Vegan: ecological diversity

Jari Oksanen

processed with vegan 2.7-8 in R version 4.6.2 (2026-10-29) on November 18, 2026

Abstract

This document explains diversity related methods in **vegan**. The methods are briefly described, and the equations used them are given often in more detail than in their help pages. The methods discussed include common diversity indices and rarefaction, families of diversity indices, species abundance models, species accumulation models and beta diversity, extrapolated richness and probability of being a member of the species pool. The document is still incomplete and does not cover all diversity methods in **vegan**.

Contents

1 Diversity indices **1**

2 Rarefaction **2**

3 Taxonomic and functional diversity **3**
- 3.1 Taxonomic diversity: average distance of traits 3
- 3.2 Functional diversity: the height of trait tree 4

4 Species abundance models **4**
- 4.1 Fisher and Preston 4
- 4.2 Ranked abundance distribution . . . 5

5 Species accumulation and beta diversity **6**
- 5.1 Species accumulation models 6
- 5.2 Beta diversity 6

6 Species pool **8**
- 6.1 Number of unseen species 8
- 6.2 Pool size from a single site 9
- 6.3 Probability of pool membership . . . 10

The **vegan** package has two major components: multivariate analysis (mainly ordination), and methods for diversity analysis of ecological communities. This document gives an introduction to the latter. Ordination methods are covered in other documents. Many of the diversity functions were written by Roeland Kindt, Bob O'Hara and Péter Sólymos.

Most diversity methods assume that data are counts of individuals. The methods are used with other data types, and some people argue that biomass or cover are more adequate than counts of individuals of variable sizes. However, this document mainly uses a data set with counts: stem counts of trees on 1 ha plots in the Barro Colorado Island. The following steps make these data available for the document:

```
> library(vegan)
> data(BCI)
```

1 Diversity indices

Function `diversity` finds the most commonly used diversity indices (Hill, 1973):

$$H = -\sum_{i=1}^{S} p_i \log_b p_i \qquad \text{Shannon–Weaver} \qquad (1)$$

$$D_1 = 1 - \sum_{i=1}^{S} p_i^2 \qquad \text{Simpson} \qquad (2)$$

$$D_2 = \frac{1}{\sum_{i=1}^{S} p_i^2} \qquad \text{inverse Simpson}\,, \qquad (3)$$

where p_i is the proportion of species i, and S is the number of species so that $\sum_{i=1}^{S} p_i = 1$, and b is the base of the logarithm. It is most common to use natural logarithms (and then we mark index as H'), but $b = 2$ has theoretical justification. The

default is to use natural logarithms. Shannon index is calculated with:

```
> H <- diversity(BCI)
```

which finds diversity indices for all sites.

Vegan does not have indices for evenness (equitability), but the most common of these, Pielou's evenness $J = H'/\log(S)$ is easily found as:

```
> J <- H/log(specnumber(BCI))
```

where `specnumber` is a simple **vegan** function to find the numbers of species.

vegan also can estimate series of Rényi and Tsallis diversities. Rényi diversity of order a is (Hill, 1973):

$$H_a = \frac{1}{1-a} \log \sum_{i=1}^{S} p_i^a \,, \tag{4}$$

and the corresponding Hill number is $N_a = \exp(H_a)$. Many common diversity indices are special cases of Hill numbers: $N_0 = S$, $N_1 = \exp(H')$, $N_2 = D_2$, and $N_\infty = 1/(\max p_i)$. The corresponding Rényi diversities are $H_0 = \log(S)$, $H_1 = H'$, $H_2 = -\log(\sum p_i^2)$, and $H_\infty = -\log(\max p_i)$. Tsallis diversity of order q is (Tóthmérész, 1995):

$$H_q = \frac{1}{q-1} \left(1 - \sum_{i=1}^{S} p^q \right) . \tag{5}$$

These correspond to common diversity indices: $H_0 = S - 1$, $H_1 = H'$, and $H_2 = D_1$, and can be converted to Hill numbers:

$$N_q = (1 - (q-1)H_q)^{\frac{1}{1-q}} \,. \tag{6}$$

We select a random subset of five sites for Rényi diversities:

```
> k <- sample(nrow(BCI), 6)
> round(renyi(BCI[k,]), 3)
```

```
       0  0.25   0.5     1     2     4     8    16
31 4.344 4.186 4.025 3.725 3.313 2.988 2.805 2.695
24 4.554 4.418 4.274 3.979 3.488 2.987 2.660 2.489
43 4.454 4.286 4.105 3.736 3.146 2.594 2.283 2.135
13 4.533 4.403 4.265 3.982 3.480 2.941 2.635 2.496
17 4.533 4.361 4.165 3.737 3.090 2.665 2.464 2.346
46 4.454 4.299 4.135 3.810 3.343 2.975 2.756 2.629
      32    64   Inf
31 2.626 2.587 2.546
24 2.409 2.371 2.334
43 2.066 2.033 2.001
13 2.425 2.387 2.350
17 2.276 2.240 2.205
46 2.558 2.520 2.480
```

We can really regard a site more diverse if all of its Rényi diversities are higher than in another site.

Finally, the α parameter of Fisher's log-series can be used as a diversity index (Fisher *et al.*, 1943):

```
> alpha <- fisher.alpha(BCI)
```

2 Rarefaction

Species richness increases with sample size, and differences in richness actually may be caused by differences in sample size. To solve this problem, we may try to rarefy species richness to the same number of individuals. Expected number of species in a community rarefied from N to n individuals is (Hurlbert, 1971):

$$\hat{S}_n = \sum_{i=1}^{S} (1 - q_i) \,, \quad \text{where } q_i = \frac{\binom{N-x_i}{n}}{\binom{N}{n}} \,. \tag{7}$$

Here x_i is the count of species i, and $\binom{N}{n}$ is the binomial coefficient, or the number of ways we can choose n from N, and q_i give the probabilities that species i does *not* occur in a sample of size n. This is positive only when $N - x_i \geq n$, but for other cases $q_i = 0$ or the species is sure to occur in the sample. The variance of rarefied richness is (Heck *et al.*, 1975):

$$s^2 = q_i(1 - q_i) + 2 \sum_{i=1}^{S} \sum_{j>i} \left[\frac{\binom{N-x_i-x_j}{n}}{\binom{N}{n}} - q_i q_j \right] . \tag{8}$$

Equation 8 actually is of the same form as the variance of sum of correlated variables:

$$\mathsf{VAR}\left(\sum x_i\right) = \sum \mathsf{VAR}(x_i) + 2 \sum_{i=1}^{S} \sum_{j>i} \mathsf{COV}(x_i, x_j) \,. \tag{9}$$

The number of stems per hectare varies in our data set:

```
> quantile(rowSums(BCI))
```

```
   0%   25%   50%   75%  100%
340.0 409.0 428.0 443.5 601.0
```

To express richness for the same number of individuals, we can use:

```
> Srar <- rarefy(BCI, min(rowSums(BCI)))
```

Rarefaction curves often are seen as an objective solution for comparing species richness with different sample sizes. However, rank orders typically differ among different rarefaction sample sizes, rarefaction curves can cross.

As an extreme case we may rarefy sample size to two individuals:

```
> S2 <- rarefy(BCI, 2)
```

This will not give equal rank order with the previous rarefaction richness:

```
> all(rank(Srar) == rank(S2))
[1] FALSE
```

Moreover, the rarefied richness for two individuals is a finite sample variant of Simpson's diversity index (Hurlbert, 1971) – or more precisely of $D_1 + 1$, and these two are almost identical in BCI:

```
> range(diversity(BCI, "simp") - (S2 -1))
[1] -0.002868298 -0.001330663
```

Rarefaction is sometimes presented as an ecologically meaningful alternative to dubious diversity indices (Hurlbert, 1971), but the differences really seem to be small.

3 Taxonomic and functional diversity

Simple diversity indices only consider species identity: all different species are equally different. In contrast, taxonomic and functional diversity indices judge the differences of species. Taxonomic and functional diversities are used in different fields of science, but they really have very similar reasoning, and either could be used either with taxonomic or functional traits of species.

3.1 Taxonomic diversity: average distance of traits

The two basic indices are called taxonomic diversity Δ and taxonomic distinctness Δ^* (Clarke and Warwick, 1998):

$$\Delta = \frac{\sum\sum_{i<j} \omega_{ij} x_i x_j}{n(n-1)/2} \tag{10}$$

$$\Delta^* = \frac{\sum\sum_{i<j} \omega_{ij} x_i x_j}{\sum\sum_{i<j} x_i x_j}\,. \tag{11}$$

These equations give the index values for a single site, and summation goes over species i and j, and ω are the taxonomic distances among taxa, x are species abundances, and n is the total abundance for a site. With presence–absence data, both indices reduce to the same index called Δ^+, and for this it is possible to estimate standard deviation. There are two indices derived from Δ^+: it can be multiplied with species richness[1] to give $s\Delta^+$, or it can be used to estimate an index of variation in taxonomic distinctness Λ^+ (Clarke and Warwick, 2001):

$$\Lambda^+ = \frac{\sum\sum_{i<j} \omega_{ij}^2}{n(n-1)/2} - (\Delta^+)^2\,. \tag{12}$$

We still need the taxonomic differences among species (ω) to calculate the indices. These can be any distance structure among species, but usually it is found from established hierarchic taxonomy. Typical coding is that differences among species in the same genus is 1, among the same family it is 2 etc. However, the taxonomic differences are scaled to maximum 100 for easier comparison between different data sets and taxonomies. Alternatively, it is possible to scale steps between taxonomic level proportional to the reduction in the number of categories (Clarke and Warwick, 1999): if almost all genera have only one species, it does not make a great difference if two individuals belong to a different species or to a different genus.

Function `taxondive` implements indices of taxonomic diversity, and `taxa2dist` can be used to convert classification tables to taxonomic distances either with constant or variable step lengths between successive categories. There is no taxonomic table for the BCI data in **vegan**[2] but there is such a table for the Dune meadow data (Fig. 1):

```
> data(dune)
> data(dune.taxon)
> taxdis <- taxa2dist(dune.taxon, varstep=TRUE)
> mod <- taxondive(dune, taxdis)
```

[1] This text normally uses upper case letter S for species richness, but lower case s is used here in accordance with the original papers on taxonomic diversity

[2] Actually I made such a classification, but taxonomic differences proved to be of little use in the Barro Colorado data: they only singled out sites with Monocots (palm trees) in the data.

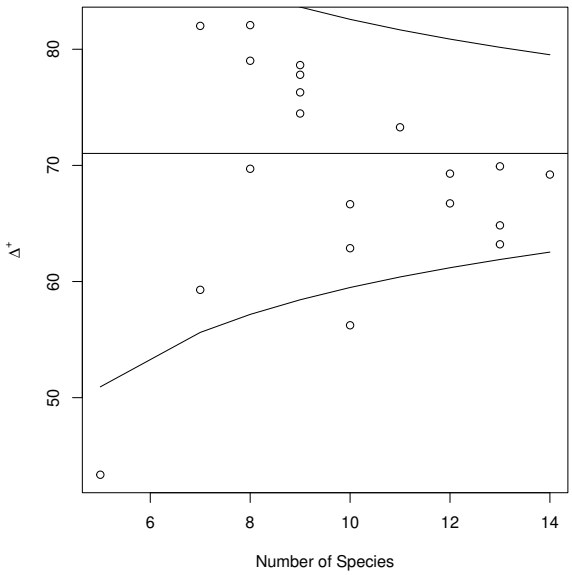


Figure 1: Taxonomic diversity Δ^+ for the dune meadow data. The points are diversity values of single sites, and the funnel is their approximate confidence intervals (2× standard error).

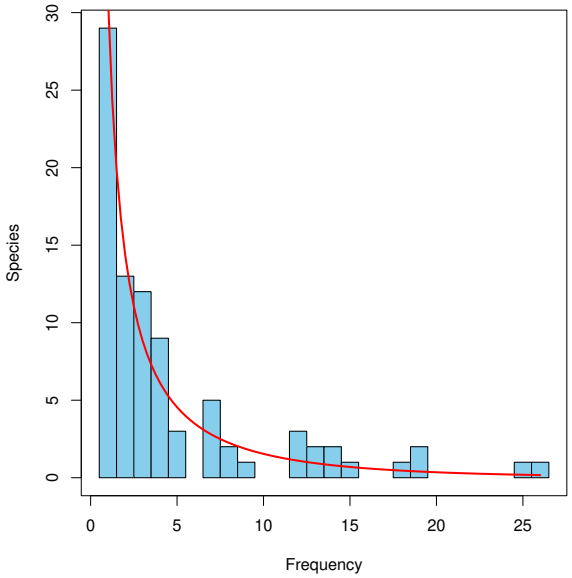


Figure 2: Fisher's log-series fitted to one randomly selected site (42).

3.2 Functional diversity: the height of trait tree

In taxonomic diversity the primary data were taxonomic trees which were transformed to pairwise distances among species. In functional diversity the primary data are species traits which are translated to pairwise distances among species and then to clustering trees of species traits. The argument for using trees is that in this way a single deviant species will have a small influence, since its difference is evaluated only once instead of evaluating its distance to all other species (Petchey and Gaston, 2006).

Function `treedive` implements functional diversity defined as the total branch length in a trait dendrogram connecting all species, but excluding the unnecessary root segments of the tree (Petchey and Gaston, 2002, 2006). The example uses the taxonomic distances of the previous chapter. These are first converted to a hierarchic clustering (which actually were their original form before `taxa2dist` converted them into distances)

```
> tr <- hclust(taxdis, "aver")
> mod <- treedive(dune, tr)
```

4 Species abundance models

Diversity indices may be regarded as variance measures of species abundance distribution. We may wish to inspect abundance distributions more directly. **Vegan** has functions for Fisher's log-series and Preston's log-normal models, and in addition several models for species abundance distribution.

4.1 Fisher and Preston

In Fisher's log-series, the expected number of species $\hat{f}$ with n individuals is (Fisher *et al.*, 1943):

$$\hat{f}_n = \frac{\alpha x^n}{n}\,, \tag{13}$$

where α is the diversity parameter, and x is a nuisance parameter defined by α and total number of individuals N in the site, $x = N/(N-\alpha)$. Fisher's log-series for a randomly selected plot is (Fig. 2):

```
> k <- sample(nrow(BCI), 1)
> fish <- fisherfit(BCI[k,])
> fish

Fisher log series model
No. of species: 87
Fisher alpha:   33.59822
```

We already saw α as a diversity index.

Preston's log-normal model is the main challenger to Fisher's log-series (Preston, 1948). Instead of plotting species by frequencies, it bins species into frequency classes of increasing sizes. As a result, upper bins with high range of frequencies become more common, and sometimes the result looks similar to Gaussian distribution truncated at the left.

There are two alternative functions for the log-normal model: `prestonfit` and `prestondistr`. Function `prestonfit` uses traditionally binning approach, and is burdened with arbitrary choices of binning limits and treatment of ties. It seems that Preston split ties between adjacent octaves: only half of the species observed once were in the first octave, and half were transferred to the next octave, and the same for all species at the octave limits occurring 2, 4, 8, 16... times (Williamson and Gaston, 2005). Function `prestonfit` can either split the ties or keep all limit cases in the lower octave. Function `prestondistr` directly maximizes truncated log-normal likelihood without binning data, and it is the recommended alternative. Log-normal models usually fit poorly to the BCI data, but here our random plot (number 42):

```
> prestondistr(BCI[k,])

Preston lognormal model
Method: maximized likelihood to log2 abundances
No. of species: 87

     mode     width        S0
 1.242040  1.626518 23.296555

Frequencies by Octave
                0        1        2        3         4
Observed 14.50000 21.00000 23.00000 13.50000 10.000000
Fitted   17.40485 23.04004 20.89953 12.99063  5.533046
                5
Observed 5.000000
Fitted   1.614874
```

4.2 Ranked abundance distribution

An alternative approach to species abundance distribution is to plot logarithmic abundances in decreasing order, or against ranks of species (Whittaker, 1965). These are known as ranked abundance distribution curves, species abundance curves, dominance–diversity curves or Whittaker plots. Function `radfit` fits some of the most popular models (Wilson, 1991) using maximum likelihood estimation:

$$\hat{a}_r = \frac{N}{S}\sum_{k=r}^{S}\frac{1}{k} \qquad \text{brokenstick} \tag{14}$$

$$\hat{a}_r = N\alpha(1-\alpha)^{r-1} \qquad \text{preemption} \tag{15}$$

$$\hat{a}_r = \exp\left[\log(\mu) + \log(\sigma)\Phi\right] \qquad \text{log-normal} \tag{16}$$

$$\hat{a}_r = N\hat{p}_1 r^{\gamma} \qquad \text{Zipf} \tag{17}$$

$$\hat{a}_r = Nc(r+\beta)^{\gamma} \qquad \text{Zipf–Mandelbrot} \tag{18}$$

In all these, $\hat{a}_r$ is the expected abundance of species at rank r, S is the number of species, N is the number of individuals, Φ is a standard normal function, $\hat{p}_1$ is the estimated proportion of the most abundant species, and α, μ, σ, γ, β and c are the estimated parameters in each model.

It is customary to define the models for proportions p_r instead of abundances a_r, but there is no reason for this, and `radfit` is able to work with the original abundance data. We have count data, and the default Poisson error looks appropriate, and our example data set gives (Fig. 3):

```
> rad <- radfit(BCI[k,])
> rad

RAD models, family poisson
No. of species 87, total abundance 414

           par1      par2     par3    Deviance
Null                                   33.7036
Preemption  0.044636                   21.8063
Lognormal   1.0671    1.0084           19.6278
Zipf        0.11343  -0.74891          49.1963
Mandelbrot  40.114   -2.1294   19.864   4.9471
           AIC      BIC
Null       292.2468 292.2468
Preemption 282.3495 284.8154
Lognormal  282.1710 287.1028
Zipf       311.7395 316.6713
Mandelbrot 269.4903 276.8881
```

Function `radfit` compares the models using alternatively Akaike's or Schwartz's Bayesian information criteria. These are based on log-likelihood, but penalized by the number of estimated parameters. The penalty per parameter is 2 in AIC, and $\log S$ in BIC. Brokenstick is regarded as a null model and has no estimated parameters in **vegan**. Preemption model has one estimated parameter (α), log-normal and Zipf models two (μ, σ, or

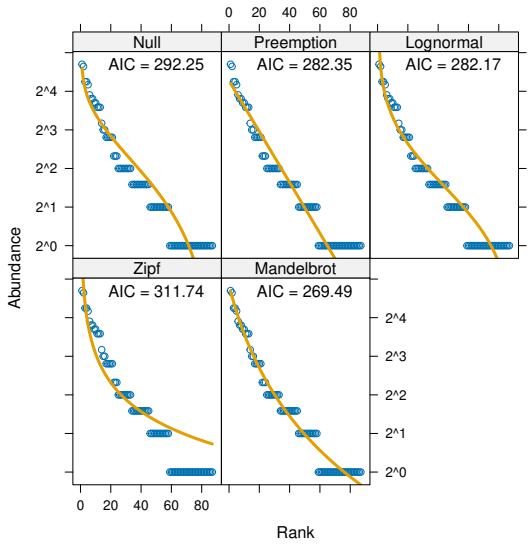


Figure 3: Ranked abundance distribution models for a random plot (no. 42). The best model has the lowest AIC.

$\hat{p}_1, \gamma$, resp.), and Zipf–Mandelbrot model has three (c, β, γ).

Function `radfit` also works with data frames, and fits models for each site. It is curious that log-normal model rarely is the choice, although it generally is regarded as the canonical model, in particular in data sets like Barro Colorado tropical forests.

5 Species accumulation and beta diversity

Species accumulation models and species pool models study collections of sites, and their species richness, or try to estimate the number of unseen species.

5.1 Species accumulation models

Species accumulation models are similar to rarefaction: they study the accumulation of species when the number of sites increases. There are several alternative methods, including accumulating sites in the order they happen to be, and repeated accumulation in random order. In addition, there are three analytic models. Rarefaction pools individuals together, and applies rarefaction equation (7) to these individuals. Kindt's exact accumulator resembles rarefaction (Ugland *et al.*, 2003):

$$\hat{S}_n = \sum_{i=1}^{S}(1 - p_i), \quad \text{where } p_i = \frac{\binom{N-f_i}{n}}{\binom{N}{n}}, \quad (19)$$

and f_i is the frequency of species i. Approximate variance estimator is:

$$s^2 = p_i(1 - p_i) + 2\sum_{i=1}^{S}\sum_{j>i}\left(r_{ij}\sqrt{p_i(1-p_i)}\sqrt{p_j(1-p_j)}\right), \quad (20)$$

where r_{ij} is the correlation coefficient between species i and j. Both of these are unpublished: eq. 19 was developed by Roeland Kindt, and eq. 20 by Jari Oksanen. The third analytic method was suggested by Coleman *et al.* (1982):

$$S_n = \sum_{i=1}^{S}(1 - p_i), \quad \text{where } p_i = \left(1 - \frac{1}{n}\right)^{f_i}, \quad (21)$$

and the suggested variance is $s^2 = p_i(1 - p_i)$ which ignores the covariance component. In addition, eq. 21 does not properly handle sampling without replacement and underestimates the species accumulation curve.

The recommended is Kindt's exact method (Fig. 4):

```
> sac <- specaccum(BCI)
> plot(sac, ci.type="polygon", ci.col="yellow")
```

5.2 Beta diversity

Whittaker (1960) divided diversity into various components. The best known are diversity in one spot that he called alpha diversity, and the diversity along gradients that he called beta diversity. The basic diversity indices are indices of alpha diversity. Beta diversity should be studied with respect to gradients (Whittaker, 1960), but almost everybody understand that as a measure of general heterogeneity (Tuomisto, 2010a,b): how many more species do you have in a collection of sites compared to an average site.

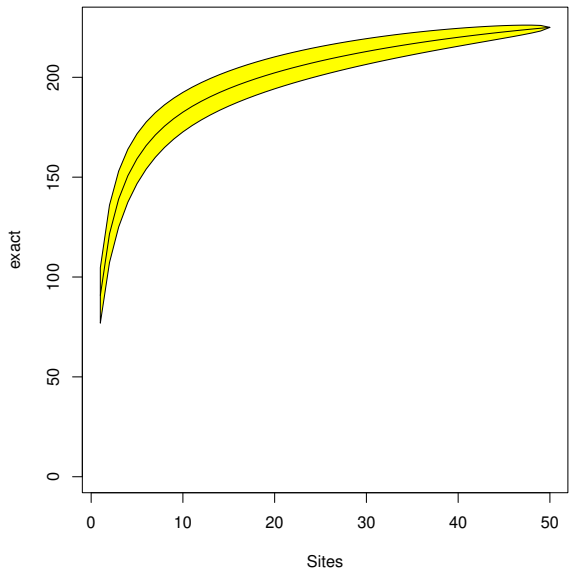


Figure 4: Species accumulation curve for the BCI data; exact method.

The best known index of beta diversity is based on the ratio of total number of species in a collection of sites S and the average richness per one site $\bar{\alpha}$ (Tuomisto, 2010a):

$$\beta = S/\bar{\alpha} - 1\,. \tag{22}$$

Subtraction of one means that $\beta = 0$ when there are no excess species or no heterogeneity between sites. For this index, no specific functions are needed, but this index can be easily found with the help of **vegan** function `specnumber`:

```
> ncol(BCI)/mean(specnumber(BCI)) - 1
```

```
[1] 1.478519
```

The index of eq. 22 is problematic because S increases with the number of sites even when sites are all subsets of the same community. Whittaker (1960) noticed this, and suggested the index to be found from pairwise comparison of sites. If the number of shared species in two sites is a, and the numbers of species unique to each site are b and c, then $\bar{\alpha} = (2a + b + c)/2$ and $S = a + b + c$, and index 22 can be expressed as:

$$\beta = \frac{a+b+c}{(2a+b+c)/2} - 1 = \frac{b+c}{2a+b+c}\,. \tag{23}$$

This is the Sørensen index of dissimilarity, and it can be found for all sites using **vegan** function `vegdist` with binary data:

```
> beta <- vegdist(BCI, binary=TRUE)
> mean(beta)
```

```
[1] 0.3399075
```

There are many other definitions of beta diversity in addition to eq. 22. All commonly used indices can be found using `betadiver` (Koleff *et al.*, 2003). The indices in `betadiver` can be referred to by subscript name, or index number:

```
> betadiver(help=TRUE)
```

```
1 "w" = (b+c)/(2*a+b+c)
2 "-1" = (b+c)/(2*a+b+c)
3 "c" = (b+c)/2
4 "wb" = b+c
5 "r" = 2*b*c/((a+b+c)^2-2*b*c)
6 "I" = log(2*a+b+c) - 2*a*log(2)/(2*a+b+c) -
((a+b)*log(a+b) + (a+c)*log(a+c)) / (2*a+b+c)
7 "e" = exp(log(2*a+b+c) - 2*a*log(2)/(2*a+b+c)
- ((a+b)*log(a+b) + (a+c)*log(a+c)) /
(2*a+b+c))-1
8 "t" = (b+c)/(2*a+b+c)
9 "me" = (b+c)/(2*a+b+c)
10 "j" = a/(a+b+c)
11 "sor" = 2*a/(2*a+b+c)
12 "m" = (2*a+b+c)*(b+c)/(a+b+c)
13 "-2" = pmin.int(b,c)/(pmax.int(b,c)+a)
14 "co" = (a*c+a*b+2*b*c)/(2*(a+b)*(a+c))
15 "cc" = (b+c)/(a+b+c)
16 "g" = (b+c)/(a+b+c)
17 "-3" = pmin.int(b,c)/(a+b+c)
18 "l" = (b+c)/2
19 "19" = 2*(b*c+1)/(a+b+c)/(a+b+c-1)
20 "hk" = (b+c)/(2*a+b+c)
21 "rlb" = a/(a+c)
22 "sim" = pmin.int(b,c)/(pmin.int(b,c)+a)
23 "gl" = 2*abs(b-c)/(2*a+b+c)
24 "z" = (log(2)-log(2*a+b+c)+log(a+b+c))/log(2)
```

Some of these indices are duplicates, and many of them are well known dissimilarity indices. One of the more interesting indices is based on the Arrhenius species–area model

$$\hat{S} = cX^z\,, \tag{24}$$

where X is the area (size) of the patch or site, and c and z are parameters. Parameter c is uninteresting, but z gives the steepness of the species area curve and is a measure of beta diversity. In islands typically $z \approx 0.3$. This kind of islands can be regarded as subsets of the same community, indicating that we really should talk about gradient differences if $z \gtrapprox 0.3$. We can find the value of z for a pair of plots using function `betadiver`:

```
> z <- betadiver(BCI, "z")
> quantile(z)

      0%       25%       50%       75%      100%
0.2732845 0.3895024 0.4191536 0.4537180 0.5906091
```

The size X and parameter c cancel out, and the index gives the estimate z for any pair of sites.

Function `betadisper` can be used to analyse beta diversities with respect to classes or factors (Anderson, 2006; Anderson *et al.*, 2006). There is no such classification available for the Barro Colorado Island data, and the example studies beta diversities in the management classes of the dune meadows (Fig. 5):

```
> data(dune)
> data(dune.env)
> z <- betadiver(dune, "z")
> mod <- with(dune.env, betadisper(z, Management))
> mod

        Homogeneity of multivariate dispersions
        squared Beta.z distance

Call: betadisper(d = z, group = Management)

No. of Positive Eigenvalues: 12
No. of Negative Eigenvalues: 7

Average distance to median:
    BF     HF     NM     SF
0.2532 0.2512 0.4406 0.3635

Eigenvalues for PCoA axes:
(Showing 8 of 19 eigenvalues)
  PCoA1   PCoA2   PCoA3   PCoA4   PCoA5   PCoA6
1.65466 0.88696 0.53336 0.37435 0.28725 0.22445
  PCoA7   PCoA8
0.16128 0.08099
```

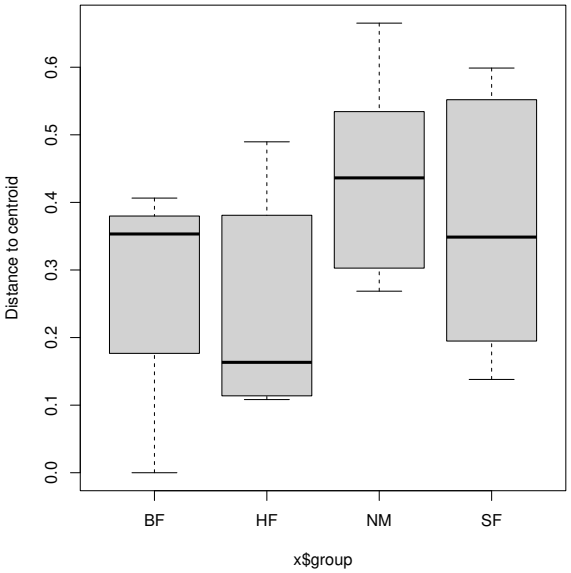


Figure 5: Box plots of beta diversity measured as the average steepness (z) of the species area curve in the Arrhenius model $S = cX^z$ in Management classes of dune meadows.

6 Species pool

6.1 Number of unseen species

Species accumulation models indicate that not all species were seen in any site. These unseen species also belong to the species pool. Functions `specpool` and `estimateR` implement some methods of estimating the number of unseen species. Function `specpool` studies a collection of sites, and `estimateR` works with counts of individuals, and can be used with a single site. Both functions assume that the number of unseen species is related to the number of rare species, or species seen only once or twice.

The incidence-based functions group species by their number of occurrences $f_i = f_0, f_1, \ldots, f_N$, where f is the number of species occurring in exactly i sites in the data: f_N is the number of species occurring on every N site, f_1 the number of species occurring once, and f_0 the number of species in the species pool but not found in the sample. The total number of species in the pool S_p is

$$S_p = \sum_{i=0}^{N} f_i = f_0 + S_o\,, \tag{25}$$

where $S_o = \sum_{i>0} f_i$ is the observed number of species. The sampling proportion i/N is an estimate for the commonness of the species in the community. When species is present in the community but not in the sample, $i = 0$ is an obvious under-estimate, and consequently, for values $i > 0$ the species commonness is over-estimated (Good, 1953). The models for the pool size estimate the number of species missing in the sample f_0.

Function `specpool` implements the following models to estimate the number of missing species f_0. Chao estimator is (Chao, 1987; Chiu *et al.*,

2014):

$$\hat{f}_0 = \begin{cases} \frac{f_1^2}{2f_2}\frac{N-1}{N} & \text{if } f_2 > 0 \\ \frac{f_1(f_1-1)}{2}\frac{N-1}{N} & \text{if } f_2 = 0\,. \end{cases} \quad (26)$$

The latter case for $f_2 = 0$ is known as the bias-corrected form. Chiu *et al.* (2014) introduced the small-sample correction term $\frac{N}{N-1}$, but it was not originally used (Chao, 1987).

The first and second order jackknife estimators are (Smith and van Belle, 1984):

$$\hat{f}_0 = f_1\frac{N-1}{N} \quad (27)$$

$$\hat{f}_0 = f_1\frac{2N-3}{N} + f_2\frac{(N-2)^2}{N(N-1)}\,. \quad (28)$$

The bootstrap estimator is (Smith and van Belle, 1984):

$$\hat{f}_0 = \sum_{i=1}^{S_o}(1-p_i)^N\,. \quad (29)$$

The idea in jackknife seems to be that we missed about as many species as we saw only once, and the idea in bootstrap that if we repeat sampling (with replacement) from the same data, we miss as many species as we missed originally.

The variance estimaters only concern the estimated number of missing species $\hat{f}_0$, although they are often expressed as they would apply to the pool size S_p; this is only true if we assume that $\mathsf{VAR}(S_o) = 0$. The variance of the Chao estimate is (Chiu *et al.*, 2014):

$$\mathsf{VAR}(\hat{f}_0) = f_1\left(A^2\frac{G^3}{4} + A^2G^2 + A\frac{G}{2}\right), \quad \text{where } A = \frac{N-1}{N} \text{ and } G = \frac{f_1}{f_2}\,. \quad (30)$$

For the bias-corrected form of eq. 26 (case $f_2 = 0$), the variance is (Chiu *et al.*, 2014, who omit small-sample correction in some terms):

$$\mathsf{VAR}(\hat{f}_0) = \tfrac{1}{4}A^2f_1(2f_1-1)^2 + \tfrac{1}{2}Af_1(f_1-1) - \tfrac{1}{4}A^2\frac{f_1^4}{S_p}\,. \quad (31)$$

The variance of the first-order jackknife is based on the number of "singletons" r (species occurring only once in the data) in sample plots (Smith and van Belle, 1984):

$$\mathsf{VAR}(\hat{f}_0) = \left(\sum_{i=1}^{N} r_i^2 - \frac{f_1}{N}\right)\frac{N-1}{N}\,. \quad (32)$$

Variance of the second-order jackknife is not evaluated in `specpool` (but contributions are welcome).

The variance of bootstrap estimator is(Smith and van Belle, 1984):

$$\mathsf{VAR}(\hat{f}_0) = \sum_{i=1}^{S_o} q_i(1-q_i) + 2\sum_{i\neq j}^{S_o}\left[(Z_{ij}/N)^N - q_iq_j\right] \quad \text{where } q_i = (1-p_i)^N\,, \quad (33)$$

and Z_{ij} is the number of sites where both species are absent.

The extrapolated richness values for the whole BCI data are:

```
> specpool(BCI)

    Species     chao chao.se  jack1 jack1.se    jack2
All     225 236.3732 6.54361 245.58 5.650522 247.8722
        boot  boot.se  n
All 235.6862 3.468888 50
```

If the estimation of pool size really works, we should get the same values of estimated richness if we take a random subset of a half of the plots (but this is rarely true):

```
> s <- sample(nrow(BCI), 25)
> specpool(BCI[s,])

    Species     chao  chao.se  jack1 jack1.se    jack2
All     203 222.5323 10.73448 225.08 6.046222 234.7783
        boot  boot.se  n
All 213.7013 3.619807 25
```

6.2 Pool size from a single site

The `specpool` function needs a collection of sites, but there are some methods that estimate the number of unseen species for each single site. These functions need counts of individuals, and species seen only once or twice, or other rare species, take the place of species with low frequencies. Function `estimateR` implements two of these methods:

```
> estimateR(BCI[k,])
```

```
                 42
S.obs     87.000000
S.chao1  116.000000
se.chao1  14.313533
S.ACE    113.102796
se.ACE     5.223075
```

In abundance based models a_i denotes the number of species with i individuals, and takes the place of f_i of previous models. Chao's method is similar as the bias-corrected model eq. 26 (Chao, 1987; Chiu *et al.*, 2014):

$$S_p = S_o + \frac{a_1(a_1-1)}{2(a_2+1)}\,. \tag{34}$$

When $f_2 = 0$, eq. 34 reduces to the bias-corrected form of eq. 26, but quantitative estimators are based on abundances and do not use small-sample correction. This is not usually needed because sample sizes are total numbers of individuals, and these are usually high, unlike in frequency based models, where the sample size is the number of sites (Chiu *et al.*, 2014).

A commonly used approximate variance estimator of eq. 34 is:

$$s^2 = \frac{a_1(a_1-1)}{2} + \frac{a_1(2a_1+1)^2}{(a_2+1)^2} + \frac{a_1^2 a_2 (a_1-1)^2}{4(a_2+1)^4}\,. \tag{35}$$

However, **vegan** does not use this, but instead the following more exact form which was directly derived from eq. 34 following Chiu *et al.* (2014, web appendix):

$$\begin{aligned} s^2 = \frac{1}{4}\frac{1}{(a_2+1)^4 S_p}[a_1(&S_p a_1^3 a_2 + 4S_p a_1^2 a_2^2 \\ &+ 2S_p a_1 a_2^3 + 6S_p a_1^2 a_2 + 2S_p a_1 a_2^2 - 2S_p a_2^3 \\ &+ 4S_p a_1^2 + S_p a_1 a_2 - 5S_p a_2^2 - a_1^3 - 2a_1^2 a_2 \\ &- a_1 a_2^2 - 2S_p a_1 - 4S_p a_2 - S_p)]\,. \end{aligned} \tag{36}$$

The variance estimators only concern the number of unseen species like previously.

The ACE is estimator is defined as (O'Hara, 2005):

$$\begin{aligned} S_p &= S_{\text{abund}} + \frac{S_{\text{rare}}}{C_{\text{ACE}}} + \frac{a_1}{C_{\text{ACE}}}\gamma^2\,, \quad \text{where} \\ C_{\text{ACE}} &= 1 - \frac{a_1}{N_{\text{rare}}} \\ \gamma^2 &= \frac{S_{\text{rare}}}{C_{\text{ACE}}} \sum_{i=1}^{10} i(i-1)a_1 \frac{N_{\text{rare}}-1}{N_{\text{rare}}}\,. \end{aligned} \tag{37}$$

Now a_1 takes the place of f_1 above, and means the number of species with only one individual. Here S_{abund} and S_{rare} are the numbers of species of abundant and rare species, with an arbitrary upper limit of 10 individuals for a rare species, and N_{rare} is the total number of individuals in rare species. The variance estimator uses iterative solution, and it is best interpreted from the source code or following O'Hara (2005).

The pool size is estimated separately for each site, but if input is a data frame, each site will be analysed.

If log-normal abundance model is appropriate, it can be used to estimate the pool size. Log-normal model has a finite number of species which can be found integrating the log-normal:

$$S_p = S_\mu \sigma \sqrt{2\pi}\,, \tag{38}$$

where S_μ is the modal height or the expected number of species at maximum (at μ), and σ is the width. Function `veiledspec` estimates this integral from a model fitted either with `prestondistr` or `prestonfit`, and fits the latter if raw site data are given. Log-normal model may fit poorly, but we can try:

```
> veiledspec(prestondistr(BCI[k,]))
```

```
Extrapolated     Observed       Veiled
   94.981844    87.000000     7.981844
```

```
> veiledspec(BCI[k,])
```

```
Extrapolated     Observed       Veiled
   103.55167    87.00000     16.55167
```

6.3 Probability of pool membership

Beals smoothing was originally suggested as a tool of regularizing data for ordination. It regularizes data too strongly, but it has been suggested as a method of estimating which of the missing species could occur in a site, or which sites are suitable for

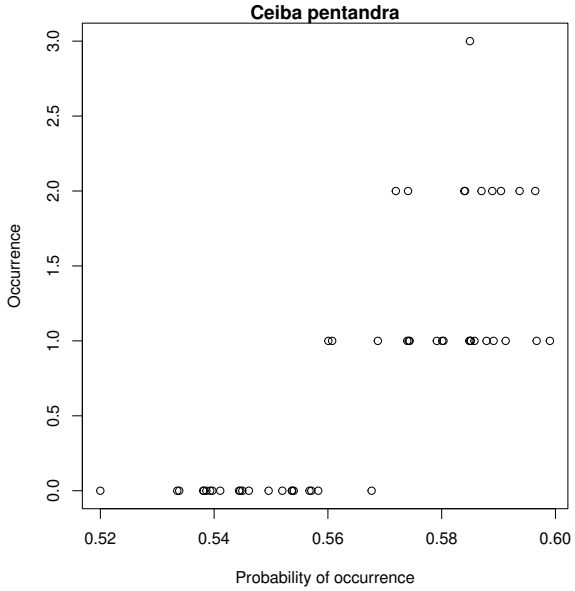


Figure 6: Beals smoothing for *Ceiba pentandra*.

a species. The probability for each species at each site is assessed from other species occurring on the site.

Function `beals` implement Beals smoothing (McCune, 1987; De Cáceres and Legendre, 2008):

```
> smo <- beals(BCI)
```

We may see how the estimated probability of occurrence and observed numbers of stems relate in one of the more familiar species. We study only one species, and to avoid circular reasoning we do not include the target species in the smoothing (Fig. 6):

```
> j <- which(colnames(BCI) == "Ceiba.pentandra")
> plot(beals(BCI, species=j, include=FALSE), BCI[,j],
       ylab="Occurrence", main="Ceiba pentandra",
       xlab="Probability of occurrence")
```

References

Anderson MJ (2006). "Distance-based tests for homogeneity of multivariate dispersions." *Biometrics*, **62**, 245–253.

Anderson MJ, Ellingsen KE, McArdle BH (2006). "Multivariate dispersion as a measure of beta diversity." *Ecology Letters*, **9**, 683–693.

Chao A (1987). "Estimating the population size for capture-recapture data with unequal catchability." *Biometrics*, **43**, 783–791.

Chiu CH, Wang YT, Walther BA, Chao A (2014). "An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula." *Biometrics*, **70**, 671–682.

Clarke KR, Warwick RM (1998). "A taxonomic distinctness index and its statistical properties." *Journal of Applied Ecology*, **35**, 523–531.

Clarke KR, Warwick RM (1999). "The taxonomic distinctness measure of biodiversity: weighting of step lengths between hierarchical levels." *Marine Ecology Progress Series*, **184**, 21–29.

Clarke KR, Warwick RM (2001). "A further biodiversity index applicable to species lists: variation in taxonomic distinctness." *Marine Ecology Progress Series*, **216**, 265–278.

Coleman BD, Mares MA, Willis MR, Hsieh Y (1982). "Randomness, area and species richness." *Ecology*, **63**, 1121–1133.

De Cáceres M, Legendre P (2008). "Beals smoothing revisited." *Oecologia*, **156**, 657–669.

Fisher RA, Corbet AS, Williams CB (1943). "The relation between the number of species and the number of individuals in a random sample of animal population." *Journal of Animal Ecology*, **12**, 42–58.

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

Heck KL, van Belle G, Simberloff D (1975). "Explicit calculation of the rarefaction diversity measurement and the determination of sufficient sample size." *Ecology*, **56**, 1459–1461.

Hill MO (1973). "Diversity and evenness: a unifying notation and its consequences." *Ecology*, **54**, 427–473.

Hurlbert SH (1971). "The nonconcept of species diversity: a critique and alternative parameters." *Ecology*, **52**, 577–586.

Koleff P, Gaston KJ, Lennon JJ (2003). "Measuring beta diversity for presence-absence data." *Journal of Animal Ecology*, **72**, 367–382.

McCune B (1987). "Improving community ordination with the Beals smoothing function." *Ecoscience*, **1**, 82–86.

O'Hara RB (2005). "Species richness estimators: how many species can dance on the head of a pin." *Journal of Animal Ecology*, **74**, 375–386.

Petchey OL, Gaston KJ (2002). "Functional diversity (FD), species richness and community composition." *Ecology Letters*, **5**, 402–411.

Petchey OL, Gaston KJ (2006). "Functional diversity: back to basics and looking forward." *Ecology Letters*, **9**, 741–758.

Preston FW (1948). "The commonness and rarity of species." *Ecology*, **29**, 254–283.

Smith EP, van Belle G (1984). "Nonparametric estimation of species richness." *Biometrics*, **40**, 119–129.

Tóthmérész B (1995). "Comparison of different methods for diversity ordering." *Journal of Vegetation Science*, **6**, 283–290.

Tuomisto H (2010a). "A diversity of beta diversities: straightening up a concept gone awry. 1. Defining beta diversity as a function of alpha and gamma diversity." *Ecography*, **33**, 2–22.

Tuomisto H (2010b). "A diversity of beta diversities: straightening up a concept gone awry. 2. Quantifying beta diversity and related phenomena." *Ecography*, **33**, 23–45.

Ugland KI, Gray JS, Ellingsen KE (2003). "The species-accumulation curve and estimation of species richness." *Journal of Animal Ecology*, **72**, 888–897.

Whittaker RH (1960). "Vegetation of Siskiyou mountains, Oregon and California." *Ecological Monographs*, **30**, 279–338.

Whittaker RH (1965). "Dominance and diversity in plant communities." *Science*, **147**, 250–260.

Williamson M, Gaston KJ (2005). "The lognormal distribution is not an appropriate null hypothesis for the species-abundance distribution." *Journal of Animal Ecology*, **74**, 409–422.

Wilson JB (1991). "Methods of fitting dominance/diversity curves." *Journal of Vegetation Science*, **2**, 35–46.