ECOGRAPHY

Research

Estimating the number of species shared by incompletely sampled communities

Yi Zou and Jan Christoph Axmacher

Y. Zou (https://orcid.org/0000-0002-7082-9258) ✉ (yi.zou@xjtlu.edu.cn), Dept of Health and Environmental Sciences, Xi'an Jiaotong-Liverpool Univ., Suzhou, China. – J. C. Axmacher (https://orcid.org/0000-0003-1406-928X), UCL Dept of Geography, Univ. College London, London, UK.

Ecography
44: 1098–1108, 2021
doi: 10.1111/ecog.05625

Subject Editor: Luis Mauricio Bini
Editor-in-Chief: Miguel Araújo
Accepted 21 March 2021

There are numerous ways to estimate the true number of species in a community based on incomplete samples. Nonetheless, comparable approaches to estimate the number of species shared between two incompletely sampled communities are scarce. Here, we introduce the 'total expected species shared' (TESS) measure and provide the R function for its calculation. Based on parametric asymptotic models, TESS provides estimates of the true number of species shared between incompletely sampled communities based on abundance data. We compare TESS results with abundance-based non-parametric methods in terms of precision and accuracy, using different simulated sampling scenarios. We further calculate TESS using an empirical dataset, highlighting changes in accuracy and precision with increasing sample size. We also demonstrate how TESS values can be combined with species richness estimators in turnover estimates using traditional β-diversity indices. Our results show that mean values of TESS reliably approximate the true shared species number for varying sample completeness scenarios, with both accuracy and precision increasing with increasing sample completeness. Overall, we demonstrate the viability of TESS in estimations of the true number of species shared between two incompletely sampled communities. We also stress the importance of a sufficient sample size for the accuracy of estimates – requiring sampling designs that carefully balance sampling effort per site with the number of sampling sites.

Keywords: ACE-shared, β-diversity, Chao1-shared, sample size, TESS, under-sampling

Introduction

β-Diversity describes the change in the assemblage composition between different communities (Whittaker 1960) and is commonly measured as the change in species composition between pairs of samples, resulting in (dis)similarity matrices (Tuomisto 2010). Such (dis)similarity matrices are calculated using a variety of approaches that compare the observed number of species shared by, and unique to, the paired samples (Koleff et al. 2003, Baselga 2010).

Biodiversity studies commonly rely on highly incomplete samples (i.e. samples that do not contain all species present in the local community). Such 'undersampling'

represents a general problem for biodiversity assessments of hyper-diverse taxa and mega-diverse regions (Coddington et al. 2009, Schroeder and Jenkins 2018). For species richness assessments, a plethora of methods to estimate the number of species in a community from the number of species observed in an incomplete sample have been developed (Chao and Chiu 2016). Nonetheless, comparable approaches to estimate the number of species shared between two incompletely sampled communities are scarce (Beck et al. 2013). According to our knowledge, the only available approach has been a non-parametric method that uses frequencies of shared rare species (Chao et al. 2000), with subsequent suggested improvements (Pan et al. 2009). These calculations form extensions of the abundance-based coverage estimator (ACE, Chao and Lee 1992) that estimates the number of species in a community. Parametric asymptotic models based on curve fitting, which have been used to estimate species richness within a single community (Flather 1996, Rosenzweig et al. 2003), have nonetheless never been explored to estimate shared species between two communities.

We now introduce a novel approach to estimate the number of species shared between two communities based on abundance data and simple probability calculations that works for low sampling completeness scenarios. This approach is based on expected number of species shared (ESS) between two communities represented by samples of standardized size m (Morisita 1959, Grassle and Smith 1976, Trueblood et al. 1994, Zou and Axmacher 2020). It assumes that the probability for a species to be shared by two communities, but to be missing from randomly drawn samples of these communities, follows a hypergeometric distribution. In this regard, it resembles rarefaction approaches used to estimate the species richness for a standardized sample size (Hurlbert 1971). The measure of ESS can be expressed as

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

where m represents the standardized sample size (number of individuals) used for the comparison; S represents the total species number; and for samples i and j representing the compared communities, N_{i*} and N_{j*} are the total sample sizes, and N_{ik} and N_{jk} represent the abundances of the kth species in samples i and j. To calculate this measure, the user needs to specify a standardized sample size (m) for which the expected number of species shared between the two samples is estimated. The ESS-concept has been expanded to create further measures such as the chord-normalized expected species shared (CNESS) index that work robustly across sample sizes in estimating the relative dissimilarity between samples of standardized size (Zou and Axmacher 2020). Variation of the standardized sample size m in ESS-based dissimilarity measures allows for an emphasis to be put either on dominant species (using low values of m) or more strongly on the overall composition of the sampled communities (for high values of m). Nonetheless, the maximum value for m has remained limited to the size of the smallest sample.

In the following sections, we demonstrate a way to expand the value-range for m in probabilistic calculation-based ESS measures. This expansion allows us to estimate the total number of species shared by two communities that are represented by incomplete samples, and we call this estimate the 'total expected species shared' (TESS). TESS is generated by combing the ESS with asymptotic extrapolation models. After introducing this measure, we show changes in accuracy and precision of TESS with increasing sample size based on simulated data, and we compared the TESS results with two abundance based non-parametric methods – the abundance-based coverage estimator (ACE)-shared species (Chao et al. 2000), and the Chao1-shared species estimator (Pan et al. 2009). We further calculate the TESS for an empirical dataset, again showing changes in accuracy and precision with increasing sample size. We also briefly demonstrate how TESS values can be combined with standard species richness estimators to estimate true similarity based on traditional measures such as the Jaccard index. We finally discuss performance and applications of TESS.

Material and methods

For this study, we started with creating four simulated communities following specific distribution models: one baseline or 'control' community and three different 'scenario' communities that shared either 25 (scenario S25), 50 (scenario S50) or 75 (scenario S75) species with the 'control' community. In a second step, we then developed the probability-based models for the calculation of TESS. In a third step, we evaluated the performance of TESS in comparison with other indices based on the simulated communities and on empirical data. Finally, we demonstrated the application of TESS in estimations of the true similarity in the species pool between two incompletely sampled communities.

Simulated communities

The four simulated communities used in this study each contained approximately 100 000 individuals distributed among 100 species. Their structure followed a log-normal distribution model ($\log_e$ SD = 1, $\log_e$ mean = 6.5). We parameterized the model so that the minimum abundance of any species in any of the resulting four communities exceeded 50 individuals to fit the assumed threshold for the minimum population size required for the long-term survival of a species (Franklin 1980). After creating the first (baseline or 'control') community, we selected a set numbers of species (25, 50 and 75 species) from this community that were to be shared with the three 'scenario' communities (S25, S50 and S75, respectively). To determine which species specifically are shared, we

randomly selected individuals from the 'control' community and recorded their respective species names until this process reached the required number of species. In both the 'control' and respective 'scenario' community, we then ranked all species from the most to least abundant, and we ensured that the shared species selected from the 'control' population occupied the same abundance rank in the 'scenario' community, while all other species in the 'scenario' communities were allocated new species names that differed from the names of the 'control' community.

In addition to the log-normal distribution model, we repeated the generation of communities containing 100 species and approximately 100 000 individuals with a 50 individuals-minimum threshold for each species, for communities following negative binomial (mean = 1000, size = 3) and geometric abundance distribution (probability of success = 0.001) models. We created two communities following each model, one a 'control' and the other as S50 (50 shared species) to gain further insights into the performance of the TESS approach for these different community structures (Supporting information).

The conceptual framework for TESS

The relationship between ESS and the (log-) sample size (m) for each simulation scenario can be approximated by the following model (Fig. 1):

$$\mathrm{ESS} = a - b \times e^{-c \times M^d}, \text{ where } M = \log(m) \quad (1)$$

In this model, the asymptotic value a represents the estimated overall number of species shared between the two communities. A robust estimation of this number based on the Weibull model requires a high sampling completeness, as four parameters that determine the curve need to be estimated. Nonetheless, when the sampling completeness is relatively low, the asymptotic value can also be estimated using a three3-parameter logistic regression model (Fig. 1; Supporting information):

$$\mathrm{ESS} = \frac{a'}{1 + b' \times e^{-c' \times M}}, \text{ where } M = \log(m) \quad (2)$$

where a' again represents the total number of expected species shared (TESS) between the two communities. We developed an R function to calculate these asymptotic values (a and a'). The function was based on the increase of ESS values with increasing sample size (m), calculated for a specific number of values for (m) (default: 40) that were equally distributed between the value of $\log_e(m) = 1$ and the actual size of the randomly drawn sample. The values of ESS for the increasing values of m created 40 'knots' used to fit the respective curves (Fig. 1). We then used the resulting relationship to approximate the value of the parameters defining our four-parameter Weibull model and estimated the parameters of the model

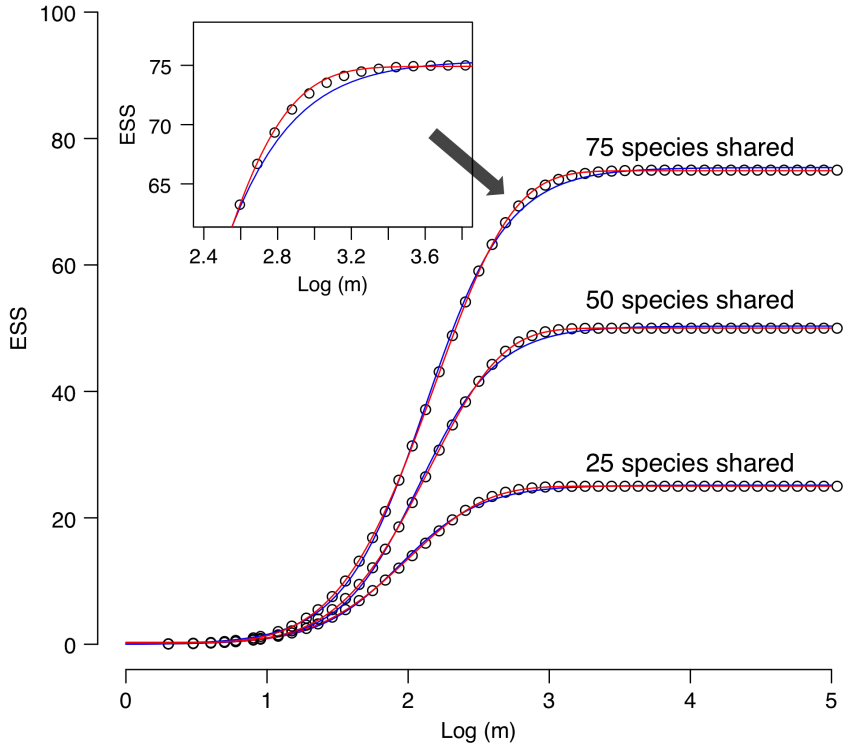

Figure 1. Change of the ESS value along with the change of m on a logarithmic scale for different 'scenario' groups. Points represent observed values, while the lines refer to asymptotic regression models based on 4 (red) and 3 (blue) parameters used in the parametrization of the respective models.

equation by nonlinear least-squares (Crawley 2012), in turn obtaining the value of TESS. Initial simulations for scenario S50 (log normal-distributed abundance patterns) showed that the four-parameter Weibull model effectively estimates the true number of shared species for sample sizes > 400 individuals, whereas three-parameter logistic regression models tend to slightly overestimate the shared species number for large sample sizes (Supporting information). In the TESS function that we developed (Supporting information), the calculation of a four-parameter Weibull model is therefore set as default option. Nonetheless, in cases where the sample size was too low to calculate the Weibull model, the curve was fitted using the three-parameter logistic regression model.

While we present results for communities with log normal-distributed abundance patterns in the main text, the curve-fitting methods were also shown to perform well for negative binomial or geometric species abundance distribution models (Supporting information).

Evaluation of TESS

To evaluate the overall performance of TESS, we compared its estimations with estimations for shared species based on the ACE-shared species estimator (Chao et al. 2000) and the Chao1-shared species estimator (Pan et al. 2009). We assumed two sampling scenarios: 'equal sampling' and 'unequal sampling'. For the 'equal sampling' scenario, samples taken from both 'scenario' and 'control' groups were of equal size, with sample sizes of the two samples varying from 25 to 1000 individuals. For 'unequal sampling', the control sample was set constant at 200 individuals, while the 'scenario' samples varied from 25 to 1000 individuals.

We repeated this approach for 1000 iterations (randomly drawn sub-samples from the underlying simulated communities/empirical samples) to evaluate the precision and accuracy of TESS, ACE-shared and Chao1-shared species estimators. Precision was calculated as coefficient of variation (CV) and accuracy was calculated as scaled root mean square error (SRMSE) (Walther and Moore 2005), expressed as:

$$\mathrm{CV} = \frac{1}{\bar{E}}\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(E_j - \bar{E}\right)^2}, \text{ and}$$

$$\mathrm{SRMSE} = \frac{1}{A}\sqrt{\frac{1}{n}\sum_{j=1}^{n}\left(E_j - A\right)^2}$$

where E_j is the estimated shared number of species for the jth simulation, $\bar{E}$ is the mean value of the estimated shared species richness, n is the number of simulations and A is the actual shared number of species.

Testing TESS using an empirical dataset

We selected a dataset for oribatid mite communities from a spruce forest stands in the Tharandter Wald, Germany (Zaitsev et al. 2002) as empirical dataset (available from <https://doi.org/10.5519/0066354>, dataset Source_ID 'CM1_2002_Zaitsev', Hudson et al. 2016, 2017). We selected three samples, with different sample sizes (135 150, 64 906 and 50 048 individuals representing 53, 51 and 38 species, respectively). In total, the three samples contained 71 species, and the lowest abundance of any species within these three samples was 34 individuals. We considered these samples to be complete, i.e. we assumed that they contained all species found in the local mite communities. We then took random subsamples of set sizes from each of these three samples and calculated the TESS for each pair of subsamples, before comparing these TESS values with the known number of species shared by the samples. We repeated this approach for varying sub-sample sizes to again generate insights into the sample size-specific performance of TESS.

When evaluating this performance of TESS on the empirical data, we used the same two sampling scenarios as for the simulated data, i.e. an 'equal sampling' and an 'unequal sampling' scenario, with sample sizes ranging from 25 to 1000 individuals. While the minimum sample size of 25 is too low to realistically estimate the species shared between the underlying communities, we set this low value in order to obtain insights into the overall performance of TESS across a wide range of sample sizes. Resampling and calculation was repeated for 1000 iterations again to generate values for the mean and 95% confidence intervals (CI). We plotted mean and 95% CI against the sample size for TESS.

Extending the use of TESS to estimate the true similarity of communities

The calculation of measures describing (dis)similarity (β-diversity) such as the Jaccard or Sørensen indices are currently based on ratios between the number of species shared between, and unique to, two samples that were randomly taken from underlying communities (Koleff et al. 2003). Normally, these measures rely on the species actually observed in each sample (but see Chao et al. 2006). They are therefore heavily influenced by sample completeness, and they commonly heavily underestimate similarity where sampling completeness is low (Chao et al. 2006). Nonetheless the same formulae used in these measures can also be employed in the context of estimated values for both, the number of shared and unique species. In this case, the species richness of a community can be estimated from an incomplete sample using a variety of species richness estimators, such as the abundance-based coverage estimator (ACE, Chao and Lee 1992), the Chao1 estimator (Chao 1984) or iNEXT (iNterpolation/EXTrapolation) techniques (Chao and Jost 2012, Chao et al. 2014, Hsieh et al. 2016). These values can now be combined with the TESS value. This approach allows us to estimate the similarity between incompletely sampled communities. To exemplify this approach, we use the ACE estimator in combination with the TESS measure to estimate the similarity based on

Jaccard similarities for incomplete samples. Other species richness estimators can easily replace ACE in such calculations. In the example we illustrate here, we furthermore used the simulated data to estimate the similarity both for the 'equal sampling' and 'unequal sampling' scenarios, with sample sizes ranging from 25 to 1000 individuals, and repeated these calculations for 1000 iterations to calculate mean similarity values and their 95% confidence intervals.

All calculations were based on the R software (<www.r-project.org>). The functions 'SSweibull()' and 'SSlogis()', from the 'stats' package, were used to calculate the Weibull and logistic regression models. The function 'nls()' was used to determine the non-linear least-square estimates of the model parameters. The package 'SpadeR' (Chao et al. 2016) was used to calculate ACE-shared and Chao1-shared species estimators and the package 'vegan' (Oksanen et al. 2018) to calculate ACE. We provide all R functions and scripts required for the calculation of TESS (Supporting information) and for our simulations (Supporting information).

Results

For two samples randomly drawn from the simulated communities, the 'total expected number of species shared' (TESS) between control and respective scenario community was generally much closer to the true number of species shared between the two underlying communities than the number of species shared between the two samples (Fig. 2). Crucially, this was true even for a small sample completeness. Comparing the two sampling scenarios, TESS had a larger variance, but an estimated mean closer to the real value in equal as compared to unequal sampling (Fig. 2). TESS values increased strongly with increasing sample size for very low sample completeness (0.2–0.5, i.e. from 25 to 100 individuals randomly taken from the two baseline communities of approximately 100 000 individuals) and rapidly approached the true number of species shared when sample sizes increased further (Fig. 2). The variance of TESS exceeded that of the observed number of shared species across all measurements (see 95% CI ranges in Fig. 2). Nonetheless, TESS variance

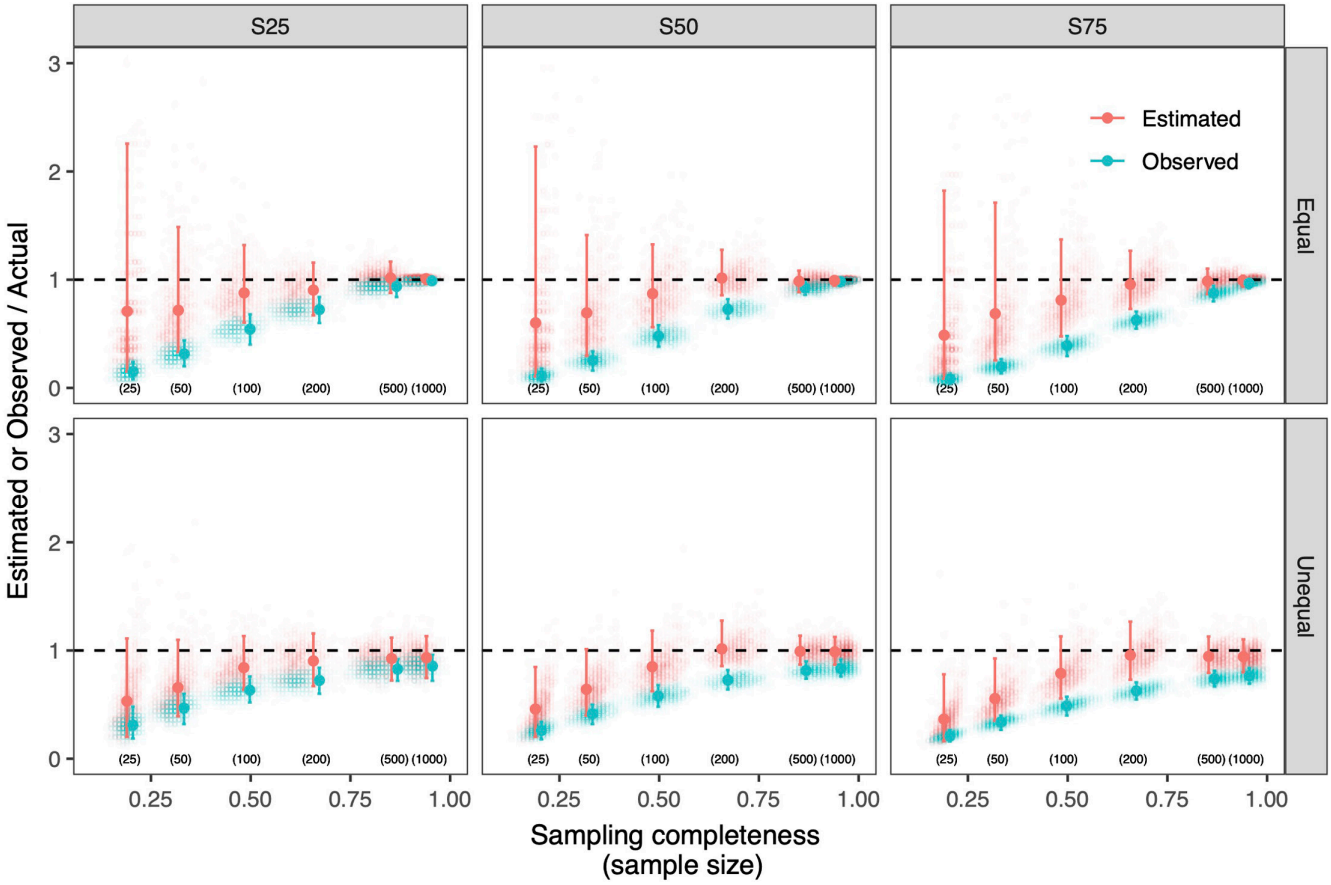

Figure 2. Ratio between the estimated or observed number and the actual number of species shared between two communities, plotted against sampling completeness (as total number of species contained in the treatment sample/total number of species in the underlying community – actual sample sizes are provided in brackets below the data points for further guidance). Panels refer to plots for three different sharing scenarios (S25: 25% of species shared; S50: 50% of species shared; S75: 75% species shared) and two sampling scenarios – equal and unequal – of the simulated data; error bars represent 95% confidence intervals (CIs); dots represent individual simulations.

decreases dramatically with increasing sample size (Fig. 2). Overall, our results indicate that TESS can robustly estimate the actual number of species shared between two communities for a sampling completeness > 0.5 (in our case, this is reached at ~ 100 individuals, Fig. 2). When the sampling completeness is higher than 0.7 (in our case at ~ 200 individuals), TESS is a highly accurate estimator (Fig. 2), with results for scenario 'S25' being slightly more accurate than results for the other two scenarios ('S50' and 'S75').

TESS had a similar precision (CV) to both ACE-shared and Chao1-shared estimators for equal sampling scenarios. TESS performed worse than these two measures only for scenario S25 when the sample size was extremely low (25 individuals) (Fig. 3). For the unequal sampling scenarios, TESS had a higher precision than both the ACE-shared and Chao1-shared estimators across almost all scenarios and sample sizes (Fig. 3). For accuracy (SRMSE), TESS again showed a similar performance to ACE-shared and Chao1-shared estimators for equal sampling scenarios (Fig. 4). For the unequal sampling scenario, TESS had a higher accuracy than both, Chao1-shared and ACE-shared estimations, particularly once the sampling size exceeded 100 individuals (Fig. 4).

For the empirical data, TESS values showed a larger variance, but a more accurate mean value when compared to the observed number of species shared by the samples from the underlying 'communities' (Fig. 5). Comparisons of results for the two sampling scenarios showed the same trends as for the simulated data, with equal sampling generally resulting in a larger variance, but a mean estimated value that is slightly closer to the real value than for unequal sampling (Fig. 5).

When combining TESS with the ACE-based species richness estimator to obtain the estimated similarity for the Jaccard index (Fig. 6), the mean similarity values particularly for small sample sizes provide a greatly superior approximation of the real similarity when compared to similarity measures based on the observed species' distributions in incomplete samples. Nonetheless, the variance of estimated similarity values is much higher than for the observed values. Furthermore, reflecting the clear trends observed in TESS values, equal sampling results in a higher variance of the resulting estimation, but a mean value that is closer to the real similarity value than unequal sampling scenarios for small sample sizes (i.e. < 200 individuals). Both estimated and observed similarities closely approximate the real similarity for large sample sizes (n > 500 individuals, Fig. 6).

Discussion

Performance of TESS

The results we obtained support the viability of our new measure to estimate the total number of expected species shared (TESS) between two incompletely sampled communities,

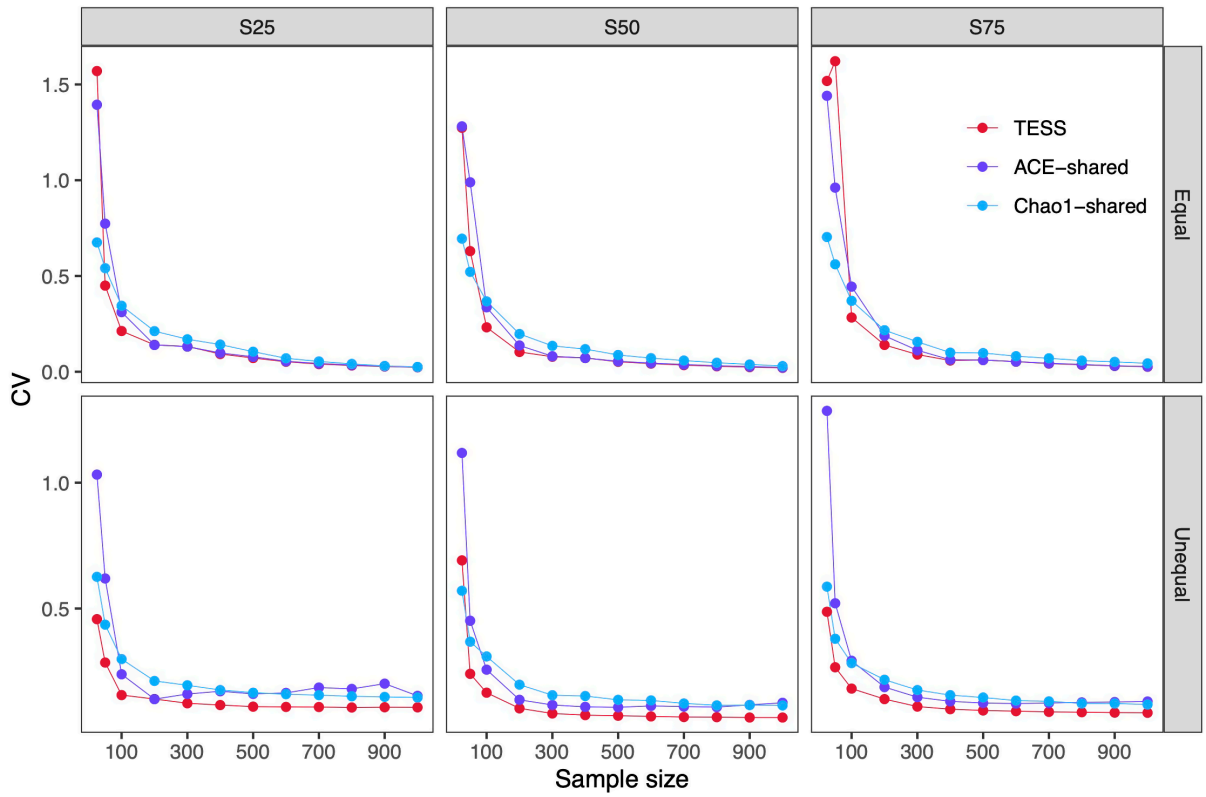

Figure 3. Relationship between sample size and the coefficient of variation (CV) for TESS, ACE-shared and Chao1-shared estimators for three different sharing scenarios (S25: 25% of species shared; S50: 50% of species shared; S75: 75% species shared) and two sampling scenarios of the simulated data.

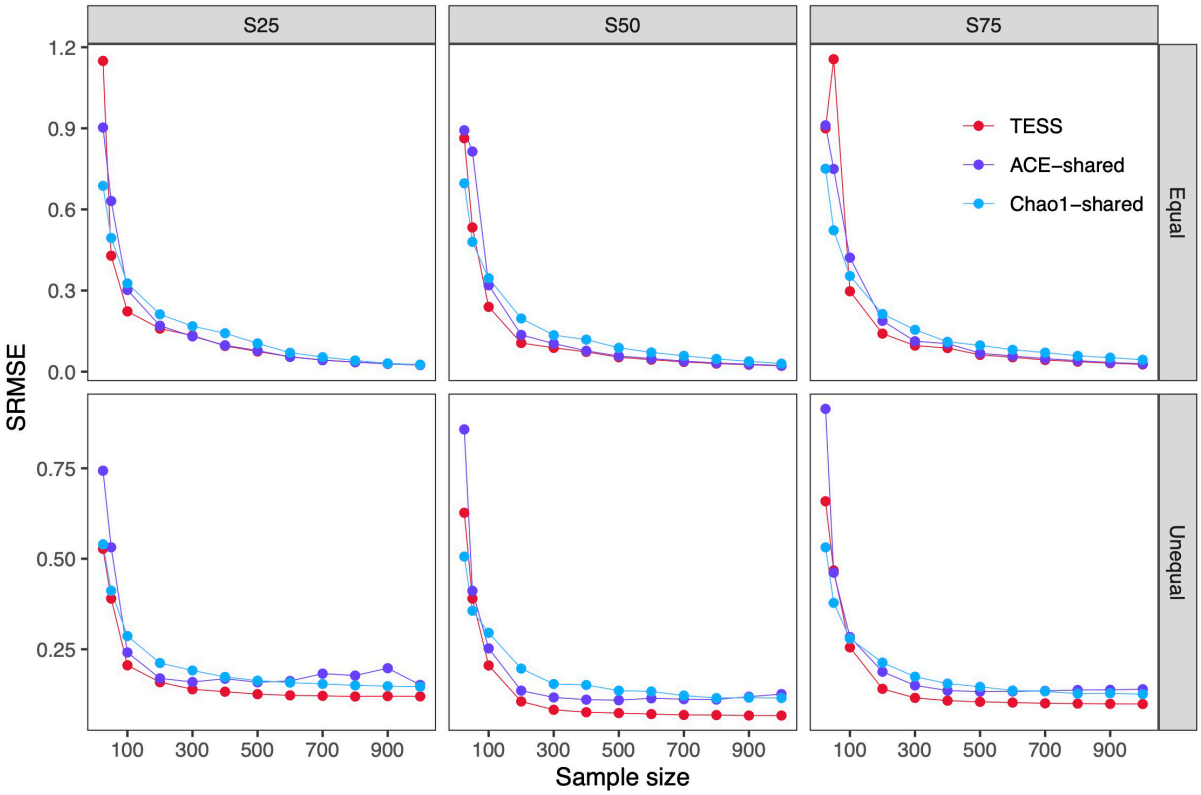

Figure 4. Relationship between sample size and the scaled root mean square error (SRMSE) for TESS, ACE-shared and Chao1-shared estimators for three different sharing scenarios (S25: 25% of species shared; S50: 50% of species shared; S75: 75% species shared) and two sampling scenarios of the simulated data.

with mean TESS values robustly approximating the true number of species shared for a wide range of sample sizes, particular for scenarios based on samples of equal sizes. It needs to be emphasized that TESS was specifically developed for abundance-based datasets and cannot be used for presence/absence data.

To our knowledge, TESS represents the first parametric curve-fitting method allowing for estimations of the true number of species shared between two incomplete communities, using probability estimations combined with asymptotic models. Similar to other parametric curve-fitting methods, the variance for the asymptote of TESS cannot be mathematically generated (Chao and Chiu 2016), preventing rigorous statistical comparisons between the two paired samples. While variance values can be generated for such non-parametric approaches (Chao et al. 2000, Pan et al. 2009, Chao and Chiu 2012), the high uncertainties in estimated results usually meant researchers showed little interest to draw direct statistical conclusions. TESS provides a clear pathway that is both easy to understand and to calculate, while providing results that – for intermediate to high sample completeness – are superior to existing approaches such as the ACE-shared and Chao1-shared estimators. TESS uses 'knots' that describe distinct points of the curve that shows the changes in ESS with increasing sample sizes (m). The resulting 'smooth' curve through the 'knots' is then expressed by a mathematical formula describing the total number of species shared between two communities as its asymptote. In the R script (Supporting information), we also provide a curve-fitting routine of the model that provides an R^2 value of the model fit.

Parametric asymptotic models based on curve-fitting have already been used to estimate species richness for abundance data before (Flather 1996, Rosenzweig et al. 2003), but not to estimate the shared number of species. In TESS, both Weibull and three-parameter logistic regression models provide accurate results. While the four-parameter Weibull model estimations are slightly more accurate (Supporting information), it is not mathematically meaningful to use these models for small sample sizes. The three-parameter model in contrast is applicable for small sample sizes, too, but tends to overestimate the actual number of shared species for large sample sizes. In our R script for the TESS function (Supporting information), the default setting is to calculate the four-parameter Weibull model, while the script automatically shifts to the alternative, three-parameter logistic regression model if the sample size is insufficient to calculate a Weibull model. Our function additionally allows users to override this approach and manually select a three- or four-parameter model approach.

With ESS already accounting for effects of sample size and completeness through the inclusion of the standardized

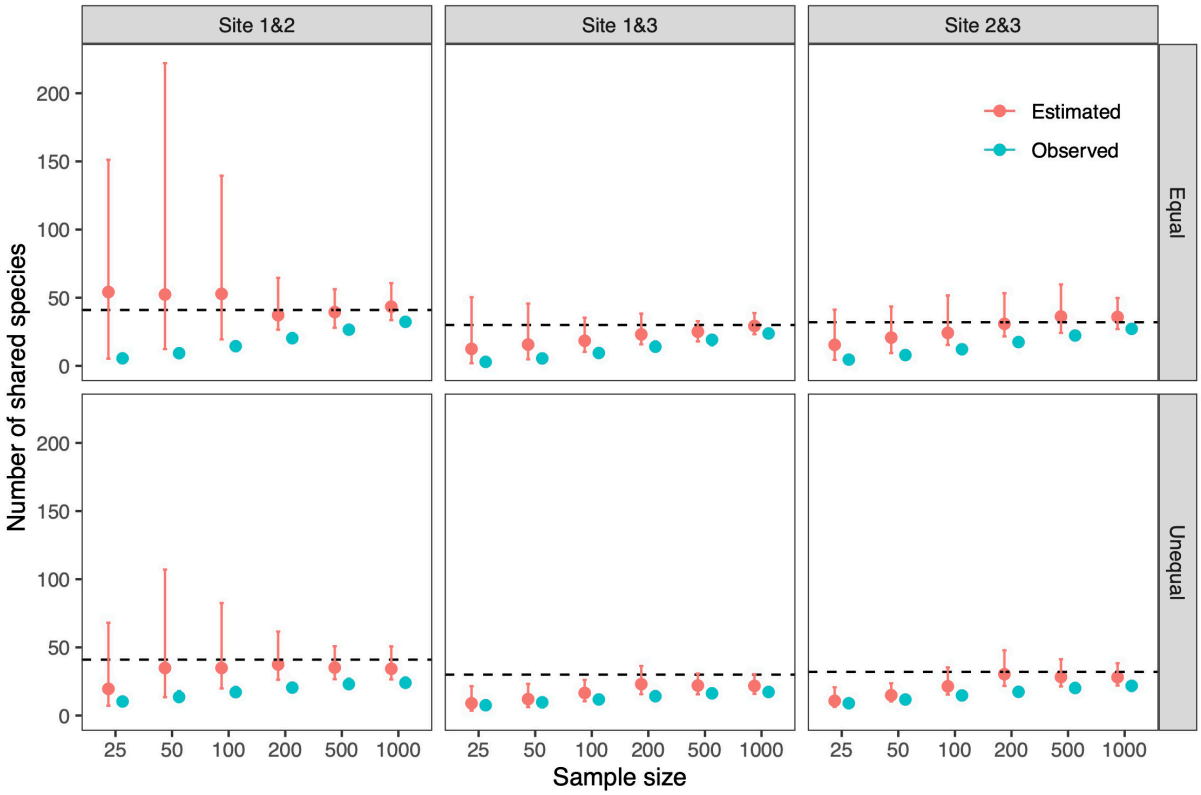

Figure 5. Mean values and 95% CIs for the estimated (TESS) and observed number of shared species for varying sample sizes between the paired samples for the empirical data. The dashed lines refer to the actual number of shared species between the two sites; error bars represent 95% confidence intervals (CIs are very small for observed values and hence are invisible in these graphs).

sample size parameter *m* (Grassle and Smith 1976), TESS uniquely works across a wide range of sample completeness scenarios. The high standard deviation of TESS values for highly incomplete samples, however, means that precise approximations of the true number of shared species requires a sample completeness > 50%. In other words, TESS works well when at least half of all species present in the underlying community are also present in the respective samples. Beyond this threshold, we have shown that TESS provides a superior measure of the number of shared species compared to the observed species richness for two randomly taken incomplete samples. We can therefore significantly decrease the threshold of sample completeness required for a meaningful analysis.

When the sample completeness is very small (< 0.2), samples are generally not informative. In these cases, mean TESS results strongly underestimate the true number of species shared. It might be possible to include a correction factor accounting for this trend, for example introducing an additional function related to estimated sampling completeness. However, the sampling completeness of empirical field data usually cannot be known and the uncertainty of estimating sampling completeness will always remain extremely high for very small sample sizes (Colwell et al. 2012). In addition, the accuracy of TESS further depends on the species abundance distribution patterns of the shared species, with a higher sampling completeness being required when communities chiefly share rare rather than common species, which could explain the slightly different performances for the empirical data between paired samples.

With regards to our different sampling scenarios, we assumed that one of the communities was already almost completely sampled for the unequal sampling scenarios. In this case, TESS strongly underestimates the real value of shared species for small sample sizes of the second sample. Equal sampling scenarios in contrast generate a larger overall variance in the estimated shared species values. Given the crucial role of a sufficiently large sample size for TESS calculations, field study sampling designs still need to carefully balance sampling effort at each individual site – to reach the 50% threshold – with the number of replicates or, more generally, sites sampled.

Applications of TESS

While TESS could be used for example to investigate questions relating to species losses from fragmented habitats in comparison to large, unfragmented areas of the same habitat, we see one key application of this measure when it is combined with a species richness estimator to estimate the true species (dis)similarity between communities that are represented by incomplete samples. We already demonstrate that combining TESS with a species diversity estimator (here

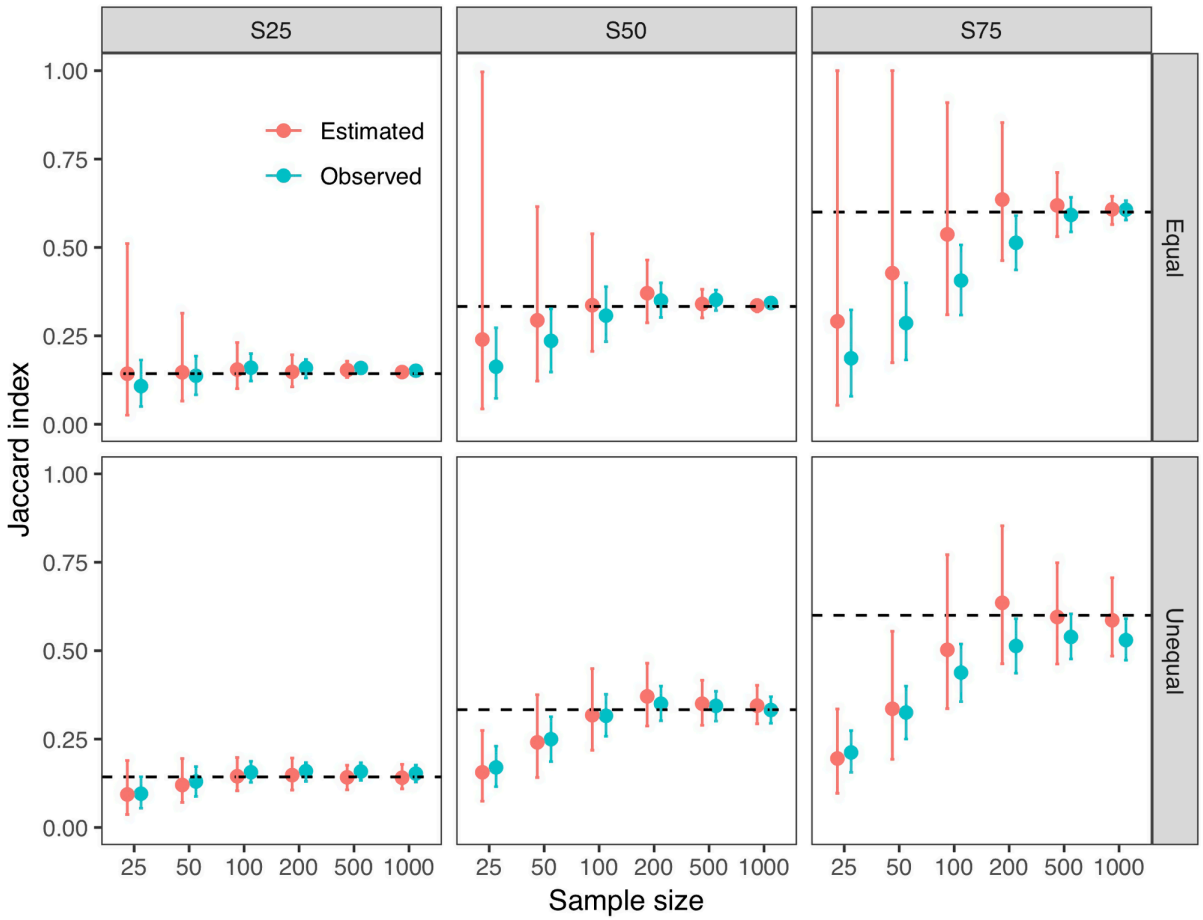

Figure 6. Mean Jaccard similarity values based on the estimated and observed number of species for varying sample sizes and the three different sharing scenarios (S25: 25% species shared; S50: 50% species shared; S75: 75% species shared) and the two sampling scenarios described in the Methods chapter for simulated data. Dashed lines refer to the real proportion of species similarity (for example, the total number of species is 175 for the S25 scenario, resulting in a Jaccard similarity of 25/175 = 0.14), while error bars represent the 95% CI for each index. Note: this approach combined TESS with a well-established species richness estimator (in this example, we use ACE), and a more accurate estimator will likely strongly decrease the range of the 95% CIs for the estimated similarities.

ACE) greatly increases the accuracy, despite at a relatively low precision, of the Jaccard similarity index when compared to results based on observed numbers of shared and unique species in incomplete samples. While we only demonstrated this approach for Jaccard similarities, TESS allows us to conduct similar calculations for all β-diversity measures whose calculations are based on the numbers of unique and shared species (Koleff et al. 2003). The high variance observed in our estimated β-diversity values is due to the combination of three main sources of variation occurring during the calculations of TESS and of the two 'true' species richness values for each of the incomplete samples. The combination of multiple errors means that even in cases where the total shared number of species can be estimated accurately, the tendency of established species richness estimators to underestimate total species richness in incomplete samples results in a tendency of similarity overestimation, which is particularly visible in intermediate sampling completeness scenarios While species richness estimators differ in terms of their accuracy (Walther and Moore 2005), an evaluation of the 'best' species richness-estimator remains beyond the scope of this study. Nonetheless, our results show that estimated Jaccard indices are useful in cases where communities are only represented by samples of small sizes especially when the underlying communities can be assumed to share a high proportion of species. In such cases, (dis)similarity measures based on estimated values are greatly superior to measures using observed values in approximating the true turnover rate. In contrast, where two communities share only few species, both estimated and observed (dis)similarity indices generate values close to the true value even for relatively small sample sizes.

Underlying calculations of TESS can also be used across entire sets of samples, with multiple pairwised comparisons allowing users to generate pairwised 'total shared species' matrices. From an applied biodiversity-conservation perspective, again combined with species richness estimators, TESS can provide an indication of the uniqueness of an assemblage, with low values of estimated shared species indicative of a high uniqueness that in many cases can be interpreted as a high conservation value (Barlow et al. 2010,

Ejrnæs et al. 2018). When evaluating the uniqueness, it is nonetheless important to consider potential site-specific factors and associated contributions of unusual species combination, for example in relation to the presence and impact of invasive alien species (Legendre and De Cáceres 2013). TESS can furthermore allow better insights into the degree of 'true' nestedness (Baselga 2010) of species pools in degraded habitats, or between habitat types in the context of land-use change. Overall, knowing the total number of species shared between two habitats from a subset of incomplete samples will therefore be helpful for both conservation management and biodiversity assessments, offering additional important insights for community changes.

Acknowledgements – We thank Johannes Knops and Jan Beck providing comments for the manuscript. We also thank two anonymous reviewers and the editor for their helpful comments. The authors declare no conflict of interest.
Funding – This study was financially supported by the Jiangsu Science and Technology Department (BK20181191) and the National Natural Science Foundation of China (31700363).

Transparent Peer Review

The peer review history for this article is available at <https://publons.com/publon/10.1111/ecog.05625>.

Data availability statement

All simulation scripts are available in electronic supporting information. Empirical dataset is available from the PREDICTS database (<https://doi.org/10.5519/0066354>) (Hudson et al. 2016).

References

Barlow, J. et al. 2010. Measuring the conservation value of tropical primary forests: the effect of occasional species on estimates of biodiversity uniqueness. – PLoS One 5: e9609.
Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. – Global Ecol. Biogeogr. 19: 134–143.
Beck, J. et al. 2013. Undersampling and the measurement of beta diversity. – Methods. Ecol. Evol. 4: 370–382.
Chao, A. 1984. Non-parametric estimation of the number of classes in a population. – Scand. J. Stat. 11: 265–270.
Chao, A. and Chiu, C. 2012. Estimation of species richness and shared species richness. – In: Balakrishnan, N. (ed.), Methods and applications of statistics in the atmospheric and earth sciences. Wiley, pp. 76–111.
Chao, A. and Chiu, C. 2016. Species richness: estimation and comparison. – In: Wiley StatsRef: statistics reference online, pp. 1–26.
Chao, A. and Jost, L. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. – Ecology 93: 2533–2547.
Chao, A. and Lee, S.-M. 1992. Estimating the number of classes via sample coverage. – J. Am. Stat. Assoc. 87: 210–217.
Chao, A. et al. 2000. Estimating the number of shared species in two communities. – Stat. Sin. 10: 227–246.
Chao, A. et al. 2006. Abundance-based similarity indices and their estimation when there are unseen species in samples. – Biometrics 62: 361–371.
Chao, A. et al. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. – Ecol. Monogr. 84: 45–67.
Chao, A. et al. 2016. SpadeR: species-richness prediction and diversity estimation with R. R package ver. 0.1.1. – <https://CRAN.R-project.org/package=SpadeR>
Coddington, J. A. et al. 2009. Undersampling bias: the null hypothesis for singleton species in tropical arthropod surveys. – J. Anim. Ecol. 78: 573–584.
Colwell, R. K. et al. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. – J. Plant Ecol. 5: 3–21.
Crawley, M. J. 2012. The R book. – Wiley.
Ejrnæs, R. et al. 2018. Uniquity: a general metric for biotic uniqueness of sites. – Biol. Conserv. 225: 98–105.
Flather, C. 1996. Fitting species–accumulation functions and assessing regional land use impacts on avian diversity. – J. Biogeogr. 23: 155–168.
Franklin, I. R. 1980. Evolutionary change in small populations. – In: Soulé, M. E. and Wilcox, B. A. (eds), Conservation biology: an evolutionary-ecological perspective. Sinauer, pp. 135–149.
Grassle, J. F. and Smith, W. 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. et al. 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). – Methods Ecol. Evol. 7: 1451–1456.
Hudson, L. N. et al. 2016. Dataset: the 2016 release of the PREDICTS database. Natural History Museum Data Portal (data.nhm.ac.uk). – <https://doi.org/10.5519/0066354>.
Hudson, L. N. et al. 2017. The database of the PREDICTS (projecting responses of ecological diversity in changing terrestrial systems) project. – Ecol. Evol. 7: 145–188.
Hurlbert, S. H. 1971. The nonconcept of species diversity: a critique and alternative parameters. – Ecology 52: 577–586.
Koleff, P. et al. 2003. Measuring beta diversity for presence–absence data. – J. Anim. Ecol. 72: 367–382.
Legendre, P. and De Cáceres, M. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. – Ecol. Lett. 16: 951–963.
Morisita, M. 1959. Measuring of interspecific association and similarity between communities. – Memoirs Facul. Sci. Kyushu Univ. Ser. E Biol. 3: 65–80.
Oksanen, J. et al. 2018. vegan: community ecology package. R package ver. 2.5-6. – <http://CRAN.R-project.org/package=vegan>.
Pan, H.-Y. et al. 2009. A nonparametric lower bound for the number of species shared by multiple communities. – J. Agric. Biol. Environ. Stat. 14: 452–468.
Rosenzweig, M. L. et al. 2003. Estimating diversity in unsampled habitats of a biogeographical province. – Conserv. Biol. 17: 864–874.
Schroeder, P. J. and Jenkins, D. G. 2018. How robust are popular beta diversity indices to sampling error? – Ecosphere 9: e02100.
Trueblood, D. D. et al. 1994. Three stages of seasonal succession on the Savin Hill Cove mudflat, Boston Harbor. – Limnol. Oceanogr. 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 Moore, J. L. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. – Ecography 28: 815–829.
Whittaker, R. H. 1960. Vegetation of the Siskiyou mountains, Oregon and California. – Ecol. Monogr. 30: 279–338.
Zaitsev, A. S. et al. 2002. Oribatid mite diversity and community dynamics in a spruce chronosequence. – Soil Biol. Biochem. 34: 1919–1927.
Zou, Y. and Axmacher, J. C. 2020. The Chord-normalized expected species shared (CNESS)-distance represents a superior measure of species turnover patterns. – Methods. Ecol. Evol. 11: 273–280.