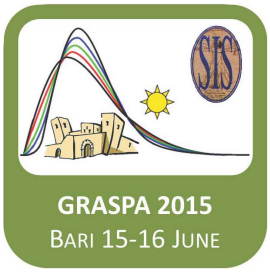

Rarefaction and extrapolation with Hill numbers: a study of diversity in the Ross Sea

C. Ghiglione[1], C. Carota[2], C. R. Nava[2,*], I. Soldani[3] and S. Schiaparelli[1]

[1] *DiSTAV, University of Genova and Italian National Anctartic Museum (section of Genova); claudio.ghiglione@riftia.eu, stefano.schiaparelli@unige.it*
[2] *Department of Economics and Statistics "Cognetti de Martiis", University of Torino; cinzia.carota@unito.it, consuelorubina.nava@unito.it*
[3] *aizoOn Technology Consulting; irene.soldani@aizoon.it*
[*] *Corresponding author*

Abstract. *The Ross Sea can be considered, in a biological sense, one of the better-known areas in Antarctica due to the high number of expeditions engaged since 1899. Hundreds of mollusc species have been collected and classified along years in a unique database which is now available for study. The possibility to access such impressive information offers the opportunity to apply important results in the study of biodiversity for that area. Recent influential scientific contributions induce us to study species diversity by means of accumulation curves based on Hill numbers, i.e. the effective number of equally frequent species.*

Keywords. *Accumulation curves; Biodiversity; Extrapolation; Hill numbers; Rarefaction.*

1 Introduction

The construction of a complex and unique database, containing information related to the species richness in the Ross Sea (Antarctica) since 1899, is the result of the intensive work of an international team of ecologists. The information collected, standardized and stored along years with respect to the Phylum Mollusca in the Ross Sea, creates the biological context in which we perform the statistical biodiversity analysis described below.

Before the illustration of Hill numbers and diversity accumulation curves, a brief description of available data and their finding techniques is provided. A partial but meaningful difference, regarding the way and the aim that move researchers during expeditions, arises. Indeed, 2004 can be considered the turning point in this respect. Expeditions prior to this date were primarily focused on the realization of species inventories, without taking into account species abundances and without recording zeros for stations where any species was found. Even if this approach positively changed with the new century, these aspects generates some limitations. For instance the variable identifying species richness, i.e. the number of species found in a given sample, has only positive integer values.

Antarctica is a key location to monitor trends in biodiversity. This is a critical issue under a global warming scenario which likely would have a major impact in polar areas by introducing species from

warmer latitudes and by extinguishing stenothermal ones. Geographical shifts in species distribution patterns and temporal trends could be recognized and studied in data sets as the one here considered.

However, as already mentioned, the treatment of this type of data is not trivial. In ecology, moreover, the measurement of biodiversity is itself a complex issue deeply discussed in the literature.
Many shortcomings in the quantification of biodiversity can be defused by resorting to Hill numbers, first introduced by [3]. For a complete review over the advantages of this approach and a unification of the most important related results see [1], where main techniques for sample rarefaction and extrapolation are discussed. Here we apply some of the methods described in [1, 2] for sample-based incidence data. We draw a picture of the whole set of collected data accordingly to different gears: grab, towed and Rauschert, a specific type of dredge having a fine mesh size.

2 Methods

Traditionally, Hill numbers have been used for individual-based abundance data. Here we apply to our sample-based incidence data the methods presented in [1, 2], where a comprehensive statistical framework for the analysis of biodiversity data is provided. Therefore, our main results consist of unified diversity accumulation curves (Figures 1.a, 1.b, 1.c and 1.d). The latter are based on empirical estimates of the principal Hill numbers extended in order to incorporate information on the incidence probabilities. All these concepts are made clear in the next paragraphs.

Our data consist of a species-by-sampling-unit incidence matrix (W_{ij}) with S rows (S denotes the total number of species present in the assemblage) and $T = 456$ columns (the number of independent sampling units, i.e. discrete sampling events). Entries of the incidence matrix record the presence or absence of each species within each sampling unit: $W_{ij} = 1$ if species i is detected in the sampling unit j, $W_{ij} = 0$ otherwise. The row sum of the incidence matrix, $Y_i = \sum_{j=1}^{T} W_{ij}$, denotes the incidence-based frequency of species i, for $i = 1,..,S$. The frequencies Y_i represent the incidence reference sample to be rarefied or extrapolated in the diversity accumulation curves detailed in Figures 1.a, 1.b and 1.c. Species non detected in any sampling unit but present in the assemblage yield $Y_i = 0$. Only species with $Y_i > 0$ contribute to the total number of species observed in the reference sample denoted by S_{obs}.

Under the assumption that each species i has its own unique incidence probability π_i and that π_i is constant for any randomly selected sampling unit, each element of the incidence matrix can be viewed as a Bernoulli random variable with probability π_i that $W_{ij} = 1$ and probability $1 - \pi_i$ that $W_{ij} = 0$. This implies a Bernoulli product model for the incidence matrix,

$$P(W_{ij} = w_{ij} | \forall i = 1,2,\ldots,S,\ j = 1,2,\ldots,T) = \prod_{j=1}^{T}\prod_{i=1}^{S} \pi_i^{w_{ij}}(1-\pi_i)^{1-w_{ij}} = \prod_{i=1}^{S} \pi_i^{y_i}(1-\pi_i)^{T-y_i}.$$

Note that the likelihood under this model for W_{ij} is proportional to the likelihood under a Binomial model for $Y_1,\ldots,Y_S$. Note also that the sum of the incidence probabilities, $\sum_{i=1}^{S} \pi_i$, may be greater than 1. In [1] each parameter π_i is normalized (divided by the sum $\sum_{i=1}^{S} \pi_i$), to obtain the *relative incidence* of each species i in the assemblage; then, under the Bernoulli product model specified above, the Hill number of order q is defined as

$${}^{q}\Delta = \left(\sum_{i=1}^{S} \left[\frac{\pi_i}{\sum_{j=1}^{S} \pi_i} \right]^{q} \right)^{\frac{1}{(1-q)}}, \qquad q \geq 0,\ q \neq 1.$$

The sensitivity of ${}^{q}\Delta$ to differences in the incidence probabilities increases with q. In all cases a value ${}^{q}\delta$ of ${}^{q}\Delta$ is interpreted as the effective number of equally frequent species in the assemblage from which the sampling units are drawn. In other words, the diversity of the assemblage is the same as an ideal assemblage with ${}^{q}\delta$ species all with equal probability of incidence.

If all the incidence probabilities $(\pi_1, .., \pi_S)$ are identical, than the Hill number of all orders reduces to the species richness ${}^0\Delta$.

Hill numbers have to be regarded as theoretical or asymptotic diversities at an infinite sample size for which the true relative incidences of each i species are known. In contrast, diversity accumulation curves are based on estimates of Hill numbers of order $q = 0, 1^1, 2$, yielding the species richness, the exponential of Shannon entropy and the inverse Simpson concentration, respectively. Given the sufficient statistics $Y_0, Y_1, .., Y_S$, estimates of Hill numbers for a sample of size m are based on the incidence frequency counts $Q_k = \sum_{i=1}^{S} I(Y_i = k)$, i.e. the number of species each represented exactly $Y_i = k$ times in the incidence matrix sample $0 \leq t \leq T$.

3 Results and Conclusions

For our data we plot the main diversity accumulation curves. First, we provide the *sample-size-based rarefaction and extrapolation sampling curve* (Figure 1.a) which shows the trend of Hill numbers when the number of sampling units increases. Then the *sample completeness curve* (Figure 1.b) describes the sample completeness as a function of the sample size; it is useful to derive the sample size needed to reach a prefixed population coverage.

Finally, we show the *coverage-based rarefaction and extrapolation sampling curve* (Figure 1.c) which points out the behaviour of the species diversity as the sample coverage increases.

Finally, in order to compare the efficiency of the different gears used to collect sampling units in different expeditions, we plot the sample-size-based rarefaction and extrapolation sampling curve for grab, towed and Rauschert (Figure 1.d).

Although only eighteen sampling units are picked up with the Rauschert (due to its only recent use in the Ross Sea), which disproportionately increases the length of the 95% confidence interval (the shaded area about the curve), the superior efficiency of such a gear is apparent. Given the same sample size, the Rauschert allows to find a greater number of different species than grab and towed. Related results can be found in [4].

References

[1] Chao, A., Gotelli, N.J., Hsieh, T., Sander, E.L., Ma, K., Colwell, R.K., and Ellison, A.M. (2014). Rarefaction and extrapolation with hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs*, **84**(1), 45–67.

[2] Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.Y., Mao, C.X., Chazdon, R.L., and Longino, J.T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, **5**(1),3–21.

[3] MacArthur, R. H. (1965). Patterns of species diversity. *Biological Reviews* **40**, 510–533.

[4] Schiaparelli, S., Ghiglione, C., Alvaro, M.C., Griffiths, H.J., and Linse, K. (2014). Diversity, abundance and composition in macrofaunal molluscs from the Ross sea (Antarctica): results of fine-mesh sampling along a latitudinal gradient. *Polar biology* **37**(6), 859–877.

[1] We shortly denote with 1 the limit $q \to 1$

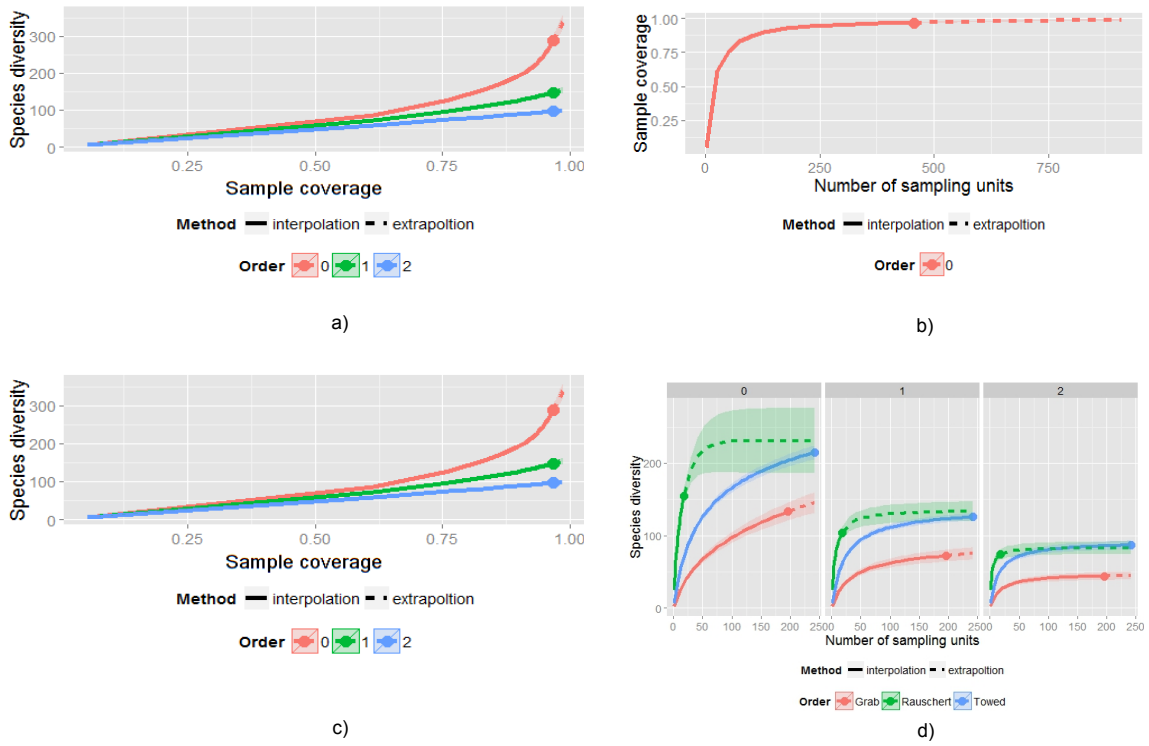

Figure 1: **a)** Sample-size-based rarefaction and extrapolation sampling curves: estimates of the number of species found in a random set of t $(t < T)$ sampling units (solid curves/rarefaction) or in an augmented set of $(T + t^*)$ $t^* > 0$ sampling units from the assemblage (dashed curves/extrapolation). **b)** Sample completeness curves. **c)** Coverage-based rarefaction and extrapolation sampling curves. **d)** Sample-size-based rarefaction and extrapolation sampling curves: grab, Rauschert and towed comparison.