Assessing the efficiency of non-parametric estimators of species richness for marine microplankton

Miguel Branco^{1,*}, Francisco G. Figueiras², Pedro Cermeño³

¹ Departamento de Ecología e Biología Animal, Universidade de Vigo, E-36310, Vigo, Spain

² Departamento de Oceanología, Instituto de Investigaciones Mariñas, Consejo Superior de Investigaciones Científicas, Eduardo Cabello, 6, Vigo E-36208, Spain

³ Departamento de Biología Marina y Oceanografía, Institut de Ciències del Mar, Consejo Superior de Investigaciones Científicas, Passeig Marítim de la Barceloneta, 39-47, Barcelona E-08003, Spain

*Corresponding author: mbranco@uvigo.es

Abstract

Non-parametric asymptotic estimators rely on the assumption that rare species are indicative of the degree of undersampling. We evaluate the performance of 11 non-parametric asymptotic species richness estimators and an individual-based rarefaction and extrapolation (R/E) method using data of marine microplankton. These species richness estimators were evaluated by sequentially increasing sampling effort. Their bias was diminished as sample (size) completeness increased. Jackknife estimators were more accurate than others such as Chao's estimators. Underestimates were larger than 10% of species when applied to more than 30% of the individuals present in the community. The magnitude of bias varied as a function of community structure. R/E curves sorted samples by species richness but with a significant sampling error associated. We find that the use of these analytical methods lacks of validity for the comparison of marine microplankton species richness estimates from sample collections spanning large spatial and temporal scales.

Keywords: species richness, non-parametric estimators, rarefaction and extrapolation, microbial plankton, community structure

Introduction

Microplankton (20-200 μm equivalent spherical diameter) carry out cornerstone processes in marine ecosystems. The structure and the functioning of marine food webs and consequently biogeochemical cycles largely depend on the taxonomic composition of microplankton communities (Ryther, 1969; Azam *et al.*, 1983; Jiao *et al.*, 2010). Recent advances in high-throughput sequencing and exhaustive inspection of seawater samples under the microscope have revealed that microbial plankton is composed of a high number of taxa, or species. Most of them are consistently rare. Only a small number of species are abundant and dominate ecosystem functioning (Pedrós-Alió, 2006; Amaral-Zettler *et al.*, 2010; Massana, 2011; Rodríguez-Ramos *et al.*, 2014; Cermeño *et al.*, 2014). These observations open up discussions about the role of species richness in ecosystem functioning. Some classical debates were centred on how can so many species coexist and what is the benefit of such a high diversity. When multiple species share the same resource requirements over long periods of time competitive exclusion is likely to occur. To explain this plankton paradox (Hutchinson, 1961), it was suggested that species richness has a positive effect on the long-term stability of the communities as, for instance, it favours the increase in primary production under variable nutrient supply regimes (i.e. Odum, 1953; May, 1972). However, testing an hypothesis like this obligate us to quantify confidently the number of species of microplankton communities.

Rare species (those in very low abundance) are more difficult to detect than abundant ones. As microplankton communities are rich in rare species, the volume of water we need to collect to see all the species multiplies. The hydrodynamics of coastal marine environments favours dispersal and so the structure of microplankton communities shifts relatively fast (Rodríguez-Ramos *et al.*, 2014). These features of marine microplankton communities lead to a systematic undersampling of species richness (Cermeño *et al.*, 2014). It was shown that more than 40% of microplankton species are overlooked with the use of traditional sampling protocols (Cermeño *et al.*, 2014),

and therefore, obtaining representative field samples is impractical. We need to know how to optimize sampling and analytic efforts to obtain accurate estimates microplankton species richness. Specifically, it is necessary to test the reliability of different methodologies to infer and to compare species richness using relatively small sample sizes. Ultimately, the decision of which method is the most desirable will represent a trade-off between sampling effort and accuracy.

The scientific literature has seen a continuous increase in the number of statistical methods aimed at quantifying diversity, particularly the number of species present in a community. Many of these methods were proposed to facilitate the quantification of species richness when sampling effort is relatively low. However, these methods rely on different premises and there is no general consensus on which one is preferable (Hughes *et al.*, 2001; Bunge 2009). One way of determining the species richness comes from discriminating the parametric equation that best describes the distribution of species abundances in the community. This means fitting and solving which parametric distribution (such as log-series or log-normal) delimits the distribution of species in relation to individuals. First, some parametric distributions are considered and some parameters are obtained for each of them. With that, their functions can be adjusted to the species-abundance distributions observed and then scaled. In the same line, species richness can be calculated by fitting parametric curves to a species accumulation curve (SAC) from which the asymptote can be inferred. SACs are representations of the number of species observed in relation to the number of individuals or samples (i.e. traps, bottles, nets or alike) counted. These approaches are usually criticized because of the uncertainty of the parameter estimates and because of the ambiguity of discriminating which one is the distribution that best characterizes the community (Chao *et al.*, 2014). Furthermore, they do not offer estimates with an associated variance unless we impose some conditions (Chao & Chiu, 2016).

A popular method that has been used extensively over the last decades is non-parametric asymptotic estimators. They were introduced to estimate the species richness of the community based on the number of rare species present (Chao, 1984). These methods assume that the number of low abundance species found in a sample indicate how likely it is there are more undiscovered taxa. Then, they estimate the

number of species that are presumably missed. The non-parametric asymptotic estimators are founded on the principle that most of the information about the structure of the community is contained within the pool of rare species (Turing, 1939; Good, 1953). These analytical methods calculate a correction factor which is then added to the observed number of species, S_{obs} , to obtain an estimate of species richness, $\hat{S}_{est}$. They have the advantage of doing no assumption about the species-abundance distribution (Chao & Chiu, 2016). In practice, the abundance-based non-parametric estimators use some information about the number of individuals of some species in one or multiple samples. They rely on the frequencies (f) of those species collected in any sample with only one individual (singletons, S_1), species with two individuals (doubletons, S_2), or some others with higher frequencies but usually less than ten (Burnham and Overton, 1978; Burnham and Overton, 1979; Chao, 1987). For example, *Jack1* (Burnham and Overton, 1978) and *Chao1* indices (Chao, 1984) consider S_{obs} together with the number of singletons (f_1) to reduce the bias or infer a lower bound estimation of true species richness, S_{est} , respectively. This is, they are intended to either lessen the number of species missed (reduce bias) or infer a possible minimum of species present in the community. Their correction factor is simply f_1 or $f_1^2/2f_2$ (but $f_1(f_1 - 1)/2(f_2 + 1)$ if $f_2 = 0$), respectively. Others, like *Chao2* (Chao, 1987) or *ICE* (Chao and Lee, 1992), are incidence-based estimators that use the presence or absence of species in multiple samples of the same community (collected at the same time or place) to infer that number instead. For instance, *Chao2* adds $Q_1^2/2Q_2$ if $Q_2 \geq 0$ or $1/2 (Q_1(f_1 - 1))$ if $Q_2 = 0$ to S_{obs} , where Q_n is the number of species that were present n times in the different replicates (or subsamples).

Recently, non-parametric methodologies were reviewed and rarefaction (Heck et al., 1975; Coleman et al. 1982) was used as the foundation of a method designed to compare diverse communities fairly. Colwell *et al.* 2012 and Chao & Jost 2012 reckoned that studying species-rich communities had an inherent difficulty. They advanced trying not to discard information and supporting the ability to compare communities at some specific sampling effort or sampling coverage. They described how to interpolate and extrapolate diversity indices by using one sample as a reference. Colwell *et al.* described how to obtain a sample size rarefaction and

extrapolation (R/E) curve. These R/E curves were designed to compare multiple communities at the same sample size. Chao & Jost 2012 focused rather on sample completeness and designed a similar method of R/E. They standardized datasets by sample coverage instead. Coverage is understood as the fraction of the total number of individuals that the species collected in the sample represent in that community. In both cases, rarefaction interpolates diversity to smaller samples sizes or coverage and extrapolation, which is based on Chao1 for abundance data and Chao2 for incidence data, infers how much diversity the communities to bigger sample sizes or coverage.

A desirable property of species richness estimators is to be robust in their estimates and lack any bias associated with, for instance, differences in community structure. When communities are examined using fixed sample areas or volumes we obtain measures of species density (e.g. number of species per unit of volume) that are not necessarily proportional to their species richness. The number of individuals present in one sample depends on the species distribution in the habitat and the productivity of the ecosystem (Arrhenius, 1921; Colwell et al., 2004; Gotelli and Colwell, 2001; Nufio et al. 2009). The species-abundance distribution strongly affects how many samples of a community we have to count so that we collect all the species. If one community contains more rare species than another, we will have to conduct an uneven sampling to obtain unbiased richness estimates. R/E methods try to solve this problem by comparing species richness at the same degree of sampling completeness or coverage (Alroy 2010a; Chao and Jost, 2012; Jost 2010). Non-parametric asymptotic prediction capacity was indicated to rely on fieldwork (Colwell and Coddington, 1994), which, in fact, motivated the development of R/E methods. It was shown that if more than half of the species were sampled, the estimator performance would stabilize and become independent of the sample completeness (Colwell and Coddington, 1994; Chazdon *et al.*, 1998; Hughes *et al.*, 2001).

Communities change their species composition and number of individuals across space and/or through time. Predictions of estimators should be robust to environmental changes. Thus, it is important to measure the extent to which the non-parametric estimators, both asymptotic and R/E curves, provide fixed predictions of species richness, independently of the idiosyncrasy of the ecosystem.

Here, we test the predictive capacity of non-parametric asymptotic estimators of species richness, and the rarefaction and extrapolation method for comparing multiple samples using marine microplankton data. Our community dataset was collected during a full annual cycle in a dynamic coastal ecosystem (Ría de Vigo, NW Iberian Peninsula). Our strategy was, first, to compare the performance of eleven non-parametric asymptotic estimators of species richness by sequentially increasing sampling completeness (from 5 mL to >20 L). For instance, species richness estimates based on 5 mL seawater samples were compared with the number of species counted in a sample of 10 mL, and so on, until reaching the maximum volume that was sampled for that community. Second, we compared how much different microplankton community was using rarefaction and extrapolation curves. In this case, we paid attention to how they rank the community by species richness and what confidence do these estimations have. We assessed the inference capacity of each non-parametric asymptotic estimator by considering its bias, accuracy and precision at different levels of sampling completeness (understood as the amount of volume inspected), using as reference the total number of species observed at the highest level of sampling effort (as reported in Cermeño *et al.*, 2014). In addition, the predictive capacity of these methodologies was evaluated by taking into consideration community structure, which varied throughout the year in this highly dynamic temperate ecosystem.

Methods

Sampling protocol and microscopy analysis

From February to November 2012, nine oceanographic cruises were carried out on board of the R/V Mytilus to a central station in the Ría de Vigo, NW Iberian Peninsula (N42°14.09', W08°47.18'). On each visit, we recorded vertical profiles of temperature with a Sea Bird Electronics SBE 9/11 conductivity, temperature, and depth (CTD) probe attached to a rosette. Seawater samples were collected from 3, 10 and 20 m depth using 12 L Niskin bottles. These three samples were dispensed into a 20 L container and gently mixed to obtain a combined sample of the water column photic layer. From this combined sample, four subsamples of 500 mL were preserved in Lugol's iodine solution (2% final concentration). Additionally, net samples were

collected from vertical hauls (from 20 m to surface) with a plankton net (20 μ m mesh size) and preserved alike.

The samples were examined for taxonomic identification and cell counting using an inverted microscope (Nikon Eclipse TE200-S). In the case of bottle samples, increasing volumes from 5-50 mL were settled in composite sedimentation chambers and examined in the inverted microscope following the Utermöhl technique (Utermöhl, 1958). The volume of sample was fixed according to the seawater chlorophyll-a concentration (Lund et al., 1958; Sournia, 1978). Aliquots were settled until a total volume of 500-1000 mL was reached, depending on the concentration of organisms. In the case of net samples, aliquots corresponding to concentrated water column volumes of 500 to 9000 mL were inspected under the microscope. For deciding the volume of aliquots, it was taken into account the volume of seawater filtered through the net and the total volume collected. This was repeated for each net sample until examining ~25 L for each sampling day. The conventional sampling methods usually examine seawater samples in the range of 5-100 mL.

Cell counts were performed for bottle samples. For rare species, entire slides were examined. For common species, a single random transect was counted and then their total abundance estimated assuming homogeneous distribution of populations. While the marine phytoplankton bottle samples were taxonomically classified and counted (number of individuals or to abundance level), net samples were only used to detect rare species not observed in bottle samples (presence/absence or to incidence level). To avoid confusion, only diatom and dinoflagellate species, which are dominant groups, were considered for this study. Microplankton specimens of other taxonomical groups may be more susceptible to incorrect classification because of physical degradation or similar causes. All data are publicly available in Figshare (DOI: [10.6084/m9.figshare.4555456](https://doi.org/10.6084/m9.figshare.4555456))

Non-parametric species richness estimators

Non-parametric asymptotic estimates of species richness were computed at different levels of sampling completeness. The full list of non-parametric species richness estimators is detailed in *Table 1*.

First, we constructed the species accumulation curves (SAC). These curves result from the cumulative number of species observed as a function of the number of individuals counted in different subsamples (bottle samples). They showed how the rate of new species observed decreased with sampling effort (See *Figure S1*). The number of individuals of the total that it was necessary to sample in order to observe the maximum of species is defined here as the degree of sampling completeness (abbreviated as completeness). That value is the maximum of species observed in field.

Non-parametric asymptotic estimators refer here to Jackknife and Chao's coverage-based non-parametric families (see Chao and Chiu, 2016 for a review). These richness estimates were calculated for 100 random species abundance subsamples. Those estimates were averaged at each level of sampling completeness of bottle samples. Subsampling means resampling to smaller sample sizes so calculations can be performed at any degree of completeness. These subsamples were obtained using a random rarefaction, and as a result they contained a specific percentage of the total number of individuals of the community. Standardizing by completeness allows us to compare samples at reasonably similar degrees of saturation of the SACs. Calculations over multiple replicates give confidence to the estimates because processes such as spatial aggregation, for instance, becomes null. Estimates were compared against the maximum number of species richness, S_{mobs} , defined as the total number of species observed for all the volumes of seawater examined (*i.e.* bottle + net samples) (See *Figure S1*). The maximum of S_{obs} and S_{mobs} are not estimates, but the real values of the maximum number of species observed at the highest level of sampling effort (bottle and bottle + net samples, respectively). While bottle samples add up to ~1 L of seawater, net samples meant filtering around ~25 L. Abundance and incidence-based non-parametric estimates are summarized in *Tables 2* and *3*, respectively, considering the following degrees of completeness: 5%, 10%, 30% and 50% of the total number of individuals counted in bottle samples. Abundance-based estimators were calculated using all the species frequency information they need. In the case of incidence-based estimators, the sample replicate, or subsamples, were converted to presence/absence matrices before calculating the estimates. For

coverage-based estimators (i.e. *ACE* and *ICE*), the cut-off value was 10. For comparison with the previous approach, *Tables S2* and *S3* contain a summary of how estimators performed for multiple field subsamples of equal volume. Non-parametric estimators were evaluated considering asymptotic estimates of the true species richness as well.

Bias, accuracy and precision of non-parametric estimators

The reliability of each non-parametric asymptotic estimator was assessed by quantifying its bias, precision and accuracy at different degrees of completeness (Walther and Moore, 2005). The performance of estimates was tested for the nine sampling datasets by comparing species richness estimates, $\hat{S}_{est}$, to the maximum of species richness observed, S_{mobs} . Bias was measured as the proportional deviation of $\hat{S}_{est}$ from S_{mobs} , or as a scaled mean error (SME):

$$SME = \frac{1}{n} \sum_{j=1}^n \left(\frac{S_{est(j)} - S_{mobs}}{S_{mobs}} \right)$$

where $j = 1 \dots n$ are one of $n = 100$ replicates. Unidirectionality in the estimations is reflected by positive and negative bias values, that means the presence of over- or sub-estimations of the species richness.

The absolute value of the coefficient of variation of species richness estimations among replicates gives a measure of precision. The coefficient of variation was calculated as the standard deviation of estimates for each replica divided by the mean of estimates, being:

$$CV = \frac{s}{\bar{S}_{est}}$$

and where s is the corrected sample standard deviation of estimates:

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (S_{est_i} - \bar{S}_{est})^2}$$

CV informs about how much close species richness estimates are among themselves. So, the smaller the value, the higher the precision. CV values are independent of S_{mobs} and indicate the absence of random errors of the estimator.

Accuracy combines the concepts of bias and precision. It was measured as scaled mean square error (SMSE):

$$SMSE = \frac{1}{n} \sum_{j=1}^n \left(\frac{S_{est(j)} - S_{mobs}}{S_{mobs}} \right)^2$$

This quantifies how close all estimates are on average to the maximum value of species richness, S_{mobs} . Estimators that have smaller bias and higher precision show reduced SMSE values and they are more accurate than others. All these measures were used to test the inference capacity of the non-parametric asymptotic estimators. In order to compare them, we defined that a good estimator has to be unbiased and precise so it can be considered accurate. Following Walther and Moore (2005), these results can be compared with previous analysis.

Rarefaction and extrapolation, and sample coverage

We compared the species richness of microplankton samples using an abundance sample-size-based rarefaction and extrapolation (R/E) method (See Colwell et al. 2012; Chao and Jost, 2012; and Chao and Chiu 2016 for a review). Samples were standardised to a reference sample, either rarefying or extrapolating them so a sampling curve is drawn. This method extrapolates using the Chao1 estimator. Consider that we focused only on species richness, understood as the Hill number of order $q = 0$. We assessed the sample completeness or coverage, of microplankton as well. In our case the total abundances of species of around ten replicate, that were collected with seawater samples of the same volume and at the same day, were used as the samples to be compared (i.e. species frequencies in ten replicas of 5 mL in July and so on). We used the maximum number of individuals present in those samples (which was the total number of individuals of 14th May samples) as the standard sample size for creating the R/E curves. We estimated the 95% confidence intervals of the R/E curves. In *Figure S19* the entire abundance-based sample of abundances is

used to estimate the R/E curve. In *Figures S17* and *S18* these individuals-base samples were considered on their own and R/E curves drawn, extrapolating them to double.

SACs, non-parametric asymptotic species richness estimations, R/E curves and performance tests were carried out using the open-source programming language *R* (R Core Team, 2016) and with *vegan* (Oksanen *et al.*, 2014), *ChaoSpecies* (Tseng *et al.*, 2013) and *iNEXT* (Hsieh *et al.*, 2016) packages.

Results

Hydrographic conditions

The Ría de Vigo behaves as a partially mixed estuary with a two-layered circulation pattern that varies from winter mixing to summer upwelling and stratification. Phytoplankton blooms typically occur in the transition phases: the mixing-stratification periods give rise to spring blooms whereas the end of the stratification season typically culminates with autumn blooms (Crespo *et al.*, 2006). During 2012, water column vertical mixing was observed in February, October and November, and phytoplankton communities were characterized by relatively low biomass. The summer upwelling-stratification period occurred from March to September. During this time, intermittent intrusions of nutrient-rich deep waters fertilized the surface. Chlorophyll-*a* concentrations showed subsurface maxima at 15-20 m depth.

Performance of non-parametric asymptotic estimators

When all cruises were considered, SME values showed that *Jack2* was the least negative biased abundance-based estimator (See *Table 2*). Estimates of *ACE1* ranked as the second least biased. Indeed, *ACE1* outperformed *Jack2* in some cases at a completeness lower than 10%. In the case of incidence-based estimators, *Jack2* repeated as the most unbiased, whereas *ICE1* and *Chao2* were the second least biased (See *Table 3*). Incidence-based estimators were more biased than the abundance-based. Regardless of whether estimates were based on abundance or incidence of species, estimators still underestimated more than 0.1 at a completeness of 30%. At intermediate levels of sampling completeness the differences in SME were negligible

and consistent across all the seasons. Negative biases were more marked in samples collected during summer upwelling.

The most precise non-parametric estimators were the abundance and incidence-based *HM*. Their estimates had a mean CV of ~0.45 at 10 % of completeness. *Jack1* consistently occupied the second position. Both estimators were followed by *ACE* and *ICE*, which were always above 0.5 of precision at 50% of completeness. There was an exception on 6th November, when incidence-based *HM* and *Jack1* were nearly 0. *Chao1-bc* and *Chao1* were the most imprecise estimators (See *Table 2* and *3*).

According to SMSE values, the *HM* consistently underestimated species richness and can be considered the most inaccurate. In terms of accuracy, the best estimator was *Jack2* followed by *Jack1*, *ACE* and *ICE*. SMSE values were always above 0.7 at 50 % of completeness. Considering all datasets, even the SMSE of abundance- and incidence-based *Jack2* estimates were 1.05 and 1.15, respectively.

SACs and non-parametric asymptotic estimations

Predictions tended to converge towards the maximum of S_{obs} rather than S_{mobs} on 14th February, 26th July, 12th September, and 10th October (see *Figs. 1* and *2*). In other cases, $\hat{S}_{est}$ did not converge to any species richness value. As samples completeness increased, $\hat{S}_{est}$ increased steadily: the higher the degree of completeness, the larger estimates. This tendency was independent of the species richness of the community analysed (See *Fig. 3*). There were two cases in which non-parametric estimators showed positive bias at high sample completeness (i.e. $\hat{S}_{est}$ exceeded the maximum number of species observed in bottle + net samples of the community, S_{mobs}). These two cases were 14th May and 6th November. Abundance-based *Chao1* and *Jack2*, and incidence-based *Chao2* and *Jack2* showed positive bias on 14th May. *ACE*, *ACE1*, *Chao1*, *Chao1-bc*, *Jack1* and *Jack2* (abundance-based estimations), and *Chao2*, *Chao2-bc*, *ICE*, *ICE1* and *Jack2* (incidence-based estimations) showed positive bias on 6th November.

Rarefaction and extrapolation curves and sample coverage

Figure 4 and *Table 4* show how rarefaction and extrapolation differentiated our samples by species richness using the 14th May sampling day as an endpoint standard. Our results indicated that on 14th May, 26th July, 10th October and 6th November microplankton richness estimates exceeded a value of 67 species for that extrapolation point, whereas on 30th March the number of species was around 36 species, being the poorest. For the 14th May sample, the R/E curve needed more sampling effort to achieve the same saturation level as other species-rich samples such as those collected on 26th July, 10th October and 6th November. On 14th February and 14th May, the samples showed estimates with standard errors over 20%. On 26th of July, 10th of October and 6th November our S_{mobs} estimates were outside the 95% confidence intervals of the R/E curves. In every single case sample coverage was close to or equal to 1 (See *Table 4*).

Discussion

A non-parametric asymptotic methodology limited to reduce bias

Our analysis suggests that the abundance-based *Jack2* was the best of non-parametric asymptotic species richness estimators. This Jackknife class of species richness estimators was introduced by Burham and Overton to reduce bias in homogeneous communities with equal probability of species detection among sample replicates (Burnham and Overton, 1978, 1979). Jackknife estimators of order j were suggested to reduce bias in a magnitude of $1/n^j$ for a population of n individuals (Quenouilli, 1956; Miller, 1974; Colwell and Coddington, 1994). If *Jack1* is expected to reduce bias $1/n$ times, *Jack2* is expected to make it $1/n^2$ times more, and so on. In our study, *Jack2* proved to be less biased than *Jack1*. We could expect this tendency to be consistent for Jackknife estimators of higher order.

No estimator bias was diminished with increasing completeness. Contrary to expectations, non-parametric asymptotic estimators that had been designed for species-rich communities did not prevail as less biased than *Jack2*. Not even *Chao1-bc* and *Chao2-bc*, which were specifically defined as lower bound estimators to solve this limitation, performed better than Jackknife estimators (Smith and Belle, 1984; Chao, 2005).

HM reduced significantly the variance among subsamples, providing consistently the most precise estimates. Even though the degree of deviation of *HM* estimates were still considerable (mean standard deviation = 2.6). This behaviour might have resulted from variations in detection probability of species, perhaps associated with the spatial heterogeneity in the distribution of these microplankton populations (Rodríguez-Ramos *et al.*, 2014), though it could be because of rarefying the community for obtaining subsamples (See Methods and Results section and *Tables S2-S5* for comparison). The precision of *HM* was penalized with a higher bias.

Low sampling completeness led to severe underestimation of species richness. Previous analysis had suggested that long-tailed rank-abundance distributions would push bias considerably into negative values (Wagner *et al.*, 2002). Accordingly, most of our estimates were closer to S_{obs} rather than S_{mobs} (Cermeño *et al.*, 2014), independently of the degree of completeness. Even when the Jackknife estimators were less biased than the others, they were still far away from S_{mobs} . Therefore, when abundance-based and incidence-based methods are compared, there is no substantial difference among them since bias was considerable for both. As sample completeness increased, estimates reduced their differences with S_{mobs} .

Community structure and non-parametric estimators

Non-parametric asymptotic estimators were affected by community structure. Taking into account that comparisons among estimates were based on the degree of completeness (with S_{mobs} as a standard) rather than a standardized sample size, our results do not seem to depend on field sampling disparities but on differences in community structure. Indeed, the accuracy of estimators for samples collected in different seasons were quite different. Hence, if estimates for communities collected in different seasons are to be compared, then, the non-parametric asymptotic estimators must be taken with caution.

Community attributes like spatial aggregation or uneven abundance distributions have been previously reported to alter the predictive capacity of these non-parametric estimators (Wagner and Wildi, 2002; Brose *et al.*, 2003; Reese *et al.*, 2014). Likewise, communities generated randomly, i.e. from log-series or log-normal SADs,

have been also used to assess the performance of estimators (Brose *et al.*, 2003; Chao *et al.* 2014, Reese *et al.*, 2014). Model simulations showed that these non-parametric estimators are more accurate when the species in the community are more similar in abundances (i.e., even species-abundance distributions) (Brose *et al.*, 2003; Reese *et al.*, 2014). Consistent with this idea, our analysis shows that uneven communities, with a larger tail of rare species, led to more biased and imprecise estimates of species richness. The magnitude of the estimator's bias depended on the species distribution, which varied throughout the year. Bias was reduced consistently with increasing sample completeness and some estimates became close to or even exceeded S_{mobs} .

Comparing samples using R/E curves

The rarefaction/extrapolation (R/E) method was applied to the microplankton community taking as a standard the sample with the highest abundance. R/E curves ranked samples according to their species richness (See *Fig. 4*). Microplankton abundance changed significantly over time. That meant that most of the samples had to be extrapolated sizes going from two to ten times their number of individuals. If we pay attention to the standard sample, confidence intervals became broad, making difficult to discriminate by species-richness at that given sample coverage. Nonetheless, samples were ranked more unambiguously at lower coverage levels (e.g. twice or triple the number of individuals of 14th February). Sample coverage estimates were in every case equal or nearly equal to one. This indicates that most of the abundant (dominant) species were present in the analysed samples. Consistent with abundance and incidence-based species accumulation curves (See *Figure S1*), most of the species lost were in the category of rare.

The long tail of rare species meaning

Non-parametric asymptotic estimators rely on the frequency of rare species to infer the number of undetected species. These methods assume that abundant species lack any information about how many others are missed, but that rare species do so (Good, 1953; Chao 2005). Encountering a previously undetected species is expected to have a probability around f_i/n , where n is the total number of individuals; and that sample coverage information is contained in $1 - f_i/N$, where N is the total sum of individuals

(Good, 1953, 2000). Non-parametric asymptotic estimators are grounded on the assumption that populations under study are closed (immigration is zero) and that rare species are legitimate moment of the true species richness (Gotelli and Colwell, 2011). Nevertheless, the pool of rare species constitutes a changing but persistent component of microbial communities in planktonic habitats (Pedrós-Alió, 2006). Environmental variability, dispersal and species interactions lead to large fluctuations in species abundances (Finlay and Clarke, 1999; Finlay 2002; Shimadzu *et al.*, 2013; Logares *et al.*, 2014). To act as a moment, Chao1, for example, requires $f_2 > f_1^2$ and $Nf_1 > f_2$. However, if the number of singletons is remarkable, the first condition mentioned above would be false and the underlying assumptions of the estimator violated (Chao, 1984).

Good (1953) stated that the lowest possible value is the best we can expect from this inference technique when applied to communities dominated by rare species. Indeed, this is what we have obtained from this test with real data. The complexity of estimating species richness from “finite samples” motivated the development of R/E methods (Chao *et al.* 2012). Nevertheless, although some authors highlighted the complexity of elaborating an unbiased but informative non-parametric asymptotic estimator, they are still trying to implement non-parametric lower bound estimators with two-side confidence intervals (Mao and Lindsay, 2007). Others have proposed to increase the frequencies of rare species to $f_i = 4$, making it suitable for highly diverse communities (Chiu *et al.*, 2014). These arguments entail that more information about the underlying community SAD is needed to estimate with precision their species richness.

Heageman *et al.* (2013) have suggested that if there were a large proportion of rare species, the uncertainty of estimate would increase and it would be, in practice, impossible to estimate microbial species richness without considering additional assumptions about the underlying SAD. In a similar way, Pueyo (2003) highlighted that in cases where the SAD is bordering a power law, which in practice is a long-tailed SAD, any lower bound estimation would be far from the true value. He argues that any precise estimation would be unattainable by non-parametric asymptotic estimators as they can only deal with small bias reduction. Bung (2009) summarized

that bias was likely to be significant for high diverse communities too. Our observation that, in the best case, $\hat{S}_{est}$ fitted the maximum S_{obs} rather than S_{mobs} , and the magnitude of bias, support their conclusions. We have also observed that a higher precision of non-parametric asymptotic estimators could have been outweighed by a loss of capacity to reduce bias.

This led some authors to use, first, rarefaction methods for interpolation and, second, estimators like Chao1 for extrapolation. They intend to avoid false comparisons of samples at different degrees of completeness. In our case R/E curves distinguished which communities were species richer when considering a reference sample. They allowed us to sort out samples as a function of their species composition without missing sampling effort. The main pitfall was the amplitude of standard error extrapolating to that standard. We wonder whether these kinds of comparisons are applicable to microplankton time series, provided that these microbial communities exhibit relatively rapid changes in population densities. When designing an expedition, we need to decide whether to prioritize sampling effort or, for instance, measure some other environmental variables. In every case, sample coverage estimations were close to one, suggesting that the most abundant species were collected even at low sampling efforts. Most of the extra species collected with net samples might be rare, and they might be present permanently in these communities.

Our analysis shows that the assumptions of non-parametric asymptotic estimators are inadequate for microplankton communities. Even though some of them were proposed as lower bound estimators, a simple reduction of the number of species missed does not justify its applicability in our field of research. We have seen that the extent of bias depends on temporal and spatial changes in community structure. R/E solves that pitfall when comparing multiple communities. Nonetheless, this method does not necessarily improve our understanding about the origin of microbial plankton diversity. In addition to this, evaluating seasonal or geographical patterns of biodiversity still represents a challenge. The use of non-parametric asymptotic estimators might potentially lead us to misleading conclusions if they were not complemented with other analytical procedures. Analytical methods based on parametric approaches seem to emerge as an appropriate tool in this regard. Indeed

they have been suggested to perform better than coverage-based non-parametric asymptotic estimators (as for e.g. Jeon, 2006). In this line, multiple functions were proposed to fit the SADs, going from the power law to the log-series and the log-normal. These functions have been developed from pure statistical approximations or based on either population, niche or neutral models (i.e. Fisher, 1943; Caswell, 1976; Hubbell, 2001; Pueyo 2007). Chiu *et al.* 2014 or Chao *et al.* 2016 have recognised that accuracy could be a weakness of these methods, but their properties of scalability to any sample size, makes them interesting. These methods allow reducing sampling effort and, at the same time, could provide valuable information regarding the processes underlying the structure of microplankton communities.

Conclusions

Jackknife non-parametric estimators performed better than Chao's estimators for individual- and species-rich communities of marine microplankton. Nevertheless, all of them underestimated species richness and showed high imprecision despite that the community was sampled intensively. These analytical methods use the frequency of low abundance (rare) species as an index of sample completeness. This premise might be easily broken because marine microplankton communities are characterized by high levels of immigration and an enormous variability of population densities. New methods such as rarefaction and extrapolation compare communities adequately as they correct differences due to sample coverage or completeness. These methods are useful to compare diversity estimates from microplankton samples collected near in space and time, but their utility is reduced for the analysis of long-time series or across broad spatial scales.

Acknowledgements

We thank Isabel G. Teixeira for microscopy analyses. M.B. was supported by a FPI grant from the Spanish Government. P.C. was supported by a Ramon y Cajal contract from the Spanish Government. We thank three anonymous reviewers for their valuable comments and suggestions.

Funding

This research was supported by grant CTM2011-25035 from the Spanish Government (SG).

Data archiving

Data used for this research project was deposited in Figshare. It can be found in: “Branco, M., Cermeño, P., Figueiras, F.G., Marañón, E., Teixeira, I.G. (2017) Microplankton species collected in the Ría de Vigo (Galicia, NW Iberian Peninsula) in 2012 (Project DISTRAL). *figshare*. <https://dx.doi.org/10.6084/m9.figshare.4555456> “

References

Alroy, J. (2010) The shifting balance of diversity among major marine animal groups’, *Science*, **329** (5996), 1191–1194.

Amaral-Zettler, L., Artigas, L. F., Baross, J., Bharathi P.A., L., Boetius, A., Chandramohan, D., Herndl, G., et al. (2010) A global census of marine microbes. In life in the World’s oceans, 221–245. Wiley-Blackwell.

Arrhenius, O. (1921) Species and Area. *J. Ecol.*, **9**, 95–99.

Azam, F., Fenchel, T., Field, J., Gray, J., Meyer-Reil, L., and Thingstad, F. (1983) The Ecological Role of Water-Column Microbes in the Sea. *Mar. Ecol: Prog. Ser.*, **10**, 257–263.

Branco, M., Cermeño, P., Figueiras, F.G., Marañón, E., Teixeira, I.G. (2017) Microplankton species collected in the Ría de Vigo (Galicia, NW Iberian Peninsula) in 2012 (Project DISTRAL). *figshare*. <https://dx.doi.org/10.6084/m9.figshare.4555456>

Boulinier, T., Nichols, J. D., Sauer, J. R., Hines, J. E., and Pollock, K. H. (1998) Estimating species richness: The importance of heterogeneity in species detectability. *Ecology*, **79**, 1018–1028.

Brose, U., Martinez, N. D., and Williams, R. J. (2003) Estimating species richness: sensitivity to sample coverage and insensitivity to spatial pattern. *Ecology*, **84**, 2364–2377.

Bunge, J. (2009) Statistical Estimation of Uncultivated Microbial Diversity. In *Uncultivated Microorganisms*, 160–178. Ed. by S. S. Epstein. Springer Berlin Heidelberg. ISBN: 9783540854654

Bunge, J., and Fitzpatrick, M. (1993) Estimating the Number of Species: A Review. *J. Am. Stat. Assoc.*, **88**, 364–373.

Burnham, K. P., and Overton, W. S. (1978) Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, **65**, 625–633.

Burnham, K. P., and Overton, W. S. (1979) Robust Estimation of Population Size When Capture Probabilities Vary Among Animals. *Ecology*, **60**, 927–936.

Caswell, H. (1976) Community Structure: A Neutral Model Analysis. *Ecol. Monogr.*, **46**, 327–354.

Cermeño, P., Teixeira, I. G., Branco, M., Figueiras, F. G., and Marañón, E. (2014) Sampling the limits of species richness in marine phytoplankton communities. *J. Plankton Res.*, **36** (4): 1135–1139.

Chao, A. (1984) Nonparametric Estimation of the Number of Classes in a Population. *Scan J. Statist.*, **11**, 265–270.

Chao, A. (1987) Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics*, **43**, 783–791.

Chao, A., and Lee, S.M. (1992) Estimating the Number of Classes via Sample Coverage. *J. Am. Stat. Assoc.*, **87**, 210–217.

Chao, A. (2005) Species estimation and applications. *Encyclopedia of statistical sciences*. 2nd edn. (eds S. Kotz, N. Balakrishnan, C.B. Read & B. Vidakovic) **12**, 7907–7916. Wiley Press, New York.

- Chao, A., and Jost, L. (2012) Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, **93**, 2533–2547.
- Chao, A., Chiu, C.H., and Jost, L. (2014) Unifying Species Diversity, Phylogenetic Diversity, Functional Diversity, and Related Similarity and Differentiation Measures Through Hill Numbers. *Annu. Rev. Ecol. Syst.*, **45**, 297–324.
- Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R.K. and Ellison, A.M. (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecol. Monogr.*, **84**, 45–67.
- Chao, A., and Chiu, C.H. (2016) Nonparametric Estimation and Comparison of Species Richness. *eLS*. John Wiley & Sons, Ltd. Chichester.
- Chazdon, R. L., Colwell, R. K., Denslow, J. S., and Guariguata, M. R. (1998) Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of Northeastern Costa Rica. In: Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies (Dallmeier, F., Comiskey, J.A.). *Man and the Biosphere Series*, **20**, 285–309.
- Chiu, C.H., Wang, Y.T., Walther, B. A., and Chao, A. (2014) An improved nonparametric lower bound of species richness via a modified good–turing frequency formula. *Biometrics*, **70**, 671–682.
- 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. *J. Plant Ecol.*, **5**, 3–21.
- Colwell, R. K., and Coddington, J. A. (1994) Estimating Terrestrial Biodiversity through Extrapolation. *Phil. Trans. R. Soc. Lond. B*, **345**, 101–118.
- Colwell, R. K., Mao, C. X., and Chang, J. (2004) Interpolating, extrapolating, and

comparing incidence-based species accumulation curve. *Ecology*, **85**, 2717–2727.

Crespo, B. G., Figueiras, F. G., Porras, P., and Teixeira, I. G. (2006) Downwelling and dominance of autochthonous dinoflagellates in the NW Iberian margin: The example of the Ría de Vigo. *Harmful Algae*, **5**, 770–781.

Finlay, B. J. (2002) Global Dispersal of Free-Living Microbial Eukaryote Species. *Science*, **296**, 1061–1063.

Finlay, B. J., and Clarke, K. J. (1999) Ubiquitous dispersal of microbial species. *Nature*, **400**, 828–828.

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943) The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *J. Anim. Ecol.*, **12**, 42–58.

Good, I. J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Good, I. J. (2000) Turing’s anticipation of empirical bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation*, **66**, 101–111.

Gotelli, N. J., and Colwell, R. K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, **4**, 379–391.

Gotelli, N.J. & Colwell, R.K. (2011) Estimating species richness. *Frontiers in measuring biodiversity* (ed. by A.E. Magurran and B.J. McGill), pp. 39–54. Oxford University Press, New York.

Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., and Weitz, J. S. (2013) Robust estimation of microbial diversity in theory and in practice. *ISME J.*, **7**, 1092–

1101.

Hsieh TC, Ma KH and Chao A (2016) iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods Ecol Evol*, 7: 1451–1456. doi:10.1111/2041-210X.12613

Hubbell, S. P. (2001) The unified neutral theory of biodiversity and biogeography. Monographs in population biology, 32. Princeton University Press, Princeton.

Hughes, J. B., Hellmann, J. J., Ricketts, T. H., and Bohannan, B. J. M. (2001) Counting the Uncountable: Statistical Approaches to Estimating Microbial Diversity. *Appl. Environ. Microbiol.*, **67**, 4399–4406.

Hutchinson, G. E. (1961) The Paradox of the Plankton. *Am. Nat.*, **95**, 137–145.

Jeon, S.O., Bunge, J., Stoeck, T., Barger, K.J.A., Hong, S.H., and Epstein, S.S. (2006) Synthetic Statistical Approach Reveals a High Degree of Richness of Microbial Eukaryotes in an Anoxic Water Column. *Appl. Environ. Microbiol.*, **72**, 6578–6583.

Jiao, N., Herndl, G. J., Hansell, D. A., Benner, R., Kattner, G., Wilhelm, S. W., Kirchman, D. L., et al. (2010) Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nat. Rev. Micro.*, **8**, 593–599.

Jost, L. (2010) The Relation between Evenness and Diversity. *Diversity*, **2**, 207–232.

Lee, S.M., and Chao, A. (1994) Estimating Population Size Via Sample Coverage for Closed Capture-Recapture Models. *Biometrics*, **50**, 88–97.

Logares, R., Audic, S., Bass, D., Bittner, L., Boutte, C., Christen, R., Claverie, J.M., et al. (2014) Patterns of Rare and Abundant Marine Microbial Eukaryotes. *Curr. Biol.*, **24**, 813–821.

- Lund, J. W. G., Kipling, C., and Cren, E. D. L. (1958) The inverted microscope method of estimating algal numbers and the statistical basis of estimations by counting. *Hydrobiologia*, **11**, 143–170.
- Mao, C. X., and Colwell, R. K. (2005) Estimation of species richness, mixture models, the role of rare species, and inferential challenges. *Ecology*, **86**, 1143–1153.
- Mao, C. X., and Lindsay, B. G. (2007) Estimating the number of classes. *Annu. Stat.*, **35**, 917–930.
- Massana, R. (2011) Eukaryotic Picoplankton in Surface Oceans. *Annu. Rev. Microbiol.*, **65**, 91–110.
- May, R. M. (1972) Will a Large Complex System be Stable? *Nature*, **238**, 413–414.
- Miller, R. G. (1974) The jackknife - a review. *Biometrika*, **61**, 1–15.
- Nufio, C. R., McClenahan, J. L., and Thurston, E. G. (2009) Determining the effects of habitat fragment area on grasshopper species density and richness: a comparison of proportional and uniform sampling methods. *Insect Conserv.*, **2**, 295–304.
- Odum, E. P. (1954) Fundamentals of ecology. Philadelphia: W. B. Saunders Company, 1953. 383 P. *Science Education*, **38**, 314–314.
- Oksanen, J., Blanchet, F. G., Kindt, R., Legendre, P., Minchin, P. R., O’Hara, R. B., Simpson, G. L., et al. (2017) vegan: Community Ecology Package. <http://CRAN.R-project.org/package=vegan>.
- Pedrós-Alió, C. (2006) Marine microbial diversity: can it be determined? *Trends Microbiol.*, **14**, 257–263.
- Pueyo, S. (2003). Irreversibility and Criticality in the Biosphere. PhD Thesis. Universitat de Barcelona, Barcelona. ISBN: 8468830321.

779 <http://www.tdx.cat/handle/10803/1421>.

780

781 Pueyo, S., He, F., and Zillio, T. (2007) The maximum entropy formalism and the

782 idiosyncratic theory of biodiversity. *Ecol. Lett.*, **10**, 1017–1028.

783

784 R Core Team. (2016) R: A Language and Environment for Statistical Computing. R

785 Foundation for Statistical Computing, Vienna, Austria. <http://www.R-project.org/>.

786

787 Reese, G. C., Wilson, K. R., and Flather, C. H. (2014) Performance of species

788 richness estimators across assemblage types and survey parameters. *Global Ecol.*

789 *Biogeogr.*, **23**, 585–594.

790

791 Rodríguez-Ramos, T., Dornelas, M., Marañón, E., and Cermeño, P. (2014)

792 Conventional sampling methods severely underestimate phytoplankton species

793 richness. *J. Plankton Res.*, **36** (2), 334–343.

794

795 Ryther, J. H. 1969. Photosynthesis and Fish Production in the Sea. *Science*, **166**, 72–

796 76.

797

798 Shimadzu, H., Dornelas, M., Henderson, P., and Magurran, A. (2013) Diversity is

799 maintained by seasonal variation in species abundance. *BMC Biology*, **11**, 98.

800

801 Smith, E. P., and van Belle, G. (1984) Nonparametric Estimation of Species

802 Richness. *Biometrics*, **40**, 119–129.

803

804 Sournia, A. (1978) Phytoplankton manual. Monographs on Oceanographic

805 Methodology, 6. Unesco, Paris. ISBN 13: 9789231015724

806

807 Tseng, K., Lee, Y.H., and Chao, A. (2013) ChaoSpecies: Statistical package

808 “ChaoSpecies”. R package version 1.0. <http://chao.stat.nthu.edu.tw/blog/>.

809

810 Turing, A. M. (1939) Systems of Logic Based on Ordinals. *Proc. London Math Soc.*,

811 **s2-45**, 161–228.

Ugland, K. I., Gray, J. S., and Ellingsen, K. E. (2003) The species–accumulation curve and estimation of species richness. *J. Anim. Ecol.*, **72**, 888–897.

Utermöhl, H. (1958) Zur Vervollkommnung der quantitativen Phytoplankton-Methodik. Mitt. Internationale Ver. Theoretische und Angewandte Limnologie, **9**, 1–38.

Wagner, H. H., and Wildi, O. (2002) Realistic simulation of the effects of abundance distribution and spatial heterogeneity on non-parametric estimators of species richness. *Ecoscience*, **9**, 241–250.

Walther, B. A., and Moore, J. L. (2005) The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, **28**, 815–829.

Table and Figure legends

Table 1. List of non-parametric asymptotic estimators evaluated. #Estimators calculated either with their abundance and incidence-based approaches.

Table 2. Summary of abundance-based non-parametric asymptotic estimator's performance: bias, precision and accuracy results (SME, CV and SRMSE, respectively). Different sampling coverage (*Cmpl.*) values, from 5 % to 50 % of the total number of individuals in the community, are listed for each sample cruise data. The best three non-parametric estimators are shown in a decreasing order, from left to the right, considering their degree of completeness. The superscripts indicate: (a) *ACE*, (b) *ACE1*, (c) *Chao1*, (d) *Chao1-bc*, (e) *HM*, (f) *Jack1* and (g) *Jack2*.

Table 3. As in Table 2 but for incidence-based non-parametric asymptotic estimators. The superscripts indicate: (a) *Chao2*, (b) *Chao2-bc*, (c) *HM*, (d) *ICE*, (e) *ICE1*, (f) *Jack1* and (g) *Jack2*.

Table 4. Summary of individual-based interpolation with rarefaction and extrapolation. Samples are sorted by date. n is the number of individuals present in samples, S_{obs} is the total number of species detected in the field using bottle samples only. f_1 and f_2 are the number of singletons and doubletons, respectively. Estimator refers to the species-richness that R/E estimated for a given number of individuals. SE refers to the sample error and 95 % lower and 5% upper refer to the species-richness estimated at that confidence levels. S_{mobs} is the maximum number of species observed in field using bottle and net samples.

Figure 1. Abundance-based non-parametric estimators of species richness.

Number of individuals counted plotted against observed species richness (solid line) and estimates of species richness from abundance-based non-parametric asymptotic methods (dashed lines). The estimations plotted are abundance-based *HM*, *Jack1* and *Jack2* plus *ACE*, *ACE1*, *Chao1* and *Chao1-bc*. The maximum number of species observed (S_{mobs}) obtained from the analysis of bottle and net samples under the microscope (see also Cermeño *et al.*, 2014) is also shown (horizontal blue short solid line).

Figure 2. As in Figure 1 but for incidence-based non-parametric estimators.

These non-parametric estimates were calculated using microplankton incidence data, which contain information of species presence or absence. The estimations plotted are incidence-based *HM*, *Jack1*, *Jack2* and *Chao2*, *Chao2-bc*, *ICE* and *ICE1*.

Figure 3. Correlation analysis between non-parametric estimates of species richness ($\hat{S}_{est}$) and the maximum of species richness (S_{mobs}). Both abundance and incidence-based non-parametric asymptotic estimates are represented against the maximum number of species observed, S_{mobs} . $\hat{S}_{est}$ obtained for the 100% of sampling completeness were plotted. S_{mobs} is the total number of species detected in >20 L sample volumes (bottle + net samples) (see also Cermeño *et al.*, 2014). S_{mobs} provides a reference to test estimators at different degree of completeness or sampling effort. Symbols denote different sampling cruises. Considering S_{mobs} , values above the diagonal solid line are overestimates.

Figure 4. Representation of the rarefaction and extrapolation for all microplankton samples. Extrapolations were calculated using as reference (maximum) the number of individuals present in the sample collected on 14th May. The 95% confidence intervals are depicted. Solid lines represent rarefaction and dashed lines represent extrapolation.

Table 1

NPR estimator	Abbreviation	Reference
1st order Jackknife [#]	<i>Jack1</i>	Burnham and Overton 1978
2nd order Jackknife [#]	<i>Jack2</i>	Burnham and Overton 1978
Chao1	<i>Chao1</i>	Chao 1984
Chao1-bc	<i>Chao1-bc</i>	Chao 2005
Chao2	<i>Chao2</i>	Chao 1984
Chao2-bc	<i>Chao2-bc</i>	Chao 2005
Abundance-based coverage	<i>ACE</i>	<i>Eq. (2.14)</i> in Chao and Lee 1992
Abundance-based coverage	<i>ACE1</i>	Chao and Lee 1992
Incidence-based coverage or Model ^(h)	<i>ICE</i>	<i>Eq. (3.18)</i> of Lee and Chao 1994
Incidence-based coverage or Model ^(th)	<i>ICE1</i>	Lee and Chao 1994
Homogenous Model [#]	<i>HM</i>	<i>Eq. (2.3)</i> of Chao and Lee 1992

Table 2

Date	Cmpl.	sp. SD	SME			CV			SRMSE		
			<i>1st</i>	<i>2nd</i>	<i>3th</i>	<i>1st</i>	<i>2nd</i>	<i>3th</i>	<i>1st</i>	<i>2nd</i>	<i>3th</i>
14th February	5	2.76	-0.14 ^b	-0.24 ^g	-0.25 ^c	2.39 ^g	2.56 ^b	2.64 ^a	0.73 ^f	0.83 ^e	1.15 ^g
	10	2.51	-0.14 ^g	-0.16 ^b	-0.19 ^c	1.57 ^g	1.82 ^b	1.98 ^f	0.48 ^e	0.58 ^f	0.80 ^a
	30	2.00	-0.03 ^g	-0.05 ^c	-0.08 ^b	0.80 ^g	0.87 ^f	1.05 ^b	0.32 ^e	0.46 ^f	0.53 ^a
	50	1.58	-0.06 ^g	-0.07 ^f	-0.09 ^b	0.74 ^f	0.78 ^g	0.93 ^b	0.23 ^e	0.35 ^f	0.35 ^a
30th March	5	2.40	-0.32 ^b	-0.33 ^g	-0.35 ^c	3.29 ^g	3.35 ^b	3.67 ^a	0.59 ^e	0.66 ^f	1.06 ^g
	10	2.36	-0.30 ^g	-0.31 ^c	-0.34 ^f	2.91 ^g	3.26 ^f	3.27 ^c	0.43 ^e	0.56 ^f	0.62 ^a
	30	1.96	-0.15 ^c	-0.16 ^g	-0.19 ^d	1.63 ^g	1.97 ^c	2.10 ^f	0.34 ^e	0.45 ^f	0.62 ^a
	50	2.19	-0.07 ^b	-0.09 ^g	-0.11 ^c	1.13 ^g	1.23 ^b	1.41 ^a	0.37 ^e	0.49 ^f	0.71 ^a
14th May	5	2.58	-0.25 ^b	-0.28 ^c	-0.28 ^g	2.70 ^g	2.80 ^b	3.02 ^a	0.52 ^e	0.62 ^f	0.93 ^g
	10	2.43	-0.20 ^c	-0.21 ^g	-0.23 ^b	2.10 ^g	2.40 ^b	2.45 ^f	0.47 ^e	0.58 ^f	0.81 ^a
	30	1.95	-0.12 ^c	-0.12 ^g	-0.13 ^b	1.33 ^g	1.48 ^b	1.54 ^f	0.34 ^e	0.46 ^f	0.64 ^a
	50	1.98	-0.08 ^g	-0.09 ^c	-0.11 ^b	1.00 ^g	1.14 ^f	1.21 ^b	0.35 ^e	0.46 ^f	0.55 ^a
12th June	5	2.87	-0.27 ^c	-0.27 ^b	-0.27 ^g	2.72 ^g	2.83 ^b	3.20 ^c	0.57 ^e	0.69 ^f	0.99 ^a
	10	2.57	-0.20 ^g	-0.22 ^b	-0.22 ^c	2.11 ^g	2.36 ^b	2.51 ^f	0.47 ^e	0.59 ^f	0.80 ^a
	30	2.27	-0.10 ^g	-0.13 ^c	-0.14 ^b	1.28 ^g	1.49 ^f	1.55 ^b	0.37 ^e	0.51 ^f	0.65 ^a
	50	1.77	-0.04 ^g	-0.07 ^b	-0.07 ^c	0.77 ^g	0.87 ^f	0.92 ^b	0.28 ^e	0.38 ^f	0.49 ^a
11th July	5	2.69	-0.35 ^b	-0.37 ^g	-0.38 ^c	3.85 ^b	3.96 ^g	4.25 ^c	0.52 ^e	0.60 ^f	0.92 ^g
	10	2.54	-0.31 ^g	-0.31 ^b	-0.34 ^c	3.31 ^g	3.40 ^b	3.67 ^c	0.46 ^e	0.57 ^f	0.81 ^a
	30	2.29	-0.21 ^g	-0.23 ^b	-0.26 ^c	2.32 ^g	2.66 ^b	2.71 ^c	0.36 ^e	0.50 ^f	0.56 ^a
	50	1.86	-0.17 ^g	-0.19 ^c	-0.20 ^b	1.87 ^g	2.19 ^b	2.22 ^f	0.28 ^e	0.39 ^f	0.48 ^a
26th July	5	3.34	-0.29 ^g	-0.31 ^b	-0.32 ^c	3.33 ^g	3.62 ^b	3.77 ^c	0.60 ^e	0.75 ^f	1.01 ^a
	10	2.90	-0.24 ^g	-0.27 ^c	-0.29 ^f	2.78 ^g	3.21 ^f	3.23 ^c	0.43 ^e	0.58 ^f	0.66 ^a
	30	2.77	-0.14 ^g	-0.16 ^c	-0.18 ^f	1.72 ^g	2.08 ^f	2.12 ^c	0.38 ^e	0.53 ^f	0.61 ^a
	50	1.90	-0.11 ^g	-0.14 ^c	-0.15 ^f	1.36 ^g	1.66 ^f	1.72 ^b	0.27 ^e	0.38 ^f	0.42 ^a
12th September	5	2.83	-0.24 ^g	-0.25 ^b	-0.25 ^c	2.54 ^g	2.71 ^b	2.89 ^c	0.48 ^e	0.59 ^f	0.82 ^a
	10	2.87	-0.17 ^g	-0.17 ^b	-0.20 ^c	1.95 ^g	2.07 ^b	2.35 ^f	0.50 ^e	0.63 ^f	0.87 ^a
	30	2.38	-0.07 ^g	-0.09 ^b	-0.10 ^c	1.04 ^g	1.15 ^b	1.24 ^f	0.36 ^e	0.48 ^f	0.60 ^a
	50	1.93	-0.05 ^g	-0.07 ^f	-0.08 ^c	0.83 ^g	0.83 ^f	1.00 ^b	0.27 ^e	0.40 ^f	0.43 ^a
10th October	5	2.92	-0.26 ^g	-0.28 ^b	-0.29 ^c	2.75 ^g	3.03 ^b	3.12 ^f	0.51 ^e	0.66 ^f	0.84 ^a
	10	2.57	-0.17 ^g	-0.19 ^c	-0.20 ^b	1.96 ^g	2.23 ^b	2.26 ^f	0.42 ^e	0.57 ^f	0.71 ^a
	30	2.12	-0.07 ^g	-0.11 ^c	-0.11 ^b	0.99 ^g	1.24 ^f	1.29 ^b	0.31 ^e	0.42 ^f	0.49 ^a
	50	1.89	-0.07 ^g	-0.08 ^b	-0.09 ^f	0.95 ^g	0.97 ^f	1.04 ^b	0.26 ^e	0.38 ^f	0.44 ^d
6th November	5	3.25	-0.19 ^b	-0.27 ^g	-0.27 ^c	3.00 ^g	3.14 ^b	3.33 ^a	0.64 ^e	0.71 ^f	1.10 ^g
	10	3.18	-0.19 ^g	-0.19 ^b	-0.24 ^a	2.18 ^g	2.35 ^b	2.62 ^f	0.54 ^e	0.64 ^f	0.90 ^d
	30	2.55	-0.10 ^g	-0.11 ^c	-0.15 ^b	1.34 ^g	1.65 ^f	1.91 ^b	0.37 ^e	0.54 ^f	0.57 ^a
	50	2.37	-0.03 ^g	-0.03 ^c	-0.08 ^d	0.75 ^g	1.09 ^f	1.24 ^d	0.32 ^e	0.46 ^f	0.48 ^a

Table 3

Date	Cmpl.	sp. SD	SME			CV			SRMSE		
			1st	2nd	3th	1st	2nd	3th	1st	2nd	3th
14 th February	5	2.76	-0.13 ^e	-0.25 ^d	-0.26 ^g	2.60 ^g	2.65 ^e	2.68 ^d	0.71 ^f	0.82 ^c	1.07 ^g
	10	2.51	-0.15 ^e	-0.17 ^g	-0.21 ^a	1.75 ^g	1.82 ^e	2.10 ^d	0.47 ^c	0.56 ^f	0.84 ^d
	30	2.00	-0.05 ^g	-0.08 ^e	-0.08 ^a	0.84 ^g	1.03 ^e	1.06 ^f	0.33 ^c	0.45 ^f	0.56 ^d
	50	1.58	-0.07 ^g	-0.09 ^f	-0.09 ^e	0.85 ^g	0.88 ^f	0.96 ^e	0.23 ^c	0.33 ^f	0.37 ^d
30 th March	5	2.40	-0.31 ^e	-0.35 ^g	-0.37 ^a	3.36 ^e	3.47 ^g	3.72 ^d	0.59 ^c	0.64 ^f	0.99 ^g
	10	2.36	-0.32 ^g	-0.33 ^a	-0.35 ^e	3.09 ^g	3.39 ^e	3.40 ^a	0.43 ^c	0.55 ^f	0.65 ^d
	30	1.96	-0.17 ^a	-0.18 ^g	-0.18 ^e	1.81 ^g	1.95 ^a	2.25 ^e	0.34 ^c	0.44 ^f	0.65 ^d
	50	2.19	-0.06 ^e	-0.11 ^g	-0.13 ^d	1.27 ^g	1.28 ^e	1.44 ^d	0.37 ^c	0.47 ^f	0.74 ^g
14 th May	5	2.58	-0.23 ^e	-0.30 ^a	-0.30 ^g	2.82 ^e	2.89 ^g	3.05 ^d	0.52 ^c	0.61 ^f	0.86 ^g
	10	2.43	-0.23 ^a	-0.23 ^e	-0.23 ^g	2.28 ^g	2.42 ^e	2.61 ^d	0.47 ^c	0.57 ^f	0.86 ^g
	30	1.95	-0.12 ^a	-0.14 ^g	-0.14 ^e	1.46 ^g	1.51 ^e	1.67 ^a	0.34 ^c	0.44 ^f	0.69 ^d
	50	1.98	-0.10 ^g	-0.11 ^e	-0.11 ^a	1.11 ^g	1.25 ^e	1.30 ^f	0.35 ^c	0.45 ^f	0.59 ^d
12 th June	5	2.87	-0.25 ^e	-0.29 ^a	-0.29 ^g	2.76 ^e	2.94 ^g	3.22 ^d	0.57 ^c	0.68 ^f	0.97 ^g
	10	2.57	-0.21 ^e	-0.23 ^g	-0.25 ^a	2.29 ^g	2.33 ^e	2.64 ^d	0.47 ^c	0.57 ^f	0.84 ^g
	30	2.27	-0.13 ^g	-0.13 ^e	-0.15 ^a	1.41 ^g	1.56 ^e	1.66 ^f	0.37 ^c	0.49 ^f	0.70 ^g
	50	1.77	-0.06 ^e	-0.06 ^g	-0.09 ^a	0.84 ^g	0.94 ^e	1.03 ^f	0.28 ^c	0.37 ^f	0.53 ^d
11 th July	5	2.69	-0.33 ^e	-0.40 ^g	-0.40 ^a	3.74 ^e	4.20 ^g	4.31 ^d	0.51 ^c	0.58 ^f	0.83 ^g
	10	2.54	-0.30 ^e	-0.33 ^g	-0.35 ^d	3.31 ^e	3.52 ^g	3.74 ^d	0.45 ^c	0.55 ^f	0.81 ^g
	30	2.29	-0.23 ^e	-0.25 ^g	-0.26 ^a	2.52 ^g	2.75 ^e	2.82 ^a	0.36 ^c	0.49 ^f	0.60 ^g
	50	1.86	-0.19 ^g	-0.19 ^e	-0.21 ^a	2.05 ^g	2.12 ^e	2.38 ^a	0.28 ^c	0.38 ^f	0.52 ^d
26 th July	5	3.34	-0.29 ^e	-0.31 ^g	-0.34 ^a	3.51 ^e	3.55 ^g	3.94 ^a	0.59 ^c	0.72 ^f	1.06 ^g
	10	2.90	-0.26 ^g	-0.29 ^e	-0.29 ^a	2.98 ^g	3.26 ^e	3.39 ^a	0.43 ^c	0.56 ^f	0.71 ^d
	30	2.77	-0.16 ^g	-0.17 ^e	-0.18 ^a	1.89 ^g	2.06 ^e	2.25 ^f	0.38 ^c	0.51 ^f	0.66 ^d
	50	1.90	-0.13 ^g	-0.14 ^e	-0.16 ^a	1.51 ^g	1.67 ^e	1.82 ^f	0.26 ^c	0.36 ^f	0.45 ^d
12 th September	5	2.83	-0.23 ^e	-0.26 ^g	-0.27 ^a	2.63 ^e	2.76 ^g	3.04 ^a	0.48 ^c	0.58 ^f	0.81 ^g
	10	2.87	-0.16 ^c	-0.19 ^g	-0.22 ^a	2.04 ^e	2.13 ^g	2.38 ^d	0.50 ^c	0.61 ^f	0.91 ^g
	30	2.38	-0.09 ^e	-0.09 ^g	-0.12 ^a	1.13 ^g	1.26 ^e	1.31 ^d	0.36 ^c	0.46 ^f	0.64 ^g
	50	1.93	-0.07 ^g	-0.08 ^e	-0.09 ^f	0.90 ^g	0.98 ^f	1.02 ^e	0.27 ^c	0.38 ^f	0.45 ^d
10 th October	5	2.92	-0.28 ^e	-0.28 ^g	-0.31 ^a	2.92 ^g	3.00 ^e	3.29 ^f	0.51 ^c	0.64 ^f	0.89 ^d
	10	2.57	-0.19 ^g	-0.20 ^e	-0.20 ^a	2.11 ^g	2.21 ^e	2.43 ^f	0.42 ^c	0.56 ^f	0.76 ^d
	30	2.12	-0.09 ^g	-0.11 ^a	-0.12 ^e	1.11 ^g	1.28 ^e	1.39 ^f	0.31 ^c	0.41 ^f	0.52 ^d
	50	1.89	-0.08 ^e	-0.09 ^g	-0.10 ^f	1.04 ^g	1.07 ^e	1.11 ^f	0.26 ^c	0.37 ^f	0.42 ^b
6 th November	5	3.25	-0.17 ^e	-0.29 ^d	-0.29 ^g	3.22 ^g	3.24 ^e	3.34 ^d	0.64 ^c	0.69 ^f	1.02 ^g
	10	3.18	-0.18 ^e	-0.21 ^g	-0.24 ^d	2.32 ^e	2.36 ^g	2.69 ^d	0.53 ^c	0.63 ^f	0.86 ^b
	30	2.55	-0.12 ^g	-0.13 ^e	-0.17 ^a	1.47 ^g	1.81 ^e	1.89 ^f	0.37 ^c	0.52 ^f	0.61 ^d
	50	2.37	-0.05 ^a	-0.05 ^g	-0.10 ^b	0.84 ^g	1.25 ^f	1.27 ^e	0.01 ^c	0.01 ^f	1.00 ^d

Table 4

Date	n	S_{obs}	<i>f</i>₁	<i>f</i>₂	SC	Estimator	SE	95 % Lower	95% Upper	S_{mobs}
14th February	2532	36	9	2	0.9999	56.24	20.18	39.97	139.12	85
30th March	7160	34	5	5	0.9999	36.50	2.96	34.40	49.64	90
14th May	41923	45	12	3	1.0000	69.00	20.20	50.72	145.72	86
12th June	11496	41	11	4	1.0000	56.12	12.47	44.69	103.04	92
11th July	11109	48	7	5	0.9999	52.90	4.84	48.97	72.69	108
26th July	22000	58	14	6	1.0000	74.33	11.71	62.62	115.69	122
12th September	12404	47	12	5	1.0000	61.40	11.18	50.75	102.33	102
10th October	4651	57	11	6	1.0000	67.08	8.00	59.56	96.66	103
6th November	6100	56	14	8	0.9998	68.25	8.59	59.54	98.33	111

Figure 1.

Figure 2.

A scatter plot comparing estimated S values ($\hat{S}_{est}$) on the y-axis against mobility-based S values (S_{mobs}) on the x-axis. The axes range from 70 to 130. A solid blue diagonal line represents the identity line ($y=x$). Data points are categorized by date and represented by different symbols: 14th Feb (open square), 30th Mar (open circle), 14th May (open upward triangle), 12th Jun (open diamond), 11th Jul (open downward triangle), 26th Jul (filled square), 12th Sep (filled circle), 10th Oct (filled upward triangle), and 6th Nov (open square). The plot shows that for most dates, the estimated S values are close to the mobility-based values, especially for higher values. However, for the 26th of July, the estimated values are significantly higher than the mobility-based values, reaching up to approximately 106. For the 6th of November, the estimated values are also higher than the mobility-based values, reaching up to approximately 125. The data points for the 14th of February, 30th of March, 14th of May, 12th of June, 11th of July, 12th of September, and 10th of October are clustered around the identity line, indicating a good fit between the two estimation methods for these dates.

Figure 4.