Inferring total species richness and the exhaustive hierarchical structuring of species abundances in tropical Sea-Stars communities (Asteroidea), using numerical extrapolation of partial inventories

Abstract
Even when remaining substantially incomplete, the partial inventory of a species assemblage can provide much more information than could be expected first. Indeed, retrieving this information is made possible by applying a rigorous procedure of numerical extrapolation to the partial inventory. This numerical extrapolation will deliver reliable estimates of the number of the still unrecorded species and, furthermore, of the distribution of abundances of these unrecorded species. As a result, the full range of the Species Abundance Distribution is finally made available, despite dealing with data from partial sampling only. In turn, this allows to address a series of descriptive and functional aspects of the internal organization of species assemblages, which otherwise would have required disposing of exhaustive samplings. The latter being, however, often impossible to achieve in practice. Fortunately, mathematical and algorithmic basis for a reliable numerical extrapolation of incomplete samplings have been developed recently, so that partial inventories no longer remain an obstacle to gain access to the true (total) species richness and the full-range pattern of hierarchical structuring of species abundances. This approach is applied here to the previously reported partial samplings of two communities of tropical sea-stars associated to coral-reefs in the Central South China Sea. Among the main new results derived from the numerical extrapolations of these partial samplings, the following are highlighted:
- the extrapolated true (total) species richness of each of the two studied communities largely exceed the recorded figures, thereby confirming the limited completeness of these partial samplings: 53% and 67% completeness only;
- once properly completed by numerical extrapolation, the Species Abundance Distributions of both communities prove best fitting the “log-normal” than the “log-series” model, thereby suggesting that many independent factors (rather than only one dominant factor) contribute together to the hierarchical structuring of species abundances within these sea-stars communities;
- the intensity of the process of hierarchical structuration of species abundances proves being close to what would be obtained for the corresponding “broken-stick” model, which means a rather moderate level of structuration intensity, as compared to the range of values typically obtained for other kinds of marine invertebrates.

Key-words: Echinodermata, starfish, coral reef, species richness, diversity, rank abundance distribution, evenness, incomplete inventory, Malaysia
1. Introduction

Total species richness, taxonomic composition and the hierarchical structuring of species abundances are three key features that appropriately characterize species communities in the wild. Addressing properly these key features obviously requires disposing of exhaustive inventories. Yet, inventories remaining substantially incomplete are common in practice and even doomed to become still more frequent with the inevitable generalization of “rapid assessments” and “quick surveys”. This is especially true when having to deal with species-rich communities of invertebrates which often include a lot of rare, hard-to-detect species. Such incomplete inventories prevent deriving reliable inferences and, thus, may often lead to erroneous interpretations regarding the key aspects of species communities evoked above [1–3].

Fortunately, a reliable procedure of numerical extrapolation of partial inventories has been developed recently, which can overcome these difficulties and is able to provide least-biased estimations of:

(i) the number of those species remained undetected and, still further,
(ii) the respective abundances of each of these undetected species.

Thereby, reliable inferences can finally be derived regarding (i) the true (total) species richness and (ii) the completed distribution of species abundances, i.e. including the set of the still unrecorded species. Only the taxonomic identities of these undetected species escape, of course, to any attempt to highlight them. In turn, once numerically completed (and only when it is so: [4]), the Species Abundance Distribution (“S.A.D.”) can then provides synthetic pieces of information about the process (either deterministic or stochastic) that drive the hierarchical structuring of species abundances within community [5–9]. Accordingly, some light can thus be shed, both qualitatively and quantitatively, on the biological and ecological determinants of the internal structuration among species within community.

Although no further details may be extracted from this synthetic overview, the latter has, yet, the advantage of being straightforward, as it does not require the long and tedious analytical approach that would be required otherwise to go deeper in the details of structuring process. As such, this synthetic overview can serve as a convenient preliminary approach.

Hereafter, I report and discuss the results from the numerical extrapolation of the partial inventories of two sea-stars communities (Echinodermata: Asteroidea) associated to coral reefs surrounding the small neighboring islets of Terumbu Siput (“Erica reef”) and Terumbu Peninjau (“Investigator shoal”), located in the Central South China Sea, off the Malaysian coast. Such marine ecosystems, in tropical shallow waters, are of major interest to ecologists and conservationists, as they are considered as embodying remarkably high levels of biological complexity [10–12].

2. Materials and Methods

2.1 Materials

Coral reefs surrounding small islands dispersed in the Central South China Sea are home to a rich sea-stars fauna, reported as counting overall no less than 230 species [13]. Yet, inventories at the local scale of sea-stars communities in China Sea, additionally listing the abundances of the recorded species, remain very scarce at the local scale. A recent report by Kwan et al. [12], however, opportunely provides such a series of local inventories of sea-stars from the Archipelago of Beting Patinggi Ali to Pulau Layang-Layang in the Malaysian waters of Central
South China Sea. Yet, the high proportion of singletons (species detected only once during sampling) that subsist in these inventories strongly suggests that samplings remain substantially incomplete[14 – 16], thus requiring numerical extrapolation to take full advantages of the as-recorded data and avoid the risk of erroneous inference.

Hereafter, I focus on two of these investigated sea-stars communities, those having the highest number of recorded species (16 and 9 species) respectively. These two communities are located in two small neighboring coral islets: *Terumbu Siput* (“Erica reef”) and *Terumbu Peninjau* (“Investigator shoal”), both being part of “Spratly Islands”, off the coasts of the Philippines, Malaysia, and southern Vietnam. Further information regarding the precise locations, the environment and the practical details of the sampling procedure are provided in open reference [12].

### 2.2 Numerical extrapolation procedure

#### 2.2.1 Total species richness

The least-biased estimation of the number of still undetected species after partial sampling and the resulting least-biased estimation of the true, total species richness of apartially sampled community are derived according to the procedure defined in [17, 18] and briefly summarized in Appendix 1, on the basis of the numbers $f_x$ of species recorded $x$-times during partial sampling: Figures A.1 and A.2).

#### 2.2.2 Completed Species Abundance Distribution:

To accurately exploit its full potential, the as-recorded Species Abundance Distribution (“S.A.D.”) requires [19, 20]:

- first, to be corrected for the statistical sampling bias resulting from the finite size of samplings;
- second, but still more importantly, to be completed by numerical extrapolation to the extent that the sampling is suspected to be incomplete (as revealed by the subsistence of several singletons).

After being corrected and completed by numerical extrapolation, the S.A.D.:

- not only provides an overview of both the true species richness of the sampled community and the diversity of the respective abundances of member-species
- but also, can help addressing important questions regarding (i) the kind of causes that determine the hierarchical structuration of species abundances, (ii) the resulting degree of abundance unevenness and (iii) the genuine intensity of the hierarchical structuring process (which by no means identifies to the degree of unevenness, contrary to a commonly held opinion).

The appropriate procedures of correction and, then, numerical extrapolation of the “S.A.D.” beyond its as-recorded part, are described in details in reference [20] and briefly summarized in Appendix 2. Also, a concrete example of implementation of the procedure is commented in details in reference [21].

Classically, the “S.A.D.” is graphically presented according to the so-called “Ranked Abundance Distribution” (also known as “Whittaker plot”), according to which the (log-transformed) abundances $a_i$ are plotted against the rank $i$ of species, the latter being ordered by decreasing values of abundance (with, thus, $a_1$ and $a_{S_t}$ respectively standing for the highest and the lowest abundances in a community having $S_t$ species).

#### 2.2.3 species abundance structuration- (i) the apparent pattern: abundance unevenness
The “S.A.D.” (either being exhaustive or completed by numerical extrapolation) conveys all the relevant quantitative data required to address the internal organization of member-species within a local community, especially the hierarchical structuration of species abundances. In particular, it is always advisable to use such species-abundance plots to quantify the degree of evenness or, more suggestively, the degree of *unevenness* of species abundances [22]. Indeed, following [23], it is the degree of *unevenness* – rather than evenness itself – that should be preferred to address properly the hierarchical structuring of species abundances in communities. Optionally, the “S.A.D.” may be synthetically reduced to its two major descriptors: the total species richness $S_t$ and the degree $U$ of abundance unevenness.

According to the aforementioned, classical mode of representation of “S.A.D.”, it goes natural to quantify the degree $U$ of abundance unevenness as the average decreasing slope of the log-transformed abundance along the whole range of the abundance distribution, as already proposed by [24], that is:

$$U = \frac{[\log_{10}(a_1) - \log_{10}(a_{S_t})]}{(S_t - 1)} = \frac{[\log_{10}(a_1/a_{S_t})]}{(S_t - 1)}$$  \hspace{1cm} (1)

**2.2.4 species abundance structuration - (ii) origin and intensity of the underlying process**

Beyond the mere description of the pattern of hierarchical structuration, quantified by the degree of unevenness $U$, the complete “S.A.D.” can help addressing several important questions regarding (i) the kind of *mechanism* involved in the process driving the hierarchical structuration of abundances and (ii) the *intensity* of this structuring process.

As regards the kind of *mechanism* involved, it is appropriate to distinguish between two major alternative hypotheses: schematically, the hierarchical structuration of abundances may result either (i) from the major contribution of *one strongly predominant* factor or (ii) from the combined contributions of *many mutually independent factors* acting together. This can be tested by checking the conformity of the “S.A.D.” to either the log-series model or the log-normal model respectively [5, 25 – 28].

Now, as regards the genuine *intensity* of the structuring process, it is first necessary to remind that the degree of unevenness $U$ does not univocally mirror the intensity of the structuring process, since it is also mathematically dependent (negatively) upon the species richness $S_t$ [29, 30, 31]; see also Appendix 3.

One possible solution to cancel this mathematical influence is to compare the slope of the “S.A.D.” with the slope of a theoretical distribution involving a *constant* structuring process, remaining *strictly independent* of the species richness. The “broken-stick” distribution meets precisely this requirement [32]. Accordingly, an index $I_{str}$, attempting to disentangle the part of unevenness which vary *independently of species richness* can be defined by standardizing the degree of unevenness $U$ of the “S.A.D.” to the degree of unevenness $U$ of the corresponding “broken-stick” distribution, computed for the same species richness $S_t$ [33 - 35], that is:

$$I_{str} = \frac{U}{U'} = \frac{[\log_{10}(a_1/a_{S_t})/(S_t-1)]}{[\log_{10}(a_{1}'/a_{S_t}')/(S_t-1)]}$$

that is finally:

$$I_{str} = \frac{\log_{10}(a_1/a_{S_t})}{\log_{10}(a_{1}'/a_{S_t}')}$$  \hspace{1cm} (2)

with $a_1$ and $a_{S_t}$ standing for the highest and the lowest abundances in the studied community and $a_{1}'$ and $a_{S_t}'$ standing for the highest and the lowest abundances in the corresponding “broken-stick” distribution computed for the same species richness $S_t$.

As the *invariable type* of structuring process involved in the “broken-stick” distribution is the process of random allocation of abundances to species [32], the index $I_{str}$ highlights the intensity of the structuring process in the focused community by comparison to the intensity of this stochastic process, taken as a reference.
It is even possible to continue the analysis even further in the same direction. Beyond
comparing the average slopes of the actual “S.A.D.” to the corresponding “broken-stick” model,
it is additionally informative to operate the comparison separately for each of the two extremal
points that support the slopes – i.e. the maximum and the minimum abundances, $a_1$ and $a_{St}$.
Thus, the ratio $A_1 = (a_1/a')$ mirrors the intensity of those kind(s) of factors which control the
abundance $a_1$ of the more frequent species (especially the most frequent one) while the ratio
$A_{St} = (a_{St}/a'_{St})$ mirrors the intensity of those (presumably different) kind(s) of factors which
control the abundance $a_{St}$ of the less frequent species (especially the least frequent one).

Splitted that way, the comparison with the “broken-stick” model (synthetically expressed by
the index $I_{St}$), highlights even better the underlying structuring process and its genuine
intensity, beyond the immediately apparent pattern of abundance unevenness $U$. This is briefly
detailed in the following section.

2.2.5 analyzing the determinants of the species abundance structuration

Focused on quantitative terms, the “S.A.D.” can be synthetized by three main descriptive
parameters, $S_t$, $U$, $a_{St}$ (the fourth descriptive parameter, the higher abundance $a_1$, being entirely
defined by the three preceding parameters through equation (1)). Now, which “determinants”
are actually constraining the values of $S_t$, $a_1$ and $a_{St}$?

Let consider first the case where the community has hypothetically reached its saturation level
in term of species richness $S_t$ i.e. no more species may be added in the community without
causing extinction(s) [36]. This means that both the abundance unevenness $U$ and the lowest
abundance $a_{St}$ have reached their respective minimal threshold values that would necessarily
be crossed over if any additional colonization would succeed in increasing species richness
beyond the saturation level. Accordingly, in case of saturation, the species richness $S_t$ is
determined internally, the parameters $U$ and $a_{St}$ playing the role of determinants.

Now, it is widely admitted that saturation in species within a community is very uncommon
[36, 37, 38], so that non-saturation should be hypothesized first. In non-saturated communities,
species richness $S_t$ is no more determined internally but externally, the determinant being the
limitation in the colonizing flux of those species having characteristics compatible with the
habitat, a limitation which depend on both dispersal abilities and the richness of the regional
stock of species [37, 38, 39]. And, by contrast with the case where community is supposed
being saturated, here, the parameters $U$ and $a_{St}$ are no more involved as the determinants of $S_t$.

On the contrary, in this situation of limited availability of new colonizing species, it is the
resulting limitation of species richness $S_t$ which contributes to constrain the possible ranges of
values of the triplet of parameters $U$, $a_{St}$ and $a_1$, through equation (1). And, finally, it is $I_{St}$, which
is ultimately constrained, according to equation (2), as are the ratios $A_1 = (a_1/a')$ and $A_{St} =
(a_{St}/a'_{St})$.

3. Results

3.1 Estimation of the total species richness of each sea-stars community

Based on the numbers of species recorded x-times (with $x = 1$ to 5) at the end of the partial
samplings (Figures A1 and A2 in Appendix 1), the selected, least-biased estimators of the
number of undetected species are respectively Jackknife order 3 and Jackknife order 5 for the
communities at Terumbu Siput and Terumbu Peninjau (see selective key in Appendix 1). The
resulting least-biased estimation of the total species richness $S_t$ of the sampled communities
and the level of completeness of the partial samplings, $R_{0_t}/S_t$, are provided in Table 1.
Table 1 – Numerical characteristics of thesea-stars communities associated to coral reefs at Terumbu Siput, and Terumbu Peninjau: the sampling-size \( N_0 \), the number of detected species \( R_0 = R(N_0) \), the selected, least-biased estimator, the estimated number \( \Delta \) of undetected species, the resulting evaluation of the total species richness \( S_t = R_0 + \Delta \) and the level of sampling completeness \( R_0/St \).

<table>
<thead>
<tr>
<th>site</th>
<th>( N_0 )</th>
<th>( R_0 )</th>
<th>selected estimator</th>
<th>( \Delta )</th>
<th>( St )</th>
<th>( R_0/St )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terumbu Siput</td>
<td>52</td>
<td>16</td>
<td>Jack-3</td>
<td>8</td>
<td>24</td>
<td>67%</td>
</tr>
<tr>
<td>Terumbu Peninjau</td>
<td>26</td>
<td>9</td>
<td>Jack-5</td>
<td>8</td>
<td>17</td>
<td>53%</td>
</tr>
</tbody>
</table>

Due to the relatively low level of achieved sampling completeness (67% and 53%), further additional sampling could, alternatively, be considered of interest. In this perspective, the least-biased extrapolation of the Species Accumulation Curve can provide useful predictive information regarding the additional sampling efforts that would be required to obtain any desirable increase in sampling completeness. As an example, the expected increase in the number of detected species, \( R(N) \), as a function of growing sampling size \( N \), is given in Figures 1 and 2 for the sampling at Terembu Siput and Figures 3 and 4 for the sampling at Terembu Peninjau.

Figure 1 – Extrapolated part of the Species Accumulation Curve (S.A.C.) for the community at Terembu Siput, accounting for the increase of the number of detected species \( R(N) \) as a function of growing sample size \( N \), beyond the actually achieved sampling \( (N_0 = 52, R(N_0) = 16) \). Here, the selected, least-biased, nonparametric estimator of the number of undetected species is Jackknife-3, leading to a total species richness \( S_t = 16 + 8 = 24 \). The associated, least-biased extrapolation of the S.A.C. \( R(N) \) is plotted as the coarse solid line. Also plotted, for comparison, are the extrapolations of the S.A.C. associated to other, non-selected (as being more biased) estimators: Jackknife-2, Jackknife-1, Chao1. The comparison highlights the practical importance of selecting the least-biased estimator.
Figure 2 – Extrapolated part of the Species Accumulation Curve for the community at Terembu Siput, associated to the selected, least-biased, nonparametric estimator (here Jackknife-3). In practice, the least-biased extrapolation of the Species Accumulation Curve allows to predict the expected additional sampling effort required to reach higher levels of sampling completeness (for example, the sample sizes required to reach 80%, 90% and 95% completeness would bearound N = 100, 180, 350 respectively).

Figure 3 – Extrapolated part of the Species Accumulation Curve for the community at Terembu Peninjau, accounting for the increase of the number of detected species R(N) as a function of growing sample size N, beyond the actually achieved sampling (N_0 = 26, R(N_0) = 9). Here, the selected, least-biased, nonparametric estimator of the number of undetected species is Jackknife-5, leading to a total species richness S_t = 9 + 8 = 17. The associated, least-biased extrapolation of the Species Accumulation Curve R(N) is plotted as the coarse solid line. Also plotted, for comparison, are the extrapolations of the S.A.C. associated to other, here non-selected (more biased) estimators: Jackknife 4, Jackknife 3, Jackknife-2, Jackknife-1, Chao1. The comparison highlights the practical importance of selecting the least-biased estimator.
Figure 4 – Extrapolated part of the Species Accumulation Curve for the community at Terembu Peninjau, associated to the selected, least-biased, nonparametric estimator (here Jackknife-5). In practice, the least-biased extrapolation of the Species Accumulation Curve allows to predict the expected additional sampling effort required to reach higher levels of sampling completeness (for example, the sample sizes required to reach 70%, 80%, 90% and 95% completeness would be around $N = 60, 100, 220, 500$ respectively).

3.2 Correction and extrapolation of the Species Abundance Distributions

The bias-corrected and extrapolated Species Abundance Distributions are plotted in Figure 5 (recorded part: ranks $i = 1$ to 16; extrapolated part from rank 17 to 24) for the community at Terembu Siput and in Figure 6 (recorded part: ranks $i = 1$ to 9; extrapolated part from rank 10 to 17) for the community at Terembu Peninjau. Figures 5 and 6 thus provide the entire development of the Species Abundance Distribution. Note that the extrapolated part of the distribution of abundances has no less importance than the recorded part since more or less rare species may have an equal ecological importance as more common ones [40–49]. Figure 7 allows to compare directly the Species Abundance Distributions of the two communities.
Figure 5 – The completed Species Abundance Distribution for the community at Terembu Siput, including the correction of the recorded part (involving the 16 detected species: grey discs) and the least-biased extrapolation of the unrecorded part (involving the 8 undetected species: coarse double line). Note logarithmic scale for relative abundances, a classical convention of graphical representation.

Figure 6 – The completed Species Abundance Distribution for the community at Terembu Peninjau, including the correction of the recorded part (involving the 9 detected species: grey discs) and the least-biased extrapolation of the unrecorded part (involving the 8 undetected species: coarse double line). Note logarithmic scale for relative abundances, a classical convention of graphical representation.
3.3 Qualitative and quantitative characterizations of the hierarchical structuring of species abundances

3.3.1 the type of process likely involved in the hierarchical structuring of species abundances

In order to investigate which kind of structuring process is at work in these sea-stars communities, two classical models of abundance distribution – the “log-normal” distribution and the “log-series” distribution – were tentatively fitted to the completed Species Abundance Distributions provided at Figures 5 and 6. As shown in Figures 8 to 11, a fairly good fit is obtained with the log-normal model for both studied communities while the accordance with log-series model is comparatively unsatisfactory.
Figure 9 – The classical “log-normal” model (sigmoid dotted line) fitted to the completed Species Abundance Distribution of the community at Terumbu Peninjau.

Figure 10 – The two classical models: “log-normal” (sigmoid dotted line) and “log-series” (fine double line) compared to the Species Abundance Distribution of the community at Terumbu Siput. Best fit is clearly obtained with the “log-normal” distribution. Note that considering the recorded part of the Species Abundance Distribution only would have led to the opposite conclusion: a “J” shaped model such as the “log-series” model would have fit the recorded part best than a “sigmoid” shaped model such as the “log-normal” model.
Figure 11 – The two classical models: “log-normal” (coarse dotted line) and “log-series” (fine double line) compared to the Species Abundance Distribution of the community at Terumbu Peninjau. Best fit is clearly obtained with the “log-normal” distribution. Note that considering the recorded part of the Species Abundance Distribution only would have led to the opposite conclusion: a “J” shaped model such as the “log-series” model would have fit the recorded part best than a “sigmoid” shaped model such as the “log-normal” model.

Figures 12& 13 – The same as Figures 8&9, but, here, the species relative abundances are plotted untransformed (instead of being classically log-transformed: [32]) for a better visualization of the good general fit with log-normal model of the recorded part as well. The identities of the eighteen recorded species have been added.
3.3.2. The intensity of the hierarchical structuring process

According to the definitions provided in Methods section:
- the degree, $U$, of unevenness of species abundance distribution resulting from the process driving the hierarchical structuring of abundances is computed according to equation (1);
- the genuine intensity, $I_\text{str}$, of the process driving the hierarchical structuration of abundances is relevantly appreciated by comparing the “S.A.D.” of the studied community to the corresponding “broken-stick” model, computed for the same species richness. Figures 14 and 15 allows this comparison, from which the genuine intensity $I_\text{str}$ of the structuring process is derived according to equation (2).

The corresponding results are summarized in Table 2 which highlights (i) the true total species richness $S_t$, (ii) the ratio $a_1/a_{St}$ between the abundances of the commonest and the rarest species, (iii) the degree of unevenness of species abundances $U$ and, finally, (iv) the genuine intensity $I_\text{str}$ of the process driving the hierarchical structuration of species abundances. Note that the parameters $U$ and $I_\text{str}$ accounts for two complementary aspects of the hierarchical structuration of species abundances: while $U$ quantifies the apparent pattern of species abundance structuration, $I_\text{str}$ highlights the genuine intensity of the underlying process driving this structuration. Being understood that, in quantitative terms, the unevenness pattern is far from faithfully reflecting the structuring process itself, as already emphasized.

The ratios $A_1 = (a_1/a'_1)$ and $A_{St} = (a_{St}/a'_{St})$ (which mirror the intensity of those factors which control the abundances $a_1$ and $a_{St}$ of the more and the less frequent species: see section Methods) are derived accordingly: $A_1 = 0.92, A_{St} = 0.90$ at Terembu Siput and $A_1 = 0.96, A_{St} = 0.57$ at Terembu Peninjau (Figures 16 and 17).

Table 2 – A synthetic summary of the main quantitative features of the hierarchical organization of species abundances within community, as derived from each numerically completed “S.A.D.”: (i) the total species richness $S_t$ of the community; (ii) the relative abundances $a_1$ and $a_{St}$ of the most and the least abundant species (species rank 1 and $S_t$); (iii) the unevenness of abundances in the community: $U = \log(a_1/a_{St})/(S_t-1)$; (iv) the unevenness of abundances in the corresponding “broken-stick” distribution: $U' = \log(a'_{1}/a'_{St})/(S_t-1)$, (v) the genuine intensity of the structuring process $I_\text{str} = U/U'$ and, at last, the ratios $A_1 = (a_1/a'_1)$ and $A_{St} = (a_{St}/a'_{St})$.

<table>
<thead>
<tr>
<th>sites</th>
<th>$S_t$</th>
<th>$a_1$</th>
<th>$a_{St}$</th>
<th>$a_1/a_{St}$</th>
<th>$a'_1$</th>
<th>$a'_{St}$</th>
<th>$U$</th>
<th>$U'$</th>
<th>$I_\text{str}$</th>
<th>$A_1$</th>
<th>$A_{St}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>T. Siput</td>
<td>24</td>
<td>0.14422</td>
<td>0.00157</td>
<td>94</td>
<td>0.15733</td>
<td>0.00174</td>
<td>0.0854</td>
<td>0.0851</td>
<td>1.003</td>
<td>0.92</td>
<td>0.90</td>
</tr>
<tr>
<td>T. Peninjau</td>
<td>17</td>
<td>0.19341</td>
<td>0.00197</td>
<td>98</td>
<td>0.20233</td>
<td>0.00346</td>
<td>0.1246</td>
<td>0.1104</td>
<td>1.128</td>
<td>0.96</td>
<td>0.57</td>
</tr>
</tbody>
</table>
Figure 14 – The Species Abundance Distribution of the community at *Terumbu Siput*, plotted together with the corresponding “broken-stick” distribution (i.e. computed for the same species richness).

Figure 15 – The Species Abundance Distribution of the community at *Terumbu Peninjau*, plotted together with the corresponding “broken-stick” distribution (i.e. computed for the same species richness).
The variations of the maximum and minimum abundances, \(a'_{1}\) and \(a'_{St}\) of the “broken-stick” model (double lines) and the uniform abundance level (\(= 1/S_t\)) of the perfectly “even” model (dashed line) as a function of the species richness \(S_t\).

The maximum and minimum abundances, \(a_1\) and \(a_{St}\), for each of the two studied sea-stars communities plotted jointly with the maximum and minimum abundances, \(a'_1\) and \(a'_{St}\) of the “broken-stick” model and the uniform abundance level (\(= 1/S_t\)) of the perfectly “even” model.

Linckia laevigata (Linnaeus 1758) © dr.scott.mills
Linckia multifora (Lamarck 1816) © Frédéric Ducarme
**Culcita novaeguineae** Müller & Troschel 1842 © Shizhao

**Echinaster luzonicus** (Gray 1840) © Bernard Dupont

**Fromia monilis** (Perrier 1869) © Hectonichus

**Acanthaster planci** (Linnaeus 1758) © Michel Dammeron

**Echinaster callosus** Marrenzeller 1895 © Nick Hobgood

**Leiaster speciosus** von Martens 1866 © MDC Seamark
4. DISCUSSION

Trying to get a comprehensive understanding – species by species – of the internal organization of species-rich communities would normally require long and tedious programs of field investigations, often beyond the usual practical possibilities.

Limiting the scope to the mere evaluation of the total species richness and the overall, synthetic characterization of the hierarchical structuring of abundances would yet still require the extensive sampling of the studied species assemblage. And even this less demanding requirement is often difficult to reach in practice, especially when having to deal with species-rich communities including a lot of rare species, as is, for example, often the case with invertebrate faunas.

Fortunately, the implementation of appropriate methods of numerical extrapolation can “force” incomplete samplings and partial inventories to reveal much more information than one would have expected a priori. Indeed, proper numerical extrapolations of both the Species Accumulation Curve and the Species Abundance Distribution can provide an unexpectedly rich set of additional information relative to those species remaining undetected after partial sampling. This, in turn, allows to tackle the main issues relative to the evaluation of true (total) species richness and the hierarchical organization of species abundance, even when communities are only partially sampled – all subjects that otherwise would have required exhaustive inventories.

A thorough analysis of two sea-stars communities associated to coral reefs, located in the Central South China Sea, has been conducted accordingly, in compliance with this methodological approach.

4.1 Total species richness estimates and the forecasted additional sampling efforts required to improve sampling completeness

At first, the procedure of numerical extrapolation implies selecting the least-biased estimator of the number of undetected species. Here, estimators Jackknife-3 and Jackknife-5 are selected for the communities at Trembu Siput and Terembu Peninjau respectively. Accordingly, the total species richness estimated this way reaches 24 species at Trembu Siput and 17 species at Terembu Peninjau (Table 1), which substantially exceeds the recorded numbers (16 and 9 species respectively). This confirms the limited levels of sampling completeness (67% and 53% respectively) and thus justifies, a posteriori, the need for implementing numerical extrapolations of these inventories. The importance of selecting for each community
the corresponding least-biased estimator of the number of undetected species – and the associated least-biased extrapolation of the Species Accumulation Curve – is advocated at Figures 1 and 3, which both highlight the marked differences that separate the extrapolations of the Species Accumulation Curve associated to different types of estimators. In particular, here, Jackknife-1 and Chao estimators prove being strongly biased negatively (Figures 1 and 3).

Although the least-biased numerical extrapolation can provide a lot of additional interesting information regarding the set of still unrecorded species, further sampling effort, aiming at increasing the completeness of inventories, might alternatively be considered. In this perspective, reliable forecasts of the additional sampling efforts in order to meet any targeted gain in sampling-completeness would be useful for the optimal planning of the additional efforts to be implemented. The least-biased extrapolation of the Species Accumulation Curve answers appropriately this demand, as shown in Figures 2 and 4). Clearly, further improvements of sampling completeness would rapidly require very substantial additional efforts. For example, increasing completeness at Terembo Siput, from the actual 67% level up to 80%, 90%, 95% completeness levels would require multiplying the actual sample-size (No = 52) by a factor 2, a factor 3.5, a factor 7, respectively. Being able to reliably estimate the required additional efforts, as shown above, is of obvious prime interest to rationally decide whether to continue sampling operation any further or to rely only on actual partial inventories, subsequently completed by numerical extrapolation.

4.2 Correction and extrapolation of the Species Abundance Distribution

As recorded Species Abundance Distributions need correction and extrapolation because both are (i) slightly biased due to sampling stochasticity and (ii) most importantly, incomplete, as no less than eight species had remained undetected, in each studied community. After correction and extrapolation are applied, the complete development of the Species Abundance Distributions, including the estimated distribution of the abundances of the still undetected species, is made available: Figures 5, 6, 7. Note that complementation of Species Abundance Distribution to include the (undetected) less abundant species may be more important than it might seem at first glance. Indeed, less abundant species may possibly have ecological importance no less than more common species, as it has already been repeatedly emphasized [40 – 49]. In addition, considering the full range of the Species Abundance Distribution is essential, not only to deliver a full description of the pattern of abundances but, also, to question the kind of process actually involved in the hierarchical structuring of abundance distribution as well as the genuine intensity of this structuring process. Indeed, answering these questions properly requires comparing the full range of Species Abundance Distribution to different theoretical models [5], at the risk, otherwise, to provide severely erroneous inferences.

4.3 Inferring the type of process driving the hierarchical structuring of species abundances

Considered full range, the Species Abundance Distributions of both studied communities clearly fit best the “log-normal” distribution than the “log-series” distribution (Figures 8 to 11). This suggests that the process of structuration of these sea-stars communities is likely driven by the combined contributions of many independent factors, rather than by only one (or very few) dominant factor. This, in fact, might well be a rather general trend, as already argued elsewhere [5, 25, 26 – 28]. Moreover, the fairly good fit to the “log-normal” model stands as good within the range of most abundant species (i.e. for lower ranks i). This suggests that, in these sea-stars communities, no additional negatively (resp. positively) density-dependent factor actually
occurs that, otherwise, would have depressed (resp. increased) the levels of abundance of the most abundant species.

At last, it is also worth noting that relying only upon the recorded part of the Species Abundance Distribution would have led to the opposite conclusion: indeed, the “J” shape of the “log-series” model fits the recorded part best than does the “sigmoid” shape of the “log-normal” model. This is a new confirmation that relying upon incomplete distributions of abundances only (i.e. neglecting the numerical extrapolation when required) may often leads to erroneous diagnostics, as already emphasized by several authors [4, 6, 20, 27, 50, 51].

4.4 Quantifying the degree of hierarchical structuration of species abundances

Here also, considering the full range of the Species Abundance Distribution is necessary, not only to duly include the subset of the still undetected species but also, to make possible the standardization of the Species Abundance Distribution slope to the corresponding “broken-stick” reference (Figures 14 & 15). As argued above, this is a key condition to unveil the genuine intensity $I_{str}$ of the process driving the hierarchical structuration of species abundances. Here (Table 2), the intensity of the structuring process is very close to 1 (i.e. very similar to the intensity in the “broken-stick” model) for the community at Terembu Siput ($I_{str}=1.003$), while in the community at Terembu Peninjau, the intensity of the structuring process is somewhat (12%) stronger ($I_{str}=1.128$). In turn, this stronger structuring intensity is only marginally related to the level of dominance of the most abundant species (since the values of $A_1 = (a_1/a'_1)$ for both communities are very similar: Table 2, Figure 17). Instead, the stronger structuring intensity is mainly due to the lower value of $A_{St} = (a_{St}/a'_{St})$ (0.90 at Terembu Siput and 0.57 only at T. Peninjau: Table 2, Figure 17). In other words, the larger intensity of the structuring process highlighted in the community at Terembu Peninjau mainly involves the right-hand part of the abundance distribution, that is the set of less abundant species.

Another interesting question is: how the structuring intensities $I_{str}$ in these two sea-stars communities would compare with the structuring intensities in other types of marine invertebrates communities. The currently available data in this respect remains still limited [21, 34, 35], but yet suggests that tropical marine gastropod communities tend to be more strongly structured than the two studied sea-stars communities.

As regards the unevenness pattern, the difference in the degree of unevenness $U$ between the two studied communities proves being still far larger (> 46%, Table 2) than is the difference in the genuine intensity of the structuring process (12%). As already emphasized, such discrepancy between the pattern ($U$) and the underlying process ($I_{str}$) is the mere mathematical consequence of the difference in species richness between the two communities.

At last, it should be noted that if the three more abundant species are the same in both communities (labelled a, b, c: see Figures 12 & 13); the taxonomic composition of less abundant species is, on the contrary, very different – at least as regards the set of recorded species. Yet, this cannot be considered as firmly conclusive, as it might well be possible that this difference would be less pronounced if the taxonomic identities of the unrecorded species were unveiled. In this respect, numerical extrapolation finds its limit and only further sampling can actually provide a sound answer. An answer, however, at a very substantial extra-cost in term of additional sampling effort. Additional effort that the numerical extrapolation of the Species Accumulation Curves can efficiently help to predict (Figures 2 & 4).
5. Conclusion

When dealing with substantially incomplete species inventories, the numerical extrapolations
of (i) the Species Accumulation Curve and (ii) the Species Abundance Distribution offer
remarkable opportunities to unveil an unexpectedly rich sum of information relative to the set
of undetected species. In turn, thanks to the resulting access to the full range of the Species
Abundance Distribution, interesting additional information may be derived, regarding the
process driving the hierarchical organization of species abundances within a partially sampled
community. The numerical extrapolations – here applied to the partial inventories of two sea-
stars communities – demonstrate concretely the wide range of ecological questions that may be
addressed and successfully answered, even when no less than one third up to almost half of the
member-species had remained unrecorded.

In short, this clearly highlights the potential interest of numerical extrapolations applied to
partial inventories, in the context of increasingly frequent practice of "quick assessments" of
biodiversity, especially when having to deal with highly species-rich assemblages, as is often the
case with invertebrate faunas under tropical climates.

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 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 [17, 52, 53]:

$$\frac{\partial^n R(N)}{\partial N^x} = (-1)^{(x-1)} \frac{f_0(N)}{C_{N,x}} = (-1)^{(x-1)} \frac{(N!)^x}{N^x} f_0(N) \quad (\approx \text{as } N \gg x) \quad (A1.1)$$

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 $R_x(N)$ that respectively satisfy the
mathematical constraint (A1.1), considering increasing orders $x$ of derivation $\partial^n R(N)/\partial N^x$. Each
solution $R_x(N)$ is appropriate for a given range of values of $f_1$ compared to the other numbers $f_i$,
according to [17]:

* for $f_1$ up to $f_2$ \( R_1(N) = (R(N_0) + f_1) - f_1.N_0/N \)

* for $f_1$ up to $2f_2 - f_3$ \( R_2(N) = (R(N_0) + 2f_1 - f_2) - (3f_1 - 2f_2).N_0/N - (f_2 - f_1).N_0^2/N^2 \)

* for $f_1$ up to $3f_2 - 3f_3 + f_4$ \( R_3(N) = (R(N_0) + 3f_1 - 3f_2 + f_3) - (6f_1 - 8f_2 + 3f_3).N_0/N - (-4f_1 + 7f_2 - 3f_3).N_0^2/N^2 - (f_2 - f_1).N_0^3/N^3 \)

* for $f_1$ up to $4f_2 - 6f_3 + 4f_4 - f_5$ \( R_4(N) = (R(N_0) + 4f_1 - 6f_2 + 4f_3 - f_4) - (10f_1 - 20f_2 + 15f_3 - 4f_4).N_0/N - (-10f_1 + 25f_2 - 21f_3 + 6f_4).N_0^2/N^2 - (5f_1 - 14f_2 + 13f_3 - 4f_4).N_0^3/N^3 \)
The associated non-parametric estimators of the number $\Delta J$ of missing species in the sample [with $\Delta J = R(N=\infty) - R(N_0)$] are derived immediately:

* 0.6 $f_2 < f_1 \leq f_2 \rightarrow j_1 = f_1$ ; $R_1(N)$

* $f_2 < f_1 \leq 2f_2 - f_3 \rightarrow j_2 = 2f_1 - f_2$ ; $R_2(N)$

* $2f_2 - f_3 < f_1 \leq 3f_2 - 3f_3 + f_4 \rightarrow j_3 = 3f_1 - 3f_2 + f_3$ ; $R_3(N)$

* $3f_2 - 3f_3 + f_4 < f_1 \leq 4f_2 - 6f_3 + 4f_4 - f_5 \rightarrow j_4 = 4f_1 - 6f_2 + 4f_3 - f_4$ ; $R_4(N)$

* $f_1 > 4f_2 - 6f_3 + 4f_4 - f_5 \rightarrow j_5 = 5f_1 - 10f_2 + 10f_3 - 5f_4 + f_5$ ; $R_5(N)$

N.B.1: As indicated above (and demonstrated in details in Béguinot[17], 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 [2, 3, 15, 16, 54]. Also, this shows that the approach initially proposed by Brose et al.[55] – 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 [18].
Figures A1 & A2 – The recorded values of the numbers $f_x$ of species recorded $x$-times (grey discs) and the regressed values of $f_x$ (black discs) derived to reduce the consequence of stochastic dispersion during sampling.

Appendix 2

**Correction and extrapolation of the as-recorded Species Abundance Distribution (S.A.D.)**

N.B.: details regarding the derivation of the following expressions are provided in [20].

1) Correction for bias of the recorded part of the S.A.D.

The bias-corrected expression of the true abundance, $\tilde{a}_i$, of species of rank ‘$i$’ in the S.A.D. is given by:

$$\tilde{a}_i = p_i(1+1/n_i)(1-f_1/N_0)/(1+R_0/N_0) \quad (A2.1)$$

where $N_0$ is the actually achieved sample size, $R_0$ ($=R(N_0)$) the number of recorded species, among which a number $f_1$ are singletons (species recorded only once), $n_i$ is the number of recorded individuals of species ‘$i$’, so that $p_i = n_i/N_0$ is the recorded frequency of occurrence of species ‘$i$’, in the sample. The crude recorded part of the “S.A.D.” – expressed in terms of the series of as-recorded frequencies $p_i = n_i/N_0$ – should then be replaced by the corresponding series of expected true abundances, $\tilde{a}_i$, according to equation (A2.1).

2) Extrapolation of the recorded part of the S.A.D. accounting for the complementary abundance distribution of the set of unrecorded species

The following expression stands for the estimated abundance, $a_i$, of the unrecorded species of rank $i$ (thus for $i > R_0$):

$$a_i = (2/N_i)(1- [\partial R(N)/\partial N]_N)/(1+ R(N_i)/N_i) \quad (A2.2)$$

which, in practice, comes down to:

$$a_i = (2/N_i)/(1+ R(N_i)/N_i) \quad (A2.3)$$

as $f_1(N)$ already becomes quite negligible as compared to $N$ for the extrapolated part.

This equation provides the extrapolated distribution of the species abundances $a_i$ (for $i > R(N_0)$) as a function of the least-biased expression for the extrapolation of the species accumulation curve $R(N)$ (for $N > N_0$), ‘$i$’ being equal to $R(N_i)$. The key to select the least-biased expression of $R(N)$ is provided at Appendix 1.

Appendix 3

*The trivial contribution of the level of species richness to the degree of structuring of species abundances*
All things equal otherwise, the larger the species richness, the weaker is the slope of the Species Abundance Distribution. This can be easily exemplified and quantified, on a theoretical basis, by considering a theoretically constant structuring process - such as the random distribution of the relative abundances that characterises the “broken-stick” distribution model. By applying this model successively to a series of communities with increasing species richness, a steadily decrease of the slope of abundance distributions is highlighted: Figure A3

Figure A3 – The “broken-stick” distribution model applied to species communities with increasing species richness $S_t = 10, 20, 30, 60$. Although the theoretical structuring process involved in the “broken-stick” model remains unchanged (random apportionment of relative abundances among member species), the slope of the species abundance distribution strongly depends upon (and monotonously decreases with) the level of species richness $S_t$.

COMPETING INTERESTS
Author has declared that no competing interests exist.

References


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; vol. 2014 : article ID 847328, doi:10.1155/2014/847328 ; <hal-01101415>

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>
