Least-biased extrapolation of a partial Inventory of butterfly fauna in Manas Range (Royal Manas National Park, Bhutan).

Jean Béguinot, Tshering Nidup

Asian Journal of Environment & Ecology
2(2): 1-14, 2017; Article no.AJEE.32701

SCIENCEDOMAIN *international*
www.sciencedomain.org

Least-biased Extrapolation of a Partial Inventory of Butterfly Fauna in Manas Range (Royal Manas National Park, Bhutan)

Jean Béguinot[1*] and Tshering Nidup[2]

[1]Department of Biogéosciences, Université de Bourgogne, F-21000 Dijon, France.
[2]Royal Manas National Park, Ministry of Agriculture and Forests, Bhutan.

Authors' contributions

This work was carried out in collaboration between both authors. Author TN collected and published field data. Author JB conducted the extrapolation procedure applied to crude field data, discuss the results and wrote the manuscript. Both authors read and approved the manuscript.

Article Information

DOI: 10.9734/AJEE/2017/32701
Editor(s):
(1) George Tsiamis, Assistant Professor of Environmental Microbiology, Department of Environmental and Natural Resources Management, University of Patras, Agrinio, Greece.
Reviewers:
(1) Hamit Ayberk, Istanbul University, Turkey.
(2) Manoel Fernando Demétrio, Universidade Federal da Grande Dourados, Dourados, Mato Grosso do Sul, Brazil.
(3) Richa Pandey, Gujarat Forest Department, Gujarat, India.
(4) G. O. Yager, University of Agriculture, Benue State, Makurdi, Nigeria.
Complete Peer review History: http://www.sciencedomain.org/review-history/18541

Original Research Article

Received 10th March 2017
Accepted 31st March 2017
Published 6th April 2017

ABSTRACT

As a rule, most biodiversity inventories at local scales remain more or less incomplete, when dealing with relatively speciose taxonomic groups, such as butterflies in tropical regions. It remains yet possible to take maximum additional advantage of partial inventories and to develop reliable predictions by extrapolating, as accurately as possible, the species accumulation curve beyond the already achieved sampling-size. For this purpose, selecting for the less-biased estimator of the number of missing species (among the wide diversity of currently available solutions) and for the corresponding expression of the species accumulation curve is desirable. Accordingly, implementing the recently derived "least-biased extrapolation procedure" is recommended in this respect.

Least-biased extrapolation procedure was applied to an incomplete inventory of butterfly fauna (91 observed species) carried on by Tshering Nidup and coworkers in the Manas Range (Royal Manas National Park, Bhutan). The estimated total species richness of butterflies for the set of investigated

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

ecosystems reaches around 120 species; accordingly the achieved-sampling completeness is estimated around 76%. Alternative estimations, based on six empirical models of species accumulation curves (namely: Clench, Negative Exponential, Exponential, Logarithmic B, Power and Margalef) prove markedly less accurate than the selected least-biased extrapolation, with Clench model being the less worst, however.

Keywords: Lepidoptera; diversity; species richness; incomplete sampling; estimator; species accumulation; accuracy.

1. INTRODUCTION

Incomplete inventories of biodiversity are likely doomed to become increasingly frequent, as surveys progressively address new taxonomic groups more difficult to cope with, in particular those groups giving rise to species assemblages with high number of species. In addition, more commonly investigated taxonomic groups, also, are likely doomed to remain more or less incompletely surveyed at the *local scale*, due to sampling efforts often being far less intensive at these small scales than they usually are across wider areas. Accordingly, most of ongoing published inventories are admittedly *more or less incomplete* [1]. This incompleteness may be partially compensated (yet, in numerical terms only) by the estimation of the number of "missed" (i.e. unrecorded) species, thereby leading to the evaluation of the total species richness of the sampled assemblage of species. Many different (nonparametric) estimators of the number of "missing" species have been proposed in recent decades (reviewed in [2,3]). As expected, these different types of estimators provide divergent evaluations of the number of unrecorded species, without any consensus having ever been reached regarding which estimator would feature more accurate than the others [1]. And the commonly accepted suggestion to consider all these divergent estimates without being able to choose between them [4] remains rather frustrating. This, in turn, probably contributes to explain why many partial inventories are still not extrapolated numerically, in order to derive a reliable estimation of the total species richness. Yet, reliable evaluations of the richness of species assemblages would be highly desirable, at least in relative, if not in absolute terms. Note that even in relative terms, a relevant comparison of species richness between two or several assemblages requires that inventories be actually compared at a *same level of completeness*. A mandatory condition, that neither standardised sampling nor rarefaction to a same sampling size may actually secure [5], contrary to what is still too often asserted in literature (and this, simply because the level of completeness is dependent not only upon sample size but also is tightly dependent on the degree of heterogeneity of the species abundances distribution, which may usually differs between sampled assemblages).

Now, a rational method of selection in favour of the least-biased estimator, among the most commonly referenced ones, has recently been developed [6,7], enlarging the path initiated by Brose et al. [8]. This newly derived procedure avoids the above mentioned frustration of having to deal with divergent estimates without knowing how to choose the most accurate of them all.

Hereafter, advantage is taken from using this procedure to extrapolate an incomplete inventory of Butterfly fauna in the Manas Range (Royal Manas National Park, Bhutan), carried on by Tshering Nidup and coworkers [9]. Thereby, a reliable estimate of the "true" total species richness of butterfly fauna within the partially sampled ecosystems of the Royal Manas National Park is expected. Moreover, reliable predictions of the additional sampling effort required to improve the completeness of the already performed inventory are derived. This, in turn, provides a rational basis to decide whether or not it seems worth further continuing the sampling operations, putting in balance the additional effort required and the expected benefit in terms of newly recorded species.

2. MATERIALS AND METHODS

All details relative to the environmental context of the partial inventory and the list of butterfly species *with their respective abundances* are provided on-line with open access [9] and, accordingly, these details will not be recalled here. Accounting for *species abundances* is of prime interest in the perspective of the extrapolation of partial samplings, since abundance data provides estimates of the numbers f_1, f_2, f_3, f_4,…, f_x, … of those species recorded respectively 1-, 2-, 3-, …, x- times in

the realised partial sampling. These numbers are required, in turn, to reliably extrapolate the species accumulation curve, as explained below.

2.1 Numerical Extrapolation of Species Accumulation beyond the Achieved Sampling Size

As sampling size increases, the number of recorded species is monotonically growing, at first rapidly and then less and less quickly. The so-called 'Species Accumulation Curve' R(N) accounts for the growth kinetics of the number of recorded species R with increasing sampling size N (N: typically, the number of observed individuals during sampling). The mathematical expression (and thus the details of the shape) of the Species Accumulation Curve are dependent upon both the total species richness of the sampled assemblage of species and the degree of heterogeneity of the species abundance distribution within the sampled assemblage of species [1]. This would apparently make the extrapolation of the Species Accumulation Curve rather difficult to compute, since both preceding factors are unknown *a priori*. Yet, the numbers f_1, f_2, f_3, f_4,…, f_x, … of those species recorded respectively 1-, 2-, 3-, 4-, …, x- times during sampling are directly dependent also upon the total species richness *and* the degree of heterogeneity of the species abundances. This explains why these numbers f_1, f_2, f_3, f_4,…, may serve as an appropriate basis from which to extrapolate the Species Accumulation Curve, beyond the actual size of the sample under consideration. In particular, the most commonly used estimators of the number of unrecorded species (i.e. *non-parametric* estimators such as 'Chao' and the series of 'Jackknife') are computed from the recorded values of the first numbers f_x [2]. In practice, a problem remains however: as already mentioned, each of these different types of estimators provides a substantially distinct estimate and none among these estimators remains consistently the more appropriate. Accordingly the traditional practice has become to consider together all of them without making any choice [4], an admittedly frustrating situation!

Yet, it has been shown recently that although none of the available estimators consistently remains the more accurate [8], each of them may prove, in turn, being the less biased, depending on the value taken by f_1 as compared to the other $f_{x>1}$ [6]. Accordingly, in practice, the most appropriate – i.e. *the least biased* – estimator of the number of unrecorded species may be selected by comparing the value of f_1 to the values of the other f_x for $x > 1$ [6,7]. Selecting this way the least-biased type of estimator thereby provides the best possible estimate of the number Δ of "missing" species and, in turn, the best estimate of the total species richness S_t of the partially sampled assemblage. In addition, the less biased expression for the *extrapolation* of the species accumulation curve R(N) is straightforwardly derived.

In practice, the formulations summarised in Appendix 1 provide (i) the expressions of Δ, S_t and R(N), according to each of the most commonly used types of nonparametric estimators and (ii) the *key to select* among them the less biased estimator and, thereby, the less-biased expressions for Δ, S_t and R(N). Also, in order to reduce the influence of drawing stochasticity, which affects the *as-recorded* values of the f_x, it is advisable to regress the as-recorded distribution of the numbers f_x versus x.

3. RESULTS

The survey conducted by Nidup and coworkers yields $R_0 = 91$ recorded species from $N_0 = 1319$ observations. The recorded values of the numbers f_x at the end of sampling are plotted in Fig. 1 (grey points) together with their values after regression (black points) which are considered for the extrapolation of the species accumulation curve.

The extrapolations respectively associated to six types of *non-parametric* estimators – Chao and the five first Jackknife's at orders 1 to 5 – are plotted at Fig. 2. As the (regressed) values of the f_x satisfy the inequality $f_1 > 4f_2 - 6f_3 + 4f_4 - f_5$, it follows that, here, the *more accurate* extrapolation of the species accumulation curve is that associated to Jackknife-5 (*cf.* Appendix 1). Fig. 2 and Table 1 highlight the strong differences between the different extrapolations, in particular the strong difference between the selected extrapolation, associated to JK-5, and the extrapolations associated to JK-2, JK-1 and Chao (even though the latter are among the most widely used estimators however!).

The practical importance of selecting the more accurate extrapolation is obvious: for example, the estimated number of missing species differs by a factor ≈ 2 and the required sampling size to reach 90% (resp. 95%) completeness differs by a factor ≈ 3 (resp. ≈ 4) when comparing the extrapolations respectively associated to Jackknife-5 and Chao (Table 1).

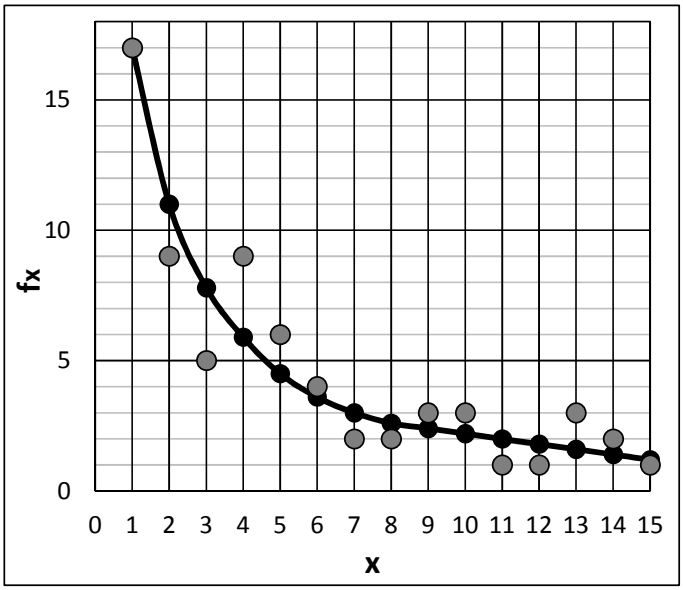

Fig. 1. The recorded values of the numbers f_x of species recorded x-times (grey discs) and the regressed values of f_x (black discs) intended to reduce the consequences of stochastic dispersion

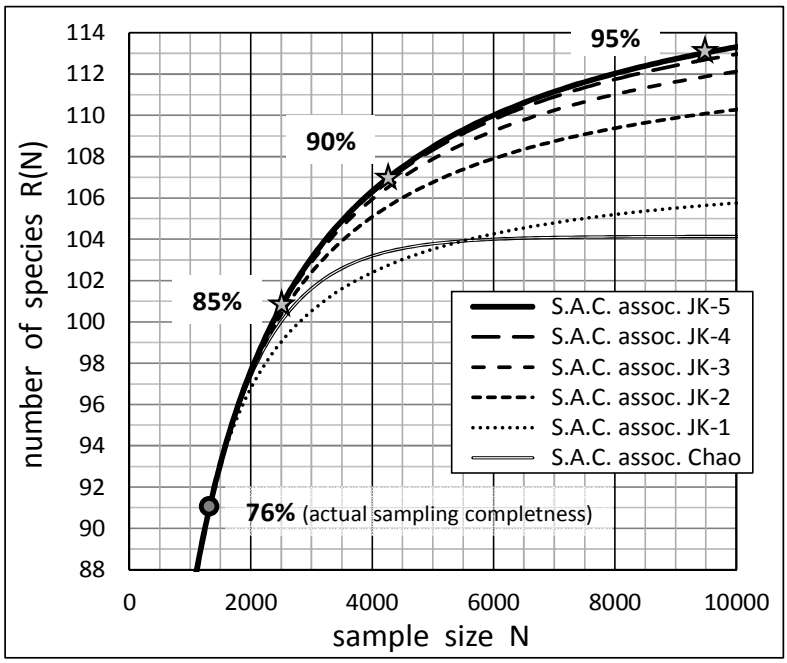

Fig. 2. Extrapolations of the Species Accumulation Curve (S.A.C.) respectively associated to each of the six estimators: Chao and the five Jackknife's at orders 1 to 5. The grey point is for the actually performed sampling: N_0 = 1319 individuals, $R(N_0)$ = 91 recorded species. Extrapolations differ markedly according to the corresponding type of estimator. Selecting the least-biased extrapolation is therefore very important, not only for a reliable estimation of the number of missing species and the total species richness, but also for a reliable prediction of the extra-sampling effort required to reach a given level of completeness. Here, the least-biased extrapolation is associated to Jackknife-5 (JK-5)

Table 1. The estimated number of missing species Δ, the resulting total species richness S_t and the sampling completeness R/S_t, according to the type of estimator involved. Here, the selected estimator, providing the less-biased extrapolation, is Jackknife at order 5 (JK-5). Also computed are the predicted sampling sizes required to reach 90% and 95% completeness, according to each kind of extrapolation

Estimator types	JK-5	JK-4	JK-3	JK-2	JK-1	Chao
nb. missing species Δ	28	27	26	23	17	13
Total species richness S_t	119	118	117	114	108	104
Completeness R/S_t	76%	77%	78%	80%	84%	87%
Sample size N for 90% compl.	≈ 4250	≈ 4000	≈ 3700	≈ 3100	≈ 2100	≈ 1550
Sample size N for 95% compl.	≈ 9500	≈ 8500	≈ 8000	≈ 6000	≈ 4100	≈ 2200

According to the selected least-biased extrapolation of the species accumulation curve, here associated to Jackknife-5, the number of missing species is estimated at 28, the total species richness at 91 + 28 = 119 species and, accordingly, the completeness reached by the inventory is estimated close to three quarters.

Although this level of sampling completeness is fair, a more thorough investigation still features desirable, since a quarter of the total number of species still remains to be recorded, among which a majority of them are expected to be comparatively rare species, thereby of particular potential interest, scientific and patrimonial. As sampling "performance" – in terms of the ratio between the number of newly discovered species and the corresponding additional effort required – consistently decreases severely, as the inventory goes on, the additional investment is expected to be heavy. This is why a reliable estimate of the additional sampling investment needed to reach a given improvement of completeness would be so useful, in term of prospective programming.

An accurate extrapolation of the species accumulation curve opportunely answers this need: Fig. 2 shows the expected additional effort required to increase the completeness, from the present 76% level up to any higher values.

Besides, it is also possible to derive the extrapolation of the numbers f_x of those species that would be recorded x- times after any additional sampling effort, by applying equation [A.1] to the selected extrapolation of the species accumulation curve (that is, here, $R_5(N)$). Accordingly, f_x is given here by:

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

with the expression of $R_{5\,(N)}$ given in Appendix 1.

The expected variations of f_1, f_2, f_3, f_4, f_5, for sampling sizes beyond the realised inventory, are plotted in Figs. 3 and 4.

Numbers f_1, f_2, f_3 have already pass their respective maxima and accordingly are consistently decreasing along continuously progressing sampling, while f_4, f_5 respectively reach their maximum values at sampling size N ≈ 1500 and ≈ 1700, respectively (Fig. 4) and then continuously decrease. As expected, the rate of decrease of the f_x, slows down consistently from f_1 to f_5.

A more thorough theoretical analysis of the regulation process that applies to the series of the f_x is given in [10].

4. DISCUSSION

To extrapolate the species accumulation curve and estimate the number of missing species, I have considered the series of the more commonly implemented types of nonparametric estimators (Chao and the five first Jackknife's). All of them are based on the values taken by the series of the number f_x of those species recorded x-times at the end of sampling. But each type of estimator is, yet, formulated differently and thus provides an estimation which is distinct from the others. Accordingly, a *procedure of selection among them all is necessary* to resolve this hardly acceptable ambiguity. Applying the procedure of selection recently developed for this purpose [6] makes possible to remove this ambiguity and, *here*, leads to retain:

(i) Jackknife-5 as the *least-biased* estimator of the number of missed (still unrecorded) species and
(ii) The expression associated to Jackknife-5 (see Appendix 1) for the *least-biased* extrapolation of the species accumulation curve.

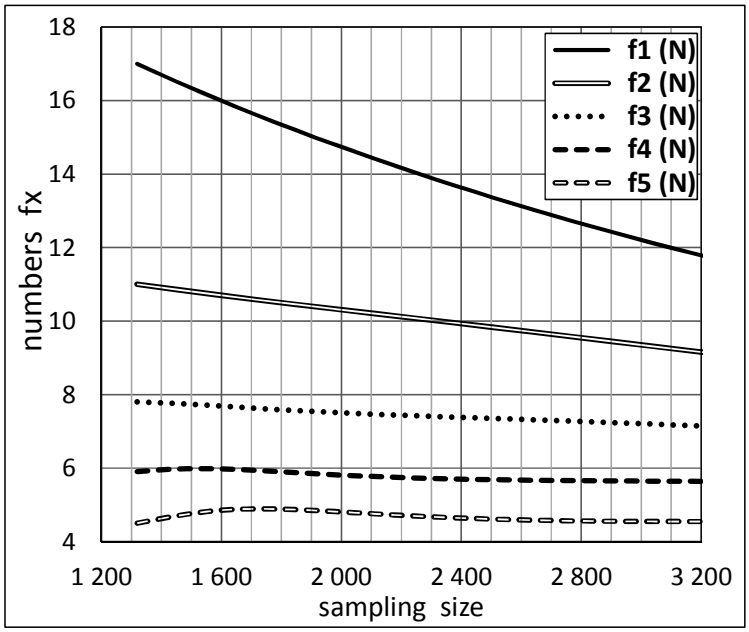

Fig. 3. Extrapolations of the numbers f_1, f_2, f_3, f_4, f_5 of those species recorded respectively 1-, 2-, 3-, 4-, 5- times, plotted against increasing additional sampling size

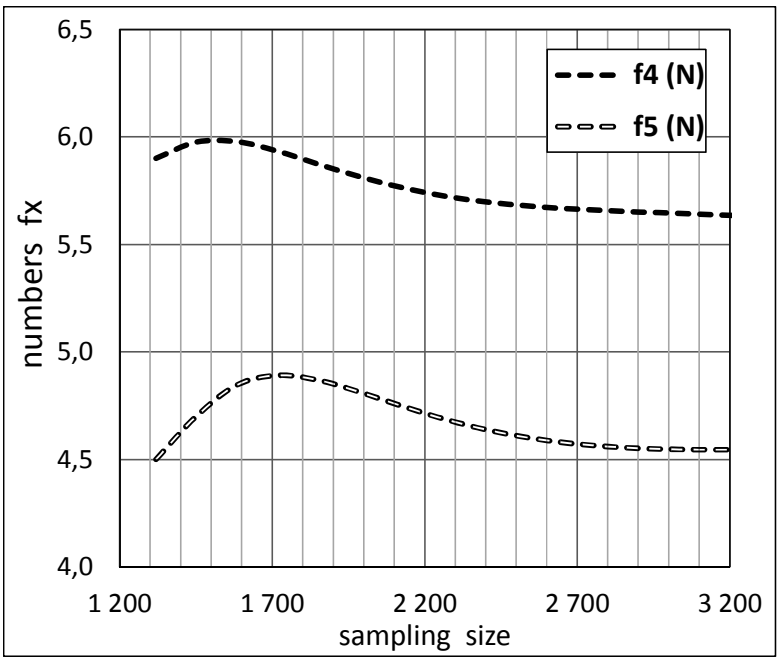

Fig. 4. A zoom on the extrapolations of the numbers f_4 and f_5 of those species recorded respectively 4- and 5- times, plotted against increasing additional sampling size

Incidentally, the selected estimator proves, here, to be the one having the highest value (Fig. 2). This, indeed, is not surprising since all the non-parametric estimators available in the literature (including the six types considered in the implemented procedure) are considered as yielding *under*-estimates of the true number of missing species [1,2]. Accordingly, it is logically expected that the less-biased, among them, should be the one leading to the highest estimate.

In fact, this trend is quite general indeed, as demonstrated directly from the inequalities defining the respective ranges of appropriate use of each of the Jackknife estimators (see Appendix 1 for more details).

Apart from the range of *non-parametric* estimators considered above; a series of purely *empirical* formulations of the species accumulation curve R(N) might also be considered alternatively. These empirical formulations are not associated to any kind of estimator of the number of missing species, but have adjustable parameters that enable them to satisfy the two following compulsory conditions:

(i) The computed number, $R(N_0)$, of recorded species at the end of the actually performed sampling (size N_0) should be equal to the number of actually recorded species R_0 (here $R_0 = 91$);
(ii) At $N = N_0$, the computed slope of the species accumulation curve, $\partial R(N)/\partial N$, must be equal to f_1/N_0, a condition originally demonstrated by TURING [11] (and subsequently generalised independently by Béguinot [10,12]: see equation (A.1) in Appendix 1).

Among empirical formulations involving two adjustable parameters, the six following are commonly considered for the extrapolation of the species accumulation curve (reviewed in [13]):

"Clench" model: $R(N) = a.N/(1+b.N)$

"Negative exponential" model: $R(N) = a.(1 – \exp(– b.N))$

"Exponential" model: $R(N) = a + b.\ln(N)$

"Logarithmic B" model: $R(N) = a.\ln(1+ b.N)$

"Power" model: $R(N) = a.(N)^b$

with a and b as the two adjustable parameters.

Also, a model with only one adjustable parameter may be easily derived from the *Margalef index*, as: $R(N) = a.\mathrm{Ln}(N) + 1$ (the derivation is based on the postulated independence of Margalef index upon sampling size N, which is implicit in the conception of this index, although this is practically never the case in practice).

As already mentioned, the adjustable parameters a and b are defined, for each model, in order to satisfy both relationships $R(N_0) = R_0$ and $\partial R(N)/\partial N = f_1/N_0$ at $N = N_0$ (see Appendix 2 for the computations of the values taken by parameters a and b in each case).

Figs. 5 and 6 and Table 2 provide representations of the extrapolated species accumulation curves at sampling sizes $N > N_0$, for each of the six empirical models and for the least-biased extrapolation associated to Jackknife-5 estimator.

All six empirical models lead to extrapolations that differ more or less markedly from the least-biased extrapolation associated to Jackknife-5.

At first, *Exponential* model, *Logarithmic B* model, *Power* model and *Margalef-index associated* model all are non-asymptotic models, thereby being inappropriate to estimate the number of missing species and the resulting total species richness. *Clench* model and *Negative exponential* model, on the contrary, are asymptotic expressions which may deliver, accordingly, finite estimations of the number of missing species and of the resulting total species richness (Table 2).

As compared to the least-biased estimate of 28 missing species, the estimates provided by *Clench* model and *Negative exponential* model are substantially lower: 21 and 6 missing species respectively. Therefore, here, *Clench* model works better than *Negative exponential* model (*Exponential*, *Logarithmic B*, *Power* and *Margalef* models being out of competition as mentioned above).

Table 2. The estimated number of missing species Δ and the expected number of newly recorded species after doubling the sampling size. As a whole, it is the extrapolation according to Clench model which deviates the least from the selected less-biased extrapolation (associated to Jackknife 5)

Extrapolation types	JK-5	Clench	NEG. Exp.	EXP/LOG	Power	Margal.
nb. missing species Δ	28	21	6	Undefined	Undefined	Undefined
nb. newly rec. species	10.5	9.4	5.8	11.8	12.6	8.7

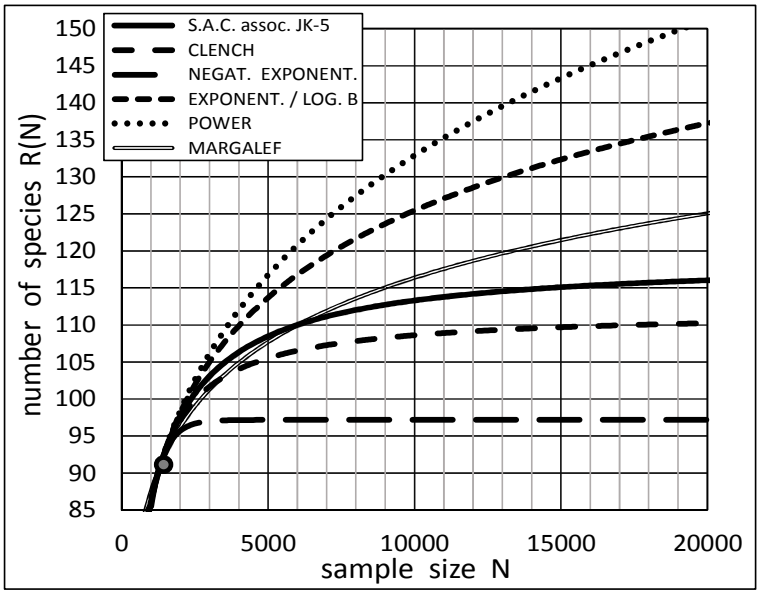

Fig. 5. Extrapolations of the Species Accumulation Curve (S.A.C.) associated (i) to the selected estimator Jackknife 5 and (ii) to six *empirical* models: "Clench", "Negative exponential", "Exponential" (or "Logarithmic-B" which is practically superimposed), "Power" and, at last, the model derived from the Margalef index

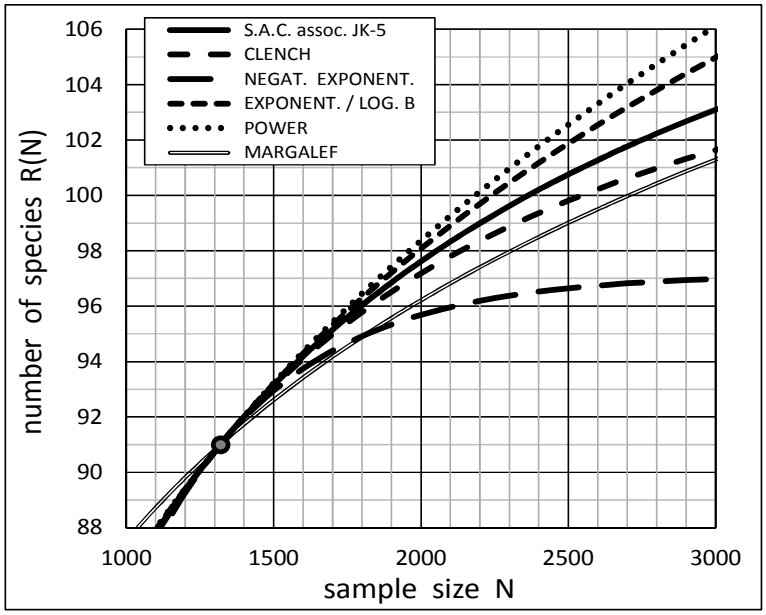

Fig. 6. Extrapolations of the Species Accumulation Curve (S.A.C.) associated (i) to the selected estimator Jackknife 5 and (ii) to six empirical models: a zoom of Fig. 5 focusing on the details of empirical models deviations from the selected extrapolation

As regards, now, the comparison between the extrapolations according to *Clench* model, on the one hand and the series of Jackknife estimators on the other hand, Fig. 7 shows that, here, the *Clench* model delivers a better prediction than Jackknife-1, but does less good than the species accumulation curves respectively associated to all the other Jackknife's : JK-2, JK-3, JK-4 and, of course, JK-5.

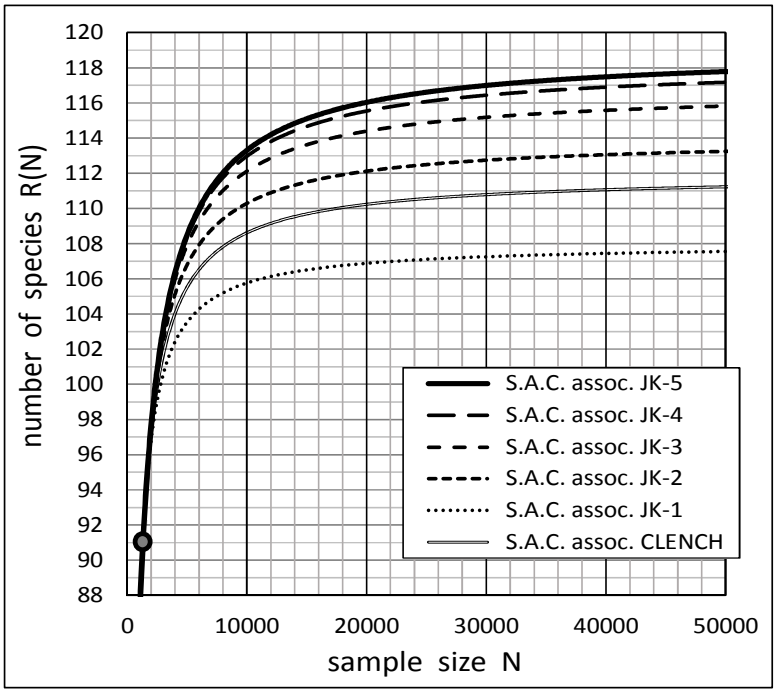

Fig. 7. Extrapolations of the Species Accumulation Curve (S.A.C.) respectively associated to the Clench model and to the series of Jackknife estimators (JK-1 to JK-5). Clench model works better than does the extrapolation associated to Jackknife-1 but less well than do the extrapolations associated to the other Jackknife estimators

5. CONCLUSION

Incomplete inventories of local biodiversity, which are doomed to become most often the ordinary rule in practice (at least for speciose taxonomic groups and/or for local investigations involving insufficient sampling efforts) may provide, however, *much more information* than would be expected from the crude consideration of the crudely recorded data. Releasing this additional information requires, however, that species inventories include not only the simple list of occurring species but also the respective abundances of each recorded species. Under this condition, extrapolating the Species Accumulation Curve, beyond the actually achieved inventory, may easily be implemented, using either *non-parametric estimators* of the number of missed species or considering alternatively, several kinds of *empirical models*. Literature provides numerous types of non-parametric estimators as well as several kind of empirical models of species accumulation function. Reliable extrapolation, however, is conditioned by the rational selection, for each inventory, of the *least-biased* estimator of the number of missing species, among the series of estimators made available in the literature. Empirical models, for their own, prove hardly appropriate, especially those models having non-asymptotic expressions. Among the asymptotic empirical models, *Clench* model performs more or less as the average of the non-selected non-parametric estimators (see Fig. 7) while the *Negative-exponential* model is very strongly negatively biased.

According to the least-biased extrapolation of the species accumulation curve (involving, for this particular inventory, the Jackknife 5 nonparametric estimator), 28 additional species would still remain unrecorded by the present inventory. The 91 recorded species thus represent about three quarters of the true species richness (≈ 119 species) of the set of investigated ecosystems within the Manas Range by Tshering Nidup and co-workers.

This, indeed, invites to add some supplementary sampling effort, at first applied to the same set of ecosystems already inventoried partially. In this perspective, the least-biased extrapolation of the species accumulation curve provide useful information that may serve to predict the level of

additional sampling effort (in term of sampling size, i.e. number of individual records) that would be necessary to reach a given increment of sampling completeness. As might be expected, the additional sampling effort needed to progress in completeness increases very rapidly, that is, the cost of recording new species becomes progressively but rapidly higher and higher. Beyond this intuitive expectation, it is the merit of a reliable extrapolation, as plotted in Fig. 2, to quantify the rapidly increasing cost required by a continuous improvement of completeness of inventory. For example, increasing the completeness from the actual 76% level to 90% completeness would require multiplying by a factor ≈ 3 the currently achieved sampling size (≈ 4250 individuals to be recorded, as compared to the 1319 presently recorded). And reaching a desirable 95% level of completeness would imply increasing the present sampling size by a factor ≈ 7 (≈ 9500 individuals to be recorded against 1319).

Finally, this opens the desirable possibility of comparing, *on a rational common basis*,

(i) The expected number of newly recorded species, many if not all of them being of potential scientific and patrimonial interest (as they are admittedly expected to be among the rarest species of the sampled assemblage) and,
(ii) The additional sampling efforts/costs that would be required to obtain this expected number of new records.

ACKNOWLEDGEMENTS

Thanks to Tshering Dorji and Tschering Ugyen, who both collaborated, along with Tshering Nidup, to the achievement of the butterfly survey at Royal Manas National Park.

Thanks also to four reviewers for their appreciated comments on the manuscript.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Rajakaruna H, Drake DAR, Chan FT, Bailey SA. Optimizing performance of nonparametric species richness estimators under constrained sampling. Ecology and Evolution. 2016;6:7311-7322.
2. Gotelli NJ, Chao A. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: Levin S.A. (ed.) Encyclopedia of Biodiversity, Second Edition, Waltham, MA: Academic Press. 2013;5:195-211.
3. 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; <hal-01238720>
4. Colwell RK, Coddington JA. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society London B. 1994;345: 101-118.
5. Lande R, DeVries PJ, Walla TR. When species accumulation curves intersect: Implications for ranking diversity using small samples. Oikos. 2000;89(3):601-605.
6. 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>
7. Béguinot J. Extrapolation of the species accumulation curve associated to “Chao” estimator of the number of unrecorded species: A mathematically consistent derivation. Annual Research & Review in Biology. 2016;11(4):1-19.
DOI: 10.9734/ARRB/2016/30522; <hal-01477263 >
8. Brose U, Martinez ND, Williams RJ. Estimating species richness: Sensitivity to sample coverage and insensitivity to spatial patterns. Ecology. 2003;84(9): 2364-2377.
9. Nidup T, Dorji T, Tshering U. Taxon diversity of butterflies in different habitat types in Royal Manas National Park. Journal of Entomology and Zoology Studies. 2014;2(6):292-298.
10. Béguinot J. On general mathematical constraints applying to the kinetics of species discovery during progressive sampling: Consequences on the

theoretical expression of the species accumulation curve. Advances in Research. 2016;8(5):1-17.
DOI: 10.9734/AIR/2016/31791

11. Good IJ. The population frequencies of species and the estimation of population parameters. Biometrika. 1953;40:237–264.
12. Béguinot J. An algebraic derivation of Chao's estimator of the number of species in a community highlights the condition allowing Chao to deliver centered estimates. ISRN Ecology; 2014. Article ID 847328:1-6.
DOI: 10.1155/2014/847328; <hal-01101415>
13. 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.
14. Béguinot J. When reasonably stop sampling? How to estimate the gain in newly recorded species according to the degree of supplementary sampling effort. Annual Research & Review in Biology. 2015;7(5):300-308.
DOI: 10.9734/ARRB/2015/18809; <hal-01228695>

APPENDIX

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

Consider the survey of an assemblage of species of size N_0 (with sampling effort N_0 typically identified either to the number of recorded individuals or to the number of sampled sites, according to the inventory being in terms of either species abundances or species incidences), including $R(N_0)$ species among which f_1, f_2, f_3, f_4, f_5, of them are recorded 1, 2, 3, 4, 5 times respectively. The following procedure, designed to select the less-biased solution, results from a general mathematical relationship that constrains the theoretical expression of *any* theoretical Species Accumulation Curves R(N) (see [6,12,14]):

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

Compliance with the mathematical constraint (equation (A.1)) warrants *reduced-bias* expression for the extrapolation of the Species Accumulation Curves R(N) (i.e. for $N > N_0$). Below are provided, accordingly, the polynomial solutions $R_{\mathbf{x}}(N)$ that respectively satisfy the mathematical constraint [1], considering increasing orders x of derivation $\partial^x R_{(N)}/\partial N^x$. Each solution $R_{\mathbf{x}}(N)$ is appropriate for a given range of values of f_1 compared to the other numbers f_x (according to [6]):

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

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

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

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

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

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

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

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

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

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

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

N.B. 1: As indicated above (and demonstrated in details in [6]), this series of inequalities define the ranges that are best appropriate, respectively, to the use of each of the five estimators, JK-1 to JK-5. That is the respective ranges within which each estimator will benefit of minimal bias for the predicted number of missing species.

Besides, it is easy to verify that another consequence of these preferred ranges is that the selected estimator will *always* provide the highest estimate, as compared to the other estimators. Interestingly, this mathematical consequence, of general relevance, is in line with the already admitted opinion that all non-parametric estimators provide *under*-estimates of the true number of missing species [1, 2]. Also, this shows that the approach initially proposed by BROSE *et al.* [8] – which has regrettably suffered from its somewhat difficult implementation in practice – might be advantageously reconsidered, now, in light of the very simple selection key above, of *far much easier practical use*.

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

N.B. 3: For f_1 falling beneath 0.6 x f_2 (that is when sampling completeness closely approaches exhaustivity), then Chao estimator may be selected: see reference [7].

Appendix 2. Computations of the adjustable parameters for each empirical model of species accumulation curve

The adjustable parameters a and b are defined, for each model, in order to satisfy both compulsory relationships: (i) $R(N_0) = R_0$ and (ii) $\partial R(N)/\partial N = f_1/N_0$, at $N = N_0$.

It follows, for the different kinds of empirical models:

***** *Clench* model: $R(N) = a.N/(1+b.N)$:

Relationship (i) imposes $a = R_0.(1+ b.N_0)/N_0$, with parameter b defined as $b = (R_0/f_1 – 1)/N_0$, so as to satisfy relationship (ii). Thus, finally, $a = R_0^2/N_0/f_1$.

***** *Negative exponential* model: $R(N) = a.(1 – \exp(– b.N))$:

Relationship (i) imposes $a = R_0/(1 – \exp(– b.N_0))$, with parameter b defined implicitly by $b = f_1/(a.N_0 .\exp(– b.N_0))$, so as to satisfy relationship (ii).

***** *Exponential* model: $R(N) = a + b.\mathrm{Ln}(N)$:

Relationship (i) imposes $a = R_0 – b.\mathrm{Ln}(N_0)$, with parameter b defined as $b = f_1$, so as to satisfy relationship (ii). Thus, finally, $a = R_0 – f_1.\mathrm{Ln}(N_0)$.

***** *Logarithmic B* model: $R(N) = a.\ln(1 + b.N)$:

In fact, this model, while formally quite different from the Exponential model, yet proves being practically identic to it, as soon as N becomes large enough, which will be usually the case in practice for the extrapolated part of the curve.

***** *Power* model: $R(N) = a.(N)^b$

Relationship (i) imposes $a = R_0/(N_0)^b$, with parameter b defined as $b = f_1/R_0$, so as to satisfy relationship (ii). Thus, finally, $a = R_0/(N_0)^{(f1/Ro)}$.

***** *Margalef-index associated* model: $R(N) = a.\ln(N) + 1$

Relationship (i) imposes $a = (R_0 – 1)/\ln(N_0)$ and, as this model has only one adjustable parameter, it is not possible to satisfy relationship (ii) in general; thus this model does not respect the slope imposed to the species accumulation curve at $N = N_0$ (this, indeed, is well apparent at Figs. 5 and 6).

Peer-review history:
The peer review history for this paper can be accessed here:
http://sciencedomain.org/review-history/18541