Diversity and Distributions

WILEY

Diversity and Distributions

METHOD OPEN ACCESS

rarestR: An R Package Using Rarefaction Metrics to Estimate α- and β-Diversity for Incomplete Samples

Yi Zou[1] | Peng Zhao[1] | Naicheng Wu[2] | Jiangshan Lai[3] | Pedro R. Peres-Neto[4] | Jan C. Axmacher[5]

[1]Department of Health and Environmental Sciences, School of Science, Xi'an Jiaotong-Liverpool University, Suzhou, China | [2]Department of Geography and Spatial Information Techniques, Ningbo University, Ningbo, China | [3]College of Ecology and Environment, Nanjing Forestry University, Nanjing, China | [4]Department of Biology and Canada Research Chair in Biodiversity and Spatial Ecology, Concordia University, Montreal, Canada | [5]UCL Department of Geography, University College London, London, UK

Correspondence: Yi Zou (yi.zou@xjtlu.edu.cn)

Received: 1 May 2024 | **Revised:** 14 November 2024 | **Accepted:** 16 November 2024

Editor: Juliano Sarmento Cabral

Funding: The authors received no specific funding for this work.

Keywords: alpha-diversity | beta-diversity | dissimilarity | expected species | sample size | species composition | species richness

ABSTRACT

Aim: Species abundance data is commonly used to study biodiversity patterns. In this context, comparing α- and β-diversity across incomplete samples can lead to biases. Therefore, it is essential to employ methods that enable standardised and accurate comparisons of α- and β-diversity across varying sample sizes. In addition, biodiversity studies also often require robust estimates of the total number of species within a community and the number of species shared by two communities.

Innovation: Rarefaction methods are commonly used to calculate α-diversity for standardised sample sizes, and they can also serve as the basis for calculating β-diversity. In this application note, we present `rarestR`, a new R package designed for calculating abundance-based α- and β-diversity measures for inconsistent samples using rarefaction-based metrics. The package also includes parametric extrapolation techniques to estimate the total expected number of species within a community, as well as the total number of species shared between two communities. Additionally, `rarestR` provides visualisation tools for curve-fitting associated with these estimators.

Main Conclusions: Overall, the `rarestR` package is a valuable tool for comparing α- and β-diversity values among incomplete samples, such as those involving highly mobile or species-rich taxa. In addition, our species estimators offer a complementary approach to non-parametric methods, including the Chao series of estimators.

1 | Introduction

Measures of biodiversity can be split into two primary dimensions: α-diversity, which typically refers to species richness within a single community, and β-diversity, which quantifies changes in the species composition between communities. However, it is important to acknowledge that alternative definitions also exist (e.g., α-diversity describing the degree of entropy or species abundance patterns and β-diversity describing the ratio between regional (γ) and local diversity [see Whittaker 1960, 1972; Jost 2006; Jurasinski, Retzer, and Beierkuhnlein 2009; Tuomisto 2010]). In this study, we adopt

Yi Zou and Peng Zhao contributed equally to this study.

the conventional definitions, treating α-diversity as a measure of species richness within a sample and β-diversity as pairwise measures of (dis)similarity between samples.

Studies on biodiversity typically rely on two types of data: abundance-based data, where each individual represents a sampling unit, and incidence-based data, where observational events such as the number of traps or quadrats serve as the sampling units (Chao and Chiu 2016). A common challenge in biodiversity assessments is estimation bias, which arises from undersampling, that is, when fewer species are observed in a sample than are present in the community's species pool (Walther and Moore 2005; Coddington et al. 2009; Schroeder and Jenkins 2018). This bias impacts the comparison of both α- and β-diversity among different samples. To address this problem, it is critical to employ robust standardisation methods when comparing α- and β-diversity across incomplete samples (Coddington et al. 2009; Beck, Holloway, and Schwanghart 2013). Additionally, biodiversity studies also often rely on accurate estimation of the total number of species within a community and the total number of shared species between two communities, both derived from incomplete samples (Magurran 2004; Chao et al. 2005). Different types of data require distinct approaches to address estimation bias (Chao and Chiu 2016); in this context, we focus specifically on abundance-based data.

Rarefaction calculates abundance-based probability distributions for standardised sample sizes, providing the expected species (ES) richness for a given standardised sample size (Sanders 1968; Hurlbert 1971; Grassle and Smith 1976; Gotelli and Colwell 2001). This approach enables comparisons of α-diversity (as standardised species richness) across samples for a common standardised sample size. Expanding on this concept, Grassle and Smith (1976) introduced a measure for the Expected number of Species Shared (ESS) between two samples, along with its normalised form, NESS (Normalised ES Shared). This index can be further standardised as the Chord-NESS (CNESS), which facilitates the estimation of β-diversity (i.e., compositional dissimilarities between samples) for standardised sample sizes (Trueblood, Gallagher, and Gould 1994; Zou and Axmacher 2020; Zou 2021).

Employing asymptotic approximation to fit rarefaction curves for variable sample sizes, Zou, Zhao, and Axmacher (2023) have introduced a parametric index to estimate the Total ES (TES) richness within a community. Building upon the same mathematical principles, fitting the ESS curve also enables the estimation of the Total ESS (TESS) by two communities (Zou and Axmacher 2021).

This application note introduces a new R package, `rarestR`, designed to calculate both abundance-based α- and β-diversity measures for incomplete and inconsistent samples using rarefaction metrics. Additionally, it incorporates parametric extrapolation tools to estimate the total species within a single community and the total number of shared species between two communities, based on incomplete samples—referred to as TES and TESS values mentioned above. The package also supports the visualisation of curve-fitting for these estimators.

2 | Mathematical Description

2.1 | ES, ESS, NESS and CNESS

Hurlbert (1971) introduced the concept of an ES richness (ES) for randomly drawn subsamples of m individuals from a larger sample, based on a hypergeometric distribution, referred to as ESa (Equation 1).

$$\mathrm{ES}a_m = \sum_{k=1}^{S}\left[1-\frac{\binom{N-N_k}{m}}{\binom{N}{m}}\right] \tag{1}$$

where S represents the number of observed species in the sample, N stands for the total sampled number of individuals, N_k denotes the number of individuals for species k and m represents the standardised subsample size.

For communities containing an infinite number of individuals, Smith and Grassle (1977) proposed that the ES richness follows a multinomial distribution, referred to as ESb (Equation 2). The results of ESb are linked to Simpson's index at $m=2$ (i.e., $\mathrm{ES}b_2 = \mathrm{Simpson}+1$).

$$\mathrm{ES}b_m = \sum_{k=1}^{S}\left[1-\left(1-\frac{N_k}{N}\right)^m\right] \tag{2}$$

Building upon ESa, Grassle and Smith (1976) proposed a measure for the ESS between two communities. This concept involves randomly drawing m individuals from each sample, assuming a hypergeometric distribution, with parameters that vary according to properties of the samples (Equation 3):

$$\mathrm{ESS}_{ij|m} = \sum_{k=1}^{S}\left[1-\frac{\binom{N_{i*}-N_{ik}}{m}}{\binom{N_{i*}}{m}}\right]\times\left[1-\frac{\binom{N_{j*}-N_{jk}}{m}}{\binom{N_{j*}}{m}}\right] \tag{3}$$

where N_{i*} and N_{j*} represent the number of individuals in the samples representing sites i and j, respectively; and N_{ik} and N_{jk} are the number of individuals for species k in site i and j, respectively. Grassle and Smith (1976) also proposed an amended form of ESS based on ESb, but this has seldom been used—likely because it underestimates the probability of shared species from two random samples.

Grassle and Smith (1976) additionally introduced a normalisation of the ESS index based on the arithmetic mean, resulting in a distance measure with values ranging from 0 to 1, known as the NESS (Equation 4). The value of NESS is identical to the Horn-Morisita index (Morisita 1959) at $m=1$ (i.e., Horn-Morisita $=1-\mathrm{NESS}_{ij|2}$).

$$\mathrm{NESS}_{ij|m} = \frac{2\times\mathrm{ESS}_{ij|m}}{\mathrm{ESS}_{ii|m}+\mathrm{ESS}_{jj|m}} \tag{4}$$

Trueblood, Gallagher, and Gould (1994) further modified this index, based on the geometric mean, termed the CNESS (Equation 5):

$$\mathrm{CNESS}_{ij|m} = \sqrt{2 \times \left[1 - \frac{\mathrm{ESS}_{ij|m}}{\sqrt{\mathrm{ESS}_{ii|m} \times \mathrm{ESS}_{jj|m}}}\right]} \quad (5)$$

Although CNESS values range from 0 and $\sqrt{2}$, making them somewhat incompatible with other dissimilarity measures, Zou and Axmacher (2020) demonstrated that the function can be modified by removing the $\sqrt{2}$ multiplier, resulting in a measure named CNESS_a (Equation 6):

$$\mathrm{CNESS}_{a(ij|m)} = \sqrt{1 - \frac{\mathrm{ESS}_{ij|m}}{\sqrt{\mathrm{ESS}_{ii|m} \times \mathrm{ESS}_{jj|m}}}} \quad (6)$$

2.2 | TES And TESS

Zou, Zhao, and Axmacher (2023) proposed curve-fitting to model the relationship between the rarefaction curve (ES) and the standardised sample size (m), using either a four-parameter Weibull model (Equation 7) or a three-parameter logistic model (Equation 8) (i.e., the Weibull-logistic model):

$$\mathrm{ES}_m = a - b * e^{-c*M^d}, \text{where } M = ln(m) \quad (7)$$

$$\mathrm{ES}_m = \frac{a'}{1 + e^{\frac{b'-M}{c'}}} \quad (8)$$

where a and a' are the horizontal intercept values of the curve asymptotes that represent the Total Estimated Species richness (TES). The variance of this value is the estimated standard deviation (σ) from the model fit (Zou, Zhao, and Axmacher 2023). As ES has two different mathematical expressions, ESa and ESb, TES can be estimated separately from these two different models, resulting in TESa and TESb. The mean value of TESa and TESb, named 'TESab', provides a third measure, with a variance of $\sigma = \frac{1}{2}\sqrt{\sigma_a^2 + \sigma_b^2}$, where σ_a and σ_b are the standard deviations of TESa and TESb.

Similarly, the relationship between ESS and the standardised sample size m can be fitted using the Weibull-logistic model, as proposed by Zou and Axmacher (2021). This allows for the estimation of a Total number of ESS (TESS) between two communities, with the variance estimated.

3 | Package Overview

The `rarestR` package can be downloaded and installed from CRAN (https://cran.r-project.org/web/packages/rarestR/index.html). The package contains four main functions, namely, `es()`, `ess()`, `tes()` and `tess()`, and a training dataset, `share`. TES and TESS can be visualised via the `plot()` function. Here are the descriptions of these functions:

1. `es(x, m, method, MARGIN)` calculates the rarefied number of species based on ES richness (ES) measures. The input `x` is a vector or a matrix representing the number of individuals for each species in one (vector) or across multiple sites (matrix). Parameter `m` represents the standardised subsample size (number of individuals randomly drawn from the sample), which can be varied according to users' requirements. For ESa, `m` cannot be larger than the sample size. Argument `method` is the estimation approach of ES used, with two options 'a' and 'b' available to calculate ESa and ESb, respectively, with the default set as 'a', returning identical values to the `rarefy()` function in the 'vegan' package (Oksanen et al. 2018), but without providing a standard error. Argument `MARGIN` is a vector giving the subscripts which the function will be applied over, inherited from the `apply()` function.
2. `ess(x, m, index)` calculates the similarity between two samples based on the ESS measure, using abundance data for the species contained in each sample. The input `x` is a community data matrix (sample × species; samples representing local communities), of which the sample name is the row name of the matrix. Argument `m` is the standardised sample size, by default set to `m=1`. Rows with a total sample size $< m$ will be excluded automatically from the analysis. Parameter `index` is the distance measure used in the calculation, as one of the four options 'CNESS_a' (formula 6), 'CNESS' (formula 5), 'NESS' (formula 4) and 'ESS' (formula 3), with the default set as 'CNESS_a'. The function returns a pairwise distance matrix.
3. `tes(x)` estimates the number of TES based on TESa, TESb and their average value TESab. The input `x` is a data vector representing the number of individuals for each species. The function returns a list with a self-defined class 'rarestr', which contains a summary dataframe of the estimated values and their standard deviations based on TESa, TESb and TESab, and the detailed results of the models used in the estimation of TES, either 'logistic' or 'Weibull'.
4. `tess(x)` estimates the number of TESS between two samples. The input `x` is a data matrix for two samples representing two communities. The function returns a list with the self-defined class 'rarestr', which contains a summary dataframe of the estimated values and their standard deviations of TESS, and the detailed results of the model used in the estimation of TES, either 'logistic' or 'Weibull'.
5. We define an S3 method, creating a generic function `plot()` for visualising the fitted curve of the models for calculating TES and TESS when the input `x` is an object with the 'rarestr' class (i.e., an object returned by the `tes()` or `tess()` function) as defined by the rarestR package.

The package includes a dataset named `'share'`, consisting of three samples randomly drawn from three simulated communities. Each community consists of 100 species and approximately 100,000 individuals, following a log-normal distribution (mean = 6.5, SD = 1). The first community serves as the reference (i.e., fully randomly generated), while the second and third communities share 25 and 50 randomly selected species, respectively, with the reference community. A detailed description of the reference and scenario communities, along with the data

generation procedure, is provided in Zou and Axmacher (2021). The 'share' dataset represents a random subsample of 100, 150 and 200 individuals, randomly drawn from these three communities, containing 58, 57 and 74 species, respectively.

4 | Performance of Rarefaction-Based α-Diversity and β-Diversity

Many biodiversity studies aim to investigate the variation between samples and explain this variation using biotic and abiotic variables. A robust diversity index must accurately capture differences in diversity among samples, irrespective of their sample size.

We briefly evaluated the performance of rarefaction-based α-diversity and β-diversity measures available in the rarestR package using simulated data. Our goal was to assess how accurately and precisely these indices capture differences between samples of varying sizes. Additionally, we emphasise that the choice of indices depends on specific sampling scenarios. Comprehensive evaluations of the biodiversity metrics is beyond the scope of this application note (but see Beck and Schwanghart 2010; Zou and Axmacher 2020). This brief demonstration, comparing its performance with other metrics, aims to help users recognise that these metrics can be accurate. It also offers examples for effectively using the package.

For α-diversity, we tested the performance of ES*a* (Hurlbert rarefaction, Equation (1) and ES*b* (Smith and Grassle rarefaction), Equation (2) for their precision and accuracy in comparison to other α-diversity indices for samples with incomplete and inconsistent sizes, and how the performance changes with sample size. Two samples were randomly drawn from the simulated reference community (i.e., 100 species of 100,000 individuals, following a log-normal distribution). The first sample contained *n* individuals, while the second sample contained twice this original sample size (i.e., *2n* individuals), with *n* increasing from 10 to 150 randomly drawn individuals.

We contrasted the performance of ES*a* and ES*b* with the following indices: Shannon diversity, which is the exponential back-transformation of Shannon entropy (Jost 2006); Fisher's alpha (Fisher, Corbet, and Williams 1943), recognised as robust against differences in sample size (Brehm, Süssenbach, and Fiedler 2003), and the observed species richness. Additionally, we compared with two commonly used species richness estimators, the (bias-corrected) Chao1 lower boundary species richness estimator (O'Hara 2005) and the Jackknife estimator (1st order) that was considered accurate for low sample coverage (Brose, Martinez, and Williams 2003). We focus on the comparison of diversity indices as our primary scope, while keeping the species richness estimator comparisons in Appendix S1. For each index, we calculated the ratio between two samples. Since both samples were drawn from the same pool (i.e., no difference in the true α-diversity), the expected true ratio between two samples should be 1. We repeated the process 1000 times to obtain the mean and 95% quantile of the ratio for each of the α-diversity measures.

Results show that, in comparison, ES*a* is both accurate and precise in capturing the true differences between samples, even for sample sizes as low as $m = 10$ individuals for the data-sets used. In contrast, ES*b* and Shannon diversity tend to underestimate these differences, while Fisher's alpha overestimates them and demonstrates low precision (Figure 1a). Observed species richness shows the lowest accuracy, consistently underestimating the true differences. Jackknife and the Chao1 estimator both underestimate this difference with low precision (Appendix S1).

For β-diversity, we evaluated the performance of $CNESS_a$ (Equation 6) and NESS (Equation 4) using two different sample sizes: the minimum value ($m = 1$) and the maximum possible value (i.e., $m =$ maximum common sample size across samples). Two samples were randomly drawn from the previously described communities—the reference community and the second community, which shared 25 species with the reference. We randomly selected *n* individuals from the first community and 2*n* individuals from the second one, with *n* increasing from 10 to 150.

As the value of the ESS series depends on the parameter *m*, with a small *m* value emphasising similarities in the composition of abundant species, while a large *m* value leading to estimations of similarities in the overall community (Zou and Axmacher 2020), accuracy cannot be accessed.

Therefore, we focused in this instance on the 'stability' analysis for CNESS and NESS and then compared our results with two established dissimilarity indices: the widely used Jaccard (incidence-based) index, and the Bray–Curtis (abundance-based) index that is considered to have low sensitivity to sample size differences (Schroeder and Jenkins 2018). Stability was calculated based on the ratio (*R*) of the pairwise result at sample sizes of n individuals (D_n) to the result of a maximum 150 individuals (i.e., D_{150}), expressed as:

$$R_n = \frac{D_n}{D_{150}}$$

Unbiased results should again return a ratio for any sample size (R_n) of 1. We repeated the process 1000 times to obtain the mean and 95% quantile values for each index.

The results indicate that CNESS, at large *m* values, demonstrates greater stability and precision compared to both Bray–Curtis and Jaccard indices, particularly when the sample size exceeds 50 individuals. In contrast, CNESS at $m = 1$ was relatively unstable and imprecise (Figure 1b). NESS shows low precision across small and large *m* values, although the stability was higher for a larger *m* value (Appendix S2).

5 | Working Examples

We demonstrate here the use of functions in the rarestR package, applying both simulated and empirical data. The simulated data was sourced from the 'share' file, described previously. For the empirical data, we used the mite dataset available in the vegan package (Oksanen et al. 2018), which is comprised of 35 species of oribatid mites across 70 sites (communities). As the entire dataset (70) is too large to allow for a clear visualisation of the results, we analysed only the first 20 sites in the dataset (3447

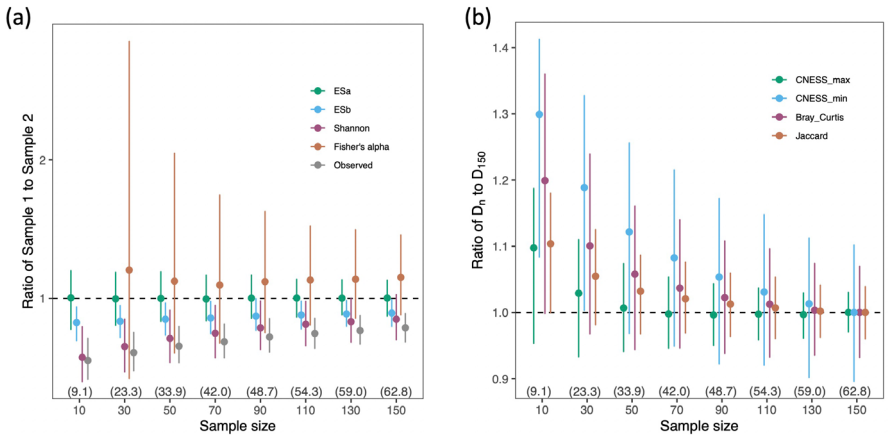

FIGURE 1 | (a) The change of the ratio between sample 1 and sample 2, both randomly collected from communities containing 100 species and approximately 100,000 individuals, for ES*a*, ES*b*, Shannon diversity, Fisher's alpha and observed species richness against the sample size (sample 2 is twice large as sample 1). The uncertainty of Fisher's alpha was too high at a sample size of 10, so it is not displayed; dashed line represents the actual ratio. (b) Stability analysis showing the ratio of dissimilarity at sample size n individuals (Dn) to the result of a maximum 150 of individuals (i.e., D150) for: CNESS at the smallest m (CNESS_min, $m=1$), largest m (CNESS_max, $m=n$), the Bray–Curtis index, and the Jaccard index. The simulation was based on two samples randomly drawn from two communities (each contains 100 species of approximately 100,000 individuals) that share a total of 25 species. Sample 1 drew n individuals from the first community, and sample 2 drew $2n$ individuals from the second community. In both cases, dots and error bars represent the mean and 95% quantiles from 1000 repetitions. Values in brackets of the x-axis refer to the mean percentage of sampling completeness, calculated as the proportion species sampled to the total number of species in the pool for a given sample size.

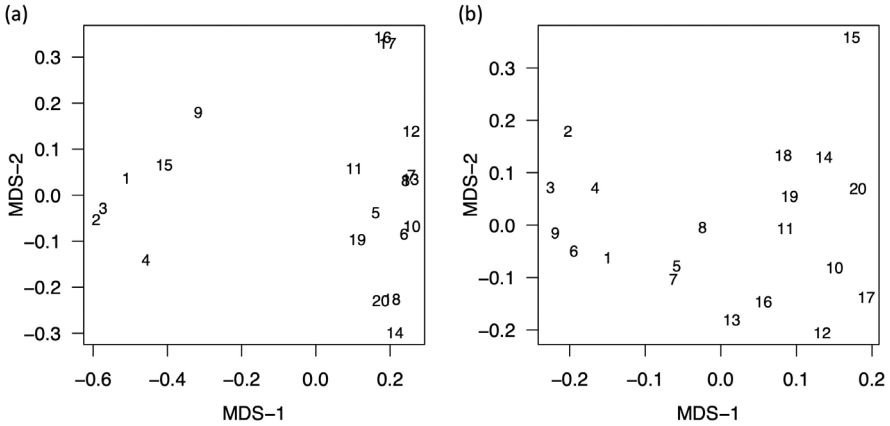

FIGURE 2 | MDS (multi-dimensional scaling) based on the CNESS (Chord-Normalised Expected Species Shared between two samples) dissimilarity measures for $m=1$ (a) and $m=90$ (b) for `mite` data in the `vegan` package; numbers represent the site ID.

individuals of 33 species) to improve the clarity of the visualisation in this demonstration.

```
#instal.packages("rarestR")
library(rarestR) # Version 1.1
library(vegan) # Version 2.6.2
data(share) # Load simulated data
data(mite) # Load empirical data
mite20<- mite[1:20,] # Only analysis the
first 20 sites.
```

5.1 | Function `es()`

We demonstrate the application of function es() for a maximum standardised value where no site (sample) is disregarded ($m=90$, i.e., the minimum sample size across all sites). When m exceeds the total sample size for a given sample, 'NA' will be returned by the software.

```
#Simulated data
es(share, m=100) # By default the method is
"ESa"
es(share, method="b", m=100) # Change the
method to "ESb"

#Empirical data
row.names(mite20)<- as.character(1:20)
min_m<- min(apply(mite20, 1, sum))
es(mite20, m=min_m) # m=90
es(mite20, m=150) # "NA" will be filled for
these sites<150 individuals.
```

5.2 | Function `ess()`

For ESS measures, we calculated the minimum standardised value, $m = 1$, and the maximum standardised value ($m = 90$) for CNESS and NESS measures. We then visualised the CNESS and NESS matrix results using classical multi-dimensional scaling (MDS, also known as principal coordinates analysis, PCoA).

MDS plots show diverging results for the two different standardised sample size values ($m = 1$ and $m = 90$) based on the CNESS dissimilarity matrices for the `mite` data. For this example, the results indicate a more homogeneous pattern when focusing on the overall community ($m = 90$) composition compared to the dominant species alone ($m = 1$) (Figure 2). The NESS matrices reflect a similar trend, showing that sites are more closely clustered when the analysis emphasises the dominant species (Appendix S3).

```
#Simulated data
ess(share) # By default the index is "CNESSa"
ess(share, index="NESS") # Change to "NESS"

#Empirical data
ess_m1<- ess(mite20, m=1) # m=1
ess_m90<- ess(mite20, m=min_m) # m=90

#NESS (Not run)
#ess_m1<- ess(mite20, m=1, index="NESS") #
m=1
#ess_m90<- ess(mite20, m=mim_m,
index="NESS") # m=90

#MDS for the CNESSa/NESS matrix and plot the
results
MDS_m1<- cmdscale(ess_m1, eig=TRUE)
df_m1<- as.data.frame(MDS_m1$points)
MDS_m90<- cmdscale(ess_m90, eig=TRUE)
df_m90<- as.data.frame(MDS_m90$points)
op <- par(mfrow=c(1, 2), mgp=c(2.5, 1, 0),
las=1, mar=c(4, 4, 2, 1))
with(df_m1, plot(x=V1, y=V2, type="n",
xlab="MDS-1", ylab="MDS-2"))
with(df_m1, text(x=V1, y=V2, labels=row.
names(df_m1)))
with(df_m90, plot(x=V1, y=V2, type="n",
xlab="MDS-1", ylab="MDS-2"))
with(df_m90, text(x=V1, y=V2, labels=row.
names(df_m90))).
```

5.3 | Function `tes()`

TES results (i.e., based on curve-fitting) for the simulated data show TES$a = 138.5$ and TES$b = 92.63$ (Figure 3), with a TES$ab = 115.56$ for the first sample. For TES measures of the empirical data, we calculated the value for pooled data of the 20 sites, which contains 33 observed species. Results show TES$a = 24.63$, TES$b = 34.14$ and TES$ab = 34.39$, which is very close to the overall species richness of the `mite` data.

```
#Simulated data (only for the first site)
Output_tes<- tes(share[1,])
Output_tes
plot(Output_tes).
#Empirical data
mite20pool<- apply(mite20,2,sum)
Output_tes_mite <- tes(mite20pool)
Output_tes_mite.
```

5.4 | Function `tess()`

The TESS value for the simulated data between the first and the second samples is 23.28, and that between the first and third sample is 40.16 (Appendix S4a,b), which are relatively close approximations of the 'real' values of 25 and 50 species. For the empirical data, to obtain a robust estimation of shared species, we calculated the estimated shared species between the two pooled groups: sites 1–10 and sites 11–20. The results indicate that the two pooled groups are expected to share 32.14 species (Appendix S4c).

```
#Simulated data
Output_tess12<- tess(share[c(1,2),])
```

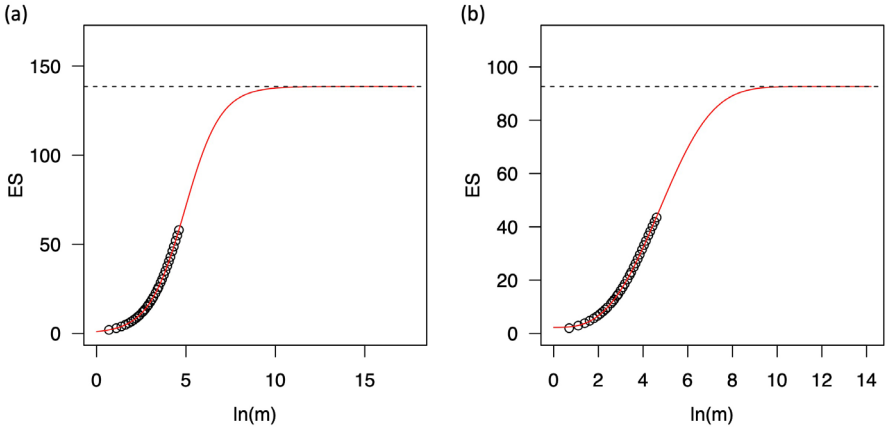

FIGURE 3 | Number of expected species (ES) based on ESa (a) and ESb (b) estimations versus the standardised sample size m for the simulated data of ~100,000 individuals split across 100 species. Solid lines refer to the model fit, and dashed lines refer to the total expected species (TES), that is, the asymptotic value of the model.

```
Output_tess13<- tess(share[c(1,3),])
Output_tess12
Output_tess13
plot(Output_tess12)
plot(Output_tess13).
#Empirical data
mite_pool <- rbind(apply(mite20[1:10,],2,-
sum),apply(mite20[11:20,],2,sum))
Output_tess_mite <- tess(mite_pool)
plot(Output_tess_mite).
```

6 | Discussion

The `rarestR` package integrates calculations of the ES number (i.e., rarefaction and extrapolation) for a single sample, as well as the ESS between communities represented by two communities, based on species abundance data. Both the `tes()` and `tess()` functions employ asymptotic approximations to extrapolate rarefaction curves, allowing for the estimation of the total number of species within single communities and the total number of species shared between pairs of communities.

As mentioned above, the `es()` function calculates a rarefaction value, where the calculation of 'ES*a*' is based on a hypergeometric distribution, which is identical to the `rarefy()` function from the `vegan` package (Oksanen et al. 2018). The calculation of 'ES*b*' is based on a multinomial distribution that is not available in the `vegan` package. However, when m exceeds the total sample size, the `es()` function returns 'NA'. This behaviour differs slightly from the `rarefy()` function, which returns the number of observed species. This distinction is intentional, highlighting the importance of excluding samples larger than the standardised sample size from comparisons. This approach aligns with the approach used in the β-diversity comparison performed by the `ess()` function. Our simulation results suggest that 'ES*a*' outperforms 'ES*b*' and other diversity indices in both precision and accuracy in detecting the true difference of species richness from incomplete samples. However, it is important to note that the rarefaction approach assumes that each individual has an equal probability of being sampled (e.g., random spatial distribution and detection probability). Biases may arise if individuals or species are distributed non-randomly in space (Engel et al. 2021) or exhibit varying activity patterns or capture rates for other reasons. In addition, here we tested a single species abundance distribution (SAD), while SAD can influence the results of rarefaction and other diversity indices (Maurer and McGill 2011; McGill 2011; Shimadzu 2018). Although the choice of a specific diversity index depends on the study's objectives and data structure, this topic has been widely discussed in the literature (Lamb et al. 2009; Beck and Schwanghart 2010; Alroy 2020; Qiao, Orr, and Hughes 2024) and is beyond the scope of this package's application.

The function `ess()` calculates the β-diversity based on an adjustable standardised sample size. Generally, β-diversity measures fall into two classes: direct calculation of the ratio between regional (γ) and local (α) diversity and multivariate measures based on pairwise (dis-)similarities (Jurasinski, Retzer, and Beierkuhnlein 2009; Anderson et al. 2011). The `ess()` function is based on the second case. Therefore, our ESS-based β-diversity estimate fundamentally differs from the recently developed β-diversity rarefaction and extrapolation methods in the package `iNEXT.beta3D` (Chao et al. 2023), as well as from the sample coverage-based rarefaction β-diversity proposed by Engel et al. (2021). Both approaches estimate β-diversity based on the ratio between estimated γ- to α- diversity, providing an average (regional) measure of β-diversity across all communities.

In contrast, our approach, implemented in the `ess()` function (index 'CNESS$_a$', 'CNESS' and 'NESS'), calculates β-diversity based on pairwise dissimilarities between communities represented by (incomplete) samples. The CNESS index, particularly for large m values, is less sensitive to variations in sample size compared to indices such as Bray–Curtis and Jaccard. Additionally, the results from NESS/CNESS can vary depending on the selected m value—smaller values emphasise dissimilarities among dominant species, while larger values increasingly reflect overall community similarities (Zou and Axmacher 2020). Therefore, we recommend using the ESS series indices, with both small and large m values, to provide a comprehensive interpretation of the results underlying community structures.

Our `tes()` function employs a parametric method to estimate the total number of species. This extrapolation approach differs from that used in the `iNEXT` package (Hsieh, Ma, and Chao 2016) and its extension, `iNEXT.3D` (to phylogenetic and functional diversity, Chao et al. 2021). The abundance-based method in `iNEXT` estimates species richness using non-asymptotic models; however, asymptotic values in `iNEXT` can be obtained based on diversity measures (see Hsieh, Ma, and Chao 2016). The `tes()` function calculates the total ES using an asymptotic parametric method that fits the rarefaction curve. Unlike observed species accumulation curves, which are often irregular and form the basis for non-parametric estimator development (Béguinot 2015), the rarefaction curve used here is smooth. This approach provides flexibility and robust applicability across various species abundance distributions models (Zou, Zhao, and Axmacher 2023), unlike traditional curve-fitting methods that rely on specific species abundance distributions (Walther and Moore 2005; Béguinot 2015). As a result, TES is comparable to non-parametric estimators such as Chao 1 and ACE (e.g., in `vegan` package) (Chao 1984; Chao and Lee 1992). Additionally, we provide visualisations of these estimations, offering a complementary approach to these non-parametric methods. For the curve-fitting, we set a default value of 40 knots (knots represent rarefied estimation points at evenly spaced intervals on a logarithmic scale between 1 and the endpoint, i.e. total number of individuals in the sample), recognising that different values may yield slightly different results. While a large number of knots might improve model fit and reduce standard error (Zou, Zhao, and Axmacher 2023), it can also increase estimation variance, and thus a trade-off that must be carefully considered.

Our `tess()` function calculates the estimated number of shared species between two communities using asymptotic parametric curve-fitting. To our knowledge, the only other shared species richness estimators available are Chao1-shared and ACE-shared in the `SpadeR` package (Chao et al. 2016). However, TESS generally outperforms these estimators in terms of precision and accuracy, particularly when dealing with unequal sample sizes

(Zou and Axmacher 2021). That said, estimating the number of shared species remains challenging and can result in significant uncertainties, especially when sample completeness is relatively low (Zou and Axmacher 2021). Similar to `tes()`, visualisations are available through the `plot()` function, allowing researchers to graphically interpret their curve-fittings. Integrating TESS with other species richness estimators can provide a more accurately estimation of true species (dis)similarities, based on both shared and unique species numbers, as outlined in Koleff, Gaston, and Lennon (2003). However, caution is advised when combining estimators, as this inadvertently reduces precision (Zou and Axmacher 2021).

Although species estimators are commonly used to account for varying sample sizes when comparing biodiversity across samples, their precision is often low, particularly for small sample sizes. As a result, we do not recommend over-relying on these estimators for comparing multiple samples. In our view, their primary utility lies in estimating sampling completeness within a specific target community. Only where estimated completeness is high should these estimators then be used to ascertain true species richness and similarity values. This reasoning informed our decision to design the `tes()` function for single-sample estimations and the `tess()` function for two-sample comparisons. For comparing multiple samples accounting for different sample sizes, we recommend using ES for α-diversity and CNESS for β-diversity. Therefore, both the `es()` and `ess()` functions are designed to handle multiple samples (communities) effectively.

In summary, the `rarestR` package is a valuable tool for ecologists studying α- and β-diversity. It is especially useful when dealing with incomplete and inconsistent sample sizes—a common issue in ecological community samples, particularly for highly mobile and species-rich taxa. The package also provides visual estimations of species richness and the number of shared species between two communities, based on individual samples. This approach complements non-parametric methods, such as the Chao series of estimators (Chao 1984; Chao et al. 2000, 2023).

Author Contributions

Y.Z. and J.C.A. conceived the ideas; Y.Z. wrote the main function and P.Z. contributed to it; P.Z. wrapped functions into the package; Y.Z. analysed the data and led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

Acknowledgements

We thank two reviewers for their constructive comments on this manuscript. We also appreciate RB Smith for his suggestion regarding the warning message incorporated within the package.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The package `rarestR` is available on CRAN (https://CRAN.R-project.org/package=rarestR). The data '`mite`' in our example is in the package `vegen`, which is available on CRAN (https://cran.r-project.org/web/packages/vegan).

Peer Review

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1111/ddi.13954.

References

Alroy, J. 2020. "On Four Measures of Taxonomic Richness." *Paleobiology* 46: 158–175.

Anderson, M. J., T. O. Crist, J. M. Chase, et al. 2011. "Navigating the Multiple Meanings of β Diversity: A Roadmap for the Practicing Ecologist." *Ecology Letters* 14: 19–28.

Beck, J., J. D. Holloway, and A. Schwanghart. 2013. "Undersampling and the Measurement of Beta Diversity." *Methods in Ecology and Evolution* 4: 370–382.

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

Béguinot, J. 2015. "When Reasonably Stop Sampling? How to Estimate the Gain in Newly Recorded Species According to the Degree of Supplementary Sampling Effort." *Annual Research & Review in Biology* 7: 300–308.

Brehm, G., D. Süssenbach, and K. Fiedler. 2003. "Unique Elevational Diversity Patterns of Geometrid Moths in an Andean Montane Rainforest." *Ecography* 26: 456–466.

Brose, U., N. D. Martinez, and R. J. Williams. 2003. "Estimating Species Richness: Sensitivity to Sample Coverage and Insensitivity to Spatial Patterns." *Ecology* 84: 2364–2377.

Chao, A. 1984. "Non-Parametric Estimation of the Number of Classes in a Population." *Scandinavian Journal of Statistics* 11: 265–270.

Chao, A., R. L. Chazdon, R. K. Colwell, and T.-J. Shen. 2005. "A New Statistical Approach for Assessing Similarity of Species Composition With Incidence and Abundance Data." *Ecology Letters* 8: 148–159.

Chao, A., and C. Chiu. 2016. *Species Richness: Estimation and Comparison*, 1–26. Wiley StatsRef: Statistics Reference Online.

Chao, A., P. A. Henderson, C.-H. Chiu, et al. 2021. "Measuring Temporal Change in Alpha Diversity: A Framework Integrating Taxonomic, Phylogenetic and Functional Diversity and the iNEXT.3D Standardization." *Methods in Ecology and Evolution* 12: 1926–1940.

Chao, A., W.-H. Hwang, Y. Chen, and C. Kuo. 2000. "Estimating the Number of Shared Species in Two Communities." *Statistica Sinica* 10: 227–246.

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

Chao, A., K. H. Ma, T. C. Hsieh, and C.-H. Chiu. 2016. "SpadeR: Species-Richness Prediction and Diversity Estimation With R." R Package Version 0.1.1.

Chao, A., S. Thorn, C.-H. Chiu, et al. 2023. "Rarefaction and Extrapolation With Beta Diversity Under a Framework of Hill Numbers: The iNEXT.beta3D Standardization." *Ecological Monographs* 93: e1588.

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: 573–584.

Engel, T., S. A. Blowes, D. J. McGlinn, et al. 2021. "Using Coverage-Based Rarefaction to Infer Non-Random Species Distributions." *Ecosphere* 12: e03745.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. "The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population." *Journal of Animal Ecology* 12: 42–58.

Gotelli, N. J., and R. K. Colwell. 2001. "Quantifying Biodiversity: Procedures and Pitfalls in the Measurement and Comparison of Species Richness." *Ecology Letters* 4: 379–391.

Grassle, J. F., and W. Smith. 1976. "A Similarity Measure Sensitive to the Contribution of Rare Species and Its Use in Investigation of Variation in Marine Benthic Communities." *Oecologia* 25: 13–22.

Hsieh, T. C., K. H. Ma, and A. Chao. 2016. "iNEXT: An R Package for Rarefaction and Extrapolation of Species Diversity (Hill Numbers)." *Methods in Ecology and Evolution* 7: 1451–1456.

Hurlbert, S. H. 1971. "The Nonconcept of Species Diversity: A Critique and Alternative Parameters." *Ecology* 52: 577–586.

Jost, L. 2006. "Entropy and Diversity." *Oikos* 113: 363–375.

Jurasinski, G., V. Retzer, and C. Beierkuhnlein. 2009. "Inventory, Differentiation, and Proportional Diversity: A Consistent Terminology for Quantifying Species Diversity." *Oecologia* 159: 15–26.

Koleff, P., K. J. Gaston, and J. J. Lennon. 2003. "Measuring Beta Diversity for Presence-Absence Data." *Journal of Animal Ecology* 72: 367–382.

Lamb, E. G., E. Bayne, G. Holloway, et al. 2009. "Indices for Monitoring Biodiversity Change: Are Some More Effective Than Others?" *Ecological Indicators* 9: 432–444.

Magurran, A. E. 2004. *Measuring Biological Diversity*. Oxford: Blackwell Publishing.

Maurer, B. A., and B. J. McGill. 2011. "Measurement of Species Diversity." In *Biological Diversity: Frontiers in Measurement and Assessment*. Edited by A. E. Magurran and B. J. McGill, 55–65. Oxford: Oxford University Press.

McGill, B. J. 2011. "Linking Biodiversity Patterns by Autocorrelated Random Sampling." *American Journal of Botany* 98: 481–502.

Morisita, M. 1959. "Measuring of Interspecific Association and Similarity Between Communities." *Memoirs of the Faculty of Science, Kyushu University Series E (Biology)* 3: 65–80.

O'Hara, R. B. 2005. "Species Richness Estimators: How Many Species Can Dance on the Head of a Pin?" *Journal of Animal Ecology* 74: 375–386.

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

Qiao, H., M. C. Orr, and A. C. Hughes. 2024. "Measuring Metrics: What Diversity Indicators Are Most Appropriate for Different Forms of Data Bias?" *Ecography* 2024: e07042.

Sanders, H. L. 1968. "Marine Benthic Diversity: A Comparative Study." *American Naturalist* 102: 243–282.

Schroeder, P. J., and D. G. Jenkins. 2018. "How Robust Are Popular Beta Diversity Indices to Sampling Error?" *Ecosphere* 9: e02100.

Shimadzu, H. 2018. "On Species Richness and Rarefaction: Size- and Coverage-Based Techniques Quantify Different Characteristics of Richness Change in Biodiversity." *Journal of Mathematical Biology* 77: 1363–1381.

Smith, W., and J. F. Grassle. 1977. "Sampling Properties of a Family of Diversity Measures." *Biometrics* 33: 283–292.

Trueblood, D. D., E. D. Gallagher, and D. M. Gould. 1994. "Three Stages of Seasonal Succession on the Savin Hill Cove Mudflat, Boston Harbor." *Limnology and Oceanography* 39: 1440–1454.

Tuomisto, H. 2010. "A Diversity of Beta Diversities: Straightening Up a Concept Gone Awry. Part 1. Defining Beta Diversity as a Function of Alpha and Gamma Diversity." *Ecography* 33: 2–22.

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: 815–829.

Whittaker, R. H. 1960. "Vegetation of the Siskiyou Mountains, Oregon and California." *Ecological Monographs* 30: 279–338.

Whittaker, R. H. 1972. "Evolution and Measurement of Species Diversity." *Taxon* 21: 213–251.

Zou, Y. 2021. "The Calculation of β-Diversity for Different Sample Sizes." *Biodiversity Science* 29: 790–797.

Zou, Y., and J. C. Axmacher. 2020. "The Chord-Normalized Expected Species Shared (CNESS)- Distance Represents a Superior Measure of Species Turnover Patterns." *Methods in Ecology and Evolution* 11: 273–280.

Zou, Y., and J. C. Axmacher. 2021. "Estimating the Number of Species Shared by Incompletely Sampled Communities." *Ecography* 44: 1098–1108.

Zou, Y., P. Zhao, and J. C. Axmacher. 2023. "Estimating Total Species Richness: Fitting Rarefaction by Asymptotic Approximation." *Ecosphere* 14: e4363.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.