An improved species richness estimator using spatial abundance data

E. W. Tekwa[1,2*], Matthew A. Whalen[1,2], Patrick T. Martone[3], Mary I. O'Connor[1]

[1]Department of Zoology, University of British Columbia, Vancouver, BC, Canada.
[2]Hakai Institute, Heriot Bay, BC, Canada
[3]Department of Botany, University of British Columbia, Vancouver, BC, Canada.

*Corresponding author email: edtekwa@gmail.com, ORCID: 0000-0003-2971-6128

Abstract

Species richness is an essential biodiversity variable indicative of ecosystem states and mass extinctions both contemporarily and in fossil records. However, limitations to sampling effort and spatial aggregation of organisms mean that surveys often fail to observe some species, making it difficult to estimate true richness and hinder the comparison of communities. Here we present a nonparametric, asymptotic, minimal-bias richness estimator, Ωo, by modelling how spatial abundance characteristics affect observation of species richness. Ωo consistently outperforms the best-established richness estimators and can detect small differences that other methods cannot. We conducted simulation tests and applied Ωo to a seaweed survey dataset. Ωo returns more consistent estimates and bootstrapped confidence bounds across years and approaches asymptote with less data than other estimators. The results provide theoretical insights on how biotic and observer variations affect species observation, identify remaining problems, and quantify possible performance gains based on Ωo.

Keywords: species richness, biodiversity, bias, estimator, observation probability, occupancy, clustering, abundance, catchability, bootstrapped confidence bounds

Introduction

A central objective of biodiversity science is understanding how species richness changes across space and time (Loreau *et al.* 2021). Not only is species richness important for today's ecosystem and human health assessments (Mora *et al.* 2011; Costello *et al.* 2013; Gopalakrishnan *et al.* 2018), it is essential to inferring paleoecological patterns and mass extinction events (Jackson & Johnson 2001). Species richness estimation techniques are among the most highly cited works in biology (Pielou 1966; Chao 1984; Magurran 2004). Further, estimating missing species or classes from limited observations is a fundamental problem in other fields, including cryptography, machine learning, and linguistics (Orlitsky *et al.* 2003).

While it is clear that species are globally experiencing elevated extinction rates (Barnosky *et al.* 2011; Kolbert 2014), synthesis of species richness timeseries have suggested that both gain and loss of local species richness are common (Dornelas *et al.* 2014; Elahi *et al.* 2015; Chase *et al.* 2019; Loreau *et al.* 2022). Observed species richness and detection of differences across communities or time are prone to biases in both data selection (Gonzalez *et al.* 2016) and false negative detection errors, which in turn are sensitive to survey effort, design, and observer skill (Chao & Jost 2012). These problems are more acute for species richness than for other biodiversity metric (Roswell *et al.* 2021). Data-intensive methods such as multi-species occupancy models can help estimate true richness of targeted species groups, but they require intensive, repeated field surveys and are sensitive to models of abiotic and biotic drivers (Welsh *et al.* 2013; Devarajan *et al.* 2020).

In most places, the best biodiversity data we can hope for are individual counts identified to the species level. The most robust asymptotic richness estimator that uses observed species abundance data – one that has finite prediction at projected infinite sampling effort – remains the Chao1 method (Chao 1984; Gotelli & Colwell 2011). Chao1 and other competing estimators are based on Alan Turing's sampling theory that relates the number of unobserved classes (codes or species) to observed rarity (Good 1953; Gotelli & Colwell 2011). These methods are expected to be inaccurate when spatial heterogeneity in species abundance is high, sampling effort is low, and sampled individuals are not independent observations. There is a need to develop new statistical methods to reduce bias that fully utilize information in common survey data, without having to separately estimate additional quantities such as survey reliability that may be less reliable than raw richness itself (Welsh *et al.* 2013).

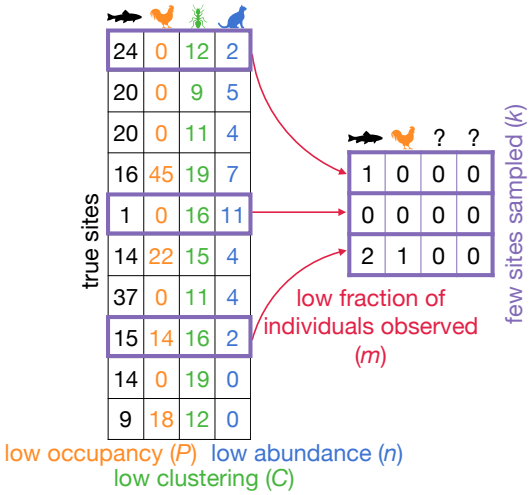

Figure 1. Five sources of downward bias to observed species richness as illustrated using an example of species abundances in space and the subsampling that occurs in a practical field survey. Factors written in colour can all lead to false negatives (unobserved species).

We first illustrate different sources of downward bias (or, underestimates of true species richness) and identify the technical problems that a new method should solve. Five sources of downward bias that we will consider are low values in: sites sampled, fraction of individuals observed, abundance, occupancy, and clustering (Figure 1). These variables will be precisely defined, which we use to mechanistic derive effects on species observation probabilities. This approach departs from both Turing's reliance on rarity, which is prone to noisy observations, and from earlier parametric extrapolations that have weak justifications (Gotelli & Colwell 2011). Using the derivations, we highlight how distributions of the variables in Figure 1 affect raw richness observations. Next, we quantify the performance of the proposed estimator relative to five other best-established asymptotic estimators. We then obtain point estimates and improve bootstrapped confidence bounds of the estimators in a multi-year, multi-site dataset of intertidal seaweed biodiversity (Whalen *et al.* 2021). Finally, we describe Matlab and R functions that can be immediately applied to other spatial abundance datasets.

Richness Estimator Derivation

Here we define variables that can be measured from a spatially explicit abundance dataset, with the aim of arriving at a minimal-bias richness estimator. Each species i has a mean

density n_i: the total number of individuals at scale M divided by the number of sites N, or equivalently the individuals per site in k randomly sampled sites. m_i is the mean fraction of individuals observed, which is also understood as the individual catchability of species i at site x (Chao 1987). Thus, $m_i n_{ix}$ is the observable abundance of species i at x.

At each site where an individual of species i occurs, there tends to be n_{ii} individuals on average because of spatial variance. Note "on average" here means per individual in the site (patch), not per occupied site, the latter being equal to or smaller than the former because some patches have more than one individual. The two subscript i's indicate that n_{ii} is the local density of species i from the perspective of an individual of species i, which contrasts from the global density n_i. The spatial variance $var_x(n_{ix})$ is the variance in the number of individuals of species i across all sites x. It can be shown that the mean local density n_{ii} is a function of spatial variance, with variance having a normalization factor k, the number of sites sampled, rather than the usual k-1 (Lloyd 1967; Tekwa *et al.* 2017).

Equation 1. $$n_{ii} = n_i + var_x(n_{ix})/n_i$$

We define the first spatial heterogeneity quantity, within-species clustering across space C_i, as:

Equation 2. $$C_i = \frac{n_{ii}}{n_i} = 1 + \frac{\mathrm{var}_x(n_{ix})}{n_i^2}$$

Note the following identity allows us to write Equation 2 in terms of the observable abundance $m_i n_{ix}$ across sites x instead of the generally unobserved true abundance n_{ix}:

Equation 3. $$\frac{\mathrm{var}_x(n_{ix})}{n_i^2} = \frac{\mathrm{var}_x(m_i n_{ix})}{(m_i n_i)^2}$$

The second spatial heterogeneity quantity is occupancy P_i, which is the number of occupied sites N_{oi} by species i divided by k:

Equation 4. $$P_i = N_{oi}/k$$

The observation probability of species i at x is the occupancy times the observation probability when the site is occupied. Observation of a species in an occupied site can be described by a Poisson sampling process (sampling with replacement) with a mean of $m_i n_{ii}$ ($=m_i C_i n_i$). Thus, the observation probability per site is D_{iX}:

Equation 5. $$D_{iX} = P_i(1 - \exp(-m_i C_i n_i))$$

On the right-hand side, the exponent term is the Poisson probability of zero observations of species i. One minus this quantity results in the observation probability of species i in a site occupied by species i. When the number of sites sampled k is much smaller than the actual number of sites N, sampling without replacement approximates the expectation from sampling with replacement.

Occupancy is related to clustering through spatial variance, but this relationship is mediated by true mean abundance (He & Gaston 2003), which we generally cannot observe from

biodiversity surveys because of imperfect observations. Thus, we treat occupancy and clustering as separate dynamic variables but, as we will see, our statistical method accounts for the effect of their covariance.

The observation probability of species i within a community M is one minus the cumulative probability of failure to observe, the latter being the probability of failure to observe within site raised to the power k. The observation probability at scale M (D_{iM}) is:

Equation 6. $$D_{iM} = 1 - (1 - D_{ix})^k$$

The raw observed species richness S can be written as the true richness Ω multiplied by the expectation of D_{iM} across extant species i as a function of $\boldsymbol{\Phi}$, which includes R variables ϕ_r. In this case R=4 and ϕ_1 to ϕ_4 are k, $m_i n_i$, C_i, and P_i:

Equation 7. $$S = \Omega \mathrm{E}_i[D_{iM}(\boldsymbol{\Phi})]$$

This expectation value can be approximated by a second-order Taylor expansion (with r and q being variable indices):

Equation 8. $$S \approx \Omega \left(D_{iM}(\bar{\varphi}) + \sum_{r=1}^{R} \left(\frac{d^2 D_{iM}(\bar{\varphi})}{2 d\varphi_r^2} \mathrm{var}_i(\varphi_r) + \sum_{q>r}^{R} \frac{d^2 D_{iM}(\bar{\varphi})}{d\varphi_q \varphi_r} \mathrm{cov}_i(\varphi_q, \varphi_r) \right) \right)$$

$\bar{\varphi}$ stands for the true average values of the R variables. Higher moments are harder to estimate from limited data and may destabilize the estimator. We thus propose an idealized estimate for Ω, the second-order Taylor approximated richness Ω_T:

Equation 9. $$\Omega_T \approx S \Big/ \left(D_{iM}(\bar{\varphi}) + \sum_{r=1}^{R} \left(\frac{d^2 D_{iM}(\bar{\varphi})}{2 d\varphi_r^2} \mathrm{var}_i(\varphi_r) + \sum_{q>r}^{R} \frac{d^2 D_{iM}(\bar{\varphi})}{d\varphi_q \varphi_r} \mathrm{cov}_i(\varphi_q, \varphi_r) \right) \right)$$

The variables ϕ are true (idealized, unbiased) estimators for sampled sites, observable abundances, clustering, and occupancy. In contrast, the observed versions of the variables, ϕ_o, are likely biased toward being underestimated, which we will revisit in the simulation tests. For intuition, observed species will likely have a higher observable abundance averaged across species ($\overline{mn}$) than the truth, because rare species tend to be missed and thus would drive down the true average observable abundance. Disregarding these biases, we propose a naïve Taylor approximated richness Ωo that uses directly measured variables ϕ_o in place of ϕ in Equation 9 and is thus immediately operational on real biodiversity survey datasets. This approximation in expanded form is:

Equation 10. $$\Omega_o \approx S \Big/ \begin{array}{l} D_{iM}(\overline{\varphi_o}) + \frac{d^2 D_{iM}(\overline{\varphi_o})}{2d(mn)^2} \mathrm{var}_i(mn_i) + \frac{d^2 D_{iM}(\overline{\varphi_o})}{2dC^2} \mathrm{var}_i(C_i) + \frac{d^2 D_{iM}(\overline{\varphi_o})}{2dP^2} \mathrm{var}_i(P_i) \\ + \frac{d^2 D_{iM}(\overline{\varphi_o})}{d(mn)dC} \mathrm{cov}_i(mn_i, C_i) + \frac{d^2 D_{iM}(\overline{\varphi_o})}{d(mn)dP} \mathrm{cov}_i(mn_i, P_i) + \frac{d^2 D_{iM}(\overline{\varphi_o})}{dCdP} \mathrm{cov}_i(C_i, P_i) \end{array}$$

$\overline{\varphi_o}$ stands for the observed average value set $(\overline{mn}, \bar{C}, \bar{P}, k)$. We assumed that k is always known and identical for all species in a survey, hence there is not a second-order correction term associated with variance in k. In contrast to this operational estimator (Equation 10), Ω_T

(Equation 9) is the theoretical best second-order richness approximation, if the variables k, $m_i n_{ix}$, C_i, and P_i can themselves be bias-corrected. We define the seven terms in the denominator of Equation 10 as "correction factors": 0th-order, V_{mn}, V_C, V_P, $C_{mn,C}$, $C_{mn,P}$, and $C_{C,P}$.

Analysis

Here we describe how sampled sites (k) and means and variances in observable abundance (mn), clustering (C), and occupancy (P) affect species observation and induce bias to the observed richness. First, we found that the effects of variable means were captured by the first correction factor D_{iM} in the estimator (Equation 10), which is the single-species observation probability (Equation 6) computed at variables' mean values (as if all species in the community share the same characteristics). Equation 6 states that observation probability increases with mean k, n_{ix}, m_{ix}, C_i, and P_i.

Variance effects depend on the signs of the second-order partial derivatives (Equation 10). The analytical expressions are long are always negative (see numerical outcomes in Figure 2C-D). Therefore, variances always decrease the overall observation probability.

These results clarify two main ways in which non-richness biodiversity components induce bias to observed richness. Clustering and occupancy are related to spatial turnover and dissimilarity, yet their mean values affect species observation (Equation 6). Variance in observable abundance, which is likely inversely proportional to community evenness (assuming m is not anti-correlated with n), reduces observed richness even though richness should ideally not be affected by evenness. These effects illustrate the need for biodiversity metrics to be bias-corrected if they are to represent their definitions.

Simulation Tests

We tested the proposed estimator (Ωo) on a simulated dataset of species abundance across a heterogeneous landscape. We compared Ωo to other leading estimators including: the unbiased Chao1 (Chao 1984; Gotelli & Colwell 2011), Chao2 (Chao 1987), the Abundance-based coverage estimator (ACE) (Chao & Lee 1992), Jackknife for abundance data, and Jackknife for incidence data (Burnham & Overton 1979). We also tested the idealized version of our proposed method (Ω_T). For incidence data versions including Chao2 and Jackknife, abundance within sites were converted to presence/absence.

All methods used data simulated from a spatially random subsampled survey. For each simulation scenario, we created 2000 independent communities with 1 to 100 species. Each community in a scenario adhered to desired mean values in quantities that lead to imperfect local observations (fraction of individuals observed E[m], abundance E[n]), spatial heterogeneity (clustering E[C], occupancy E[P]), and sampling effort (sites sampled E[k]). Further, each of the 2000 simulated communities had a mean variable value that deviated from the set means, and each species within a community also deviated from the community means (except for k, which remained constant within community).

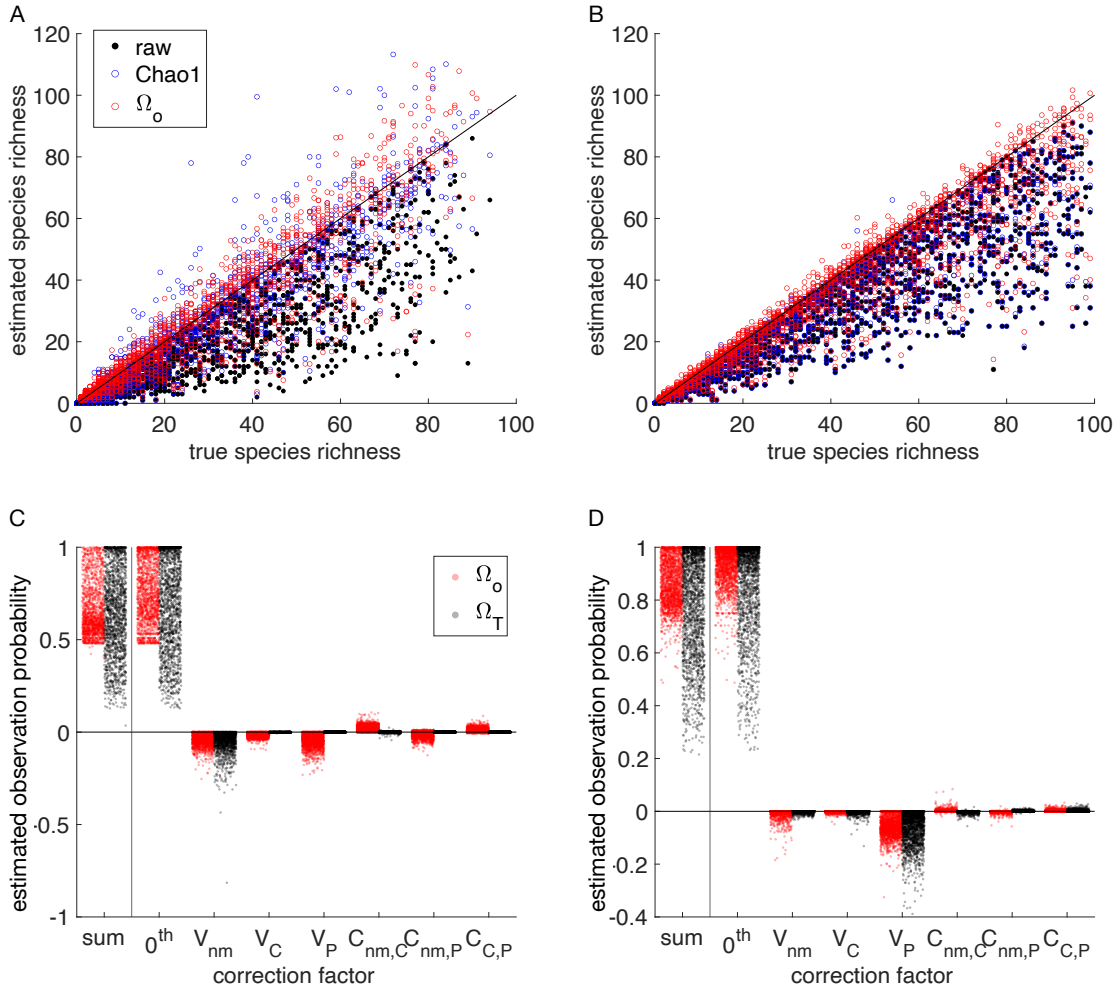

Figure 2. Simulated richness estimates. 2000 communities were populated with 1 to 200 species over 100 spatial sites, and 2 to 10 sites (k) were sampled for richness estimates, including raw (black dots), Chao1 (blue circles), and the proposed method (Ωo, red circles). Species have mean density $n = 2$ individuals per site. **A.** Imperfect local observation scenario, where the mean fraction of individuals observed m was 0.2. **B.** Spatial heterogeneity scenario, where all individuals were observed in sampled sites but with a mean clustering level (C) of 5 and a mean occupancy (P) of 0.4. **C-D.** Estimated chances of observing species by the proposed estimator (Ωo, red) and the idealized estimator (Ω_T, grey), with the sum of correction factors (first column) broken into the individual factors (Equation 10). Each dot represents an estimate from the 2000 replicates, and ideally the Ωo corrections would approach the Ω_T corrections.

Specifically, each community's mean m was ~1/(1+(E[1/m]-1)$\underline{U(0,1)}$) where U is a uniform random value. Each species in a community in turn had a 1/m value drawn from a normal distribution centered at the community mean with a variance of 10 and a minimum of 1 (perfect observation). Each community's mean n was ~E[n]U(0,1), and species n was drawn from a lognormal distribution with a S.D. of 0.5. Each community's mean C was ~E[C]U(0,1), and species C was drawn from a lognormal distribution with a S.D. of 1 and a minimum of 1 (no clustering). Each species' n and C implied a P value according to the identity provided in (He & Gaston 2003). We assigned individuals to each site by drawing from a Poisson distribution with the mean being the species' local density (Equation 1) multiplied by 1 if ~U(0,1)<P and 0

otherwise. This led to species abundance distributions that satisfied the cross-community E[n], E[C], and E[P]. Finally, subsampling biodiversity surveys were simulated by randomly picking $k \in [2,10]$ sites in each community. Within each site and for each species, individuals were observed by drawing from a Poisson distribution with the mean being the species' local abundance times m. This implied sampling individuals with replacement within sites, such that locally observed abundance may be higher than what exists as m approaches 1, but even in this case the average observed individuals will converge to n. Sampling with replacement is likely in non-destructive field surveys when individuals move and their identities are not tracked; on the other hand the scheme approximates sampling without replacement when m is small. Our main results are from two simulation scenarios with E[n]=2: imperfect local observation (E[m]<1) (Figure 2A) and spatial heterogeneity (E[C]>1 and E[P]<1) (Figure 2B), each representing unique challenges to richness estimates. The n distributions for these scenarios are shown in Figure S1A-B.

Performances were quantified in three ways. First, a simple least-squares linear regression was performed on the estimated richness vs. true richness within each estimator. A slope less than 1 indicates that more diverse communities are more severely underestimated than less diverse communities (bias), and a high standard deviation (S.D.) in slope indicates poorly resolved richness ranking across its entire range (ranking error). Second, R^{2*} was computed for the 1:1 line, which describes the variance in estimated richness explained by true richness. A small R^{2*} indicates either underestimates or overestimates, with large errors capable of generating negative values. Third, richness difference detections were checked between 2000 randomly picked community pairs that were different in richness by at most 2 or 10 but were not identical. Detection accuracies above 0.5 is better than a coin-flip.

Table 1. Estimators' performance in simulations. Performance was measured as the slope of estimated richness (± S.D.), R^{2*} of 1:1 line, and pairwise difference detection accuracy when randomly picked communities were either within a richness difference of 2 or 10. For each scenario corresponding to Figure 2A & B, the best estimator in each statistic, excluding the idealized Ω_T, was marked with †.

	Imperfect local observation			**Spatial heterogeneity**		
Estimator	**Slope ± S.D.**	**1:1 R^{2*}**	**Detection ($\leq\Delta 2$, $\leq\Delta 10$)**	**Slope ± S.D.**	**1:1 R^{2*}**	**Detection ($\leq\Delta 2$, $\leq\Delta 10$)**
Raw	0.64±0.42	0.64	0.51, 0.64	0.72±0.35	0.70	0.54, 0.66
Chao1	0.91±0.36	0.85	0.55, 0.70	0.73±0.33	0.71	0.55, 0.67
Chao2	0.83±0.81	0.73	0.32, 0.38	1.00†±0.62	0.66	0.58, 0.72
ACE	0.99†±0.45	0.80	0.56, 0.69	0.73±0.32†	0.72	0.56, 0.68
Jackknife1	0.99†±0.33†	0.89	0.53, 0.67	0.74±0.33	0.72	0.54, 0.66
Jackknife2	0.65±0.87	0.81	0.23, 0.26	1.1±0.40	0.83	0.53, 0.67
Ωo	0.93±0.46	0.87†	0.58†, 0.71†	0.85±0.32†	0.86†	0.59†, 0.74†
Ω_T	*0.92±0.23*	*0.96*	*0.60, 0.79*	*1.0±0.16*	*0.98*	*0.63, 0.83*

The results show that our proposed Ωo method outperformed the overall second-best Chao1 and ACE, leading in every measure across scenarios except for slope under imperfect local observation, where it was third-best but better than Chao1 (Figure 2, Table 1). Incidence-based estimators performed better under spatial heterogeneity (Brose *et al.* 2003), but all estimators other than Ωo performed similarly and generally no better than no correction. The

proposed Ωo gained 5-8% in detection accuracy compared to raw. The idealized version of the proposed method Ω_T performed best overall, further improving difference detection relative to Ωo to a similar degree as Ωo achieved relative to Chao1 (2-8% higher detection accuracy than Ωo and 9-17% higher than raw) while almost eliminating bias, ranking error, and large errors as measured by slope and R^{2*}.

Additional room for improvement in Ωo comes from what we call secondary biases in raw measurements of *mn*, *C*, and *P*. These biases can be observed in the differences between the Ωo and Ω_T correction factors in each simulation scenario (Figure 2C-D). Across both test sets, observed mean and variance in *mn* were biased upward because rare species were missed and observation noise added to differences (Figure S1C, E). The upward bias in mean *mn* overestimated observation probability, but the upward bias in *mn* variance underestimated observation probability (Figure 2C-D). For the imperfect local observation scenario, mean observed *P* was biased downward, and mean observed *C* and variances in *C* and *P* were all biased upward (Figure S1D). The downward bias in mean *P* and upward biases in variances in *C* and *P* led to underestimates in observation probability (Figure 2C), while the upward biases in mean *C* overestimated observation probability. For the scenario with spatial heterogeneity, the biases in means and variances of *C* and *P* were the exact opposite of those caused by low fractions of individuals observed (Figure 2D, Figure S1F). The opposing effects of biases across variables partly cancelled each other, reducing net bias in the Ωo richness estimate and causing the net correction to approach that of the idealized estimator Ω_T (Figure 2C-D). The covariance correction factors were generally smaller than the other factors and similar in Ωo and Ω_T. In comparison, other estimators depended only on the observation of rare species and did not benefit from opposing spatial or variational biases (Chao 1984). However, the net bias in our method depended on the relative magnitudes of these secondary biases, which explains why unbiased versions of the variables (inaccessible in real datasets) improved the proposed method (up to Ω_T). Consistent directions in the secondary biases associated with different biotic and abiotic sources suggest that corrections are possible but likely difficult; we identify these as future research objectives.

We also tested the estimators in scenarios with poorer local observation, higher spatial heterogeneity, and combinations of both factors at two levels of severity (Figure S2-S3). In these scenarios detecting small differences using raw observed richness was no better than coin flips. The proposed Ωo remained the top performer excluding the idealized Ω_T, never showed catastrophic failures, and was the only estimator to be consistently better than coin flips at detecting small effect sizes. Interestingly, the worst conditions led Ω_T correction factors to estimate negative observation probabilities in up to 3% of the communities. These cases indicate conditions where variances and higher moments were unstable – even though Ω_T had access to the true values of *mn*, *C*, and *P*, the data was still poor because *m* and *P* were small. Based on where outliers occurred (Figure S3C, D), we set a threshold of 10% estimated observation probability below which Ω_o and Ω_T switched to using only the 0th-order correction factor (first correction term in denominator of Equation 10), which was never negative and dominated among correction factors. This threshold did not affect Ω_o in the simulations but serves as a safeguard against unforeseen scenarios.

Empirical Application and Confidence Bounds

We test the proposed estimator against other estimators on an extensive seaweed diversity dataset collected annually from 2012 to 2019 on the coast of British Columbia, Canada, and

provide a protocol for obtaining bootstrapped confidence bounds. The dataset features relatively high standards in sampling effort and taxonomic identification (Whalen *et al.* 2021). Still, we expect uncertainty to persist because of imperfect sampling and spatial heterogeneity. We also conducted post-hoc subsampling experiments to explore how various scenarios of reduced sampling effort affect the estimators. In diversity surveys, increasing replicate samples is costly. Reduced sampling effort can reduce diversity estimation precision and undermine efforts to detect change over time, but a robust, unbiased estimator can recover missing observations to some extent. Our subsampling experiments will explore how much different estimators can recover reduced efforts.

The survey tracked seaweed species and their abundance (estimated as percent cover) across steep gradients in abiotic stress associated with rapid changes in elevation and immersion time. The researchers established nine transects running parallel to shore at three intertidal elevations. Each year, we randomly selected ten quadrats on each transect and surveyed them for percent cover of all seaweed species. Between 73 and 82 species were observed each year. Transects likely represent independent spatial samples of an underlying community, whereas quadrats were less spatially independent, so we aggregated quadrat data within transects for this analysis. A 10 x 10 grid was overlaid on the 0.5m x 0.5m quadrat, and any organism filling each square of this grid added 1 percent to the species' total cover within the quadrat (Dethier *et al.* 1993). A species that was partially present in only one square of the grid was assigned a % cover of 0.5 (or trace). We therefore divided all percent cover data by 0.5 to obtain abundance.

Abundance within the context of sampling has no real biological meaning beyond being discrete; the abundance unit does not have to coincide with the individual organism, a concept that we note is open to biological debates and subject to evolution (Heyward & Collins 1985; Tekwa *et al.* 2015) but has no consequences for richness estimation. The only important feature of count data is that the observable abundance mn, not the biological n, approximates a Poisson process; clustering C and occupancy P are not affected by a change of unit and are thus insensitive to the individual definition (Equation 3 & Equation 4). The other consideration for converting percent cover to observed local abundance (mCn) is that it is capped and thus not quite Poisson, but in practice recording abundances beyond this cap will likely not affect the estimated observation probability (eg. failure to observe, $\exp(-mCn)$, is $4\text{x}10^{-44}$ for mCn=100 in Equation 5).

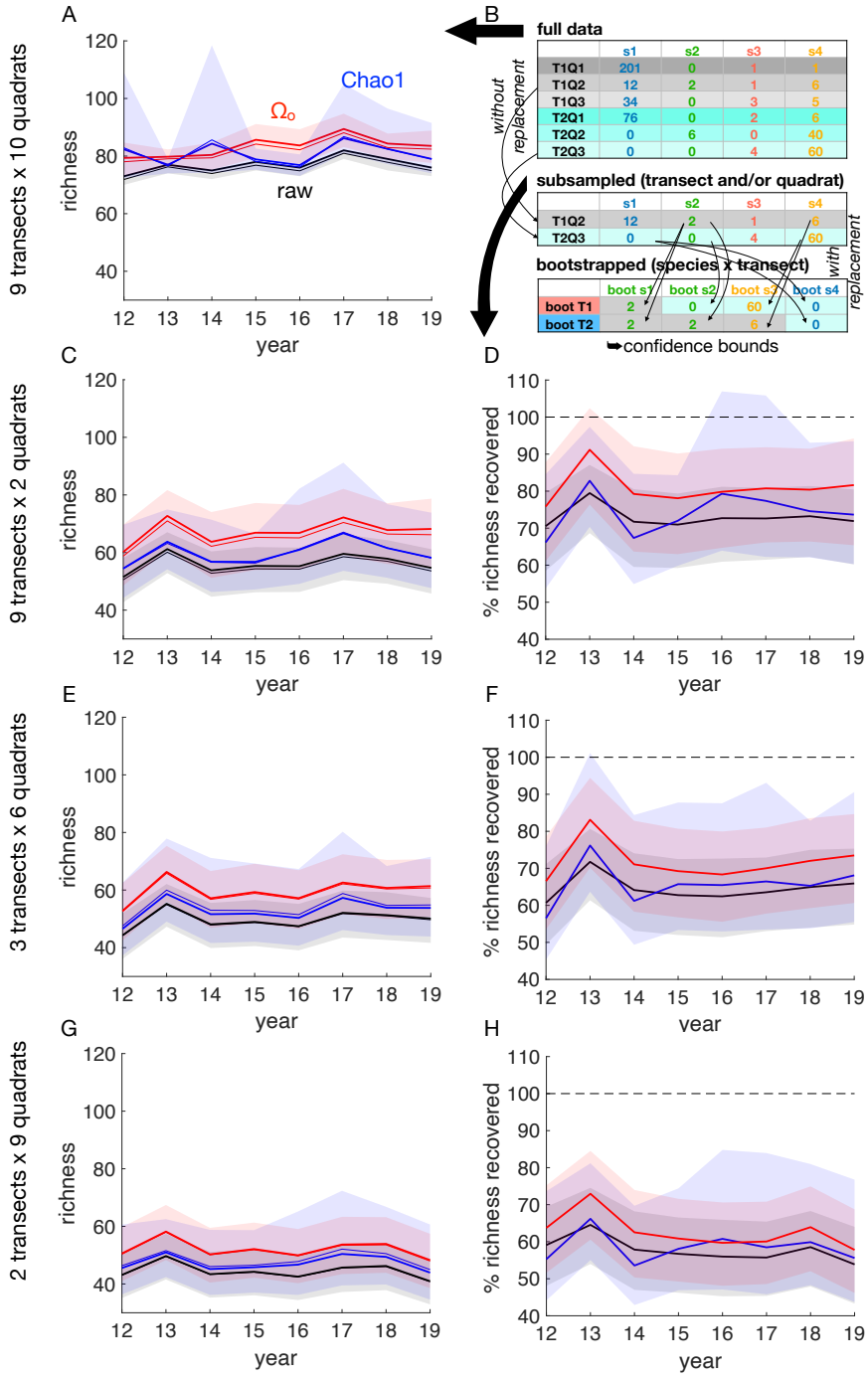

Figure 3. Richness estimates and bootstrap confidence bounds in a seaweed dataset. For each year, raw (black), Chao1 (blue), and the proposed *Ω*o (red) richness estimates were computed without **(A)** and with transect (*T*) and quadrat (*Q*) subsampling (20 times, each without replacement, **C, E, G**). Transects were considered spatially independent samples of the community of species (*s*), and quadrat data were summed within transects. Thick lines represent point estimates or mean point estimates across subsamples. 500 bootstraps were performed for each subsampled dataset each year for 95% confidence bounds. Thin lines represent mean bootstrap estimates (across subsamples when conducted) and should coincide with point estimates. Shades represent means of upper (97.5th) and lower (2.5th) percentiles across

subsamples. **(B)** Illustrations of subsampling scheme (2 transects x 2 quadrats subsampled to 2 transects x 1 quadrat) and bootstrapping. For each bootstrap, species were randomly sampled with replacement (4 times, matching subsampled richness), then within each species transect data were subsampled with replacement (2 times, matching subsampled number of transects) in order to preserve species and spatial dependences between individuals. **(D, F, H)** Same estimates as **(B-D)** but plotted as % richness recovered in subsampled data relative to richness estimates using the same method in the full dataset **(A)**.

We computed the raw observed richness, Chao1 (the second-best based on simulations), and the proposed Ωo estimates for each year using the full dataset (Figure 3A, see Figure S4A for other estimators). These are called point estimates to distinguish them from bootstrap estimates that follow later. We then subsampled transects and quadrats in different combinations while retaining the same "effort" of 18 total quadrats (Figure 3C-H), in order to assess how much each estimator recovers point estimates obtained from the full dataset. For each subsampling replicate, we randomly picked transects (without replacement) and sampled these same transects every year, but randomized the quadrats within each transect (without replacement) each year (Figure 3B). For each subsampling scheme, we replicated the subsampling 20 times since different transects and quadrats could be picked under each scheme. An ideal estimator would predict the same true richness regardless of the manner of subsampling.

We also produced 95% confidence bounds from 500 bootstraps for the full dataset and for each subsampled replicate. We recognize that individual occurrences are not independent both within species and sites because of the intraspecific tendency to cluster (Taylor 1961; Lloyd 1967). To account for these dependencies, for each bootstrap we first randomly sampled with replacement S species from the dataset, where S is the raw observed richness. Then, within each species picked, we randomly sampled with replacement k transects for local abundances (Figure 3B). This procedure was applied to all estimators, and a simple percentile method was used for confidence bounds (2.5^{th} and 97.5^{th} quantiles of bootstrapped richness estimates, Figure S4B). For subsample schemes, the confidence bounds were plotted as means of the 2.5^{th} and 97.5^{th} quantiles across replicates, which showed average confidence bounds (not uncertainties in the richness point estimates across replicates). The means of bootstrapped mean estimates were shown (thin lines, Figure 3) along with the means of point estimates (thick lines, Figure 3): these lines should coincide if bootstraps correctly accounted for the non-independence of individuals.

All estimators revealed little temporal change in seaweed species richness. The narrower confidence bounds for Ωo suggested a lower change for a false-positive trend detection (either increasing or decreasing) based on point estimates alone compared to Chao1 and other estimators (Figure 3A, Figure S4). The proposed Ωo estimator returned a higher richness than Chao1 – between 3 to 8 unobserved species – in the full dataset (Figure 3A). Different subsampling schemes showed a similar order: highest estimates for Ωo, second for Chao1, and lowest for raw measures (Figure 3C, E, G). In terms of percent of richness recovered in the subsampled schemes relative to estimates in the full dataset, Ωo converged on the asymptote faster than other estimators (Figure 3D, F, H). However, all estimators exhibited downward biases under subsampling. Bootstrapping showed that Ωo generated consistent confidence bounds across all years and sampling schemes, in contrast to Chao1 especially in the full dataset where uncertainty was sometimes very high and sometimes very low (Figure 3A). The lower bounds of Ωo always coincided with the mean raw richness, whereas the Chao1 lower bounds reached below the mean raw richness by as much as 8 species (Figure 3A, C, E, G). Lower bounds of corrected richness

estimates should ideally not dip below raw estimates when considering only false negatives (the true number of species cannot be lower than what is observed).

A note on bootstrapping: previous methods assumed that sampled individuals are independent (Chao 1984). This assumption is also necessary for rarefaction and extrapolation techniques that control for sampling effort when comparing communities (Chao & Jost 2012). When we bootstrapped by individuals instead of species and sites, both Ωo and Chao1 yielded bootstrapped means that strongly deviated from the point richness estimates (Figure S5). These results suggest that estimation and detection methods based on rarefaction and extrapolation by individual (or coverage derived from sampled individuals) can be unreliable. Typically, previous estimators used bootstrapping to first fit a parametric standard error, then projected it around the point estimate to create centered confidence bounds (Chao & Jost 2012). This centering inadvertently obscured the individual non-independence problem. We leave it for future research to address how non-independence between individuals can be corrected for in rarefaction/extrapolation techniques.

Statistical Packages

The data and code used in this paper are available on a Github/Zenodo repository at https://github.com/EWTekwa/RichnessEstimator. In it, Matlab and R statistical packages are available with two functions. One function, *RichnessEstsCov*, takes in an abundance matrix with columns indicating species, rows indicating sites, and entries being counts. The function returns raw richness, Chao1, Chao2, ACE, Jackknife abundance, Jackknife incidence, Ωo, and a vector containing the 7 Ωo correction terms. The second function, *BootRichnessEsts,* takes in the same abundance matrix described already, and an additional variable indicating the number of bootstraps to perform. The function returns raw richness, Chao1, Ωo, and vectors containing the bootstrapped raw richness, Chao1, and Ωo. The bootstrapped vectors can be plotted as histograms to visualize the estimates' uncertainty distributions, and simple percentiles can be used for confidence bounds (eg. 2.5^{th} and 97.5^{th} quantiles for 95% confidence bounds). Simple modifications to the function can return bootstrapped estimates for the other estimators.

Conclusion

The proposed richness estimator for species richness Ωo based on spatial ecology and sampling departs from Turing's non-spatial rarity approach (Good 1953), which underlies other asymptotic, non-parametric estimators commonly used in biodiversity studies (Magurran 2004; Gotelli & Colwell 2011). We demonstrated that the proposed method consistently outperformed Chao1 (Chao 1984) and the Abundance-based coverage estimator (ACE) (Chao & Lee 1992), the previous-best estimators, as well as other methods including Chao2 (Chao 1987) and Jackknife (Gotelli & Colwell 2011). The proposed method accounts for the mean and variance effects of sites sampled (effort), fractions of individuals observed within each species (catchability), abundance (rarity), clustering (spatial variance), and occupancy, all readily computed from spatial abundance data and without the need to independently assess survey effectiveness. Consequently, the proposed Ωo can be used to study spatial and temporal trends as well as their relationships with other environmental covariates, without the confounding effects of these covariates on richness estimates (i.e. richness and other covariates may spuriously influence each other through endogeneities introduced if richness is first estimated with the aid of those covariates) (Welsh *et al.* 2013; Devarajan *et al.* 2020).

Bootstrapped confidence bounds that account for non-independence of individual observations due to species and site associations showed that the proposed Ωo is more consistent, better behaved, and converges faster to asymptote than other estimators. Ωo is the only asymptotic estimator that detects richness differences between very similar communities at a higher rate than no correction under all scenarios. These results indicate that Ωo, when independent estimates of observation probability are not available, will be useful for identifying biodiversity hotspots and temporal trends when the underlying change is small (Gonzalez *et al.* 2016). Through simulations and real-world applications, we did not observe any instances of large catastrophic error. These results give confidence to immediate deployment of the proposed method in biodiversity science.

The proposed Ωo suffers from secondary biases in the observed spatial abundance distributions besides richness itself, but the magnitudes and directions of these biases are sensitive to underlying non-spatial and spatial heterogeneities in the community, which always partially cancel each other, explaining why the sum outperforms existing estimators. There continues to be a downward bias to richness estimates as sampling effort and sampling schemes lead to poorer representation of the underlying community, but the proposed method is less biased than other estimators. Nevertheless, we identified that if these secondary biases can be corrected, the proposed approach would come very close to eliminating bias under explored scenarios. As it stands, the currently operational estimator Ωo, provided through Matlab and R statistical packages, is the best available asymptotic species richness estimator. We note that the method is not restricted to counting species and individuals – the number of any taxonomic unit or non-biological class can be estimated instead of species (e.g., order, family, functional, or morphological richness, alphabets, words, etc.), and any discrete abundance unit can replace the biological individual counts. The method only relies on individual units being sampled within-site through a Poisson process, which is likely satisfied even by survey methods that appear to carry continuous values.

Ultimately, any richness estimator will be sensitive to whether the observed species are representative of the underlying community; in fact, the meaning of the underlying community in terms of taxonomic and geographic inclusion will be defined by the organisms that we choose to observe and record in the first place. Thus, even when armed with an ideal richness estimator, biological and survey methodological considerations are indispensable to improving the accuracy of estimated richness and interpreting its meaning across different communities.

Acknowledgement

We thank the Hakai Biodiversity Synthesis group for discussions. EW and MW were funded by the Tula Foundation. EW was funded by a Mitacs Accelerate Internship. PM was funded by Discovery Grant from the Natural Sciences and Engineering Research Council (RGPIN 2019-06240).

Statement of authorship

ET and MO designed the study, ET performed the derivations and simulations, MW and PM collected the data, ET performed the analyses, ET and MW wrote the code, ET wrote the first draft of the manuscript, and all authors contributed to editing.

Data accessibility statement

All code and data are provided on Github at https://github.com/EWTekwa/RichnessEstimator and will be assigned a DOI through Zenodo upon journal acceptance.

References

Barnosky, A.D., Matzke, N., Tomiya, S., Wogan, G.O.U., Swartz, B., Quental, T.B., *et al.* (2011). Has the Earth's sixth mass extinction already arrived? *Nature*, 471, 51–57.

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

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

Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics*, 11, 265–270.

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

Chao, A. & Jost, L. (2012). Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology*, 93, 2533–2547.

Chao, A. & Lee, S.-M. (1992). Estimating the Number of Classes via Sample Coverage. *Journal of the American Statistical Association*, 87, 210–217.

Chase, J.M., McGill, B.J., Thompson, P.L., Antão, L.H., Bates, A.E., Blowes, S.A., *et al.* (2019). Species richness change across spatial scales. *Oikos*, 128, 1079–1091.

Costello, M.J., May, R.M. & Stork, N.E. (2013). Can We Name Earth's Species Before They Go Extinct? *Science*, 339, 413–416.

Dethier, M., Graham, E., Cohen, S. & Tear, L. (1993). Visual versus random-point percent cover estimations: "objective" is not always better. *Mar. Ecol. Prog. Ser.*, 96, 93–100.

Devarajan, K., Morelli, T.L. & Tenan, S. (2020). Multi-species occupancy models: review, roadmap, and recommendations. *Ecography*, 43, 1612–1624.

Dornelas, M., Gotelli, N.J., McGill, B., Shimadzu, H., Moyes, F., Sievers, C., *et al.* (2014). Assemblage Time Series Reveal Biodiversity Change but Not Systematic Loss. *Science*, 344, 296–299.

Elahi, R., O'Connor, M.I., Byrnes, J.E.K., Dunic, J., Eriksson, B.K., Hensel, M.J.S., *et al.* (2015). Recent Trends in Local-Scale Marine Biodiversity Reflect Community Structure and Human Impacts. *Current Biology*, 25, 1938–1943.

Gonzalez, A., Cardinale, B.J., Allington, G.R.H., Byrnes, J., Arthur Endsley, K., Brown, D.G., *et al.* (2016). Estimating local biodiversity change: a critique of papers claiming no net loss of local diversity. *Ecology*, 97, 1949–1960.

Good, I.J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika*, 40, 237–264.

Gopalakrishnan, V., Spencer, C.N., Nezi, L., Reuben, A., Andrews, M.C., Karpinets, T.V., *et al.* (2018). Gut microbiome modulates response to anti–PD-1 immunotherapy in melanoma patients. *Science*, 359, 97–103.

Gotelli, N.J. & Colwell, R.K. (2011). Estimating species richness. In: *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press, pp. 39–54.

He, F. & Gaston, K.J. (2003). Occupancy, Spatial Variance, and the Abundance of Species. *The American Naturalist*, 162, 366–375.

Heyward, A.J. & Collins, J.D. (1985). Fragmentation in Montipora ramosa: the genet and ramet concept applied to a reef coral. *Coral Reefs*, 4, 35–40.

Jackson, J.B.C. & Johnson, K.G. (2001). Measuring Past Biodiversity. *Science*, 293, 2401–2404.

Kolbert, E. (2014). *The sixth extinction: an unnatural history*. First edition. Henry Holt and Company, New York.

Lloyd, M. (1967). `Mean Crowding'. *The Journal of Animal Ecology*, 36, 1.

Loreau, M., Barbier, M., Filotas, E., Gravel, D., Isbell, F., Miller, S.J., *et al.* (2021). Biodiversity as insurance: from concept to measurement and application. *Biol Rev*, 96, 2333–2354.

Loreau, M., Cardinale, B.J., Isbell, F., Newbold, T., O'Connor, M.I. & de Mazancourt, C. (2022). Do not downplay biodiversity loss. *Nature*, 601, E27–E28.

Magurran, A.E. (2004). *Measuring biological diversity*. Blackwell Pub, Malden, Ma.

Mora, C., Tittensor, D.P., Adl, S., Simpson, A.G.B. & Worm, B. (2011). How Many Species Are There on Earth and in the Ocean? *PLoS Biology*, 9, e1001127.

Orlitsky, A., Santhanam, N.P. & Zhang, J. (2003). Always Good Turing: Asymptotically Optimal Probability Estimation. *Science*, 302, 427–431.

Pielou, E.C. (1966). The Measurement of Diversity in Different Types of Biological Colledions. *Journal of Theoretical Biology*, 13, 131–144.

Roswell, M., Dushoff, J. & Winfree, R. (2021). A conceptual guide to measuring species diversity. *Oikos*, 130, 321–338.

Taylor, L.R. (1961). Aggregation, Variance and the Mean. *Nature*, 189, 732–735.

Tekwa, E.W., Gonzalez, A. & Loreau, M. (2015). Local densities connect spatial ecology to game, multilevel selection and inclusive fitness theories of cooperation. *Journal of Theoretical Biology*, 380, 414–425.

Tekwa, E.W., Nguyen, D., Loreau, M. & Gonzalez, A. (2017). Defector clustering is linked to cooperation in a pathogenic bacterium. *Proceedings of the Royal Society B: Biological Sciences*, 284, 20172001.

Welsh, A.H., Lindenmayer, D.B. & Donnelly, C.F. (2013). Fitting and Interpreting Occupancy Models. *PLoS ONE*, 8, e52015.

Whalen, M., Starko, S., Lindstrom, S. & Martone, P. (2021). *Extreme warming restructures habitat distribution and productivity along local gradients in stress and biodiversity* (preprint). Authorea.

Supporting Information for
An improved species richness estimator using spatial abundance data
E.W. Tekwa, Matthew A. Whalen, Patrick T. Martone, Mary I. O'Connor
Supplementary Figures

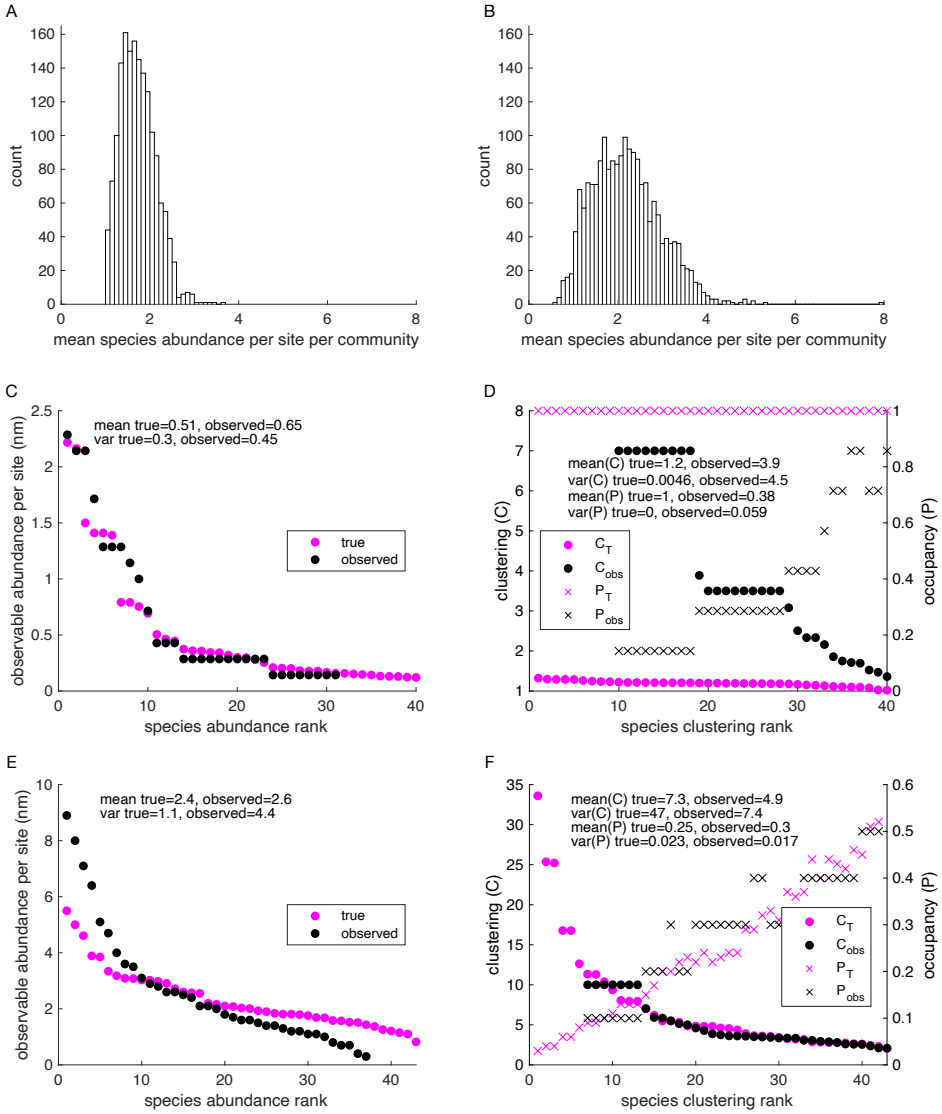

Figure S1. Simulation measurements. A-B. True mean species abundances per site among 2000 simulated communities with a mean of 2 (corresponding to Figure 2A and B parameters respectively). **C & E.** True (magenta) versus observed (black) observable abundance per site (*mn*) according to abundance rank. **D & F.** True (magenta) versus observed (black) clustering (*C*: dots) and occupancy (*P*: crosses) according to clustering rank. **C-D.** One simulated community corresponding to imperfect local observation (Figure 2A parameters). **E-F.** One simulated community corresponding to high spatial heterogeneity (Figure 2B parameters).

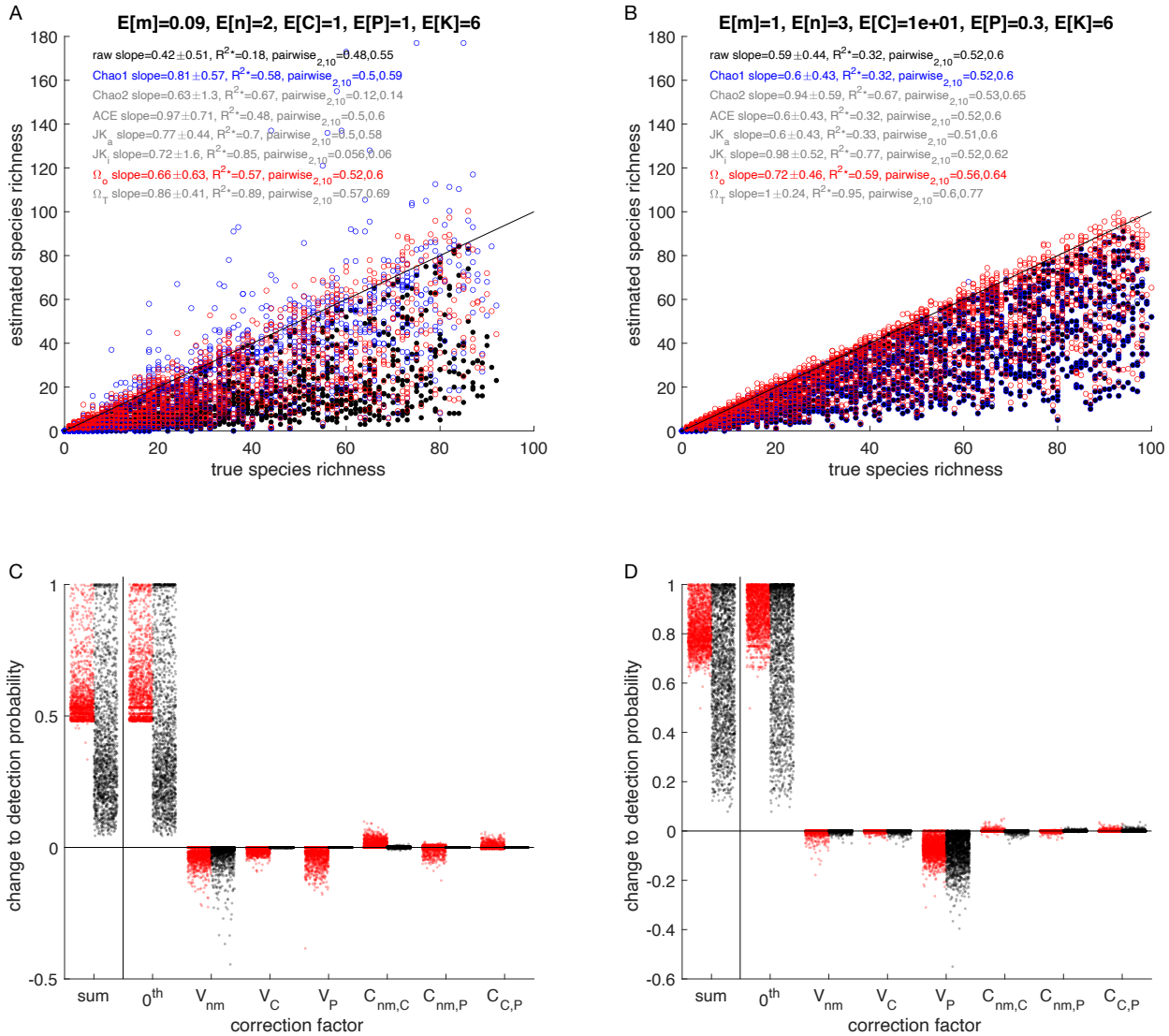

Figure S2. Simulated richness estimates under poor conditions. A. Very poor local observation scenario. **B.** Highly spatial heterogeneity scenario. Graphic descriptions are the same as in Figure 2.

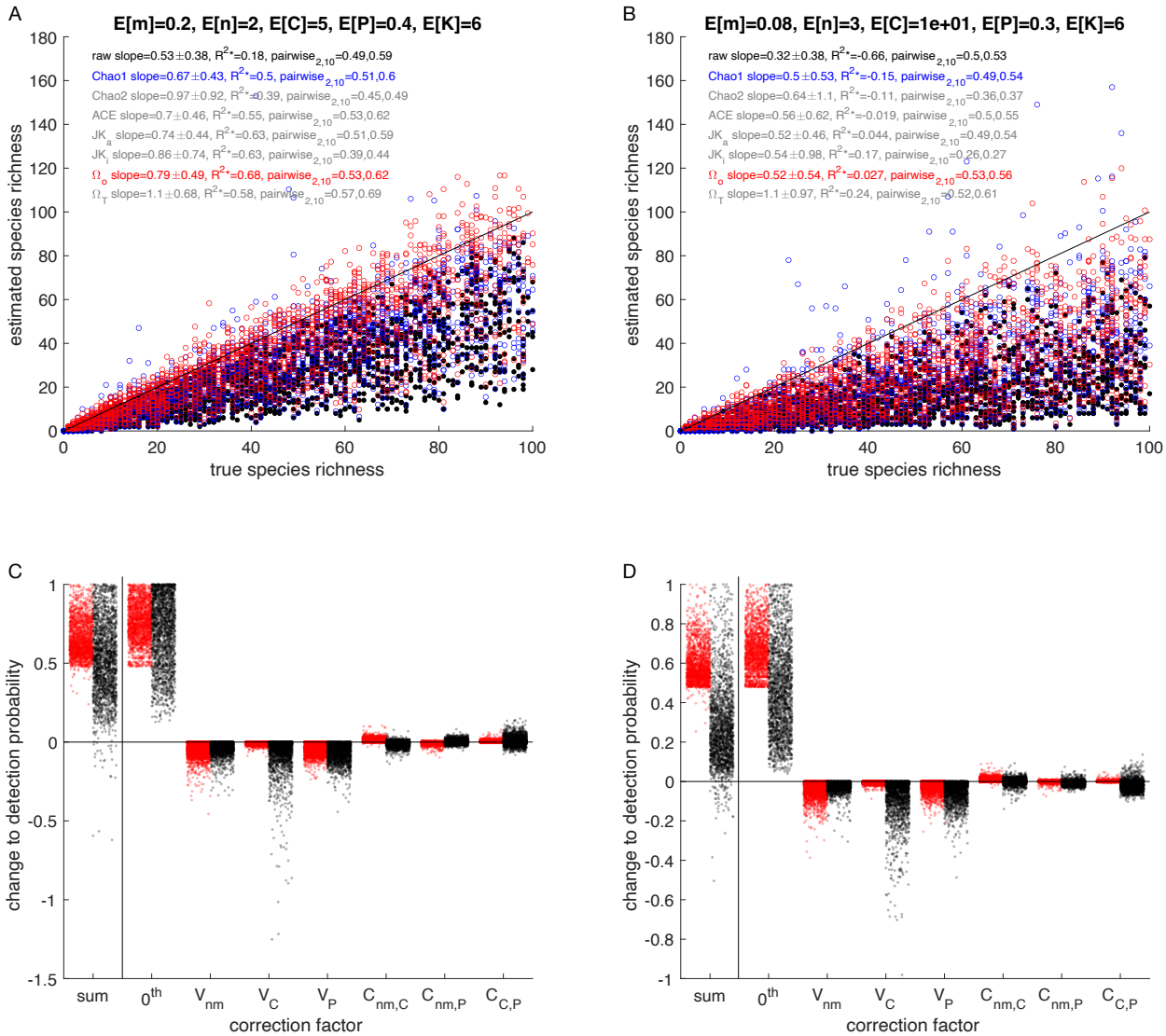

Figure S3. Simulated richness estimates under mixed conditions. A. Imperfect local observation and spatial heterogeneity scenario. **B.** Very poor local observation and highly spatial heterogeneity scenario. Graphic descriptions are the same as for Figure 2.

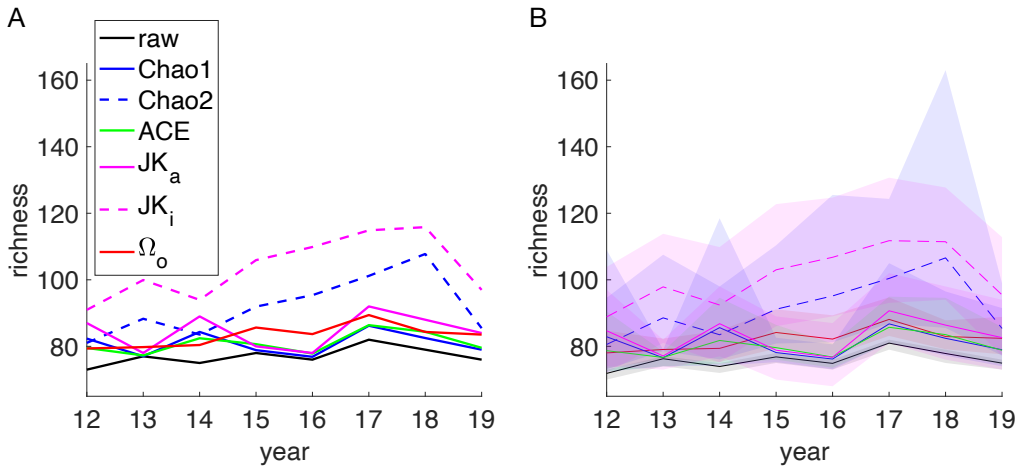

Figure S4. Richness estimates and bootstrap confidence bounds in a seaweed dataset. Graphical descriptions are the same as for Figure 3A, except the complete set of point estimates **(A)** and corresponding bootstrapped means and confidence bounds **(B)** are plotted separately for visibility.

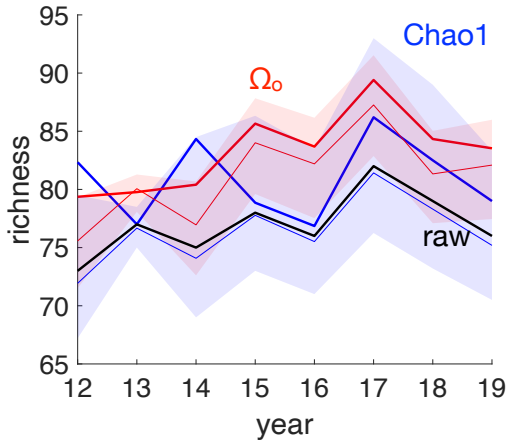

Figure S5. Boostrapped confidence bounds assuming independent individual samples in a seaweed dataset. Graphical descriptions are the same as for Figure 3A, except here bootstrapped means (thin lines) were obtained from randomly sampling individuals with replacement.