

Lab 4. Species Distribution Patterns

Smit, A. J.
University of the Western Cape


2021-01-01

 BCB743


This material must be reviewed by BCB743 students in Week 1 of Quantitative Ecology.

 This Lab Accompanies the Following Lecture

- Lecture 6: Unified Ecology

 Data For This Lab

- The Barro Colorado Island Tree Counts data (Condit et al. 2002) – load **vegan** and load the data with `data(BCI)`
- The Oribatid mite data (Borcard et al. 1992, Borcard and Legendre 1994) – load **vegan** and load the data with `data(mite)`
- The seaweed species data (Smit et al. 2017) – `SeaweedSpp.csv`
- The Doubs River species data (Verneaux 1973, Borcard et al. 2011) – `DoubsSpe.csv`

 Reading Required For This Lab

- Matthews and Whittaker (2015)
- Shade et al. (2018)

In this Lab, we will calculate the various species distribution patterns included in the paper by Shade et al. (2018).

The Data

We will calculate each for the Barro Colorado Island Tree Counts data that come with **vegan**. See `?vegan::BCI` for a description of the data contained with the package, as well as a selection

of publications relevant to the data and analyses. The primary publication of interest is Condit et al. (2002).

```
library(tidyverse)
library(vegan)
```

```
#library(vegan) # already loaded
#library(tidyverse) # already loaded
data(BCI) # data contained within vegan

# make a head-tail function
ht <- function(d) rbind(head(d, 7), tail(d, 7))

# Lets look at a portion of the data:
ht(BCI)[1:7,1:7]
```

	Abarema.macradenia	Vachellia.melanoceras	Acalypha.diversifolia
1	0	0	0
2	0	0	0
3	0	0	0
4	0	0	0
5	0	0	0
6	0	0	0
7	0	0	0

	Acalypha.macrostachya	Adelia.triloba	Aegiphila.panamensis
1	0	0	0
2	0	0	0
3	0	0	0
4	0	3	0
5	0	1	1
6	0	0	0
7	0	0	1

	Alchornea.costaricensis
1	2
2	1
3	2
4	18
5	3
6	2
7	0

Species-Abundance Distribution

The species abundance distribution (SAD) expresses one of ecology's most persistent regularities, viz., in almost any community, a few species dominate in abundance, while the majority are comparatively rare. This skewed pattern is often summarised as “few common, many rare” and has been recognised since the earliest quantitative analyses of community structure and has become

a central object of ecological theory. Graphically, SADs plot species abundance against either species identity or rank to provide a means of comparing empirical assemblages with theoretical expectations of community organisation.

The first systematic formulation of this pattern was developed by Fisher et al. (1943), who observed that species frequencies often conform to what he termed a **logarithmic series distribution**. In this representation, the expected number of species with a given number of individuals declines rapidly as abundance increases, reflecting communities where most species occur at low frequencies and only a small fraction reach very high densities. The curve of Fisher's log-series distribution thus shows the expected number of species f with n observed individuals, and the interpretation of this curve is conceptually similar across all SAD models—only the underlying mathematics and rationale differ. Fisher's model remains widely used, both as a conceptual benchmark and as a statistical model. In **R**, the **vegan** package implements this procedure through the `fisherfit()` function, which estimates the log-series parameter and compares it to observed community data (Figure 1).

```
# take one random sample of a row (site):  
# for this website's purpose, this function ensure the same random  
# sample is drawn each time the web page is recreated  
set.seed(13)  
k <- sample(nrow(BCI), 1)  
fish <- fisherfit(BCI[k,])  
fish
```

```
Fisher log series model  
No. of species: 95  
Fisher alpha: 39.87659
```

```
plot(fish)
```

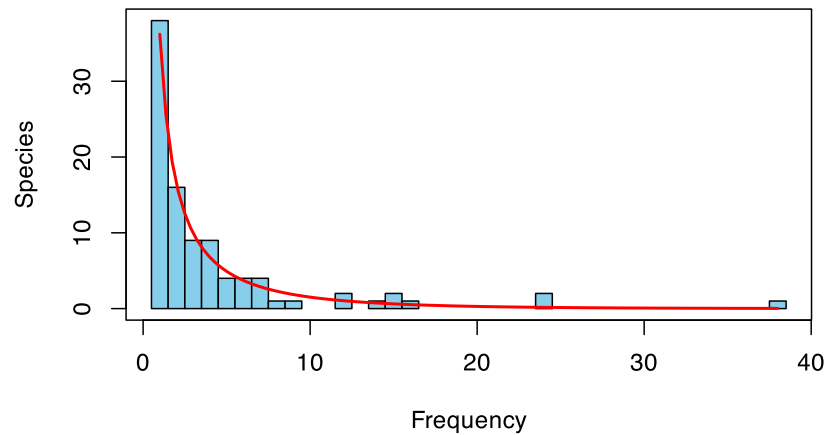


Figure 1: Fisher's log series distribution calculated for the Barro Colorado Island Tree Counts data.

Building on this foundation, Preston (1948) introduced an alternative formulation: the **log-normal distribution** of species abundances. Instead of ranking species, he grouped them into “octaves”, *i.e.*, abundance classes that double in size (1, 2, 4, 8, 16, 32, ...), and showed that, in sufficiently well-sampled communities, the frequency of species within these classes approximates a bell-shaped Gaussian curve when plotted on a logarithmic scale. Preston also emphasised the effect of incomplete sampling: rare species falling below a “veil line” remain undetected, giving the distribution its characteristic truncated form. This method captures the idea that rarity and commonness are not absolute properties but functions of sampling intensity. In **vegan**, Preston's method is implemented with mathematical modifications to handle ties more effectively via the `prestondistr()` function (Figure 2).

```
pres <- prestondistr(BCI[k,])
pres
```

```
Preston lognormal model
Method: maximized likelihood to log2 abundances
No. of species: 95
```

```
      mode      width      S0
0.9234918 1.6267630 26.4300640
```

```
Frequencies by Octave
```

```
      0      1      2      3      4      5      6
Observed 19.00000 27.00000 21.50000 17.00000 7.000000 2.500000 1.0000000
Fitted   22.49669 26.40085 21.23279 11.70269 4.420327 1.144228 0.2029835
```

```
plot(pres)
```

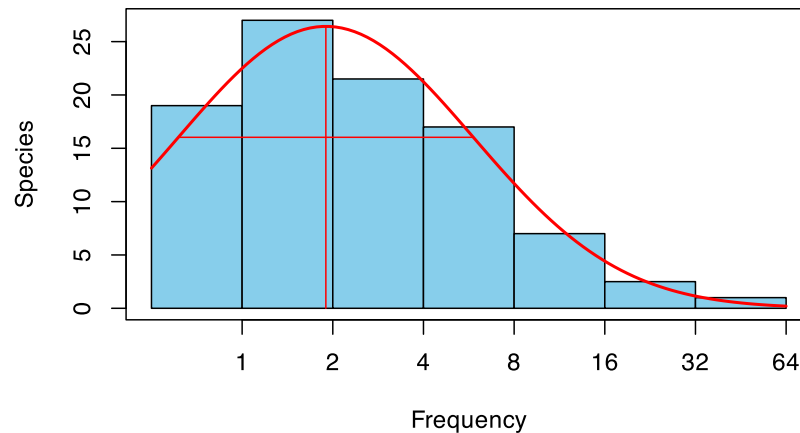


Figure 2: Preston's log-normal distribution demonstrated for the BCI data.

A further step was taken by Whittaker (1965), who reformulated SADs as **rank-abundance distributions** (RADs), also known as dominance–diversity curves or Whittaker plots. Here, species are ordered from most to least abundant along the x -axis, and their relative abundances (often log-transformed) are plotted on the y -axis. The slope and curvature of the resulting profile reveal both evenness and the extent of dominance: steep curves indicate communities dominated by a few species, whereas long, shallow tails point to greater species equity. Unlike Fisher's or Preston's frequency-based models, Whittaker's approach highlights the shape of the assemblage as a whole and facilitates comparison across communities with different richness levels. In **vegan**, this is achieved via the `radfit()` function, which overlays multiple fitted models—broken-stick, preemption, log-normal, Zipf, and Zipf–Mandelbrot—onto the observed ranked data (Figure 3).

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

```
RAD models, family poisson
No. of species 95, total abundance 392
```

	par1	par2	par3	Deviance	AIC	BIC
Null				56.3132	324.6477	324.6477
Preemption	0.042685			55.8621	326.1966	328.7504
Lognormal	0.84069	1.0912		16.1740	288.5085	293.6162

Zipf	0.12791	-0.80986		21.0817	293.4161	298.5239
Mandelbrot	0.66461	-1.2374	4.1886	6.6132	280.9476	288.6093

```
plot(rad)
```

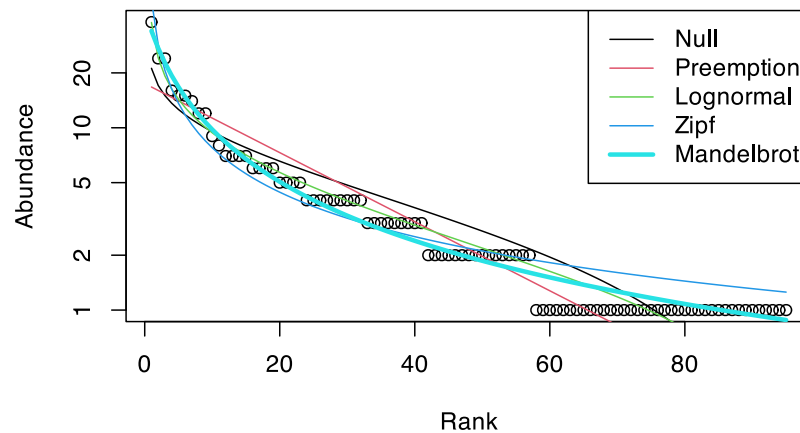


Figure 3: Whittaker's rank abundance distribution curves demonstrated for the BCI data.

We can also fit the rank abundance distribution curves to several sites at once (previously we have done so on only one site) (Figure 4):

```
m <- sample(nrow(BCI), 6)
rad2 <- radfit(BCI[m, ])
rad2
```

Deviance for RAD models:

	3	37	10	13	6	22
Null	86.1127	93.5952	77.2737	52.6207	72.1627	114.1747
Preemption	58.9295	104.0978	62.7210	57.7372	54.7709	110.5156
Lognormal	29.2719	19.0653	20.4770	15.8218	19.5788	26.2510
Zipf	50.1262	11.3048	39.7066	22.8006	32.4630	15.5222
Mandelbrot	5.7342	8.9107	9.8353	12.1701	5.5973	9.6047

```
plot(rad2)
```

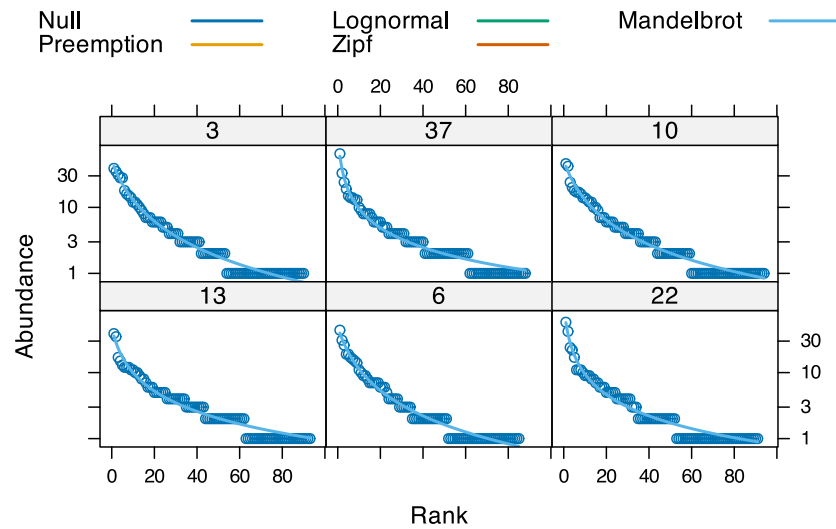


Figure 4: Rank abundance distribution curves fitted to several sites.

Above, we see that the model selected for capturing the shape of the SAD is the Mandelbrot, and it is plotted individually for each of the randomly selected sites. Model selection works through Akaike's or Schwartz's Bayesian information criteria (AIC or BIC; AIC is the default—select the model with the lowest AIC).

BiodiversityR (and here and here) also offers options for rank abundance distribution curves; see `rankabundance()` (Figure 5):

```
library(BiodiversityR)
rankabund <- rankabundance(BCI)
rankabunplot(rankabund, cex = 0.8, pch = 0.8, col = "indianred4")
```

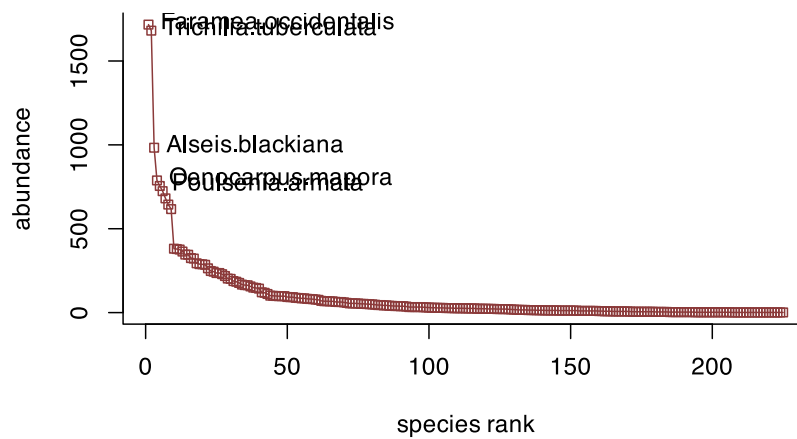


Figure 5: Rank-abundance curves for the BCI data.

Refer to the help files for the respective functions to see their differences.

Occupancy-Abundance Curves

Occupancy refers to the number or proportion of sites in which a species is detected. Occupancy-abundance relationships are used to infer niche specialisation patterns in the sampling region. The hypothesis (almost a theory) is that species that tend to have high local abundance within one site also tend to occupy many other sites (Figure 6).

```
library(ggpubr)

# A function for counts:
# count number of non-zero elements per column
count_fun <- function(x) {
  length(x[x > 0])
}

BCI_OA <- data.frame(occ = apply(BCI, MARGIN = 2, count_fun),
                    ab = apply(BCI, MARGIN = 2, mean))

ggplot(BCI_OA, aes(x = ab, y = occ/max(occ))) +
  geom_point(colour = "indianred3") +
  scale_x_log10() +
  # scale_y_log10() +
  labs(title = "Barro Colorado Island Tree Counts",
       x = "Log (abundance)", y = "Occupancy") +
  theme_linedraw()
```

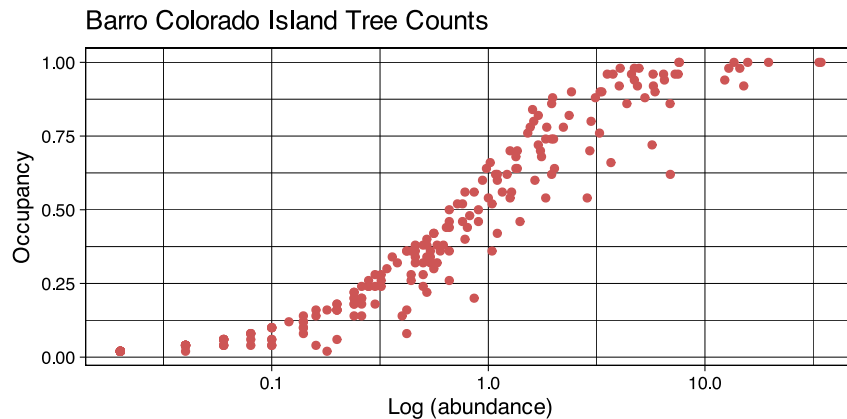



Figure 6: Occupancy-abundance relationships seen in the BCI data.

Species-Area (Accumulation)

Species accumulation curves (species area relationships, SAR) try and estimate the number of unseen species. These curves can be used to predict and compare changes in diversity over increasing spatial extent. Within an ecosystem type, one would expect that more and more species would be added (accumulates) as the number of sampled sites increases (i.e. extent increases). This continues to a point where no more new species are added as the number of sampled sites continues to increase (i.e. the curve plateaus). Species accumulation curves, as the name suggests, accomplishes this by adding (accumulation or collecting) more and more sites and counting the average number of species along y each time a new site is added. See Roeland Kindt's description of how species accumulation curves work (on p. 41). In the community matrix (the sites \times species table), we can do this by successively adding more rows to the curve (seen along the x -axis). The `specaccum()` function has many different ways of adding the new sites to the curve, but the default 'exact' seems to be a sensible choice. **BiodiversityR** has the `accumresult()` function that does nearly the same. Let's demonstrate using **vegan**'s function (Figure 7, Figure 8, and Figure 9):

```
sp1 <- specaccum(BCI)
sp2 <- specaccum(BCI, "random")

# par(mfrow = c(2,2), mar = c(4,2,2,1))
# par(mfrow = c(1,2))
plot(sp1, ci.type = "polygon", col = "indianred4", lwd = 2, ci.lty = 0,
      ci.col = "steelblue2", main = "Default: exact",
      ylab = "No. of species")
```

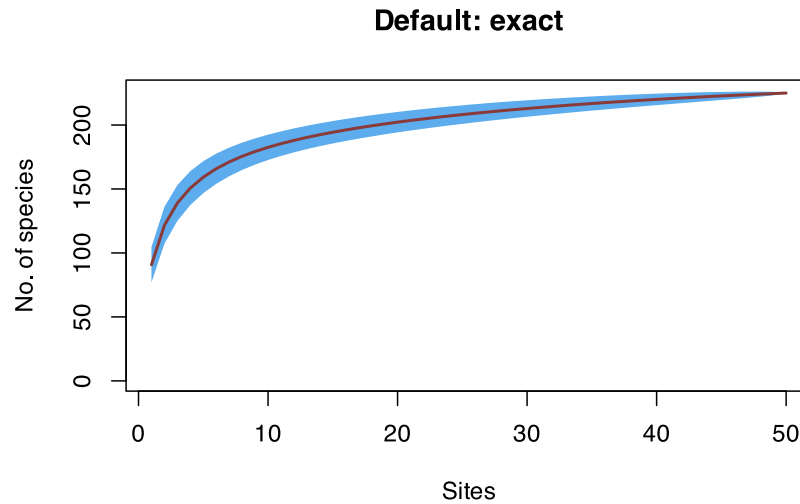


Figure 7: Species-area accumulation curves seen in the BCI data.

```
mods <- fitspecaccum(sp2, "arrh")
plot(mods, col = "indianred", ylab = "No. of species")
boxplot(sp2, col = "yellow", border = "steelblue2", lty = 1, cex = 0.3, add =
TRUE)
sapply(mods$models, AIC)
```

```
[1] 311.4642 303.7835 346.3668 320.0786 338.7978 320.2538 325.6968 346.2671
[9] 320.3900 343.8570 318.2509 369.8303 335.9936 350.8711 327.9831 348.1287
[17] 328.2393 347.8133 324.3837 314.8555 333.1390 340.5678 332.6836 360.5208
[25] 335.3660 325.3150 347.4324 336.7498 336.6374 276.1878 349.9283 295.0268
[33] 308.4656 315.8304 303.0776 329.8425 356.2393 368.4302 318.0514 359.5975
[41] 327.4228 335.7604 259.8340 318.0063 335.7753 285.8790 323.5174 300.3546
[49] 327.1448 355.2747 288.2583 366.5995 287.4120 327.5877 362.6487 323.5904
[57] 339.5650 321.2264 336.6331 353.1295 317.9578 311.6528 336.3613 337.8327
[65] 328.4787 311.6842 345.8035 367.5620 319.0269 305.6546 338.7805 321.8859
[73] 330.6029 326.7097 345.8923 338.4755 352.8710 355.8038 307.7327 329.2355
[81] 341.6628 340.1687 333.4771 348.3144 321.4417 317.4331 339.2211 313.1990
[89] 305.3069 342.4581 318.0308 299.7067 294.7851 324.3237 333.5849 349.2749
[97] 369.8287 323.0041 332.6820 329.3875
```

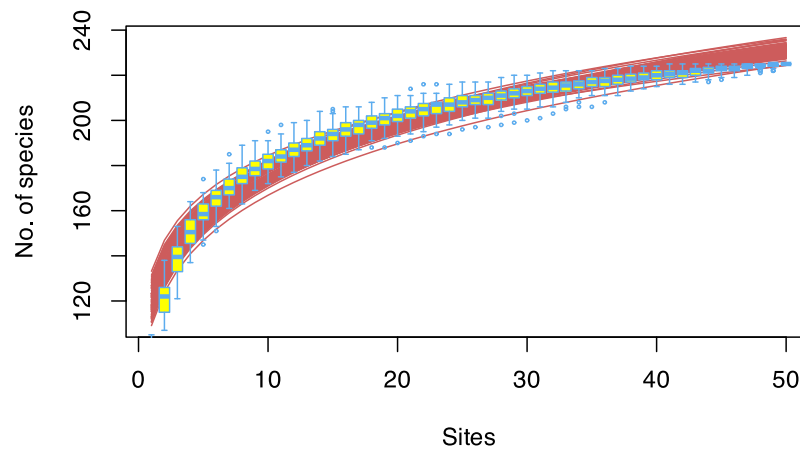


Figure 8: Fit Arrhenius models to all random accumulations

```
accum <- accumresult(BCI, method = "exact", permutations = 100)
accumplot(accum)
```

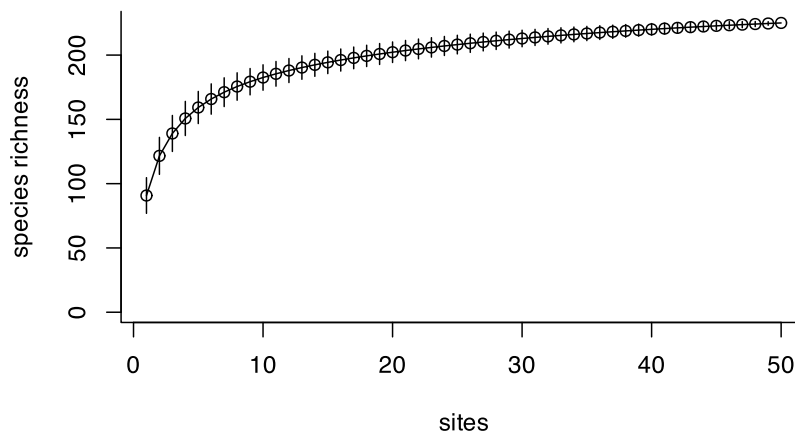


Figure 9: A species accumulation curve.

Species accumulation curves can also be calculated with the `alpha.accum()` function of the **BAT** package (Figure 10). In addition, the **BAT** package can also apply various diversity and species distribution assessments to **phylogenetic** and **functional** diversity. See the examples provided by Cardoso et al. (2015).

```

library(BAT)
BCI.acc <- alpha accum(BCI, prog = FALSE)

par(mfrow = c(1,2))
plot(BCI.acc[,2], BCI.acc[,17], col = "indianred",
     xlab = "Individuals", ylab = "Chao1P")
plot(BCI.acc[,2], slope(BCI.acc)[,17], col = "indianred",
     xlab = "Individuals", ylab = "Slope")

```

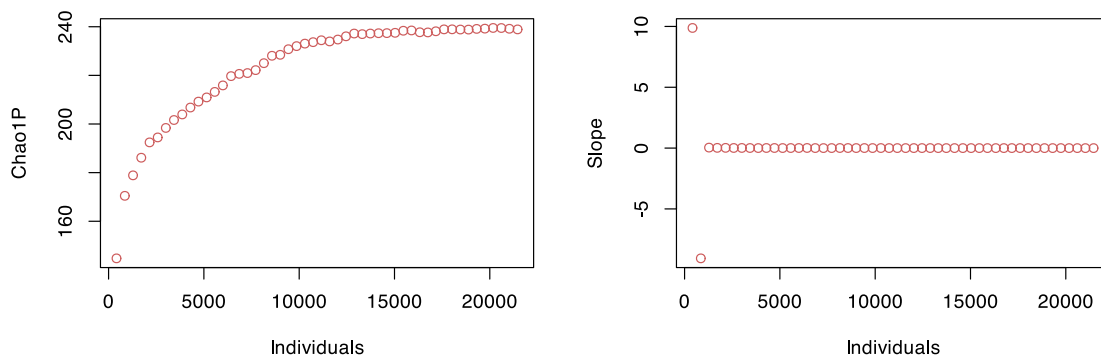


Figure 10: A species accumulation curve made with the `alpha.accum()` function of **BAT**.

Rarefaction Curves

Like species accumulation curves, rarefaction curves also try to estimate the number of unseen species. Rarefaction, meaning to scale down (Heck Jr et al. 1975), is a statistical technique used by ecologists to assess species richness (represented as S , or diversity indices such as Shannon diversity, H' , or Simpson's diversity, λ) from data on species samples, such as that which we may find in site \times species tables. Rarefaction can be used to determine whether a habitat, community, or ecosystem has been sufficiently sampled to fully capture the full complement of species present.

Rarefaction curves may seem similar to species accumulation curves, but there is a difference as I will note below. Species richness, S , accumulates with sample size *or* with the number of individuals sampled (across all species). The first way that rarefaction curves are presented is to show species richness as a function of number of individuals sampled. Here the principle demonstrated is that when only a few individuals are sampled, those individuals may belong to only a few species; however, when more individuals are present more species will be represented. The second approach to rarefaction is to plot the number of samples along x and the species richness along the y -axis (as in SADs too). So, rarefaction shows how richness accumulates with the number of individuals counted or with the number of samples taken. Rarefaction curves rise rapidly at the start when few species have been sampled and the most common species have been found; the slope then decreases and eventually plateaus suggesting that the rarest species remain to be sampled.

But what really distinguishes rarefaction curves from SADs is that rarefaction randomly re-samples the pool of N samples (that is equal or less than the total community size) a number of times, n , and plots the average number of species found in each resample ($1, 2, \dots, n$) as a function of individuals or samples. The `rarecurve()` function draws a rarefaction curve for each row of the species data table. All these plots are made with base R graphics Figure 11, but it will be a trivial exercise to reproduce them with **ggplot2**.

```
# Example provided in ?vegan::rarefy
# observed number of species per row (site)
S <- specnumber(BCI)

# calculate total no. individuals sampled per row, and find the minimum
(raremax <- min(rowSums(BCI)))
```

```
[1] 340
```

```
Srare <- rarefy(BCI, raremax, se = FALSE)
par(mfrow = c(1,2))
plot(S, Srare, col = "indianred3",
      xlab = "Sample size\n(observed no. of individuals)", ylab = "No. species found")
rarecurve(BCI, step = 20, sample = raremax, col = "indianred3", cex = 0.6,
          xlab = "Sample size\n(observed no. of individuals)", ylab = "No. species found")
```

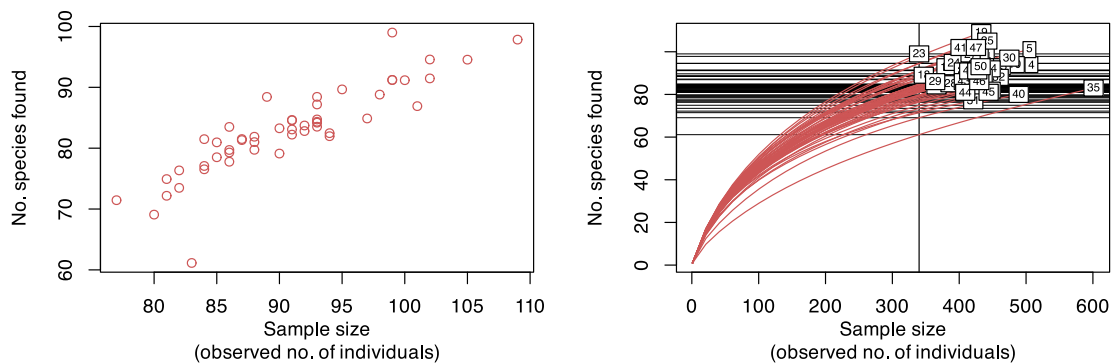


Figure 11: Rarefaction curves for the BCI data.

i Note

iNEXT

We can also use the **iNEXT** package for rarefaction curves. From the package's Introduction Vignette:

iNEXT focuses on three measures of Hill numbers of order q : species richness ($q = 0$), Shannon diversity ($q = 1$, the exponential of Shannon entropy) and Simpson diversity ($q = 2$, the inverse of Simpson concentration). For each diversity measure, iNEXT uses the observed sample of abundance or incidence data (called the "reference sample") to compute diversity estimates and the associated 95% confidence intervals for the following two types of rarefaction and extrapolation (R/E):

1. Sample-size-based R/E sampling curves: iNEXT computes diversity estimates for rarefied and extrapolated samples up to an appropriate size. This type of sampling curve plots the diversity estimates with respect to sample size.
2. Coverage-based R/E sampling curves: iNEXT computes diversity estimates for rarefied and extrapolated samples with sample completeness (as measured by sample coverage) up to an appropriate coverage. This type of sampling curve plots the diversity estimates with respect to sample coverage.

iNEXT also plots the above two types of sampling curves and a sample completeness curve. The sample completeness curve provides a bridge between these two types of curves.

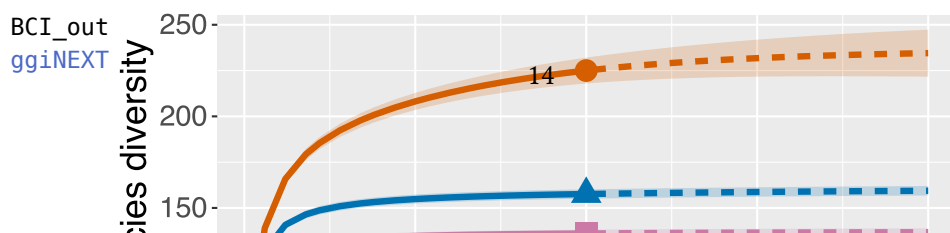
For information about Hill numbers see David Zelený's Analysis of community data in R and Jari Oksanen's coverage of diversity measures in **vegan**.

There are four datasets distributed with iNEXT and numerous examples are provided in the Introduction Vignette. iNEXT has an 'odd' data format that might seem foreign to **vegan** users. To use iNEXT with dataset suitable for analysis in **vegan**, we first need to convert BCI data to a species \times site matrix (Figure 12):

```
library(iNEXT)

# transpose the BCI data:
BCI_t <- list(BCI = t(BCI))
str(BCI_t)
```

```
List of 1
 $ BCI: int [1:225, 1:50] 0 0 0 0 0 0 2 0 0 0 ...
 .. attr(*, "dimnames")=List of 2
 .. ..$ : chr [1:225] "Abarema.macradenia" "Vachellia.melanoceras"
 "Acalypha.diversifolia" "Acalypha.macrostachya" ...
 .. ..$ : chr [1:50] "1" "2" "3" "4" ...
```



Distance-Decay Curves

The principles of distance decay relationships are clearly captured in analyses of β -diversity—see specifically **turnover**, β_{sim} . Distance decay is the primary explanation for the spatial pattern of β -diversity along the South African coast in Smit et al. (2017). A deeper dive into distance decay calculation can be seen in Deep Dive into Gradients.

Elevation and Other Gradients

In one sense, an elevation gradient can be seen as a specific case of distance decay. The Doubs River dataset offers a nice example of data collected along an elevation gradient. Elevation gradients have many similarities with depth gradients (e.g. down the ocean depths) and latitudinal gradients.

! Lab 4

(To be reviewed by BCB743 student but not for marks)

1. Produce the following figures for the species data indicated in [square brackets]:
 - a. species-abundance distribution [mite];
 - b. occupancy-abundance curves [mite];
 - c. species-area curves [seaweed]—note, do not use the **BAT** package's `alpha.accum()` function as your computer might fall over;
 - d. rarefaction curves [mite].

Answer each under its own heading. For each, also explain briefly what the purpose of the analysis is (i.e. what ecological insights might be provided), and describe the findings of your own analysis as well as any ecological implications that you might be able to detect.

2. Using the **biodiversityR** package, find the most dominant species in the Doubs River dataset.
3. Discuss how elevation, depth, or latitudinal gradients are similar in many aspects to distance decay relationships.

! Submission Instructions

The Lab 4 assignment is due at **08:00 on Monday 21 August 2025**.

Provide a **neat and thoroughly annotated** R file which can recreate all the graphs and all calculations. Written answers must be typed in the same file as comments.

Please label the R file as follows:

- BDC334_<first_name>_<last_name>_Lab_4.R

(the < and > must be omitted as they are used in the example as field indicators only).

Submit your appropriately named R documents on iKamva when ready.

Failing to follow these instructions carefully, precisely, and thoroughly will cause you to lose marks, which could cause a significant drop in your score as formatting counts for 15% of the final mark (out of 100%).

Bibliography

Borcard D, Gillet F, Legendre P, others (2011) Numerical ecology with R. Springer

Borcard D, Legendre P (1994) Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). Environmental and Ecological statistics 1:37–61.

Borcard D, Legendre P, Drapeau P (1992) Partialling out the spatial component of ecological variation. Ecology 73:1045–1055.

Cardoso P, Rigal F, Carvalho JC (2015) BAT–Biodiversity Assessment Tools, an R package for the measurement and estimation of alpha and beta taxon, phylogenetic and functional diversity. Methods in Ecology and Evolution 6:232–236.

Condit R, Pitman N, Leigh Jr EG, Chave J, Terborgh J, Foster RB, Núñez P, Aguilar S, Valencia R, Villa G, others (2002) Beta-diversity in tropical forest trees. Science 295:666–669.

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

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

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

Smit AJ, Bolton JJ, Anderson RJ (2017) Seaweeds in two oceans: beta-diversity. Frontiers in Marine Science 4:404.

Verneaux J (1973) Cours d'eau de Franche-Comté (Massif du Jura). Recherches écologiques sur le réseau hydrographique du Doubs.

Whittaker RH (1965) Dominance and Diversity in Land Plant Communities: Numerical relations of species express the importance of competition in community function and evolution. *Science* 147:250–260.