Received: 21 September 2022 | Accepted: 7 October 2022

DOI: 10.1002/ecs2.4363

ECOSPHERE
AN ESA OPEN ACCESS JOURNAL

ARTICLE

Methods, Tools, and Technologies

Estimating total species richness: Fitting rarefaction by asymptotic approximation

Yi Zou[1] | **Peng Zhao**[1] | **Jan Christoph Axmacher**[2]

[1]Department of Health and Environmental Sciences, Xi'an Jiaotong-Liverpool University, Suzhou, China

[2]UCL Department of Geography, University College London, London, UK

Correspondence
Yi Zou
Email: yi.zou@xjtlu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 31700363

Handling Editor: Debra P. C. Peters

Abstract

Estimating the number of species in a community is important for assessments of biodiversity. Previous species richness estimators are mainly based on nonparametric approaches. Although parametric asymptotic models have been applied, they received limited attention due to specific limitations. Here, we introduce parametric models fitting the probability-based rarefied species richness curve that allow us to estimate the "Total Expected Species" (TES) in a community based on species' abundance data. We develop two approaches to calculate TES (termed "TESa" and "TESb"), based on two slightly different mathematical assumptions regarding Expected Species (ES) models. We provide R functions to calculate both these estimation approaches and their standard deviation. The function also enables users to visualize the estimation. We test the performance of TESa, TESb, and their average (TESab) across simulated and empirical data, and compare their bias, precision, and accuracy with other commonly used, nonparametric species richness estimators: the bias-corrected (bc-)Chao1 and the abundance-based coverage estimator (ACE). Simulation reveals that in small samples TESa shows a tendency to overestimate and TESb to underestimate overall species richness. TESab performs well in bias, precision, and accuracy when compared with (bc-)Chao1 and ACE estimators. Results from empirical data show that the variance generated from TES estimates is comparable with that for (bc-)Chao1 and ACE. Our study demonstrates that rarefaction theory in combination with parametric approximation models provides a valuable new approach to estimate the species richness of incompletely sampled communities. Robust estimates are likely to be obtained where the observed number of species is greater than half of the TES estimation. When the ratio of TESa to the observed richness is $\gg 2$, we suggest the use of TESb or TESab. Although more comprehensive comparisons with other estimators are suggested, we encourage researchers to consider the TES approach in their biodiversity studies as a complement to current existing estimators.

Yi Zou and Jan Christoph Axmacher contributed equally to this study.

KEYWORDS
ACE, Chao1, species estimator, Total Expected Species, α-diversity

INTRODUCTION

Knowledge of a community's species richness is a crucial prerequisite for the effective assessment and conservation of biodiversity. Samples of mobile organisms routinely provide only a partial picture of the complete species pool representing the sampled communities. Such under-sampling represents a problem in biodiversity assessments (Coddington et al., 2009). Estimating the number of species in a community or at a specific sample site has attracted great interest from ecologists for a long time (Bunge & Fitzpatrick, 1993; Chao & Chiu, 2016; Clench, 1979; Palmer, 1990).

There are a variety of statistical methods used to estimate the total number of species (Bunge & Fitzpatrick, 1993), which can generally be grouped into nonasymptotic and asymptotic approaches (Chao & Chiu, 2016). In this study, we focus on asymptotic estimations, which can be further divided into nonparametric and parametric approaches.

Nonparametric asymptotic approaches consider the frequencies in the number of rare species, for example, looking at species only present with one or two individuals (singletons and doubletons) in a sample. Examples of this approach are the Chao1 estimator (Chao, 1984) and its improved, bias-corrected versions (Chiu et al., 2014; O'Hara, 2005), the abundance-based coverage estimator (ACE) (Chao & Lee, 1992), as well as the Jackknife estimator family (Cormack, 1989). Such nonparametric estimations generally generate "lower boundary" estimations, providing estimates of the minimum species richness that can be expected to exist in the base community, although they have been widely used in diversity studies as a measure of expected total species richness (e.g., McGeoch et al., 2007; Vester et al., 2007). These nonparametric measures require very large sample sizes and associated sampling efforts in order to generate reliable estimations of the true species pool (Brose et al., 2003; Hortal et al., 2006; Reese et al., 2014).

Parametric asymptotic models that use fitted curves to extrapolate saturation points reflecting a community's species richness have been used for many decades (Preston, 1948). A majority of these curve-fitting methods use sample-based (or quadrat-based in botanical studies) accumulations, where sampling census points (e.g., specimens recorded in traps or standardized areas) represent the observation unit and only presence–absence data are required (Flather, 1996; Jiménez-Valverde & Lobo, 2006; Keating & Quinn, 1998; Rosenzweig et al., 2003). In contrast, few parametric curve-fitting approaches are based on species' abundance accumulation models, where each individual specimen represents an observation unit (Fisher, 1999; Jiménez-Valverde & Lobo, 2006; Palmer, 1990). While some studies report that parametric methods based on resulting species' abundance distributions give a better estimate of the true species richness than nonparametric estimates (O'Hara, 2005), this approach has so far received very limited attention.

There are three main limitations for species' abundance-based accumulation curves. Firstly, abundance-based accumulation curves are per se discrete and not smooth, and direct curve-fitting may result in high inaccuracies (Colwell & Coddington, 1994; Gotelli & Colwell, 2001). Secondly, a good fit of the curve depends on the underlying species' abundance distribution pattern, but this pattern is commonly unknown and context-dependent (McGill et al., 2007). In this context, different mathematical models may fit curves equally well, but their predictions can differ dramatically (Colwell & Coddington, 1994). Finally, parametric curve-fitting methods usually do not provide a variance for their estimations (Colwell & Coddington, 1994). Therefore, parametric specimen-based curve-fitting approaches have been regarded as generally poorly suited to approximate the total number of species (Chao & Chiu, 2016; Walther & Moore, 2005).

Rarefaction curves estimate the expected number of species for a given sample size (that has to be equal to or smaller than the actual sample) based on a hypergeometric distribution model (Hurlbert, 1971; Sanders, 1968). Rarefaction curves are smooth and therefore overcome the first limitation in parametric curve-fitting methods. These curves have been widely used to compare the biodiversity of incompletely sampled communities that are represented by samples of varying sizes (see the review by Gotelli & Colwell, 2001). However, as the curve per se does not provide an asymptotic value, it cannot be used to estimate the total species richness (Tipper, 1979). Rarefaction curves can also be extrapolated to a standardized sample size that exceeds that of the original sample, for example, using the iNEXT approach (Chao et al., 2014; Chao & Jost, 2012; Colwell et al., 2012), but these extrapolations do not generate asymptotic values either. To date, studies using asymptotic models to estimate the total species richness based on rarefaction curves are surprisingly rare (see, e.g., Mauffrey et al., 2007; O'Hara, 2005). Mauffrey et al. (2007) tested several models to extend rarefaction curves and estimate a

value for the resulting asymptote. They found that, while model performance varied when compared against known community compositions, none of the tested models showed a general superior performance, even for large sample sizes (Mauffrey et al., 2007).

Zou and Axmacher (2021) recently proposed an approach to estimate the total number of species shared between two incompletely sampled communities (i.e., the intrinsic element of β-diversity) based on parametric rarefaction curve-fitting. This method has been shown to be robust for different species' abundance distribution models. In this study, we follow a similar approach, by applying parametric curve-fitting to two different "Expected Species" (ES) functions proposed by Hurlbert (1971) and Smith and Grassle (1977) to estimate the species richness (i.e., α-diversity). We refer to this estimate as "Total Expected Species" (TES) and provide the standard deviation of the estimation, which potentially allows us to overcome all limitations in species' abundance-based accumulation parametric curve fittings. We furthermore test the performance of TES across different simulated species distribution models commonly observed in natural communities, and we compare its bias, precision, and accuracy with other commonly used species richness estimators for different sample sizes and associated sample completeness scenarios. We finally apply the TES calculations to two empirical datasets. The functions to compute TES values and their respective standard deviations are provided in R language (R Core Team, 2018) in Appendix S1.

METHODS

The Expected Species concept

The Total Expected Species model we present in this paper is fundamentally based on the concept of the Expected Species (ES) richness, calculated as number of species when randomly selecting m individuals from a collection based on a hypergeometric distribution (Hurlbert, 1971), hereafter referred to as "ESa":

$$\mathrm{ESa}_m = \sum_{i=1}^{S}\left[1-\frac{\binom{N-N_i}{m}}{\binom{N}{m}}\right], \quad (1)$$

where S represents the total number of (observed) species in the sample collection, while N represents the total number of individuals, N_i represents the number of individuals of species i in the sample, and m is the standardized sample size that is rarefied to.

Under a multinomial sampling model based on the assumption of a community containing an infinite number of individuals, where each individual is sampled independently from all others (Smith & Grassle, 1977) without assuming any sampling or detection bias, this function can alternatively (referred to as "ESb") be formulated as:

$$\mathrm{ESb}_m = \sum_{i=1}^{S}\left[1-\left(1-\frac{N_i}{N}\right)^m\right]. \quad (2)$$

For $m = 2$, the resulting value of ESb is linked to Simpson's concentration index as: $\mathrm{ESb}_2 = \mathrm{Simpson} + 1$. For larger values of m, the value for ES in both formats can furthermore be expected to converge toward the true species richness of the underlying community. Once m approaches infinity, it can hence be assumed that the resulting value does represent the total number of species that are contained in the community. In practice, it is also possible to calculate the change in ES values for a sample containing N_t specimens drawn randomly from an underlying community N for increasing values of m from $m = 1$ to $m = N_t$, where $N_t \leq N$.

From Expected Species to Total Expected Species

Once the change in ES values in response to an increasing sample size m has been calculated, a curve can be fitted to represent this change. For the curve fitting, we applied a Weibull-logistic model based on the nonlinear least square estimation described by Zou and Axmacher (2021). In brief, we calculated values of ESa (formula 1) and ESb (formula 2) at regular integer intervals of sample size m varying between 1 and N_t, initially resulting in a curve representing the relationship between the estimated number of species ES and m, that is, the rarefaction curve. This curve can in turn be extrapolated for values of $m > N_t$ to generally describe how the estimated species richness changes as a function of changes in the sample size m ($\mathrm{ES} = f[\ln(m)]$) using a four-parameter Weibull model:

$$\mathrm{ES}_m = a - b \times e^{-c \times M^d}, \text{ where } M = \ln(m). \quad (3)$$

Here, the parameter a, which is the horizontal intercept of asymptotes of the curve, represents the estimated overall number of species (Figure 1).

While the calculation of the Weibull model requires substantial sample sizes, the curve can alternatively be fitted using a three-parameter logistic regression model:

$$\mathrm{ES}_m = \frac{a'}{1 + e^{\frac{b'-M}{c'}}}. \quad (4)$$

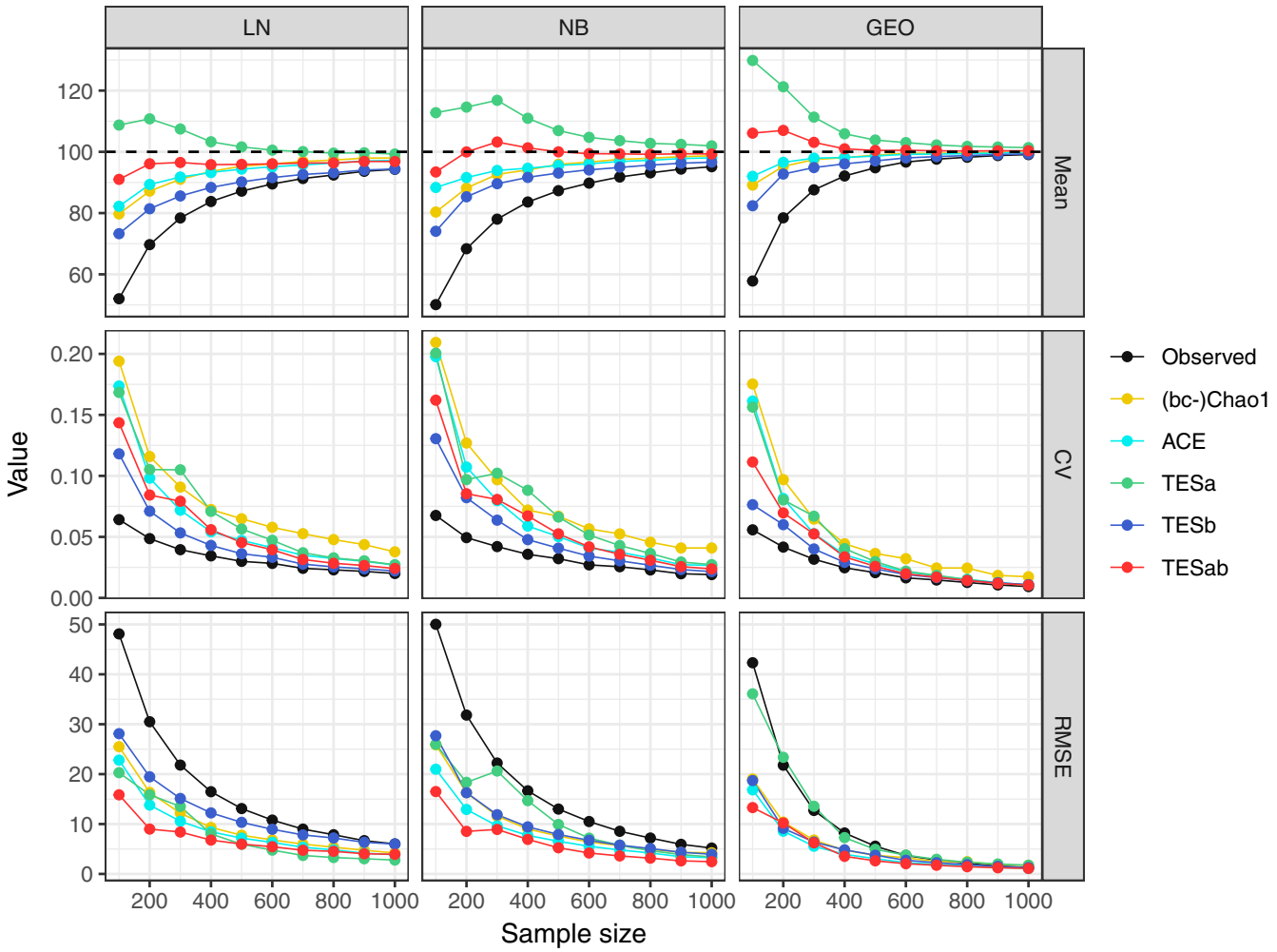

FIGURE 1 Bias (expressed as the mean), precision (expressed as the coefficient of variance [CV]), and accuracy (expressed as the root of mean square error [RMSE]), for different species richness estimators based on samples randomly taken from the three simulated baseline communities. Each community contains 100 species (the dashed line), following a lognormal (LN), negative binomial (NB), or geometric (GEO) distribution. ACE, Abundance-based Coverage Estimator; TES, Total Expected Species.

The curve of the logistic regression model can be fitted either based on equations ESa or ESb—in both cases leading to an asymptotic value representing TES, hereafter called "TESa" and "TESb," respectively, in accordance with the use of formula 1 or 2.

The variance of TESa and TESb was calculated as the estimated standard deviation (σ) following O'Hara's (2005) proposal, which can be expressed as:

$$\sigma = \sqrt{s_{\text{est}}{}^2 \times (n-p)}, \quad (5)$$

where s_{est} is the standard error of either of the estimations (i.e., TESa or TESb), n is the number of knots that are used in the model fitting and p is the number of estimated coefficients, that is, 3 for the logistic model and 4 for the Weibull model. We did not use the estimated standard error generated from the resampling procedure of the curve fitting (e.g., Colwell & Coddington, 1994; Keating & Quinn, 1998), as the standard error in this context refers to the error related to resampling processes, as pointed out by O'Hara (2005).

In practice, our approach uses nonlinear regression models to fit the change of ES values to the increases in the theoretical sample size m. The regression initially uses the four-parameter Weibull model. If the overall sample size proves too low to calculate the Weibull model, then the regression is refitted using the three-parameter logistic regression model. Although ESa and ESb values are very similar, particularly for large sample sizes (Appendix S1: Figure S1a,b), the shape of their curves differs slightly, resulting in different values for TESa and TESb (Appendix S1: Figure S1c,d). Therefore, we also calculated the mean value of TESa and TESb, hereafter called "TESab," with TESab = (TESa + TESb)/2. Here we

consider the value of TESab as the mathematical average of TESa and TESb (i.e., these two estimators are generated independently from two curves), and hence its standard deviation is $\frac{1}{2}\sqrt{\sigma_a^2+\sigma_b^2}$, where σ_a and σ_b represent the standard deviation of TESa and TESb, respectively.

In addition to the Weibull-logistic model, we also fit the curve with the Michaelis–Menten model, which has been extensively used in curve-fitting methods to estimate species richness (Butler & Chazdon, 1998; Clench, 1979; Keating & Quinn, 1998; Rosenzweig et al., 2003; Soberon & Llorente, 1993). The asymptote of the Michaelis–Menten model has been reported to be a good estimator of species richness (e.g., Rosenzweig et al., 2003), although this assumption has also been questioned (e.g., Keating & Quinn, 1998). The respective equation can be written as:

$$\mathrm{ES}_m = \frac{a'' \times m}{b'' + m}, \tag{6}$$

where the asymptote, a'', represents the value of TES. Again, we calculated TESa and TESb based on formulae 1 and 2, and the average value, TESab.

Simulated dataset

We created three simulated baseline communities based on three different abundance distribution models. In order to allow for direct comparisons of species richness estimator performances between different distribution models, we created a separate community so that it contained a total of 100 species that were distributed across ~100,000 individuals. The three abundance distribution models we used are based on models suggested in the literature as approximations of species' abundance patterns in natural populations: a lognormal distribution (LN), a negative binomial distribution (NB), and a geometric distribution (GEO) (Appendix S1: Figure S2). The abundance of each species in each community exceeded 50 individuals, which was chosen as the minimum threshold for a self-sustainable population size (Franklin, 1980).

Comparisons of the Weibull-logistic model and Michaelis–Menten model

The first step in our analysis was to compare the estimated species richness generated by the Weibull-logistic versus the Michaelis–Menten model. In both cases, we established respective regression models for TESa, TESb, and TESab for the three abundance distribution datasets (LN, NB, and GEO) for different, randomly selected samples ranging in size between 100 and 1000 individuals, at intervals of 100 individuals. The TES values were calculated for 1000 independent, randomly selected samples for each distribution dataset and sample size, and we compared our TES estimations generated from Weibull-logistic and Michaelis–Menten models for their respective bias, precision, and accuracy, following suggestions from Walther and Moore (2005). As the actual richness of the underlying community was known and stable at 100 species, our "bias" represents the mean value rather than the mean value of mean error, in order to give a direct picture. Precision was calculated as the coefficient of variation (CV), while accuracy was calculated as the root of the mean square error (RMSE). Further calculation details can be found in Appendix S1.

Our analysis showed that results based on the Michaelis–Menten model had a consistent positive estimation bias (i.e., overestimate) for the geometric distributions for both TESa and TESb, resulting in a low accuracy even for large sample sizes (>500 individuals). The Weibull-logistic model had a much higher accuracy for geometric distributions, while both estimation methods showed similar performance for lognormal and negative binomial-distributed communities (Appendix S1: Figure S3). These results indicate that the Michaelis–Menten model works well only for communities with specific abundance distribution patterns. We therefore only used TES values based on Weibull-logistic models in our subsequent evaluations and applications.

Evaluation of the TES performance

The performance of TESa, TESb, and TESab based on the Weibull-logistic model were tested in comparison with two commonly used species richness estimators—the bias-corrected form of Chao1 (Chao, 1984; O'Hara, 2005) and the ACE (Chao & Lee, 1992), which are two default estimators used for example in the "vegan" package (Oksanen et al., 2018) in R. The performance of TESa, TESb, TESab, (bc-)Chao1, and ACE were assessed for samples of sizes ranging from 100 to 1000 individuals, again using intervals of 100 individuals and 1000 replications. The abovementioned bias (mean), precision (CV), and accuracy (RMSE) were again compared for the different estimated species richness values.

Applying TES to empirical dataset

We finally applied and tested the TES estimation on two empirical datasets, both provided in the context of the vegan package (Oksanen et al., 2018). The first dataset is

the mite dataset comprising data on oribatid mites (Acari, Oribatei) at an ecological station in Quebec, Canada, comprised of 9800 individuals spread across 35 species based on 70 individual observations (Borcard & Legendre, 1994). The least abundant species in this dataset contains 11 individuals (i.e., no singleton and doubleton). We hence consider this dataset to be widely complete (Chao, 1984) in the sense that the 35 recorded species likely represent the total species pool of the underlying community. To use TES in the context of this dataset, we pooled all observations. We then randomly created subsamples from the pooled data accounting for 1% (98 individuals), 5%, 10%, 50%, and 100% of the total dataset. The resampling procedure allowed for multiple sampling of the same individual under the assumption of an indefinite local population. For each resampled community, we calculated TESa, TESb, TESab, (bc-)Chao1, and ACE. As before, we then compared the mean, CV, and RMSE between these estimators based on 1000 simulated replications for those indices.

The second dataset we used in our testing of TES is the BCI data from Barro Colorado Island, reporting on trees sampled on 50 plots (1 ha each) that resulted in a total of 21,457 recorded trees representing 225 species (Condit et al., 2002). Unlike mite, 19 species in the BCI dataset are represented by a single individual (singleton) only, indicating that the samples are incomplete (Chao, 1984). In this case, we therefore do not know the real species richness of the community from which this dataset was collected. We calculated the total species richness, estimated using TESa, TESb, and TESab, and their standard deviation, and we compared these values with estimates and standard deviation for (bc-)Chao1 and ACE. The calculation was done in different steps: (1) the first sampling plot in the dataset containing 448 individuals across 93 species, (2) then pooling the first 5 plots (2359 individuals and 152 species), (3) followed by pooling the first 10 plots (4510 individuals and 170 species), and finally (4) pooling all plots.

All simulations and calculations were completed in the R software (R Core Team, 2018). The R functions for the calculation of ESa and ESb, as well as TESa and TESb, are presented at Dryad (Zou et al., 2022). We provide the "ES()" function to calculate the Expected Species with parameter "a" and "b" to calculate ESa (formula 1) and ESb (formula 2), of which results from ESa are identical to "rarefy()" function in the vegan package (Oksanen et al., 2018). We provide the "TES()" function to calculate Total Expected Species based on ESa and ESb according to the Weibull-logistic model. The function returns the mean value and estimated standard deviation of TESa, TESb, and TESab. We also provide the "plot.TES()" function that allows users to visualize the quality of their curve-fitting. The "estimateR()" function in the vegan package (Oksanen et al., 2018) was used to calculate the (bc-)Chao1 and ACE estimators.

RESULTS

Simulated dataset

For the samples of different sizes randomly drawn from the three simulated datasets, TESa estimator consistently overestimates the true species richness, which was particularly pronounced for small sample sizes (Figure 1). For the overall estimation of mean, this measure still performs well when compared with the other estimators for the geometric distribution model community, but shows a worse performance for negative binomial and log-normal distribution populations. TESb, in contrast, consistently underestimates the overall richness and performs generally worse than the established estimators included in the analysis (Figure 1).

For precision, TESb performs better than all other species richness estimators. In contrast, TESa shows levels of precision that are very similar to the established measures we used in this comparison. With regards to accuracy, TESb shows a performance that is similar to the other species richness estimators, with ACE commonly performing slightly better, while (bc-)Chao1 performs slightly worse (Figure 1). Nonetheless, TESa shows consistently the worst accuracy for the NB community as well as for a wide range of under-sampling scenarios for the LN community, while it proves widely superior in the GEO community (Figure 1).

When comparing the performance of TESab with its two foundation measures, TESa and TESb, but also with both, (bc-)Chao1 and ACE, this measure shows qualities that are promising to estimate the true species richness from the incomplete samples. TESab approaches the true value of 100 species much more rapidly than all other measures, while its precision is only slightly worse than that of TESb, but better than that of all other species richness estimators, and its accuracy is also generally similar or better than that of all the other species richness estimators we used in this analysis (Figure 1).

Empirical dataset

For the mite dataset, TESa and TESab both overestimate the total species richness for sample sizes <5% of the total number of individuals (Figure 2), while the value estimated by TESb are much closer to the assumed true

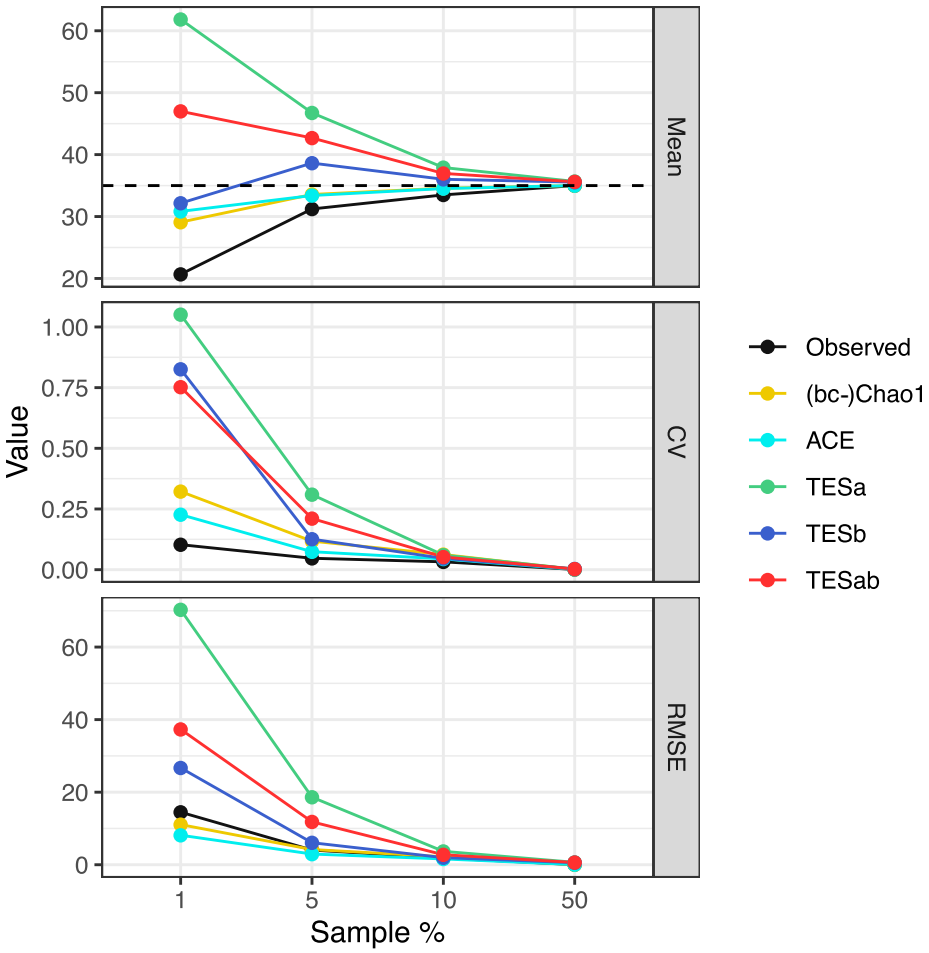

FIGURE 2 Bias (expressed as the mean), precision (expressed as the coefficient of variance, CV), and accuracy (expressed as the root of mean square error, RMSE), for the different species richness estimators based on samples randomly taken from the mite dataset that contains a total of 9800 individuals spread across 35 species (the dashed line). Samples were randomly taken to represent 1%, 5%, 10%, 50%, and 100% of the number of individuals in this dataset. ACE, abundance-based coverage estimator; TES, Total Expected Species.

species richness (Figure 2). Both precision and accuracy from TES estimates are worse than these measures for (bc-)Chao1 and ACE (Figure 2).

For the BCI dataset, values estimated by TESa and TESab are consistently slightly higher, while TESb values are similar to the species richness values estimated by (bc-)Chao1 and ACE, based on the first sampling plot (Figure 3a), the first five sampling plots (Figure 3b), and the first 10 sampling plots (Figure 3c). When pooling all plots, TESa estimates are similar to (bc-)Chao1 and ACE and slightly higher than for TESb (Figure 3d). Here, the standard deviations of the TES series values are similar to those calculated by (bc-)Chao1 and ACE (Figure 3).

DISCUSSION

Our study demonstrates that probability theory can be combined with parametric approximation models to calculate meaningful estimates of the total number of species in a community represented by incomplete samples. Our parametric approaches are novel in three ways: (1) our tests across simulated communities representing different species abundance distribution models show that our novel approach generates results that are widely applicable across these different models; (2) we provide a way to calculate the variance for this parametric estimator; and (3) we present an approach to easily visualize the curve-fitting results that allow a direct evaluation of the respective shape of the curve. This information can, for example, be used to establish the saturation predicted by the model, providing indications for the reliability of the predicted species saturation point.

While the theoretical basis of TES resulted in two distinct, mathematically slightly different formulas to approximate total species richness (TESa and TESb), both of these show some strong qualities, but also some performance issues, when compared with established species

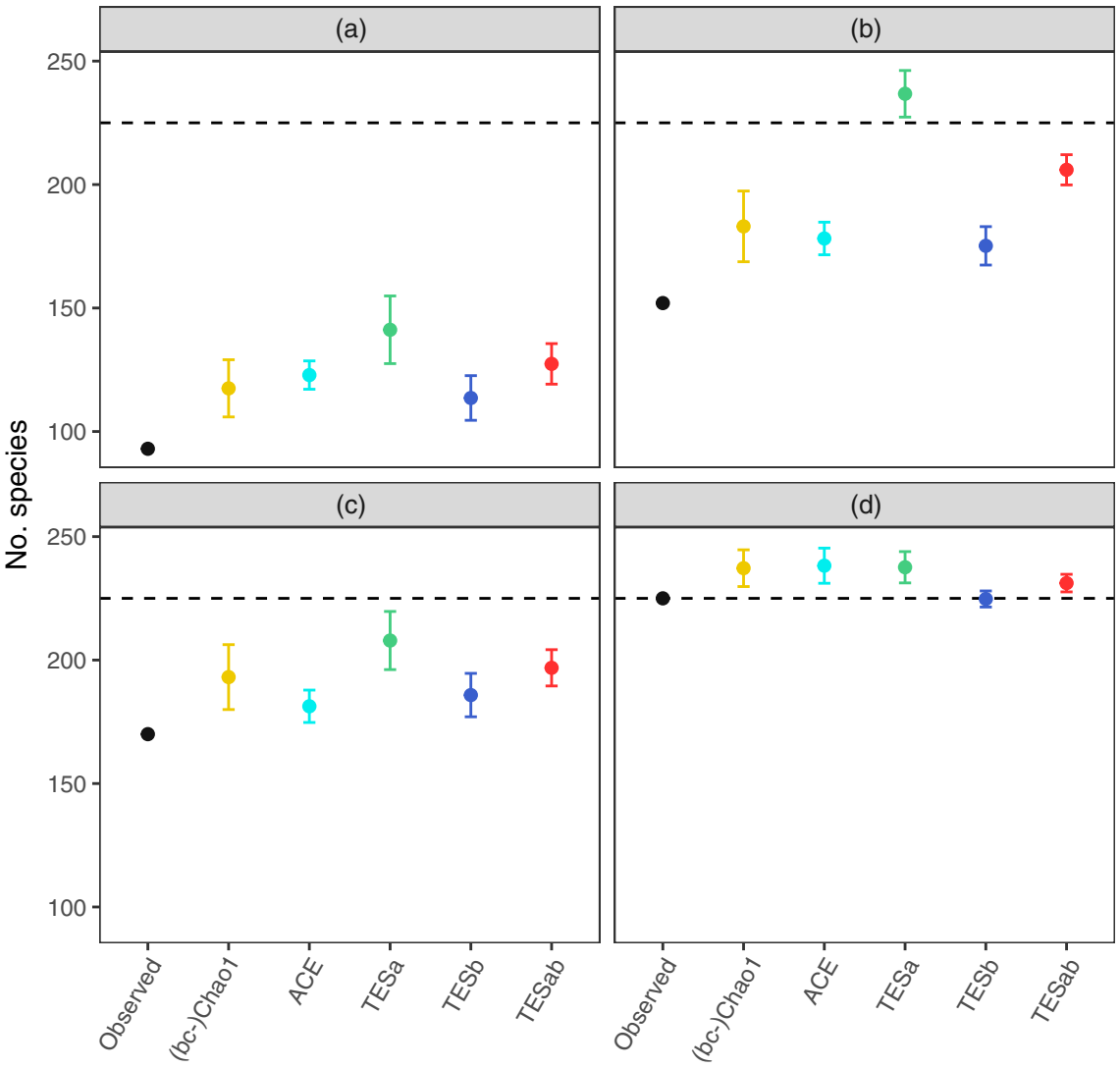

FIGURE 3 The observed and estimated number of species richness for different estimators for the BCI dataset based on the pooling of the first (a; 448 individuals and 93 species), the first five (b; 2359 individuals and 152 species), the first 10 (c; 4510 individuals and 170 species), and all (d; 21,457 individual and 225 species) sampling plots. The dashed lines refer to the total number of species. ACE, abundance-based coverage estimator; TES, Total Expected Species.

richness estimators. TESa appears to represent a promising estimator of the total species richness in communities following a geometric abundance distribution, while TESb appears to be a useful measure for samples reflecting the other distribution models. Results furthermore indicated a tendency for TESa to overestimate and for TESb to chiefly underestimate the overall species richness of a community represented by small, highly incomplete samples. We are not able to correct these biases from TESa and TESb, and we encourage further studies to investigate optimization methods for the mathematical curve-fitting that will further reduce these biases. The average value of these two approximation models, TESab, showed a good performance in all three criteria, bias, precision, and accuracy, when compared with commonly used (bc-)Chao1 and ACE estimators. Given the symmetric property for the "S" shape of the TES curve, a robust estimate more likely can be obtained when the observed value is greater than half of the model asymptotic value. We therefore suggest users to use TESb or TESab when the ratio of TESa to the observed richness is $\gg 2$.

To calculate TES, we used a curve-fitting method based on established rarefaction curves, that is, calculating the expected species for variations of given sample sizes, m, with saturation parameters. This leaves our approach fundamentally different from already widespread rarefaction and extrapolation curve-fitting approaches (Chao et al., 2014; Chao & Jost, 2012). In these latter cases, rarefaction and extrapolation do not generally result in saturation estimates, with estimates instead approximating infinity for

large sample sizes (Cayuela et al., 2015). In addition, our approach also inherently differs from other parametric curve-fitting methods that are based on sample-based species accumulation curves (either with asymptotic or not) (e.g., Flather, 1996; Rosenzweig et al., 2003). Although bootstrap subsampling could be used to smoothen the respective curves in these instances (Flather, 1996; Rosenzweig et al., 2003), this approach might result in a strong increase in uncertainty during subsampling (Ulrich et al., 2020).

Traditional individual-based curve-fitting approaches, either based on individual-accumulation curves (Palmer, 1990) or rarefaction curves (Mauffrey et al., 2007; O'Hara, 2005), have always strongly depended on prior decisions relating to the specific species distribution model (Chao & Chiu, 2016; Walther & Moore, 2005). Here we use the Weibull-logistic model based on the logarithm of the rarefied parameter (m) for the curve-fitting, showing that this Weibull-logistic model is more robust than the Michaelis–Menten model. Our simulation results demonstrate that the TES approach can be effectively used to estimate total species richness across all three common species' abundance distribution models even where samples are highly incomplete. These results mean that the TES approach has a relatively high tolerance for the species' abundance distribution, which indicates the potential of this new approach to serve as a good estimator, or, at least, as a viable alternative to nonparametric approaches.

Resampling from the pooled mite dataset shows an overestimation of TES, particularly for TESa, in comparison to the overall species richness. Nonetheless, it needs to be noted that here we assume the observed total number of species represents the true value and then conducted a resampling procedure. By pooling all study plots and conducting the resampling procedure, we also assume that subsamples are randomly collected from the community, which might not be the case. As the real total species richness is unknown and could indeed be higher than the total observed number, we are not able to fully judge whether the higher TES estimates show a true positive bias.

A further important novelty of our study is that we provide the variance for our parametric estimates. Parametric curve-fitting methods usually do not provide the variance, which is one of the reasons why they received more limited attention than nonparametric approaches (Chao & Chiu, 2016). Traditional bootstrapping methods to estimate the variance might work for nonparametric methods (e.g., Chao & Jost, 2012), but to take a subsample (even sampling with replacement at the same size as N_t) is problematic in this instance, as the resulting variance reflects the estimates from a subsample, and hence cannot be used for the calculation of the variance of the total species richness estimation (O'Hara, 2005). It needs to be noted that the variance of our estimates is generated based on the estimation of standard deviation (i.e., residual mean square) of the curve-fitting, which can be viewed as a prediction question (O'Hara, 2005). Our results show that the variance from TES estimates is comparable to the variance of nonparametric estimates such as (bc-)Chao1 and ACE. This robust estimated variance supports a wider potential application of the TES estimates in rigorous comparisons.

While our TES approach is based on individual-based rarefaction, TES for sample-based rarefaction (Pielou, 1975) might also be developed. Here we compared TES only with two commonly used species estimators, (bc-)Chao1 and ACE. We are aware that a wide variety of further estimators is available, whose performances have been the focus of a variety of previous studies (Beck & Schwanghart, 2010; Chiu et al., 2014; Colwell & Coddington, 1994; Hortal et al., 2006; Mao & Colwell, 2005; Reese et al., 2014; Rosenzweig et al., 2003). A comprehensive comparison of TES with the majority of existing estimators, however, is beyond the scope of our current study, which is chiefly based on the formulation and preliminary evaluation of the performance of TES. We now strongly encourage such wider comparisons based both on simulated and empirical data to be conducted in the near future.

The estimation of a community's species richness in target taxa from an incomplete sample in terms of the species pool remains a strong challenge. Overall, we showed that rarefaction theory in combination with parametric approximation models provides a valuable new approach to estimate the species richness of incompletely sampled communities. Our parametric method, TES, demonstrates a high tolerance to different species' abundance distribution models. Although TES is not impressively superior to traditional nonparametric approaches, the R functions we developed allow users to visualize the performance of the estimation. Therefore, we encourage researchers to consider this approach as a complement to current existing estimators in their biodiversity studies.

AUTHOR CONTRIBUTIONS

Yi Zou and Jan Christoph Axmacher conceived the idea. Yi Zou and Peng Zhao wrote the R script. Yi Zou did the analysis. Yi Zou and Jan Christoph Axmacher wrote the manuscript with contributions from Peng Zhao.

ACKNOWLEDGMENTS

This study was financially supported by the National Natural Science Foundation of China (31700363). We thank X. Y. Chen for advice on the graphic design of the manuscript; we thank Jan Beck for suggestions on the

draft manuscript and two anonymous reviewers for their helpful comments and feedback.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Data are already published, publicly available, and cited herein. Datasets "mite" and "BCI" from the "vegan" package (Oksanen et al., 2018, http://CRAN.R-project.org/package=vegan) are utilized for this research. Novel functions used in this study (Zou et al., 2022) are available from Dryad: https://doi.org/10.5061/dryad.2jm63xst2.

ORCID

Yi Zou https://orcid.org/0000-0002-7082-9258
Peng Zhao https://orcid.org/0000-0001-5267-9797
Jan Christoph Axmacher https://orcid.org/0000-0003-1406-928X

REFERENCES

Beck, J., and W. Schwanghart. 2010. "Comparing Measures of Species Diversity from Incomplete Inventories: An Update." *Methods in Ecology and Evolution* 1: 38–44.

Borcard, D., and P. Legendre. 1994. "Environmental Control and Spatial Structure in Ecological Communities: An Example Using Oribatid Mites (Acari, Oribatei)." *Environmental and Ecological Statistics* 1: 37–61.

Brose, U., N. D. Martinez, and R. J. Williams. 2003. "Estimating Species Richness: Sensitivity to Sample Coverage and Insensitivity to Spatial Patterns." *Ecology* 84: 2364–77.

Bunge, J., and M. Fitzpatrick. 1993. "Estimating the Number of Species: A Review." *Journal of the American Statistical Association* 88: 364–73.

Butler, B. J., and R. L. Chazdon. 1998. "Species Richness, Spatial Variation, and Abundance of the Soil Seed Bank of a Secondary Tropical Rain Forest." *Biotropica* 30: 214–22.

Cayuela, L., N. J. Gotelli, and R. K. Colwell. 2015. "Ecological and Biogeographic Null Hypotheses for Comparing Rarefaction Curves." *Ecological Monographs* 85: 437–55.

Chao, A. 1984. "Non-Parametric Estimation of the Number of Classes in a Population." *Scandinavian Journal of Statistics* 11: 265–70.

Chao, A., and C. Chiu. 2016. "Species Richness: Estimation and Comparison." In *Wiley StatsRef: Statistics Reference Online* 1–26. Hoboken, NJ: John Wiley & Sons.

Chao, A., N. Gotelli, T. C. Hsieh, E. Sander, K. H. Ma, R. K. Colwell, and A. M. Ellison. 2014. "Rarefaction and Extrapolation with Hill Numbers: A Framework for Sampling and Estimation in Species Diversity Studies." *Ecological Monographs* 84: 45–67.

Chao, A., and L. Jost. 2012. "Coverage-Based Rarefaction and Extrapolation: Standardizing Samples by Completeness Rather than Size." *Ecology* 93: 2533–47.

Chao, A., and S.-M. Lee. 1992. "Estimating the Number of Classes Via Sample Coverage." *Journal of the American Statistical Association* 87: 210–7.

Chiu, C.-H., Y.-T. Wang, B. A. Walther, and A. Chao. 2014. "An Improved Nonparametric Lower Bound of Species Richness Via a Modified Good–Turing Frequency Formula." *Biometrics* 70: 671–82.

Clench, H. K. 1979. "How to Make Regional Lists of Butterflies: Some Thoughts." *Journal of the Lepidopterists' Society* 33: 216–31.

Coddington, J. A., I. Agnarsson, J. A. Miller, M. Kuntner, and G. Hormiga. 2009. "Undersampling Bias: The Null Hypothesis for Singleton Species in Tropical Arthropod Surveys." *Journal of Animal Ecology* 78: 573–84.

Colwell, R. K., A. Chao, N. J. Gotelli, S. Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino. 2012. "Models and Estimators Linking Individual-Based and Sample-Based Rarefaction, Extrapolation and Comparison of Assemblages." *Journal of Plant Ecology* 5: 3–21.

Colwell, R. K., and J. A. Coddington. 1994. "Estimating Terrestrial Biodiversity through Extrapolation." *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* 345: 101–18.

Condit, R., N. Pitman, E. G. Leigh, Jr., J. Chave, J. Terborgh, R. B. Foster, P. Núnez, S. Aguilar, R. Valencia, and G. Villa. 2002. "Beta-Diversity in Tropical Forest Trees." *Science* 295: 666–9.

Cormack, R. M. 1989. "Log-Linear Models for Capture-Recapture." *Biometrics* 45: 395–413.

Fisher, B. L. 1999. "Improving Inventory Efficiency: A Case Study of Leaf-Litter Ant Diversity in Madagascar." *Ecological Applications* 9: 714–31.

Flather, C. 1996. "Fitting Species–Accumulation Functions and Assessing Regional Land Use Impacts on Avian Diversity." *Journal of Biogeography* 23: 155–68.

Franklin, I. R. 1980. "Evolutionary Change in Small Populations." In *Conservation Biology: An Evolutionary-Ecological Perspective*, edited by M. E. Soulé and B. A. Wilcox, 135–49. Sunderland: Sinauer.

Gotelli, N. J., and R. K. Colwell. 2001. "Quantifying Biodiversity: Procedures and Pitfalls in the Measurement and Comparison of Species Richness." *Ecology Letters* 4: 379–91.

Hortal, J., P. A. V. Borges, and C. Gaspar. 2006. "Evaluating the Performance of Species Richness Estimators: Sensitivity to Sample Grain Size." *Journal of Animal Ecology* 75: 274–87.

Hurlbert, S. H. 1971. "The Nonconcept of Species Diversity: A Critique and Alternative Parameters." *Ecology* 52: 577–86.

Jiménez-Valverde, A., and J. M. Lobo. 2006. "Establishing Reliable Spider (Araneae, Araneidae and Thomisidae) Assemblage Sampling Protocols: Estimation of Species Richness, Seasonal Coverage and Contribution of Juvenile Data to Species Richness and Composition." *Acta Oecologica* 30: 21–32.

Keating, K. A., and J. F. Quinn. 1998. "Estimating Species Richness: The Michaelis-Menten Model Revisited." *Oikos* 81: 411–6.

Mao, C. X., and R. K. Colwell. 2005. "Estimation of Species Richness: Mixture Models, the Role of Rare Species, and Inferential Challenges." *Ecology* 86: 1143–53.

Mauffrey, J.-F., C. Steiner, and F. Catzeflis. 2007. "Small-Mammal Diversity and Abundance in a French Guianan Rain Forest: Test of Sampling Procedures Using Species Rarefaction Curves." *Journal of Tropical Ecology* 23: 419–25.

McGeoch, M. A., M. Schroeder, B. Ekbom, and S. Larsson. 2007. "Saproxylic Beetle Diversity in a Managed Boreal Forest:

Importance of Stand Characteristics and Forestry Conservation Measures." *Diversity and Distributions* 13: 418–29.

McGill, B. J., R. S. Etienne, J. S. Gray, D. Alonso, M. J. Anderson, H. K. Benecha, M. Dornelas, B. J. Enquist, J. L. Green, and F. He. 2007. "Species Abundance Distributions: Moving beyond Single Prediction Theories to Integration within an Ecological Framework." *Ecology Letters* 10: 995–1015.

O'Hara, R. B. 2005. "Species Richness Estimators: How Many Species Can Dance on the Head of a Pin?" *Journal of Animal Ecology* 74: 375–86.

Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, et al. 2018. "Vegan: Community Ecology Package." R Package Version 2.5-6. http://CRAN.R-project.org/package=vegan.

Palmer, M. W. 1990. "The Estimation of Species Richness by Extrapolation." *Ecology* 71: 1195–8.

Pielou, E. C. 1975. *Ecological Diversity*. New York: John Wiley and Sons.

Preston, F. W. 1948. "The Commonness, and Rarity, of Species." *Ecology* 29: 254–83.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing. Version 3.5.2*. Vienna: R Foundation for Statistical Computing.

Reese, G. C., K. R. Wilson, and C. H. Flather. 2014. "Performance of Species Richness Estimators across Assemblage Types and Survey Parameters." *Global Ecology and Biogeography* 23: 585–94.

Rosenzweig, M. L., W. R. Turner, J. G. Cox, and T. H. Ricketts. 2003. "Estimating Diversity in Unsampled Habitats of a Biogeographical Province." *Conservation Biology* 17: 864–74.

Sanders, H. L. 1968. "Marine Benthic Diversity: A Comparative Study." *The American Naturalist* 102: 243–82.

Smith, W., and J. F. Grassle. 1977. "Sampling Properties of a Family of Diversity Measures." *Biometrics* 33: 283–92.

Soberon, M. J., and B. J. Llorente. 1993. "The Use of Species Accumulation Functions for the Prediction of Species Richness." *Conservation Biology* 7: 480–8.

Tipper, J. C. 1979. "Rarefaction and Rarefiction—The Use and Abuse of a Method in Paleoecology." *Paleobiology* 5: 423–34.

Ulrich, W., B. Kusumoto, S. Fattorini, and Y. Kubota. 2020. "Factors Influencing the Precision of Species Richness Estimation in Japanese Vascular Plants." *Diversity and Distributions* 26: 769–78.

Vester, H. F. M., D. Lawrence, J. R. Eastman, B. L. Turner, II, S. Calmé, R. Dickson, C. Pozo, and F. Sangermano. 2007. "Land Change in the Southern Yucatán and Calakmul Biosphere Reserve: Effects on Habitat and Biodiversity." *Ecological Applications* 17: 989–1003.

Walther, B. A., and J. L. Moore. 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: 815–29.

Zou, Y., and J. C. Axmacher. 2021. "Estimating the Number of Species Shared by Incompletely Sampled Communities." *Ecography* 44: 1098–108.

Zou, Y., P. Zhao, and J. C. Axmacher. 2022. "Estimating Total Species Richness: Fitting Rarefaction by Asymptotic Approximation." Dryad, Dataset. https://doi.org/10.5061/dryad.2jm63xst2.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Zou, Yi, Peng Zhao, and Jan Christoph Axmacher. 2023. "Estimating Total Species Richness: Fitting Rarefaction by Asymptotic Approximation." *Ecosphere* 14(1): e4363. https://doi.org/10.1002/ecs2.4363