VEGAN: ECOLOGICAL DIVERSITY

JARI OKSANEN

Contents

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 and Bob O'Hara.

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 1ha plots in the Barro Colorado Island. The following steps make these data available for the document:

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

Date: Id: diversity-vegan.Rnw 1036 2009-10-07 07:06:36Z jarioksa processed with vegan 1.17-3 in R version 2.11.1 (2010-05-31) on June 16, 2010.

1. Diversity indices

Function `diversity` finds the most commonly used diversity indices:

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

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

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

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 Rényi diversities of order a:

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

or the corresponding Hill numbers $N_a = \exp(H_a)$. 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 $H_\infty = -\log(\max p_i)$. We select a random subset of five sites for Rényi diversities:

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

We can really regard a site more diverse if all of its Rényi diversities are higher than in another site. We can inspect this graphically using the standard `plot` function for the `renyi` result (Fig. 1).

Finally, the α parameter of Fisher's log-series can be used as a diversity index:

```
> 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:

$$\hat{S}_n = \sum_{i=1}^{S} (1 - p_i), \text{ where } \quad p_i = \binom{N - x_i}{n} \Big/ \binom{N}{n} \tag{5}$$

where 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 p_i give the probabilities that species i does not occur in a sample of size n. This is defined only when $N - x_i > n$, but for other

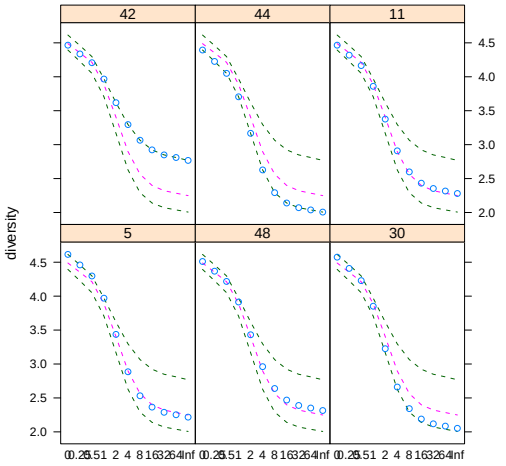

FIGURE 1. Rényi diversities in six randomly selected plots. The plot uses Trellis graphics with a separate panel for each site. The dots show the values for sites, and the lines the extremes and median in the data set.

cases $p_i = 0$ or the species is sure to occur in the sample. The variance of rarefied richness is:

$$s^2 = p_i(1 - p_i) + 2\sum_{i=1}^{S}\sum_{j>i}\left[\binom{N - x_i - x_j}{n} \Big/ \binom{N}{n} - p_i p_j\right] \tag{6}$$

Equation 6 actually is of the same form as the variance of sum of correlated variables:

$$\operatorname{var}\left(\sum x_i\right) = \sum \operatorname{var}(x_i) - 2\sum_{i=1}^{S}\sum_{j>i}\operatorname{cov}(x_i, x_j) \tag{7}$$

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.

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 (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, 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 see how different two different species are. 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 properties of species.

3.1. **Taxonomic diversity: average distance of properties.** The two basic indices are called taxonomic diversity (Δ) and taxonomic distinctness (Δ^*):

$$\Delta = \frac{\sum\sum_{i<j}\omega_{ij}x_ix_j}{n(n-1)/2} \tag{8}$$

$$\Delta^* = \frac{\sum\sum_{i<j}\omega_{ij}x_ix_j}{\sum\sum_{i<j}x_ix_j} \tag{9}$$

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^+:

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

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: 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. 2):

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

3.2. **Functional diversity: the height of property 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 properties which are translated to pairwise distances among species and then to clustering trees of species properties. The argument for trees is that in this way a single

[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.

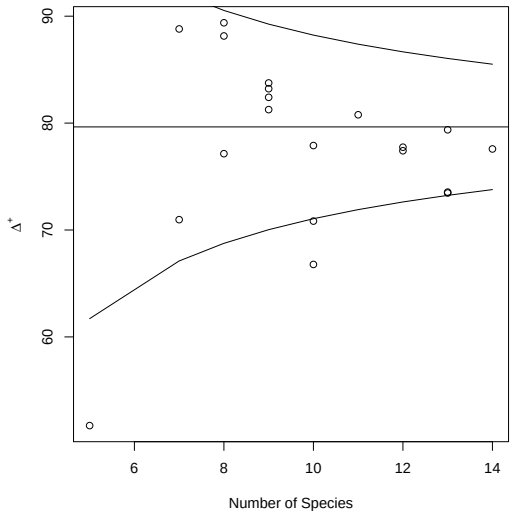

FIGURE 2. Taxonomic diversity Δ^+ for the dune meadow data. The points are diversity values of single sites, and the funnel is their approximate confidence intervals (2× standard error).

deviant species will have a small influence, since its difference is evaluated only once instead of evaluating its distance to all other species.

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. 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:

$$\hat{f}_n = \frac{\alpha x^n}{n} \tag{11}$$

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. 3):

```
> k <- sample(nrow(BCI), 1)
> fish <- fisherfit(BCI[k, ])
> fish
Fisher log series model
No. of species: 92

      Estimate Std. Error
alpha   35.888     4.6777
```

We already saw α as a diversity index. Now we also obtained estimate of standard error of α (these also are optionally available in `fisherfit`). The standard errors are based on the second derivatives (curvature) of log-likelihood at the solution of

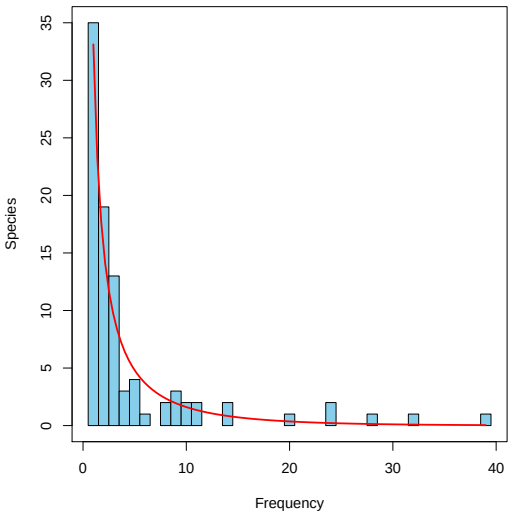

FIGURE 3. Fisher's log-series fitted to one randomly selected site (36).

α. The distribution of α is often non-normal and skewed, and standard errors are of not much use. However, `fisherfit` has a `profile` method that can be used to inspect the validity of normal assumptions, and will be used in calculations of confidence intervals from profile deviance:

```
> confint(fish)
```

```
   2.5 %   97.5 %
27.64699 46.10228
```

Preston's log-normal model is the main challenger to Fisher's log-series. 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. 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 36):

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

```
Preston lognormal model
Method: maximized likelihood to log2 abundances
No. of species: 92

      mode      width         S0
 0.8352028  1.7913322 24.1844800

Frequencies by Octave
                0        1        2        3        4        5
Observed 35.00000 19.00000 16.00000  7.00000 9.000000 5.000000
Fitted   21.69362 24.08235 19.57600 11.65218 5.078656 1.620871
                 6
Observed 1.0000000
Fitted   0.3787968
```

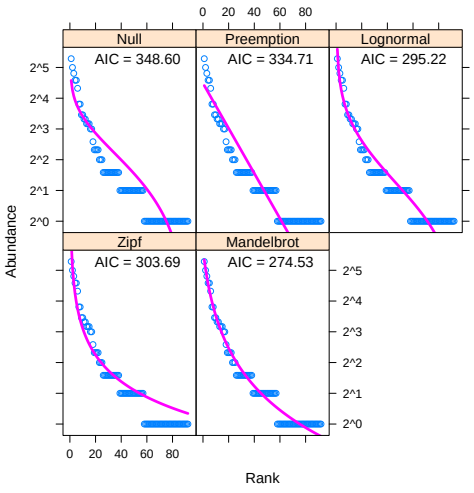

FIGURE 4. Ranked abundance distribution models for a random plot (no. 36). The best model has the lowest AIC.

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. 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 using maximum likelihood estimation:

$$\hat{a}_r = \frac{N}{S}\sum_{k=r}^{S}\frac{1}{k} \qquad \text{brokenstick} \tag{12}$$

$$\hat{a}_r = N\alpha(1-\alpha)^{r-1} \qquad \text{preemption} \tag{13}$$

$$\hat{a}_r = \exp\left[\log(\mu) + \log(\sigma)\Phi\right] \qquad \text{log-normal} \tag{14}$$

$$\hat{a}_r = N\hat{p}_1 r^{\gamma} \qquad \text{Zipf} \tag{15}$$

$$\hat{a}_r = Nc(r+\beta)^{\gamma} \qquad \text{Zipf–Mandelbrot} \tag{16}$$

Where $\hat{a}_r$ 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. 4):

```
> rad <- radfit(BCI[k, ])
> rad
RAD models, family poisson
No. of species 92, total abundance 430

            par1      par2     par3    Deviance AIC      BIC
Null                                    85.9601 348.6007 348.6007
Preemption  0.049495                    70.0645 334.7051 337.2269
Lognormal   0.86697    1.1832           28.5832 295.2238 300.2674
Zipf        0.14615   -0.86285          37.0519 303.6925 308.7361
Mandelbrot  1.6203    -1.4888   5.7543   5.8926 274.5332 282.0986
```

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 $\hat{p}_1, \gamma$, 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 (5) to these individuals. Kindt's exact accumulator resembles rarefaction:

$$\hat{S}_n = \sum_{i=1}^{S}(1 - p_i), \text{ where } \quad p_i = \binom{N - f_i}{n} \Big/ \binom{N}{n} \tag{17}$$

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

$$s^2 = p_i(1 - p_i) + 2\sum_{i=1}^{S}\sum_{j>i}\left(r_{ij}\sqrt{p_i(1 - p_i)}\sqrt{p_j(1 - p_j)}\right) \tag{18}$$

where r_{ij} is the correlation coefficient between species i and j. Both of these are unpublished: eq. 17 was developed by Roeland Kindt, and eq. 18 by Jari Oksanen. The third analytic method was suggested by Coleman:

$$S_n = \sum_{i=1}^{S}(1 - p_i), \text{ where } \quad p_i = \left(1 - \frac{1}{n}\right)^{f_i} \tag{19}$$

and he suggested variance $s^2 = p_i(1 - p_i)$ which ignores the covariance component. In addition, eq. 19 does not properly handle sampling without replacement and underestimates the species accumulation curve.

The recommended is Kindt's exact method (Fig. 5):

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

5.2. **Beta diversity.** Whittaker 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, but almost everybody understand that as a measure of general heterogeneity: how many more species do you have in a collection of sites compared to an average site.

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 ($\bar{\alpha}$):

$$\beta = S/\bar{\alpha} - 1 \tag{20}$$

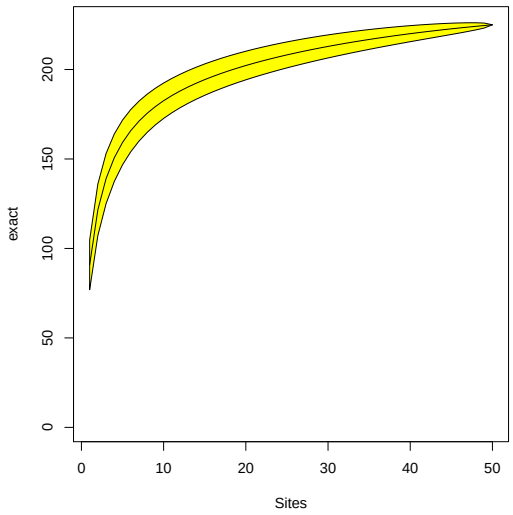

FIGURE 5. Species accumulation curve for the BCI data; exact method.

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. 20 is problematic because S increases with the number of sites even when sites are all subsets of the same community. Whittaker 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 20 can be expressed as:

$$\beta = \frac{a+b+c}{(2a+b+c)/2} - 1 = \frac{b+c}{2a+b+c} \tag{21}$$

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. 20. All commonly used indices can be found using `betadiver`. 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(b,c)/(pmax(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(b,c)/(a+b+c)
18 "l" = (b+c)/2
19 "19" = 2*(b*c+1)/((a+b+c)^2+(a+b+c))
20 "hk" = (b+c)/(2*a+b+c)
21 "rlb" = a/(a+c)
22 "sim" = pmin(b,c)/(pmin(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

$$\hat{S} = cX^z \tag{22}$$

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, z is typically about 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 $z > 0.3$. 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. 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. 6):

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

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

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

Average distance to centroid:
    BF     HF     NM     SF
0.3080 0.2512 0.4406 0.3635

Eigenvalues for PCoA axes:
  PCoA1   PCoA2   PCoA3   PCoA4   PCoA5   PCoA6   PCoA7   PCoA8   PCoA9
```

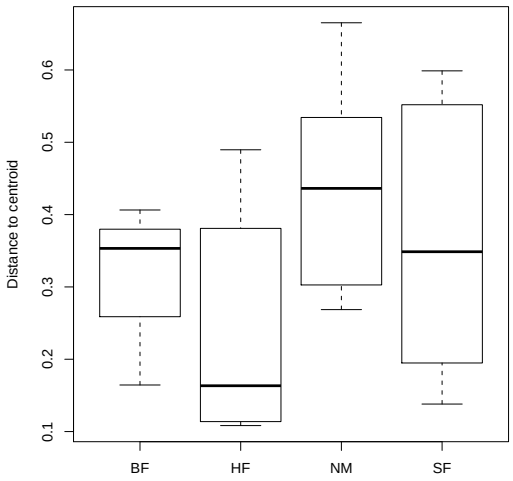

FIGURE 6. Box plots of beta diversity measured as the average steepness (z) of the species area curve in the Arrhenius model $S = cX^z$ in Management classes of dune meadows.

```
 1.6547  0.8870  0.5334  0.3744  0.2873  0.2245  0.1613  0.0810  0.0652
 PCoA10  PCoA11  PCoA12  PCoA13  PCoA14  PCoA15  PCoA16  PCoA17  PCoA18
 0.0353  0.0183  0.0040 -0.0042 -0.0194 -0.0369 -0.0429 -0.0536 -0.0602
 PCoA19
-0.0828
```

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.

Function `specpool` implements the following models to estimate the pool size S_p:

$$S_p = S_o + \frac{f_1^2}{2f_2} \qquad \text{Chao} \tag{23}$$

$$S_p = S_o + f_1\frac{N-1}{N} \qquad \text{1st order Jackknife} \tag{24}$$

$$S_p = S_o + f_1\frac{2N-3}{N} + f_2\frac{(N-2)^2}{N(N-1)} \qquad \text{2nd order Jackknife} \tag{25}$$

$$S_p = S_o + \sum_{i=1}^{S_o}(1-p_i)^N \qquad \text{Bootstrap} \tag{26}$$

Here S_o is the observed number of species, f_1 and f_2 are the numbers of species observed once or twice, N is the number of sites, and p_i are proportions of species. 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 any many species as we missed originally.

The variance estimators of Chao is:

$$s^2 = f_2\left(\frac{G^4}{4} + G^3 + \frac{G^2}{2}\right), \text{ where } \quad G = \frac{f_1}{f_2} \tag{27}$$

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:

$$s^2 = \left(\sum_{i=1}^{N} r_i^2 - \frac{f_1}{N}\right)\frac{N-1}{N} \tag{28}$$

Variance of the second-order jackknife is not evaluated in `specpool` (but contributions are welcome). For the variance of bootstrap estimator, it is practical to define a new variable $q_i = (1-p_i)^N$ for each species:

$$s^2 = \sum_{i=1}^{S_o} q_i(1-q_i) + 2\sum\sum Z_p, \quad \text{where} \quad Z_p = \ldots \tag{29}$$

The extrapolated richness values for the whole BCI data are:

```
> specpool(BCI)
    Species     chao  chao.se  jack1 jack1.se    jack2     boot
All     225 236.6053 6.659395 245.58 5.650522 247.8722 235.6862
     boot.se  n
All 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:

```
> s <- sample(nrow(BCI), 25)
> specpool(BCI[s, ])
    Species     chao  chao.se  jack1 jack1.se   jack2     boot  boot.se
All     209 231.5333 11.59737 233.96 7.394701 244.655 220.9073 4.106949
     n
All 25
```

These typically are even lower than the observed richness (225 species) at the whole data set.

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, ])
                  36
S.obs     92.000000
S.chao1  121.750000
se.chao1  14.342232
S.ACE    136.318712
se.ACE     6.669613
```

Chao's method is similar as above, but uses another, "unbiased" equation. ACE is based on rare species also:

$$\begin{aligned} S_p &= S_{\text{abund}} + \frac{S_{\text{rare}}}{C_{\text{ACE}}} + \frac{a_1}{C_{\text{ACE}}}\gamma^2 \quad \text{where} \\ C_{\text{ACE}} &= 1 - \frac{a_1}{N_{\text{rare}}} \\ \gamma^2 &= \frac{S_{\text{rare}}}{C_{\text{ACE}}}\sum_{i=1}^{10} i(i-1)a_1\frac{N_{\text{rare}}-1}{N_{\text{rare}}} \end{aligned} \tag{30}$$

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 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:

$$S_p = S_\mu \sigma \sqrt{2\pi} \tag{31}$$

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 fits badly, and `prestonfit` is particularly poor. Therefore the following explicitly uses `prestondistr`, although this also may fail:

```
> veiledspec(prestondistr(BCI[k, ]))
Extrapolated     Observed       Veiled
   108.59325     92.00000     16.59325
> veiledspec(BCI[k, ])
Extrapolated     Observed       Veiled
    212901.8         92.0     212809.8
```

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

Function `beals` implement Beals smoothing:

```
> 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. 7):

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

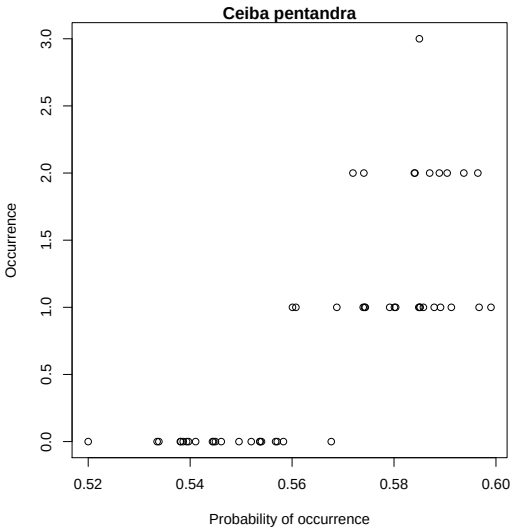

FIGURE 7. Beals smoothing for *Ceiba pentandra*.