partial R_{(N)}/\partial N = A/N_0^2 + 2B/N_0^3 + 3C/N_0^4 + 4C/N_0^5 + 5E/N_0^6 = f_1/N_0$

$\partial^2 R_{(N)}/\partial N^2 = -2A/N_0^3 - 6B/N_0^4 - 12C/N_0^5 - 20D/N_0^6 - 30E/N_0^7 = -2f_2/N_0^2$
$\partial^3 R_{(N)}/\partial N^3 = 6A/N_0^4 + 24B/N_0^5 + 60C/N_0^6 + 120D/N_0^7 + 210E/N_0^8 = 6f_3/N_0^3$
$\partial^4 R_{(N)}/\partial N^4 = -24A/N_0^5 - 120B/N_0^6 - 360C/N_0^7 - 840D/N_0^8 - 1680/N_0^9 = -24f_4/N_0^4$
$\partial^5 R_{(N)}/\partial N^5 = 120A/N_0^6 + 720B/N_0^7 + 2520C/N_0^8 + 6720D/N_0^9 + 15120/N_0^{10} = -120f_5/N_0^5$

Solving the system above yields $A = (15f_1 - 40f_2 + 45f_3 - 24f_4 + 5f_5).N_0$; $B = (-20f_1 + 65f_2 - 81f_3 + 46f_4 - 10f_5).N_0^2$; $C = (15f_1 - 54f_2 + 73f_3 - 44f_4 + 10f_5).N_0^3$; $D = (-6f_1 + 23f_2 - 33f_3 + 21f_4 - 5f_5).N_0^4$; $E = (f_1 - 4f_2 + 6f_3 - 4f_4 + f_5).N_0^5$;

$S = R_0 + 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5$

Accordingly:

$R_5(N) = (R_0 + 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5) - (15f_1 - 40f_2 + 45f_3 - 24f_4 + 5f_5).N_0/N - (-20f_1 + 65f_2 - 81f_3 + 46f_4 - 10f_5).N_0^2/N^2 - (15f_1 - 54f_2 + 73f_3 - 44f_4 + 10f_5).N_0^3/N^3 - (-6f_1 + 23f_2 - 33f_3 + 21f_4 - 5f_5).N_0^4/N^4 - (f_1 - 4f_2 + 6f_3 - 4f_4 + f_5).N_0^5/N^5$

- *Generalisation : derivation at the order m*

More generally, the estimation Δ_m of the number of missing species, computed at order m is:

$$\Delta_m = \Sigma_{x=1 \text{ to } m} [(-1)^{(x-1)}.C_{(m, x)}.f_x] \quad (A3.1)$$

where $\Sigma_{x=1 \text{ to } m}$ stands for the summation from x = 1 to x = m and $C_{(m, x)} = m!/x!/(m-x)!$ is the number of combinations of x objects among m. Alternatively, accounting for equation (1), the estimation Δ_m of the number of missing species may also be written as:

$$\Delta_m = \Sigma_{x=1 \text{ to } m} [C_{(m, x)}.C_{(N, x)}.\partial^x R_{(N)}/\partial N^x] \quad (A3.2)$$

with N as the number N_0 of individuals already recorded in the sample.

In both formulations, the estimation Δ_m of the number of unrecorded species involves (i) the recorded data characteristic of the actually realised sampling (either the series f_x or the series of $\partial^x R_{(N)}/\partial N^x$) and (ii) combinatorial analysis (*via* $C_{(m, x)}$, $C_{(N, x)}$).

The same holds true for the general, polynomial expression $R_m(N)$ of the extrapolated 'SAC' (equation (3)): as for Δ_m above, the coefficients A, B, C,… are in terms of the series of f_x (or the series of $\partial^x R_{(N)}/\partial N^x$), in connection with combinatorial analysis. Thus, it may be easily controlled that the coefficients A_m, B_m, …, of the extrapolation $R_m(N)$ at order m (equation (3)) are defined as follows.

For A_m :

$$A_m = N_0.\Sigma_{x=1 \text{ to } m} [(-1)^{(x-1)}.x.C_{(m+1, x+1)}.f_x] \quad (A3.3)$$

For B_m :

$$B_m = N_0^2.\Sigma_{x=1 \text{ to } m} [B_m(x).f_x] \quad (A3.4)$$

the terms $B_m(x)$ being defined by recurrence as:

$$B_m(x) = B_{m-1}(x) + (-1)^x.C_{(m-1, x-1)}.C_{(m, 2)} \quad (A3.5)$$

$$\text{with } B_{m-1}(x) = 0 \text{ when } x = m \text{ and also } B_1(1) = 0 \quad (A3.6)$$

etc …

The limits of the range of preferred use of estimators at order m (i.e. Δ_m and $R_m(N)$) are defined by the values of f_1 which satisfy respectively: $\Delta_m = \Delta_{m-1}$ and $\Delta_m = \Delta_{m+1}$. Accordingly, the limits of the range of preferred use of the estimator Δ_m and the extrapolation of the 'SAC' $R_m(N)$ are respectively:

$$f_1 = \Sigma_{x=2 \text{ to } m-1} [(-1)^x.(C_{(m, x)} - C_{(m-1, x)}).f_x] + (-1)^m.f_m \quad \text{(A3.7)}$$

$$f_1 = \Sigma_{x=2 \text{ to } m} [(-1)^x.(C_{(m+1, x)} - C_{(m, x)}).f_x] + (-1)^{m+1}.f_{m+1} \quad \text{(A3.8)}$$

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/14875