VEGAN: ECOLOGICAL DIVERSITY

JARI OKSANEN

Contents

The `vegan` packages 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 units than counts of individuals of variable sizes. However, this document only 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)
```

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

Date: Id: diversity-vegan.Rnw 205 2008-02-15 06:08:56Z jarioksa processed with vegan 1.11-0 in R version 2.6.2 (2008-02-08) on February 20, 2008.

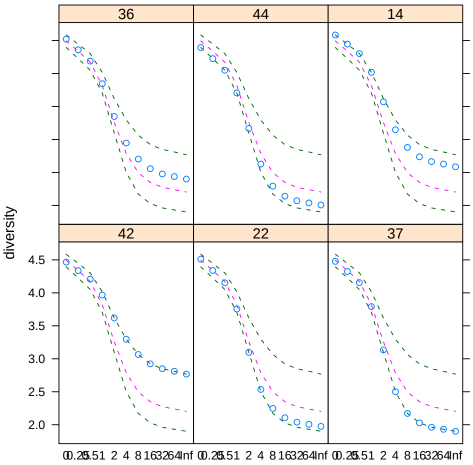

FIGURE 1. Rényi diversities in six randomly selected plots. The plot uses Trellis graphics with a separate panel for each site. The dots show the values for sites, and the lines the extremes and median in the data set.

logarithms (and then we mark index as H'), but $b = 2$ has theoretical justification. 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.

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 $N_\infty = 1/(\max p_i)$. 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. p_i give the probabilities that species i does not occur in a sample of size n. This is only defined for $N - x_i > n$, but for other 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:

$$\mathrm{var}\left(\sum x_i\right) = \sum \mathrm{var}(x_i) - 2\mathrm{cov}(x_i, x_j) \tag{7}$$

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, and rarefaction richness often shares the problems of Rényi diversities.

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 only is a finite sample variant of Simpson's diversity index (or, more precisely of $D_1 + 1$), and almost identical with sample sizes in BCI:

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

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

3. 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.

3.1. **Fisher and Preston.**

In Fisher's log-series, the expected number of species with n individuals is:

$$\hat{f}_n = \frac{\alpha x^n}{n} \tag{8}$$

where 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
```

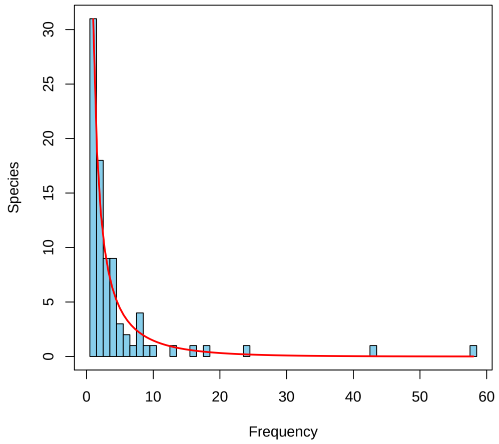

FIGURE 2. Fisher's log-series fitted to one randomly selected site (28).

```
Fisher log series model
No. of species: 85

      Estimate Std. Error
alpha   33.654     4.5788
```

We already saw this model as a diversity index. Now we also obtained estimate of standard error of α (these also are optionally available in `fisher.fit`). The standard errors are based on the second derivatives (curvature) of the partial derivatives of log-likelihood at the solution of α. The distribution of α often is very 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 %
25.62719 43.70514
```

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:

```
> prestondistr(BCI[k, ])
Preston lognormal model
Method: maximized likelihood to log2 abundances
No. of species: 85

      mode      width         S0
 0.9394031  1.6444133 23.4100353
```

```
Frequencies by Octave
                0        1        2        3        4        5
Observed 31.00000 18.00000 18.00000 10.00000 4.000000 2.000000
Fitted   19.88549 23.39415 19.01393 10.67653 4.141737 1.110014
                 6
Observed 2.0000000
Fitted   0.2055267
```

3.2. **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, among other names, as ranked abundance distribution curves, dominance–diversity curves and Whittaker plots. Function `radfit` fits some of the most popular models using maximum likelihood estimation:

$$\hat{a}_r = \frac{N}{S} \sum_{k=r}^{S} \frac{1}{k} \qquad \text{brokenstick} \tag{9}$$

$$\hat{a}_r = N\alpha(1-\alpha)^{r-1} \qquad \text{preemption} \tag{10}$$

$$\hat{a}_r = \exp\left[\log(\mu) + \log(\sigma)\Phi\right] \qquad \text{log-normal} \tag{11}$$

$$\hat{a}_r = N\hat{p}_1 r^{\gamma} \qquad \text{Zipf} \tag{12}$$

$$\hat{a}_r = Nc(r+\beta)^{\gamma} \qquad \text{Zipf–Mandelbrot} \tag{13}$$

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. 3):

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

RAD models, family poisson
No. of species 85, total abundance 387

           par1      par2     par3     Deviance AIC      BIC
Null                                   111.8736 353.0672 353.0672
Preemption  0.053337                   121.0869 364.2806 366.7232
Lognormal   0.76046   1.255             28.3779 273.5715 278.4568
Zipf        0.17283  -0.93043            8.2282 253.4219 258.3072
Mandelbrot  0.25035  -1.0368   0.58633   6.2294 253.4230 260.7510
```

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.

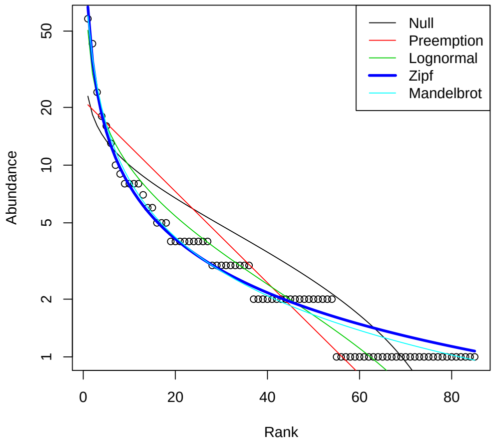

FIGURE 3. Ranked abundance distribution models for a random plot (no. 28). The best model is chosen by the AIC, and displayed with a thick line.

4. Species accumulation and species pool

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

4.1. 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:

$$\hat{S}_n = \sum_{i=1}^{S}(1-p_i), \text{ where } \quad p_i = \binom{N-f_i}{n} \Big/ \binom{N}{n} \tag{14}$$

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

$$s^2 = p_i(1-p_i) + 2\sum_{i=1}^{S}\sum_{j>i}\left(r_{ij}\sqrt{p_i(1-p_i)}\sqrt{p_j(1-p_j)}\right) \tag{15}$$

where r_{ij} is the correlation coefficient between species i and j. Both of these are unpublished: eq. 14 was developed by Roeland Kindt, and eq. 15 by Jari Oksanen. The third analytic method was suggested by Coleman:

$$S_n = \sum_{i=1}^{S}(1-p_i), \text{ where } \quad p_i = \left(1-\frac{1}{n}\right)^{f_i} \tag{16}$$

and he suggested variance $s^2 = p_i(1-p_i)$ which ignores the covariance component. In addition, eq. 16 does not properly handle sampling without replacement and underestimates the species accumulation curve.

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

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

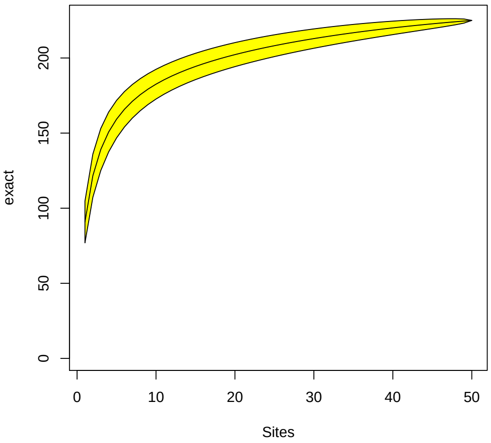

FIGURE 4. Species accumulation curve for the BCI data; exact method.

4.2. **Number of unseen species.** Species accumulation models indicate that not all potential species are seen in any sites. These unseen species also belong to the species pool of the site. Functions specpool and estimateR implement some methods of estimating the number of unseen species. Function specpool studies a collection of sites, and assumes how many species may be unobserved. Function estimateR works with counts of individuals, and also 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:

$$S_p = S_o + \frac{f_1^2}{2f_2} \qquad \text{Chao} \tag{17}$$

$$S_p = S_o + f_1 \frac{N-1}{N} \qquad \text{1st order Jackknife} \tag{18}$$

$$S_p = S_o + f_1 \frac{2N-3}{N} + f_2 \frac{(N-2)^2}{N(N-1)} \qquad \text{2nd order Jackknife} \tag{19}$$

$$S_p = S_o + \sum_{i=1}^{S_o} (1-p_i)^N \qquad \text{Bootstrap} \tag{20}$$

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 are of Chao is:

$$s^2 = f_2 \left(\frac{G^4}{4} + G^3 + \frac{G^2}{2} \right), \text{ where } \quad G = \frac{f_1}{f_2} \tag{21}$$

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:

$$s^2 = \left(\sum_{i=1}^{N} r_i^2 - \frac{f_1}{N} \right) \frac{N-1}{N} \tag{22}$$

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:

$$s^2 = \sum_{i=1}^{S_o} q_i(1 - q_i) + 2 \sum \sum Z_p, \quad \text{where} \qquad (23)$$
$$Z_p = \dots$$

The extrapolated richness values for the whole BCI data are:

```
> specpool(BCI)
    Species     Chao  Chao.SE Jack.1 Jack1.SE   Jack.2     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 Jack.1 Jack1.SE   Jack.2     Boot
All     212 242.0312 14.23972 241.76 8.528165 256.1733 225.7051
     Boot.SE  n
All 4.684375 25
```

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

4.3. **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, ])
                  28
S.obs      85.000000
S.chao1   109.473684
se.chao1   12.578970
S.ACE     116.301606
se.ACE      5.509697
```

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

$$\begin{aligned} S_p &= S_\text{abund} + \frac{S_\text{rare}}{C_\text{ACE}} + \frac{a_1}{C_\text{ACE}} \gamma^2 \quad \text{where} \\ C_\text{ACE} &= 1 - \frac{a_1}{N_\text{rare}} \\ \gamma^2 &= \frac{S_\text{rare}}{C_\text{ACE}} \sum_{i=1}^{10} i(i-1) a_1 \frac{N_\text{rare} - 1}{N_\text{rare}} \end{aligned} \qquad (24)$$

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.

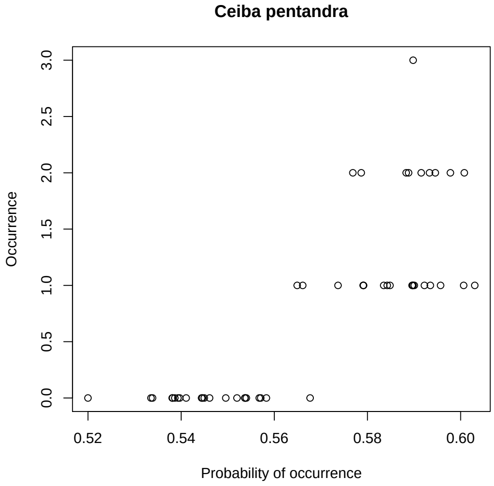

FIGURE 5. Beals smoothing for *Ceiba pentandra*.

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:

$$S_p = S_\mu \sigma \sqrt{2\pi} \tag{25}$$

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
    96.49459     85.00000     11.49459
> veiledspec(BCI[k, ])
Extrapolated     Observed       Veiled
    406.4778      85.0000     321.4778
```

4.4. **Probability of pool membership.** Beals smoothing was originally suggested as tool of regularizing data for ordination. It regularizes data too strongly for that purpose, 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 (Fig. 5):

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