On general mathematical constraints applying to the kinetics of species discovery during progressive sampling and to the resulting expression of the Species Accumulation Curve

ABSTRACT

The "Species Accumulation Curve" accounts for the rate of increase of the number of recorded species during progressive sampling of an assemblage of species. Due to the usual incompleteness of samplings, the accurate extrapolation of the Species Accumulation Curve has become an essential tool to estimate the total species richness of a sampled assemblage and to predict the additional sampling effort required to obtain a given increase of sample completeness. In this perspective, important efforts have been devoted to improve the accuracy of the extrapolation of the Species Accumulation Curves. Substantial progress in this respect was achieved recently by considering a general mathematical relationship that constrains the theoretical expression of any kind of Species Accumulation Curves. Moreover, this general relationship proves having interesting corollaries applying specifically to the detailed process of species accumulation during progressive sampling.

Hereafter, I first derive these correlative relationships and then I show how they link together the variations of the numbers of species respectively recorded 1-, 2-, 3- …, x- times and their cumulative contributions to the Species Accumulation Curve. This, in turn, provides suggestive insights regarding the remarkably regulated mechanism of species discovery and accumulation along progressive sampling effort.

Key-words: species accumulation curve, extrapolation, estimation, species richness, incomplete sampling, regulation, Jackknife estimator

1. INTRODUCTION

The process of continuous discovery of new species during progressive sampling of an assemblage of species is expressed graphically in term of the so called "Species Accumulation Curve", also formerly designed as "Discovery Curve" or "Collector Curve" [1]. The Species Accumulation Curve is the basic tool, which is systematically referred to in inventories of biodiversity [2].

Species Accumulation Curves are quite polymorphic, apart from some basic and intuitive common traits shared by all of them (monotonic increase of the number of recorded species with sampling size, at consistently decreasing rate, see Figure 1 for an example). This is so because the detailed shape of the Species Accumulation Curves is entirely dependent upon the particular species abundance distribution within the sampled assemblage of species under consideration. Accordingly, there are virtually as many different shapes of Species Accumulation Curves as there are species assemblages differing from each other by either their species richness and/or their particular distribution of species abundances.

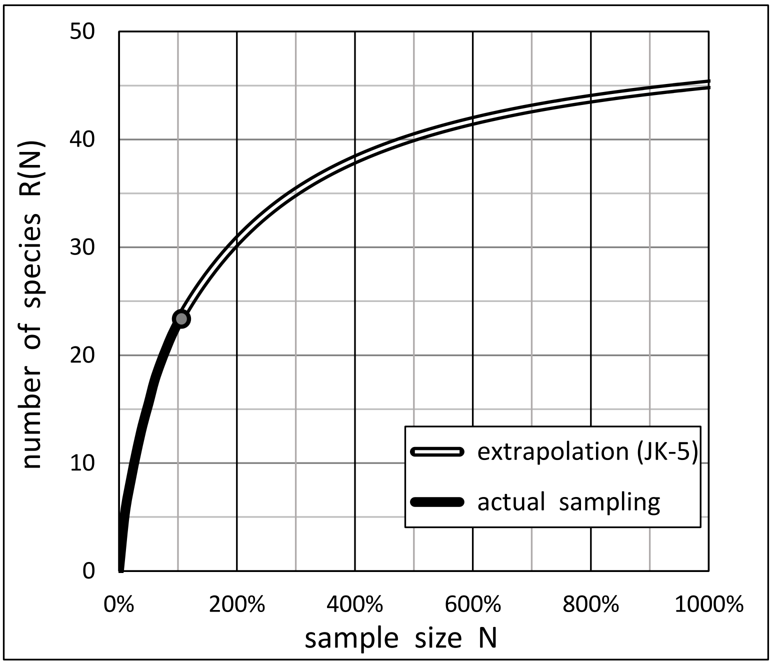

Figure 1 – Typical shape of a Species Accumulation Curve, showing the basic common features pertaining to any kind of S.A.C.: monotonic increase of the number of recorded species R(N) with sampling size N, while the rate of growth is monotonically decreasing. Here is plotted the S.A.C for a partial inventory of land snails fauna in a xerothermic grassland at 'Cersot', south Burgundy (France) [from BÉGUINOT, *unpublished data*]. Sampling size is expressed in % of the size of the actually achieved sampling. Extrapolation is estimated according to Jackknife-5 estimator, selected as being the less biased for this particular inventory (see reference [3]).

In spite of these causes of polymorphism, the theoretical expressions of all Species Accumulation Curves are compelled to satisfy a common constraining mathematical relationship which applies to the whole series of its successive derivatives. This constraining relationship explicitly determines the boundaries of the yet wide range of polymorphism mentioned above for the Species Accumulation Curves. On a more practical point of view, accounting for this constraining relationship reveals also of major importance to *improve the accuracy of extrapolations* of species accumulation beyond actually achieved samplings. Thereby, more precise estimations of total species richness and more reliable predictions of the additional sampling effort needed to achieve a given increase in sample completeness are made possible (details in reference [3]).
Now, coming back to more *theoretical* ground, several corollaries which can be derived from this fundamental relationship also provide useful insights into the details of the complex process of species discovery during progressive sampling.
Let R(N) be the number of recorded species after sampling of N individuals (N thus quantify the sampling size). Obviously, R(N) results from the additive contributions 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 at the end of this sampling of size N:

$$R(N) = \Sigma_x f_x(N) \qquad (1)$$

Thereby, the Species Accumulation Curve reveals its "composite" dependence upon the whole series of the $f_x(N)$. A composite dependence which is made still more complex by the fact that each function $f_x(N)$ has its own dependence upon N. Yet, this mutual independence of the fx (N) is not total: a kind of regulation links, step by step, the respective variations of the successive functions $f_x(N)$, as will be shown later. This regulation, indeed, is at the hearth of the mechanism of species progressive discovery and accumulation, which plays, of course, a decisive role in shaping the Species Accumulation Curve.

The main purpose of this article is precisely to highlight the mathematics underlying this regulation by mutual linkage between the successive fx (N). This, in turn, will provide a more deep understanding of the fundamentals of Species Accumulation during progressive sampling.

Preliminary: the fundamental mathematical relationship constraining the theoretical expression of all Species Accumulation Curves

The successive derivatives $\partial^x R(N)/\partial N^x$, of the Species Accumulation Curve R(N) satisfy the following equation:

$$[\partial^x R_{(N)}/\partial N^x] = (-1)^{x-1} f_{x\,(N)} / C_{N,\,x} \qquad (2)$$

with $f_{x\,(N)}$ as the number of species recorded x-times in the sample of size N and $C_{N,\,x} = N!/X!/(N-x)!$ is the number of combinations of x items among N. A detailed proof of this general theorem is provided in Appendix.

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

$$[\partial^x R_{(N)}/\partial N^x] = (-1)^{x-1} (x!/N^x)\, f_{x\,(N)} \qquad (3)$$

This relation has a *general relevance* because its derivation does not require any specific assumption relative to the particular shape of the distribution of species abundances in the sampled assemblage of species. Accordingly, equations (2), (3) actually constrain the theoretical expressions of *any kind* of Species Accumulation Curves.

One particular consequence of this relationship is that the successive derivatives of the Species Accumulation Curve have alternating signs, since the numbers $f_{x\,(N)}$ of species recorded x-times are necessarily positive or nil. More precisely, the derivatives of even and odd orders are respectively negative and positive.

2. THE MATHEMATICS UNDERLYING THE REGULATION PROCESS APPLYING TO THE NUMBERS f_x OF SPECIES RECORDED x-TIMES

From equation (3) it comes:

$$f_{x\,(N)} = (-1)^{x-1} (N^x/x!)\, [\partial^x R_{(N)}/\partial N^x] \qquad (4)$$

The derivation of equation (4) according to sample size N then gives:

$$[\partial f_{x\,(N)}/\partial N] = (-1)^{x-1}/x!\, \{x.\, N^{x-1}.[\partial^x R_{(N)}/\partial N^x] + N^x.[\partial^{x+1} R_{(N)}/\partial N^{x+1}]\}$$

Accounting for the expression (3), applied to $[\partial^x R_{(N)}/\partial N^x]$ and $[\partial^{x+1} R_{(N)}/\partial N^{x+1}]$, it comes:

$$[\partial f_{x\,(N)}/\partial N] = (1/N).[\, x.f_{x\,(N)} - (x+1).f_{x+1\,(N)}\,] \qquad (5)$$

which may be written as well as:

$$x.f_{x\,(N)} - N.[\partial f_{x\,(N)}/\partial N] = (x+1).f_{x+1\,(N)} \qquad (6)$$

Note that an alternative, independent demonstration of the equation (6) is provided at Appendix A.2, equation A2.1.

As a corollary of relationship (3) above, equation (6) benefits consequently from the same *general relevance* and, thus, is valid for *all kinds* of Species Accumulation Curves.

Equation (6) establishes a *mathematical linkage* between the variations of $f_{x+1\,(N)}$ with N and the variations of $f_{x\,(N)}$ with N. Thereby, all the $f_{x\,(N)}$ are ultimately *linked together* by this “iterative chaining”. In other words, although each function $f_{x\,(N)}$ has its own dependence upon sampling size N, the series of $f_{x\,(N)}$ nevertheless admits a kind of connection which, if one may say, "propagates" from each $f_{x\,(N)}$ to the next one, $f_{x+1\,(N)}$.

Mathematical "chaining" between the successive numbers $f_{x\,(N)}$

The consequence of this regulation may be more easily grasped graphically, by considering how the maximum of each $f_{x\,(N)}$ is linked to the value of $f_{x+1\,(N)}$ at the same sample size N. When $f_{x\,(N)}$ reaches its maximum value, its first derivative, $\partial f_{x\,(N)}/\partial N$, falls to zero and, accordingly, from equation (6), it comes:

$$f_{x+1\,(N)} = [x/(x+1)].f_{x\,(N)} \qquad (7)$$

Thus, when $f_{x\,(N)}$ reaches its maximum, in the course of progressive sampling, the corresponding value taken by $f_{x+1\,(N)}$ is then exactly $[x/(x+1)]$ times the (maximum) value taken by $f_{x\,(N)}$. By reiteration of this relationship, a kind of "linkage pattern" is generated, that constrains the relative locations of the successive curves $f_{x\,(N)}$. Figure 2 exemplifies graphically this "chaining" linkage, *propagating successively, step by step*, from $f_{1\,(N)}$ to $f_{2\,(N)}$, to $f_{3\,(N)}$, to $f_{4\,(N)}$, to $f_{5\,(N)}$, etc...

As a consequence, the maxima of $f_{1\,(N)}$, of $f_{2\,(N)}$, of $f_{3\,(N)}$, of $f_{4\,(N)}$, of $f_{5\,(N)}$, ..., respectively succeed each other sequentially, as shown in Figure 2. The corresponding positions of these succeeding maxima are located along the Species Accumulation Curve at Figure 3, and it is worth noting that the regulating linkage between the successive $f_{x\,(N)}$ is such that no peculiarity is affecting the Species Accumulation Curve at any of these locations (in spite of the series of bumps constituted by the successive maxima of $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{3\,(N)}$, $f_{4\,(N)}$, $f_{5\,(N)}$,...)

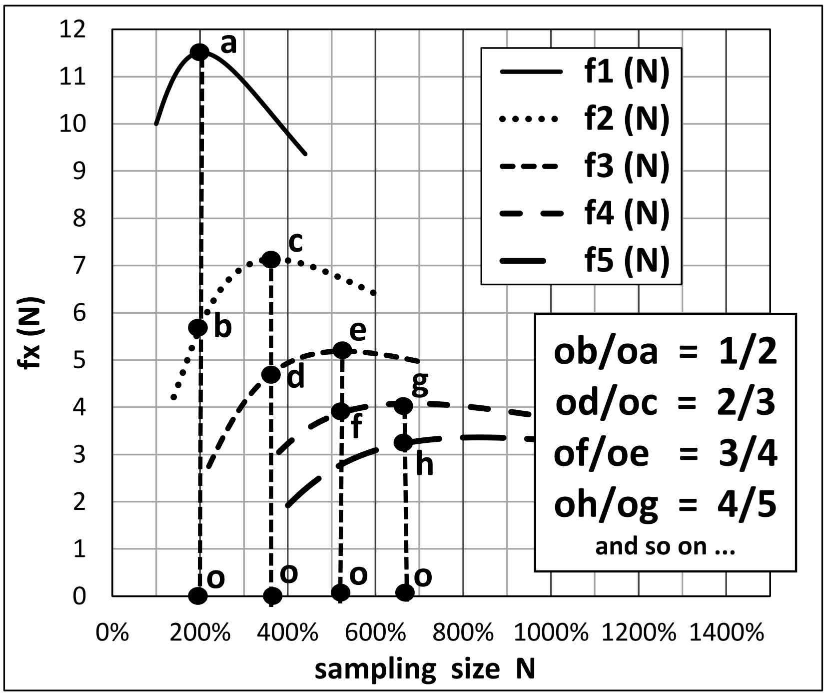

Figure 2 – Extrapolations of the five first $f_{x\,(N)}$ ($f_{1\,(N)}$, $f_{2\,(N)}$, $f_{3\,(N)}$, $f_{4\,(N)}$, $f_{5\,(N)}$) for increasing sampling size N beyond the size of the actually achieved sampling (sampling size N is expressed in % of the size of the actually achieved sampling). Here, the maxima of $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{3\,(N)}$, $f_{4\,(N)}$, $f_{5\,(N)}$ happens to be located at sample size ≈ 200%, 360%, 510%, 680%, 810%, respectively. Same inventory as in Figure 1 ; extrapolations according to Jackknife-5 estimator, selected as being the less biased for this particular inventory (see reference [3]). This figure highlights the "linkage pattern" between the successive curves $f_{x\,(N)}$ imposed by the constraining relationship (7) (i.e.: $f_{x+1\,(N)} = [x/(x+1)].f_{x\,(N)}$). That is when $f_{x\,(N)}$ reaches its maximum, the corresponding value taken by $f_{x+1\,(N)}$ is then exactly $[x/(x+1)]$ times the value of $f_{x\,(N)}$.

Mathematical “chaining” between the successive numbers $x.f_{x\,(N)}$

Alternatively, equation (7) may be written equivalently as:

$$x.f_{x\,(N)} = (x+1).f_{x+1\,(N)} \qquad (8)$$

Equation (8), as equation (7), stands for $\partial f_{x\,(N)}/\partial N = 0$, and thus stands as well for $\partial(x.f_{x\,(N)})/\partial N = 0$. It follows that the curve $(x+1).f_{x+1\,(N)}$ *intersects* the curve $x.f_{x\,(N)}$ exactly when the latter *reaches its maximum value* (i.e. when $\partial(x.f_{x\,(N)})/\partial N = 0$) : Figure 4. Keeping in mind the significance of $x.f_{x\,(N)}$ which is the total number of recorded individuals belonging to either of those species recorded x-times.

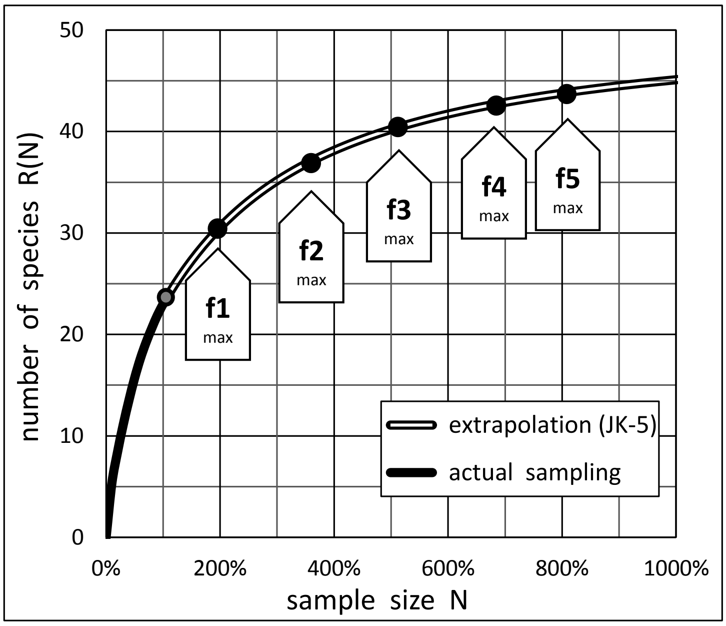

Figure 3 – The locations, along the Species Accumulation Curve, of the successive maximum of $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{3\,(N)}$, $f_{4\,(N)}$, $f_{5\,(N)}$, according to Figure 2. Same inventory as in Figure 1.

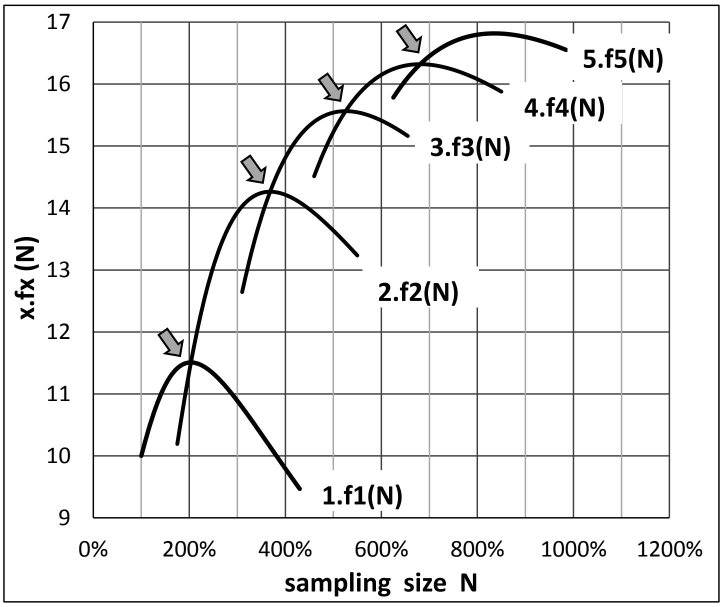

Figure 4 - The computed variations, with increasing sampling size N, of the five first numbers $x.f_{x\,(N)}$ of recorded individuals belonging to species recorded x-times ($1.f_{1\,(N)}$, $2.f_{2\,(N)}$, $3.f_{3\,(N)}$, $4.f_{4\,(N)}$, $5.f_{5\,(N)}$). As prescribed by the constraining equation (8), *for any value of x*, the curve $(x+1).f_{x+1\,(N)}$ intersects the curve $x.f_{x\,(N)}$ exactly when the latter reaches its maximum value. Sampling size is expressed in % of the size of the actually achieved sampling. Same inventory as in Figure 1.

The regularly repetitive shift from any one curve, $x.f_{x\,(N)}$, to the next one, $(x+1).f_{x+1\,(N)}$, resulting from this regulating process (Figure 4) is particularly demonstrative. This, indeed, likely offers the best visual evidence of the sequential linkage existing between each of the $f_{x\,(N)}$ successively.

Note, incidentally, that while the cumulative addition of all the $f_{x\,(N)}$ leads to the number R(N) of recorded species (*cf.* equation (1)) ; on the other hand the addition of the $x.f_{x\,(N)}$ leads "symmetrically" to the number N of recorded individuals:

$$\Sigma_x [\, f_x(N) \,] = R(N) \quad \text{and} \quad \Sigma_x [\, x.f_x(N) \,] = N \qquad (9)$$

Mathematical "chaining" between each $f_{x\,(N)}$ and the series of the first derivatives of the preceding $f_{x\,(N)}$

This is a third alternative way to express the inter-relationship within the series of the $f_x(N)$. Refering once more to equation (6), that is:

$$x.f_{x\,(N)} - N.[\partial f_{x\,(N)}/\partial N] = (x+1).f_{x+1\,(N)}$$

let now consider the successive forms taken by this equation for increasing values of x. It comes:

$$0.f_{0\,(N)} - N.[\partial f_{0\,(N)}/\partial N] = 1.f_{1\,(N)}$$
$$1.f_{1\,(N)} - N.[\partial f_{1\,(N)}/\partial N] = 2.f_{2\,(N)}$$
$$2.f_{2\,(N)} - N.[\partial f_{2\,(N)}/\partial N] = 3.f_{3\,(N)}$$

…………………………………………………

$$(x-1).f_{x-1\,(N)} - N.[\partial f_{x-1\,(N)}/\partial N] = x.f_{x\,(N)}$$

By summing these equations, the following relationship is immediately derived:

$$f_{x\,(N)} = -(N/x).\, \Sigma_i [\partial f_{i\,(N)}/\partial N] \qquad (10)$$

with the summation Σ_i extended from i = 0 to i = (x – 1)

That is, namely, the number $f_{x\,(N)}$ of species recorded x-times in a sampling of size N is proportional [via the factor – (N/x)] to the sum of the first derivatives (with respect to N) of the series of the preceding $f_{i\,(N)}$. In more practical terms, this means that the number $f_{x\,(N)}$ of species recorded x-times, in a sample of size N, is proportional to the sum of the variations of the preceding $f_{x\,(N)}$ when sampling size increases of one observation (N → N+1).

Another way to understand relation (10) results from re-writing it as follows:

$$(x.f_{x\,(N)})/N = -\Sigma_i [\partial f_{i\,(N)}/\partial N] \qquad (11)$$

with the summation Σ_i extended from i = 0 to i = (x – 1).

Which means that the proportion of sampled individuals that belong to either species recorded x-times [= $(x.f_{x\,(N)})/N$] is equals to minus the sum of the variations of the preceding $f_{x\,(N)}$ when sampling size increases of one observation.

Accordingly, the relationship (10) expresses the continuous linkage between each $f_{x\,(N)}$ and the *whole series of its predecessors*, thereby highlighting still more clearly the strong "chaining" between the successive numbers $f_{x\,(N)}$, which together rule the kinetics of species accumulation during progressive sampling.

Still another remarkable relationship may be derived from equation (10), which only involves, this time, the first derivatives of all the $f_{x\,(N)}$.

Let X be the recorded number of individuals belonging to the species most frequently met in the sampling under consideration. In other words, X is the largest value of x for which $f_{x\,(N)} \neq 0$ in this particular sampling. The sum of the numbers of sampled individuals that belong to anyone of those species recorded x-times [= $(x.f_{x\,(N)})$] for x up to its maximum value X is equal to N. Accordingly, the summation of equation (11) for x up to its maximum value X yields:

$$\Sigma_x\, [(x.f_{x\,(N)})/N] = N/N = 1 = \Sigma_x\, [-\,\Sigma_i\, (\partial f_{i\,(N)}/\partial N)]$$

with the summation Σ_x extended from x = 1 to x = X and the summation Σ_i extended from i = 0 to i = (x – 1). This finally leads to:

$$1 \; = \; -\,\Sigma_i\, [(\partial f_{i\,(N)}/\partial N).(X - i)] \qquad (12)$$

with the summation Σ_i extended from i = 0 to i = (X – 1).

3. COMPLEMENTARY ILLUSTRATIVE EXAMPLES OF THE REGULATION PROCESS GOVERNING THE NUMBERS OF SPECIES RECORDED 1-, 2-, 3-, 4-, 5-, … TIMES

The same trends demonstrated above on a theoretical basis, and illustrated by a first example at Figures 2 to 4, are illustrated again in the following complementary examples. We shall only consider, hereafter, the variations of $x.f_{x\,(N)}$ (i.e. the number of recorded individuals belonging to any one of species recorded x-times) as they provide the more graphically speaking feature, highlighting at best the specific “chaining” linkage, step by step, between the successive numbers $f_{1\,(N)}$, $f_{2\,(N)}$, $f_{3\,(N)}$, $f_{4\,(N)}$, $f_{5\,(N)}$.

3.1. Butterfly inventory on the slopes of Mount Gariwang-san (S-Korea)

Field data from reference [4]. Figure 5 relates to an inventory carried out during year 2015. The realised sampling size was N_0 = 181individuals; the number of recorded species was $R(N_0)$ = 39 species and the values of the f_x, were: f_1 = 17.0, f_2 = 8.0, f_3 = 3.7, f_4 =2.0, f_5 = 1.4 (values obtained after regression applied to the crude values of the f_x, in order to reduce the consequences of stochastic dispersion, as prescribed in [3].

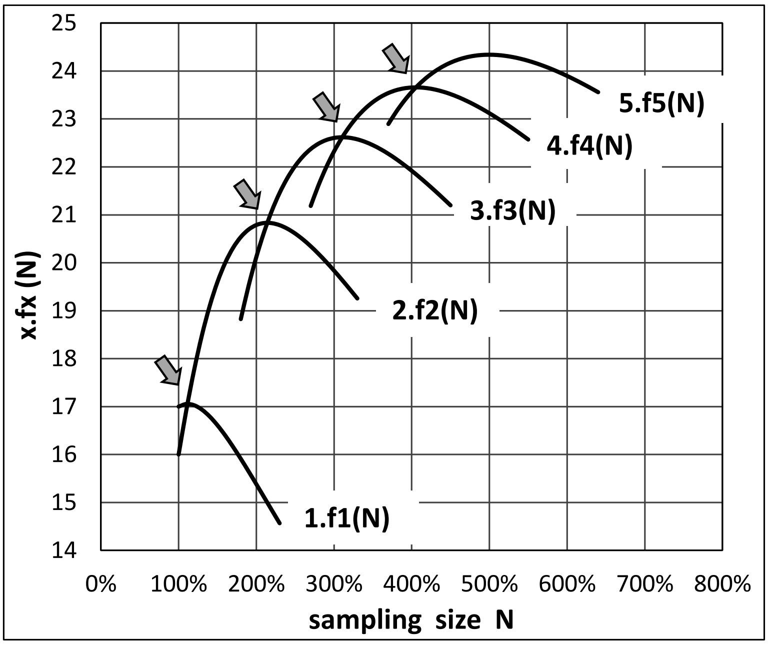

Figure 5 - The computed variations, with increasing sampling size N, of the five first numbers $x.f_{x\,(N)}$ of recorded individuals belonging to any one of species recorded x-times ($1.f_{1\,(N)}$, $2.f_{2\,(N)}$, $3.f_{3\,(N)}$, $4.f_{4\,(N)}$, $5.f_{5\,(N)}$). Butterfly inventory on the slopes of Mount Gariwang-san (field data from LEE, KIM & KWON 2016). As prescribed by the constraining equation (8), *for all values of x*, the curve $(x+1).f_{x+1\,(N)}$ intersects the curve $x.f_{x\,(N)}$ exactly when the latter reaches its maximum value. Sampling size is expressed in % of the size of the actually achieved sampling.

3.2 Butterfly inventories at Bifeng Valley (Ghansu, China)

Field data from reference [5]. Figure 6 relates to an inventory conducted at site 1, in the upper part of the valley of Bifeng (province of Ghansu), where species richness of butterfly fauna proves to be the highest in the valley, with $R(N_0) = 28$ recorded species for $N_0 = 68$ recorded individuals. The values of the f_x, were: $f_1 = 14.0$, $f_2 = 6.0$, $f_3 = 2.6$, $f_4 = 1.5$, $f_5 = 1.2$ (values obtained after regression applied to the crude values of the f_x, in order to reduce the consequences of stochastic dispersion, as prescribed [3].

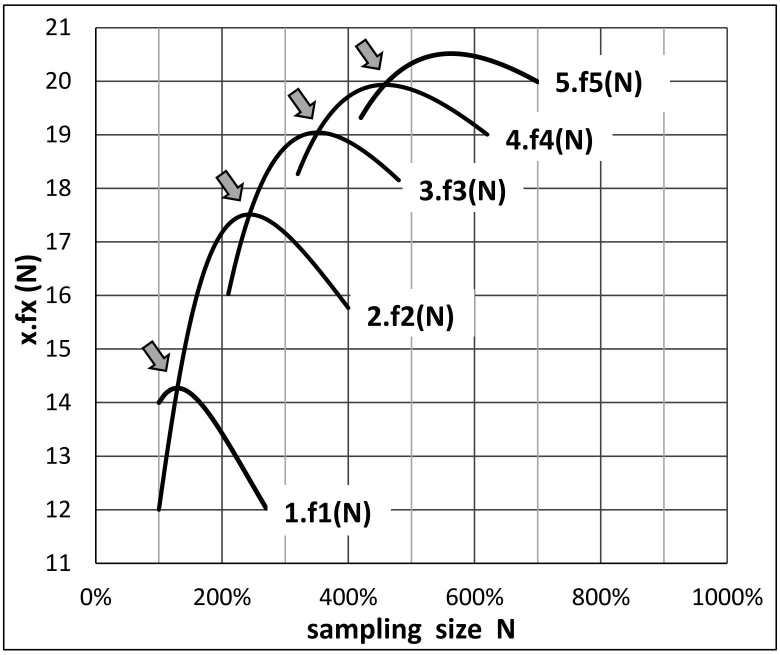

Figure 6 - The computed variations, with increasing sampling size N, of the five first numbers $x.f_{x\,(N)}$ of recorded individuals belonging to any one of species recorded x-times ($1.f_{1\,(N)}$, $2.f_{2\,(N)}$, $3.f_{3\,(N)}$, $4.f_{4\,(N)}$, $5.f_{5\,(N)}$). Butterfly inventory at Bifeng valley, site 1 (field data from Li *et al.* 2010). As prescribed by the constraining equation (8), *for all values of x*, the curve $(x+1).f_{x+1\,(N)}$ intersects the curve $x.f_{x\,(N)}$ exactly when the latter reaches its maximum value. Sampling size is expressed in % of the size of the actually achieved sampling.

4. THE GENERAL RELATIONSHIP GOVERNING THE DECREASING PROPOTION OF OBSERVATIONS PROVIDING NEWLY RECORDED SPECIES, WITH INCREASING SAMPLING SIZE

So far, we have approached the Species Accumulation Curve, R(N), in a deliberately *analytical manner*: each new observation was considered as equally informative, whether or not it gives rise to the detection of a new species. Indeed, in any case, each individual observation actually plays the same kind of role: it determines a transition of the kind $f_x(N) \rightarrow f_{x+1}(N+1)$.

Now, let consider, *alternatively*, a more usual and pragmatic approach, now paying attention to those observations *only* giving rise to the detection of a new species and neglecting accordingly all the others (in spite of their equal role in the analytical approach considered previously above). In this purely “accounting” approach, the focus is put on the proportion p(N) = R (N)/N of those observations exclusively, which have provided positive records of new species. In other words, instead of paying attention to $R(N) = \Sigma_x f_x(N)$, as previously, the focus is placed now upon:

$$R(N) = N.p(N) \qquad (13)$$

This proportion p(N) is pragmatically interesting in that it quantifies the gradual weakening of sampling efficiency, i.e. the ever-slowering rate of detection of newly recorded species, as sampling is going on further.

As for the Species Accumulation Curve, the proportion p(N) of those observations providing positive records of new species is highly polymorphic and this polymorphism, here also, is limited by a constraining relationship applying to the expression of p(N).

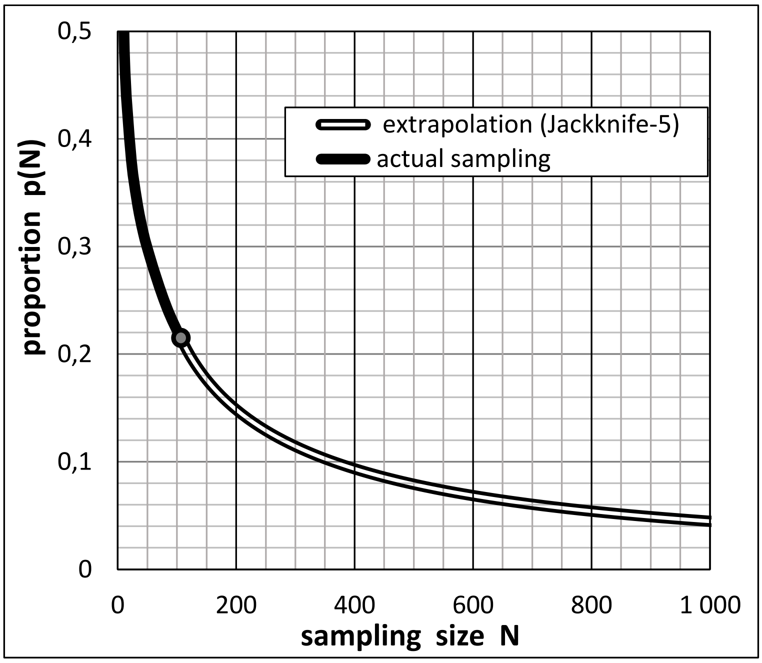

Figure 7 – Typical shape of the proportion p(N) = R (N)/N of those observations providing positive records of new species. Same inventory as in Figure 1.

I derive below this general relationship which constrains the proportion p(N).

The derivation of R(N) yields, accounting for equation (13) and then equation (3):

$$[\partial R_{(N)}/\partial N] = p(N) + N.[\partial p(N)/\partial N] = (1/N).f_{1\,(N)} \qquad (14)$$

Further derivations yield successively:

$$[\partial^2 R_{(N)}/\partial N^2] = 2.[\partial p(N)/\partial N] + N.[\partial^2 p(N)/\partial N^2] = -(2/N^2).\, f_{2\,(N)} \qquad (15)$$

$$[\partial^3 R_{(N)}/\partial N^3] = 3.[\partial^2 p(N)/\partial N^2] + N.[\partial^3 p(N)/\partial N^3] = (6/N^3).\, f_{3\,(N)} \qquad (16)$$

and more generally:

$$[\partial^x R_{(N)}/\partial N^x] = x.[\partial^{x-1} p(N)/\partial N^{x-1}] + N.[\partial^x p(N)/\partial N^x] = (-1)^{x-1}.(x!/N^x).f_{x\,(N)} \qquad (17)$$

Now, from equations (14) and (15), it follows:

$$[\partial^2 p(N)/\partial N^2] = (2/N^2).p(N) - (2/N^3).(f_{1\,(N)} + f_{2\,(N)}) \qquad (18)$$

Similarly, from equations (16) and (18):

$$[\partial^3 p(N)/\partial N^3] = -(6/N^3).p(N) + (6/N^4).(f_{1\,(N)} + f_{2\,(N)} + f_{3\,(N)}) \qquad (19)$$

and more generally:

$$[\partial^x p(N)/\partial N^x] = (-1)^x.(x!/N^x).p(N) + (-1)^{x-1}.\,(x!/N^{x+1}).\, \Sigma_{i=1 \text{to } x}\, [f_i(N)] \qquad (20)$$

At last, from equations (1) and (11), it follows:

$$[\partial^x p(N)/\partial N^x] = (-1)^x.(x!/N^{x+1}).\, \Sigma_{i>x}\, [f_i(N)] \qquad (21)$$

Note that there is part of a formal similarity between the general relationship (21) constraining the proportion p(N) of those individual observations providing positive records of new species and the general relationship (3) constraining the Species Accumulation Curve R(N). Among the differences, however, the main one is that all the $f_i(N)$ with i > x are involved in the relationship (21) constraining the proportion p(N) while it is only $f_i(N)$ for i = x which contributes in the relationship (3) constraining the Species Accumulation Curve R(N).

One particular consequence of relationship (21) is that the successive derivatives of the the proportion p(N) of observations providing positive records of new species have alternating signs, (as for the Species Accumulation Curve) since the numbers $f_{x\,(N)}$ of species recorded x-times are necessarily positive or nil. More precisely, for the proportion p(N), the derivatives of even and odd orders are respectively positive and negative, that is the inverse of what is valid for the Species Accumulation Curves.

5. DISCUSSION

Five main features are emerging from the theoretical treatment (and the corresponding illustrative examples), regarding the variations, with sampling size N, of the numbers $f_x(N)$ of species respectively recorded x-times during sampling. It should be well understood that these features, all derived on theoretical basis, are *focal tendencies*, towards which the empirical data obtained from real samplings actually converges. But may yet more or less slightly deviate, due to sampling stochasticity.

Two trends were expected, being in obvious *accordance with intuition*:

1) all the numbers $f_{x\,(N)}$ of species recorded x-times are first increasing, then pass by a maximum and finally decrease to zero. Also, in addition, the curves describing the variations of each $f_{x\,(N)}$ (and the positions of their respective maxima) are regularly shifted towards higher values of sampling size N when x takes increasing values (Figure 2) ;

2) the same holds true, *mutatis mutandis*, for the number $x.f_{x\,(N)}$ of those recorded individuals belonging to anyone of species recorded x-times, whatever the value of x.

Now, three other trends, *by no means intuitive*, were newly derived above, related to the general mathematical relationship (6) which constrains the expressions and shapes of the $f_{x\,(N)}$:

3) when $f_{x\,(N)}$ reaches its maximum, in the course of progressive sampling, the corresponding value taken by $f_{x+1\,(N)}$ is then exactly [x/(x+1)] times the maximum value taken by $f_{x\,(N)}$ (see Figure 2) ;

4) regarding now the number $x.f_{x\,(N)}$ of recorded individuals belonging to anyone of those species recorded x-times, it consistently happens that the curve describing the variations of $(x+1).f_{x+1\,(N)}$ intersects the curve of $x.f_{x\,(N)}$ exactly when the latter reaches its maximum value (see Figures 4, 5, 6) ;

5) at last, the number $f_{x\,(N)}$ of species recorded x-times in a sampling of size N is proportional to the sum of the variations of the preceding $f_{x\,(N)}$, when sampling size increases by one observation.

The three latter trends have *major importance* in that they determine the "chaining linkage" between the successive numbers $f_{x\,(N)}$ of species recorded x-times. And this is of importance because the successive numbers $f_{x\,(N)}$ regulate the process of cumulative species discovery during progressive sampling.

As already stressed, the general mathematical relationship (6)

$$x.f_{x\,(N)} - N.[\partial f_{x\,(N)}/\partial N] \;=\; (x+1).f_{x+1\,(N)}$$

which constrains the expressions and the shapes of the $f_{x\,(N)}$, is a corollary of the general relationship (A2.1), derived in Appendix A.2, which, in turn, constrains the theoretical expressions of all Species Accumulation Curves R(N). Thus, to get a full understanding of the underlying process of species accumulation during sampling, it is advisable to refer to the detailed demonstration leading to equation (A2.1) in Appendix.

6. CONCLUSION

The increasing number of newly recorded species (i.e. the "species accumulation") during progressive sampling gives rise to a rather simple shaped "Species Accumulation Curve". This apparent simplicity, indeed, does not incite to imagine the underlying complexity of the detailed process of species discovery and progressive accumulation. In fact, each new individual observation may alternatively result in one or the other of a series of different consequences. More precisely, each observation of a new individual (i.e. $N \rightarrow N + 1$) will contribute to increase by one unity either $f_{1\,(N)}$, or $f_{2\,(N)}$, or $f_{3\,(N)}$, …, $f_{x\,(N)}$, …

Now, although each of the numbers $f_{x\,(N)}$ of species recorded x-times varies with N at *its own pace* and *out of phase* with the others (Figure 2), the process of species accumulation proves to be regulated, however, due to the above mentioned "chaining linkage" between the successive $f_{x\,(N)}$. And this indeed is at the very heart of the detailed process of species accumulation during progressive sampling. A process of major practical importance since it is involved in all biodiversity surveys and, more specifically, is involved in the accurate extrapolation of the Species Accumulation Curve. Accurate extrapolation which, in turn, conditions the precise estimate of the total species richness of a partially sampled assemblage of species and the reliable prediction of the additional sampling effort required to obtain a given increase in sample completeness.

The constraining mathematical relationships highlighted above are summarized as follows :

* relationship applying to the *Species Accumulation Curve*, R(N) itself :

equations (2) & (3):

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

* relationship applying to the *proportion of efficient observations*, $p(N) = R(N)/N$:

equation (21):

$$[\partial^x p(N)/\partial N^x] = (-1)^x.(x!/N^{x+1}).\,\Sigma_{i>x}\,[f_i(N)]$$

* relationship applying to the numbers $f_1(N), f_2(N), f_3(N), \ldots, f_x(N), \ldots$ *of those species respectively recorded 1, 2, 3, .., x-times* during sampling:

equation (6):

$$x.f_{x\,(N)} - N.[\partial f_{x\,(N)}/\partial N] = (x+1).f_{x+1\,(N)}$$

and its three corollaries: equations (7), (10) and (12):

$$f_{x+1\,(N)} = [x/(x+1)].f_{x\,(N)} \quad \text{valid when} \quad \partial f_{x\,(N)}/\partial N = 0$$

$$f_{x\,(N)} = -(N/x).\,\Sigma_i\,[\partial f_{i\,(N)}/\partial N]$$

with the sum Σ_i extended from $i = 0$ to $i = (x - 1)$

and, at last:

$$-\Sigma_i\,[(\partial f_{i\,(N)}/\partial N).(X - i)] = 1$$

with X as the larger value of x for which $f_{X\,(N)} \neq 0$

and the sum Σ_i extended from $i = 0$ to $i = (X - 1)$

APPENDICES

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:

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

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 :

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

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:

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

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,

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

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:

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

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

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

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

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

In particular:

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

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

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:

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

with $[\partial^x R_{(N)}/\partial N^x]$ as the xth 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:

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

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 [6]. 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 $[i\text{-}1 \rightarrow i]$ 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 $[i \rightarrow i+1]$ 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, for $i \geq 1$:

$$\partial f_{i(N)}/\partial N = [i.f_{i(N)}/N - (i+1).f_{i+1(N)}/N](1 - f_1/N) \qquad (A2.0)$$

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

$$\partial f_{i(N)}/\partial N \;=\; i.f_{i(N)}/N - (i+1).f_{i+1(N)}/N \qquad (A2.1)$$

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

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

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):

$$\partial^2 R_{(N)}/\partial N^2 \;=\; -\,2f_{2(N)}/N^2 \qquad (A2.3)$$

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:

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

More generally:

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

REFERENCES

1. Chiarucci A, Bacaro G, Rocchini D & Fattorini L Discovering and rediscovering the sample based rarefaction formula in the ecological literature. *Community Ecology*. 2008 ; **9**(1): 121-123.

2. 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. 2013, Volume 5, pp. 195-211. Waltham, MA: Academic Press.

3. 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

4. Lee CM, Kim SS & Kwon TS Butterfly fauna in Mount Gariwang-san, Korea. Journal of Asia-Pacific Biodiversity. 2016 ; 9: 198-204.

5. Li XS, Zhang YL, Fang JH, Schweiger O & Settele J - A butterfly ‘hotspot’ in Western China, its environmental threats and conservation. Journal of Insect Conservation. 2010 ; doi 10.1007/s10841-010-9361-8

6. Lee SM & Chao A Estimating population size via sample coverage for closed capture-recapture models. Biometrics. 1994 ; 50(1): 88-97.