10

Community Composition and Diversity

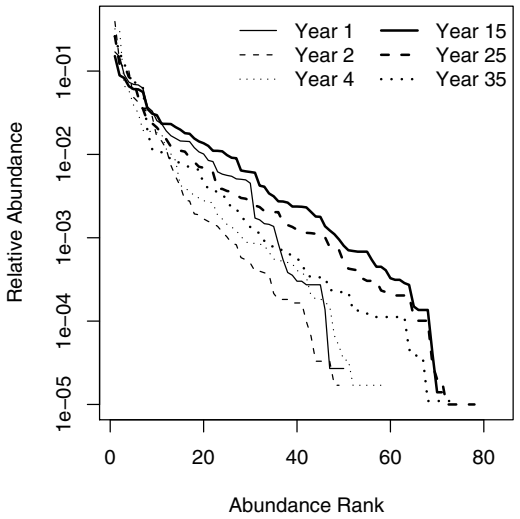


Fig. 10.1: Empirical rank–abundance distributions of successional plant communities (old-fields) within the temperate deciduous forest biome of North America. "Year" indicates the time since abandonment from agriculture. Data from the Buell-Small succession study (http://www.ecostudies.org/bss/)

It seems easy, or at least tractable, to compare the abundance of a single species in two samples. In this chapter, we introduce concepts that ecologists use to compare entire communities in two samples. We focus on two quantities: *species composition*, and *diversity*. We also discuss several issues related to this, including species–abundance distributions, ecological neutral theory, diversity partitioning, and species–area relations. Several packages in R include functions for dealing specifically with these topics. Please see the "Environmetrics" link

M.H.H. Stevens, *A Primer of Ecology with R*, Use R, DOI: 10.1007/978-0-387-89882-7_10,

within the "Task Views" link at any CRAN website for downloading Rpackages. Perhaps the most comprehensive (including both diversity and composition) is the `vegan` package, but many others include important features as well.

10.1 Species Composition

Species composition is merely the set of species in a site or a sample. Typically this includes some measure of abundance at each site, but it may also simply be a list of species at each site, where "abundance" is either presence or absence. Imagine we have four sites (A–D) from which we collect density data on two species, *Salix whompii* and *Fraxinus virga*. We can enter hypothetical data of the following abundances.

```
> dens <- data.frame(Salwho = c(1, 1, 2, 3), Fravir = c(21,
+     8, 13, 5))
> row.names(dens) <- LETTERS[1:4]
> dens
```

```
  Salwho Fravir
A      1     21
B      1      8
C      2     13
D      3      5
```

Next, we plot the abundances of both species; the plotted points then are the sites (Fig. 10.2).

```
> plot(dens, type = "n")
> text(dens, row.names(dens))
```

In Fig. 10.2, we see that the species composition in site A is most different from the composition of site D. That is, the distance between site A and D is greater than between any other sites. The next question, then, is *how* far apart are any two sites? Clearly, this depends on the scale of the measurement (e.g., the values on the axes), and also on how we measure distance through multivariate space.

10.1.1 Measures of abundance

Above we pretended that the abundances were absolute densities (i.e., 1 = one stem per sample). We could of course represent all the abundances differently. For instance, we could calculate *relative density*, where each species in a sample is represented by the *proportion* of the sample comprised of that species. For site A, we divide each species by the sum of all species.

```
> dens[1, ]/sum(dens[1, ])
```

```
   Salwho Fravir
A 0.04545 0.9545
```

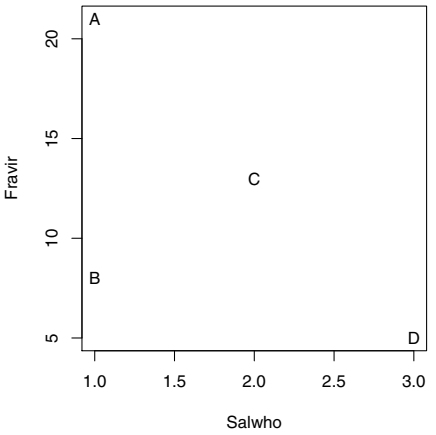


Fig. 10.2: Hypothetical species composition for four sites (A–D).

We see that *Salix* makes up about 5% of the sample for Site A, and *Fraxinus* makes up about 95% of the sample. Once we calculate relative densities for each species at each site, this eliminates differences in total density at each site because all sites then total to 1.

We could also calculate relative measures for any type of data, such as biomass or percent cover.

In most instances, *relative density* refers to the density of a species *relative* to the other species in a sample (above), but it can also be density in a sample relative to other samples. We would thus make each *species* total equal 1, and then its abundance at each site reflects the proportion of a species total abundance comprised by that site. For instance, we can make all *Salix* densities relative to each other.

```
> dens[, 1]/sum(dens[, 1])

[1] 0.1429 0.1429 0.2857 0.4286
```

Here we see that sites A and B both have about 14% of all *Salix* stems, and site D has 43%.

Whether our measures of abundance are absolute or relative, we would like to know how different samples (or sites) are from each other. Perhaps the simplest way to describe the difference among the sites is to calculate the *distances* between each pair of sites.

10.1.2 Distance

There are many ways to calculate a *distance* between a pair of sites. One of the simplest, which we learned in primary school, is *Euclidean distance*. With two species, we have two dimensional space, which is Fig. 10.2. The Euclidean distance between two sites is merely the length of the vector connecting those

sites. We calculate this as $\sqrt{x^2 + y^2}$, where x and y are the (x, y) distances between a pair of sites. The x distance between sites B and C is difference in *Salix* abundance between the two sites,

```
> x <- dens[2, 1] - dens[3, 1]
```

where `dens` is the data frame with sites in rows, and species in different columns. The y distance between sites B and C is difference in *Fraxinus* abundance between the two sites.

```
> y <- dens[2, 2] - dens[3, 2]
```

The Euclidean distance between these is therefore

```
> sqrt(x^2 + y^2)

[1] 5.099
```

Distance is as simple as that. We calculate all pairwise Euclidean distances between sites A–D based on 2 species using built-in functions in R.

```
> (alldists <- dist(dens))

       A      B      C
B 13.000
C  8.062  5.099
D 16.125  3.606  8.062
```

We can generalize this to include any number of species, but it becomes increasingly harder to visualize. We can add a third species, *Mandragora officinarum*, and recalculate pairwise distances between all sites, but now with three species.

```
> dens[["Manoff"]] <- c(11, 3, 7, 5)
> (spp3 <- dist(dens))

       A      B      C
B 15.264
C  9.000  6.481
D 17.205  4.123  8.307
```

We can plot species abundances as we did above, and `pairs(dens)` would give us all the pairwise plots given three species. However, what we really want for species is a 3-D plot. Here we load another package[1] and create a 3-D scatterplot.

```
> pairs(dens)# not shown
> library(scatterplot3d)
> sc1 <- scatterplot3d(dens,  type='h', pch="",
+       xlim=c(0,5), ylim=c(0, 25), zlim=c(0,15))
> text(sc1$xyz.convert(dens), labels=rownames(dens))
```

In three dimensions, Euclidean distances are calculated the same basic way, but we add a third species, and the calculation becomes $\sqrt{x^2 + y^2 + z^2}$. Note that we take the square root (as opposed to the cube root) because we originally *squared* each distance. We can generalize this for two sites for R species as

[1] You can install this package from any R CRAN mirror.

$$D_E = \sqrt{\sum_{i=1}^{R} (x_{ai} - x_{bi})^2} \tag{10.1}$$

Of course, it is difficult (impossible?) to visualize arrangements of sites with more than three axes (i.e., > 3 species), but we can always *calculate* the distances between pairs of sites, regardless of how many species we have.

There are many ways, in addition to Euclidean distances, to calculate distance. Among the most commonly used in ecology is *Bray–Curtis* distance, which goes by other names, including *Søorenson* distance.

$$D_{BC} = \sum_{i=1}^{R} \frac{|x_{ai} - x_{bi}|}{x_{ai} + x_{bi}} \tag{10.2}$$

where R is the number of species in all samples. Bray–Curtis distance is merely the total difference in species abundances between two sites, divided by the total abundances at each site. Bray–Curtis distance (and a couple others) tends to result in more intuitively pleasing distances in which both common and rare species have relatively similar weights, whereas Euclidean distance depends more strongly on the most abundant species. This happens because Euclidean distances are based on squared differences, whereas Bray–Curtis uses absolute differences. Squaring always amplifies the importance of larger values. Fig. 10.3 compares graphs based on Euclidean and Bray–Curtis distances of the same raw data.

Displaying multidimensional distances

A simple way to display distances for three or more species is to create a plot in two dimensions that attempts to arrange all sites so that they are *approximately* the correct distances apart. In general this is impossible to achieve precisely, but distances can be approximately correct. One technique that tries to create an optimal (albiet approximate) arrangement is *non-metric multidimensional scaling*. Here we add a fourth species (*Aconitum lycoctonum*) to our data set before plotting the distances.

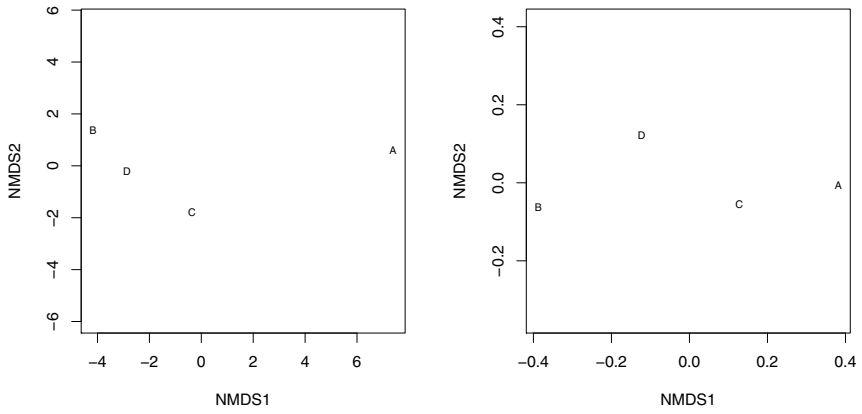
```
> dens$Acolyc <- c(16, 0, 9, 4)
```

The non-metric multidimensional scaling function is in the `vegan` package. It calculates distances for us using the original data. Here we display Euclidean distances among sites (Fig. 10.3a).

```
> library(vegan)
> mdsE <- metaMDS(dens, distance = "euc", autotransform = FALSE,
+     trace = 0)
> plot(mdsE, display = "sites", type = "text")
```

Here we display Bray–Curtis distances among sites (Fig. 10.3b).

```
> mdsB <- metaMDS(dens, distance = "bray", autotransform = FALSE,
+     trace = 0)
> plot(mdsB, display = "sites", type = "text")
```



(a) Euclidean distances (b) Bray–Curits distances

Fig. 10.3: Nonmetric multidimensional (NMDS) plots showing approximate distances between sites. These two figures display the same raw data, but Euclidean distances tend to emphasize differences due to the more abundant species, whereas Bray-Curtis does not. Because NMDS provides iterative optimizations, it will find slightly different arrangements each time you run it.

10.1.3 Similarity

Sometimes, we would like to know how *similar* two communities are. Here we describe two measures of similarity, *percent similarity*, and *Sørensen similarity* [124].

Percent similarity may be the simplest of these; it is simply the sum of the minimum percentages for each species in the community. Here we convert each species to its relative abundance; that is, its proportional abundance at each site. To do this, we treat each site (row) separately, and then divide the abundance of each species by the sum of the abundances at each site.

```
> (dens.RA <- t(apply(dens, 1, function(sp.abun) sp.abun/sum(sp.abun))))
```

```
   Salwho Fravir Manoff Acolyc
A 0.02041 0.4286 0.2245 0.3265
B 0.08333 0.6667 0.2500 0.0000
C 0.06452 0.4194 0.2258 0.2903
D 0.17647 0.2941 0.2941 0.2353
```

Next, to compare two sites, we find the minimum relative abundance for each species. Comparing sites A and B, we have,

```
> (mins <- apply(dens.RA[1:2, ], 2, min))
```

```
 Salwho  Fravir  Manoff  Acolyc
0.02041 0.42857 0.22449 0.00000
```

Finally, we sum these, and multiply by 100 to get percentages.

```
> sum(mins) * 100

[1] 67.35
```

The second measure of similarity we investigate is Sørensen's similarity,

$$S_s = \frac{2C}{A+B} \tag{10.3}$$

where C is the number of species two sites have in common, and A and B are the number of species at each site. This is equivalent to dividing the shared species by the average richness.

To calculate this for sites A and B, we could find the species which have non-zero abundances both sites.

```
> (shared <- apply(dens[1:2, ], 2, function(abuns) all(abuns !=
+     0)))

Salwho Fravir Manoff Acolyc
  TRUE   TRUE   TRUE  FALSE
```

Next we find the richness of each.

```
> (Rs <- apply(dens[1:2, ], 1, function(x) sum(x >
+     0)))

A B
4 3
```

Finally, we divide the shared species by the summed richnesses and multiply by 2.

```
> 2 * sum(shared)/sum(Rs)

[1] 0.8571
```

Sørensen's index has also been used in the development of more sophisticated measures of similarity between sites [144, 164].

10.2 Diversity

To my mind, there is no more urgent or interesting goal in ecology and evolutionary biology than understanding the determinants of biodiversity. Biodiversity is many things to many people, but we typically think of it as a measure of the variety of biological life present, perhaps taking into account the relative abundances. For ecologists, we most often think of *species diversity* as some quantitative measure of the variety or number of different species. This has direct analogues to the genetic diversity within a population [45, 213], and the connections between species and genetic diversity include both shared patterns and shared mechanisms. Here we confine ourselves entirely to a discussion of species diversity, the variety of different species present. Consider this example.

Table 10.1: Four hypothetical stream invertebrate communities. Data are total numbers of individuals collected in ten samples (sums across samples). Diversity indices (Shannon-Wiener, Simpson's) explained below.

Species	Stream 1	Stream 2	Stream 3	Stream 4
Isoperla	20	50	20	0
Ceratopsyche	20	75	20	0
Ephemerella	20	75	20	0
Chironomus	20	0	140	200
Number of species (R)	4	3	4	1
Shannon-Wiener H	1.39	1.08	0.94	0
Simpson's S_D	0.75	0.66	0.48	0

We have four stream insect communities (Table 10.1). Which has the highest "biodiversity"?

We note that one stream has only one species — clearly that can't be the most "diverse" (still undefined). Two streams have four species — are they the most diverse? Stream 3 has more bugs in total (200 *vs.* 80), but stream 1 has a more equal distribution among the four species.

10.2.1 Measurements of variety

So, how shall we define "diversity" and measure this variety? There are many mathematical expressions that try to summarize biodiversity [76, 92]. The inquisitive reader is referred to [124] for a practical and comprehensive text on measures of species diversity. Without defining it precisely (my pay scale precludes such a noble task), let us say that diversity indices attempt to quantify

- the probability of encountering different species at random, or,
- the uncertainty or multiplicity of possible community states (i.e., the entropy of community composition), or,
- the variance in species composition, relative to a multivariate centroid.

For instance, a simple count of species (Table 10.1) shows that we have 4, 3, 4, and 1 species collected from streams 1–4. The larger the number of species, the less certain we could be about the identity of an individual drawn blindly and at random from the community.

To generalize this further, imagine that we have a species pool[2] of R species, and we have a sample comprised of only one species. In a sample with only one species, then we know that the possible *states* that sample can take is limited to one of only R different possible states — the abundance of one species is 100% and all others are zero. On the other hand, if we have two species then the community could take on $R(R-1)$ different states — the first species could be any one of R species, and the second species could be any one of the other species, and all others are zero. Thus increasing diversity means increasing

[2] A *species pool* is the entire, usually hypothetical, set of species from which a sample is drawn; it may be all of the species known to occur in a region.

the possible states that the community could take, and thus increasing our uncertainty about community structure [92]. This increasing lack of information about the system is a form of *entropy*, and increasing diversity (i.e., increasing multiplicity of possible states) is increasing entropy. The jargon and topics of statistical mechanics, such as entropy, appear (in 2009) to be an increasingly important part of community ecology [71, 171].

Below we list three commonly used diversity indices: species richness, Shannon-Wiener index, and Simpson's diversity index.

- Species richness, R, the count of the number of species in a sample or area; the most widely used measure of biodiversity [82].
- Shannon-Wiener diversity.[3]

$$H' = -\sum_{i=1}^{R} p \ln(p) \tag{10.4}$$

  where R is the number of species in the community, and p_i is the relative abundance of species i.
- Simpson's diversity. This index is (i) the probability that two individuals drawn from a community at random will be different species [147], (ii) the initial slope of the species-individuals curve [98] (e.g., Fig. 10.6), and (iii) the expected variance of species composition (Fig. 10.5) [97, 194].

$$S_D = 1 - \sum_{i=1}^{R} p_i^2 \tag{10.5}$$

  The summation $\sum_{i=1}^{R} p_i^2$ is the probability that two individuals drawn at random are the same species, and it is known as Simpson's "dominance." Lande found that this Simpon's index can be more precisely estimated, or estimated accurately with smaller sample sizes, than either richness or Shannon-Wiener [97].

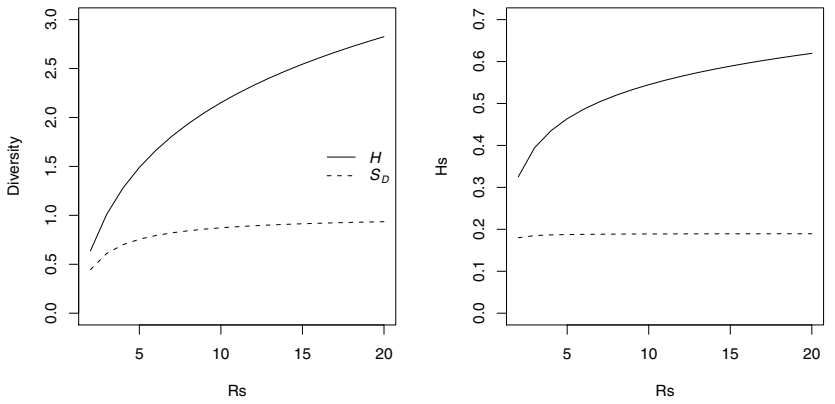
These three indices are actually directly related to each other — they comprise estimates of *entropy*, the amount of disorder or the multiplicity of possible states of a system, that are directly related via a single constant [92]. However, an important consequence of their differences is that richness depends most heavily on rare species, Simpson's depends most heavily on common species, and Shannon-Wiener stands somewhere between the two (Fig. 10.4).

Relations between number of species, relative abundances, and diversity

This section relies heavily on code and merely generates Fig. 10.4.

Here we display diversities for communities with different numbers and relative abundances of species (Fig. 10.4). We first define functions for the diversity indices.

[3] Robert May stated that this index is connected by merely an "ectoplasmic thread" to information theory [135], but there seems be a bit more connection than that.



(a) Equally abundant (b) Most common sp. = 90%

Fig. 10.4: Relations between richness, Shannon-Weiner, and Simpson's diversities (note difference in *y*-axis scale between the figures). Communities of 2–20 species are composed of either equally abundant species (a) or with the most common species equal to 90% of the community (b).

```
> H <- function(x) {
+     x <- x[x > 0]
+     p <- x/sum(x)
+     -sum(p * log(p))
+ }
> Sd <- function(x) {
+     p <- x/sum(x)
+     1 - sum(p^2)
+ }
```

Next we create a *list* of communities with from 1 to 20 *equally* abundant species, and calculate H and S_D for each community.

```
> Rs <- 2:20
> ComsEq <- sapply(Rs, function(R) (1:R)/R)
> Hs <- sapply(ComsEq, H)
> Sds <- sapply(ComsEq, Sd)
> plot(Rs, Hs, type = "l", ylab = "Diversity", ylim = c(0,
+     3))
> lines(Rs, Sds, lty = 2)
> legend("right", c(expression(italic("H")), expression(italic("S"["D"]))),
+     lty = 1:2, bty = "n")
```

Now we create a *list* of communities with from 2 to 25 species, where one species always comprises 90% of the community, and the remainder are equally abundant rare species. We then calculate H and S_D for each community.

```
> Coms90 <- sapply(Rs, function(R) {
+     p <- numeric(R)
```

```
+     p[1] <- 0.9
+     p[2:R] <- 0.1/(R - 1)
+     p
+ })
> Hs <- sapply(Coms90, H)
> Sds <- sapply(Coms90, Sd)
> plot(Rs, Hs, type = "l", ylim = c(0, 0.7))
> lines(Rs, Sds, lty = 2)
```

Simpson's diversity, as a variance of composition

This section relies heavility on code.

Here we show how we would calculate the variance of species composition. First we create a pretend community of six individuals (rows) and 3 species (columns). Somewhat oddly, we identify the degree to which each individual is comprised of each species; in this case, individuals can be only one species.[4] Here we let two individuals be each species.

```
> s1 <- matrix(c(1, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0,
+     0, 0, 0, 0, 0, 1, 1), nr = 6)
> colnames(s1) <- c("Sp.A", "Sp.B", "Sp.C")
> s1

     Sp.A Sp.B Sp.C
[1,]    1    0    0
[2,]    1    0    0
[3,]    0    1    0
[4,]    0    1    0
[5,]    0    0    1
[6,]    0    0    1
```

We can plot these individuals in community space, if we like (Fig. 10.5a).

```
> library(scatterplot3d)
> s13d <- scatterplot3d(jitter(s1, 0.3), type = "h",
+     angle = 60, pch = c(1, 1, 2, 2, 3, 3), xlim = c(-0.2,
+         1.4), ylim = c(-0.2, 1.4), zlim = c(-0.2,
+         1.4))
> s13d$points3d(x = 1/3, y = 1/3, z = 1/3, type = "h",
+     pch = 19, cex = 2)
```

Next we can calculate a *centroid*, or multivariate mean — it is merely the vector of species means.

```
> (centroid1 <- colMeans(s1))

  Sp.A   Sp.B   Sp.C
0.3333 0.3333 0.3333
```

[4] Imagine the case where the columns are traits, or genes. In that case, individuals could be characterized by affiliation with multiple columns, whether traits or genes.

Given this centroid, we begin to calculate a variance by (i) substracting each species vector (0s, 1s) from its mean, (ii) squaring each of these deviates, and (3) summing to get the sum of squares.

```
> (SS <- sum(sapply(1:3, function(j) (s1[, j] -
+     centroid1[j]))^2))
```

```
[1] 4
```

We then divide this sum by the number of individuals that were in the community (N)

```
> SS/6
```

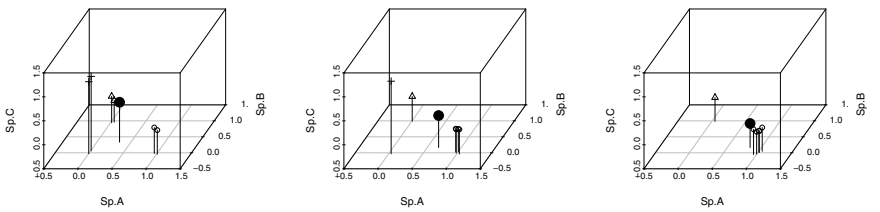
```
[1] 0.6667
```

We find that the calculation given above for Simpson's diversity returns exactly the same number. We would calculate the relative abundances, square them, add them, and subtract that value from 1.

```
> p <- c(2, 2, 2)/6
> 1 - sum(p^2)
```

```
[1] 0.6667
```

In addition to being the variance of species composition, this number is also the probability that two individuals drawn at random are different species. As we mentioned above, there are other motivations than these to derive this and other measures of species diversity, based on entropy and information theory [92] — and they are all typically correlated with each other.



(a) $S_D = 0.67$ (b) $S_D = 0.50$ (c) $S_D = 0.28$

Fig. 10.5: Plotting three examples of species composition. The centroid of each composition is a solid black dot. The third example (on right) has zero abundances of species C. Simpson's diversity is the variance of these points around the centroid. Individual points are not plotted at precisely 0 or 1 — they are plotted with a bit of jitter or noise so that they do not overlap entirely.

10.2.2 Rarefaction and total species richness

Rarefaction is the process of generating the relationship between the number of species *vs.* the number of individuals in one or more samples. It is typically determined by randomly resampling individuals [59], but could also be determined by resampling samples. Rarefaction allows direct comparison of the richness of two samples, corrected for numbers of individuals. This is particularly important because R depends heavily on the number of individuals in a sample. Thus rarefaction finds the general relation between the number(s) of species *vs.* number of individuals (Fig. 10.6), and is limited to less than or equal to the number of species you actually observed. A related curve is the *species-accumulation* curves, but this is simply a useful but haphazard accumulation of new species (a cumulative sum) as the investigator samples new individuals.

Another issue that ecologists face is trying to estimate the *true* number of species in an area, given the samples collected. This number of species would be larger than the number of species you observed, and is often referred to as *total species richness* or the *asymptotic richness*. Samples almost always find only a subset of the species present in an area or region, but we might prefer to know how many species are really there, in the general area we sampled. There are many ways to do this, and while some are better than others, none is perfect. These methods estimate minimum numbers of species, and assume that the unsampled areas are homogeneous and similar to the sampled areas.

Before using these methods seriously, the inquisitive reader should consult [59, 124] and references at http://viceroy.eeb.uconn.edu/EstimateS. Below, we briefly explore an example in R.

An example of rarefaction and total species richness

Let us "sample" a seasonal tropical rainforest on Barro Colorado Island (BCI) http://ctfs.si.edu/datasets/bci/). Our goal will be to provide baseline data for later comparison to other such studies.

We will use some of the data from a 50 ha plot that is included in the `vegan` package [36, 151]. We will pretend that we sampled every tree over 10 cm dbh,[5] in each of 10 plots scattered throughout the 50 ha area. What could we say about the forest within which the plots were located? We have to consider the scale of the sampling. Both the *experimental unit* and the *grain* are the 1 ha plots. Imagine that the plots were scattered throughout the 50 ha plot, so that the extent of the sampling was a full 50. ha[6] First, let's pretend we have sampled 10 1 ha plots by extracting the samples out of the larger dataset.

```
> library(vegan)
> data(BCI)
> bci <- BCI[seq(5, 50, by = 5), ]
```

[5] "dbh" is diameter at 1.37 m above the ground.

[6] See John Wiens' foundational paper on spatial scale in ecology [220] describing the meaning of grain, extent, and other spatial issues.

Next, for each species, I sum all the samples into one, upon which I will base rarefaction and total richness estimation.

Next we combine all the plots into one sample (a single summed count for each species present), select numbers of individuals for which I want rarefied samples (multiples of 500), and then perform rarefaction for each of those numbers.

```
> N <- colSums(bci)
> subs3 <- c(seq(500, 4500, by = 500), sum(N))
> rar3 <- rarefy(N, sample = subs3, se = T, MARG = 2)
```

Next we want to graph it, with a few bells and whistles. We set up the graph of the 10 plot, individual-based rarefaction, and leave room to graph total richness estimators as well (Fig. 10.6).

```
> plot(subs3, rar3[1, ], ylab = "Species Richness",
+     axes = FALSE, xlab = "No. of Individuals",
+     type = "n", ylim = c(0, 250), xlim = c(500,
+         7000))
> axis(1, at = 1:5 * 1000)
> axis(2)
> box()
> text(2500, 200, "Individual-based rarefaction (10 plots)")
```

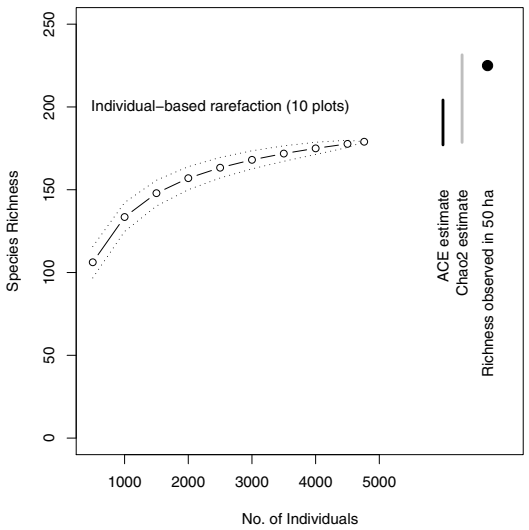


Fig. 10.6: Baseline tree species richness estimation based on ten 1 ha plots, using individual-based rarefaction, and two different total richness estimators, ACE and Chao 2. The true total tree richness in the 50 ha plot is present for comparison.

Here we plot the expected values and also ± 2 SE.

```
> lines(subs3, rar3[1, ], type = "b")
> lines(subs3, rar3[1, ] + 2 * rar3[2, ], lty = 3)
> lines(subs3, rar3[1, ] - 2 * rar3[2, ], lty = 3)
```

Next we hope to estimate the minimum total number of species (asymptotic richness) we might observe in the area around (and in) our 10 plots, if we can assume that the surrounding forest is homogeneous (it would probably be best to extrapolate only to the 50 ha plot). First, we use an *abundance-based coverage estimator*, *ACE*, that appears to give reasonable estimates [124]. We plot intervals, the expected values ± 2 SE (Fig. 10.6).

```
> ace <- estimateR(N)
> segments(6000, ace["S.ACE"] - 2 * ace["se.ACE"],
+     6000, ace["S.ACE"] + 2 * ace["se.ACE"], lwd = 3)
> text(6000, 150, "ACE estimate", srt = 90, adj = c(1,
+     0.5))
```

Next we use a frequency-based estimator, Chao 2, where the data only need to be presence/absence, but for which we also need multiple sample plots.

```
> chaoF <- specpool(bci)
> segments(6300, chaoF[1, "Chao"] - 2 * chaoF[1,
+     "Chao.SE"], 6300, chaoF[1, "Chao"] + 2 * chaoF[1,
+     "Chao.SE"], lwd = 3, col = "grey")
> text(6300, 150, "Chao2 estimate", srt = 90, adj = c(1,
+     0.5))
```

Last we add the observed number of tree species (over 10 cm dbh) found in the entire 50 ha plot.

```
> points(6700, dim(BCI)[2], pch = 19, cex = 1.5)
> text(6700, 150, "Richness observed in 50 ha",
+     srt = 90, adj = c(1, 0.5))
```

This shows us that the total richness estimators did not overestimate the total number of species within the extent of this relatively homogenous sample area (Fig. 10.6).

If we wanted to, we could then use any of these three estimators to compare the richness in this area to the richness of another area.

10.3 Distributions

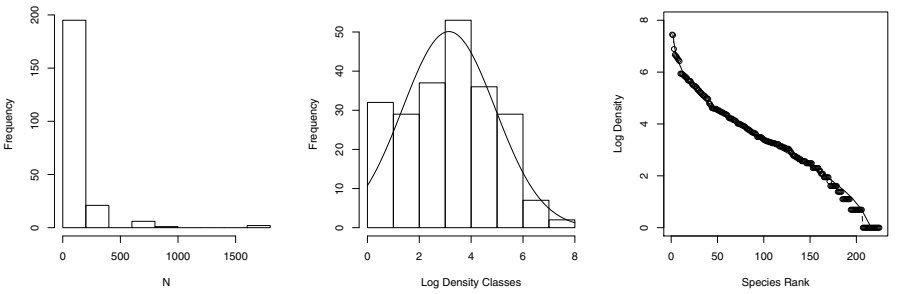
In addition to plotting species in multidimensional space (Fig. 10.3), or estimating a measure of diversity or richness, we can also examine the *distributions* of species abundances.

Like any other vector of numbers, we can make a histogram of species abundances. As an example, here we make a histogram of tree densities, where each species has its own density (Fig. 10.7a). This shows us what is patently true for nearly all ecological communities — *most species are rare.*

10.3.1 Log-normal distribution

Given general empirical patterns, that most species are rare, Frank Preston [168,170] proposed that we describe communities using the logarithms of species abundances (Fig. 10.7).[7] This often reveals that a community can be described approximately with the normal distribution applied to the log-transformed data, or the *log-normal ditribution*. We can also display this as a *rank–abundance distribution* (Fig. 10.7c). To do this, we assign the most abundant species as rank = 1, and the least abundant has rank = R, in a sample of R species, and plot log-abundance *vs.* rank.

May [129] described several ways in which common processes may drive log-normal distributions, and cause them to be common in data sets. Most commonly cited is to note that log-normal distributions arise when each observation (i.e., each random variable) results from the product of independent factors. That is, if each species' density is determined by largely independent factors which act multiplicatively on each species, the resulting densities would be log-normally distributed.



(a) Raw data (b) Species–abundance dist. (c) Rank–abundance dist.

Fig. 10.7: Three related types of distributions of tree species densities from Barro Colorado Island [36]. (a) Histogram of raw data, (b) histogram of log-transformed data; typically referred to as the "species–abundance distribution," accompanied here with the normal probability density function, (c) the "rank–abundance distribution," as typically presented with the log-transformed data, with the complement of the cumulative probability density function (1-pdf) [129]. Normal distributions were applied using the mean and standard deviation from the log-transformed data, times the total number of species.

[7] Preston used base 2 logs to make his histogram bins, and his practice remains standard; we use the natural log.

Log-normal abundance distributions (Fig. 10.7)

We can plot tree species densities from Barro Colorado Island [36], which is aailable online, or in the `vegan` package. First we make the data available to us, then make a simple histogram.

```
> data(BCI)
> N <- sort(colSums(BCI), decr = TRUE)
> hist(N, main = NULL)
```

Next we make a *species–abundance distribution*, which is merely a histogram of the log-abundances (classically, base 2 logs, but we use base e). In addition, we add the normal probability density function, getting the mean and standard deviation from the data, and plotting the expected number of species by multiplying the densities by the total number of species.

```
> hist(log(N), xlab = "Log Density Classes", main = NULL)
> m.spp <- mean(log(N))
> sd.spp <- sd(log(N))
> R <- length(N)
> curve(dnorm(x, m.spp, sd.spp) * R, 0, 8, add = T)
```

Next we create the *rank–abundance distribution*, which is just a plot of log-abundances *vs.* ranks.

```
> plot(log(N), type = "b", ylim = c(0, 8), main = NULL,
+     xlab = "Species Rank", ylab = "Log Density")
> ranks.lognormal <- R * (1 - pnorm(log(N), m.spp,
+     sd.spp))
> lines(ranks.lognormal, log(N))
```

We can think of the rank of species i as the total number of species that are more abundant than species i. This is essentially the opposite (or complement) of the integral of the species–abundance distribution [129]. That means that if we can describe the species abundance distribution with the normal density function, then 1-cumulative probability function is the rank.

10.3.2 Other distributions

Well over a dozen other types of abundance distributions exist to describe abundance patterns, other than the log-normal [124]. They can all be represented as *rank–abundance distributions*.

The *geometric distribution*[8] (or pre-emption niche distribution) reflects a simple idea, where each species pre-empts a constant fraction of the remaining niche space [129, 145]. For instance, if the first species uses 20% of the niche space, the second species uses 20% of the remaining 80%, etc. The frequency of the ith most abundant species is

[8] This probability mass function, $P_i = d(1-d)^{i-1}$, is the probability distribution of the number of attempts, i, needed for one success, if the independent probability of success on one trial is d.

$$N_i = \frac{N_T}{C} d\,(1-d)^{i-1} \tag{10.6}$$

where d is the abundance of the most common species, and C is just a constant to make $\sum N_i = N_T$, where $C = 1 - (1-d)^{S_T}$. Thus this describes the geometric rank–abundance distribution.

The *log-series distribution* [55] describes the frequency of species with n individuals,

$$F\,(S_n) = \frac{\alpha x^n}{n} \tag{10.7}$$

where α is a constant that represents diversity (greater α means greater diversity); the α for a diverse rainforest might be 30–100. The constant x is a fitted, and it is always true that $0.9 < x < 1.0$ and x increases toward 0.99 as $N/S \rightarrow 20$ [124]. x can be estimated from $S/N = [(1-x)/x] \cdot [-\ln(1-x)]$. Note that this is not described as a rank–abundance distribution, but species abundances can nonetheless be plotted in that manner [129].

The log-series rank–abundance distribution is a bit of a pain, relying on the standard exponential integral [129], $E_1(s) = \int_s^\infty exp(-t)/t\,dt$. Given a range of N, we calculate ranks as

$$F\,(N) = \alpha \int_s^\infty exp(-t)/t\,dt \tag{10.8}$$

where we can let $t = 1$ and $s = N \log\,(1 + \alpha/N_T)$.

The log-series distribution has the interesting property that the total number of species in a sample of N individuals would be $S_T = \alpha \log(1 + N/\alpha)$. The parameter α is sometimes used as a measure of diversity. If your data are log-series distributed, then α is approximately the number of species for which you expect 1 individual, because $x \approx 1$. Two very general theories predict a log-series distribution, including neutral theory, and maximum entropy. Oddly, these two theories both predict a log-series distirbution, but make opposite assumptions about niches and individuals (see next section).

MacArthur's *broken stick distribution* is a classic distribution that results in a very even distribution of species abundances [118]. The number of individuals of each species i is

$$N_i = \frac{N_T}{S_T} \sum_{n=i}^{S_T} \frac{1}{n} \tag{10.9}$$

where N_T and S_T are the total number of individuals and species in the sample, respectively. MacArthur described this as resulting from the simultaneous breakage of a stick at random points along the stick. The resulting size fragments are the N_i above. MacArthur's broken stick model is thus both a *stochastic* and a *deterministic* model. It has a simulation (stick breakage) that is the direct analogue of the deterministic analytical expression.

Other similarly tactile stick-breaking distributions create a host of different rank–abundance patterns [124, 206]. In particular, the stick can be broken *sequentially*, first at one random point, then at a random point along one of two newly broken fragments, then at an additional point along any one of the *three* broken fragments, *etc.*, with $S_T - 1$ breaks creating S_T species. The critical

difference between the models then becomes *how each subsequent fragment is selected*. If the probability of selecting each fragment is related directly to its size, then this becomes identical to MacArthur's broken stick model. On the other hand, if each subsequent piece is selected randomly, regardless of its size, then this results in something very similar to the log-normal distribution [196, 205]. Other variations on fragment selection generate other patterns [206].

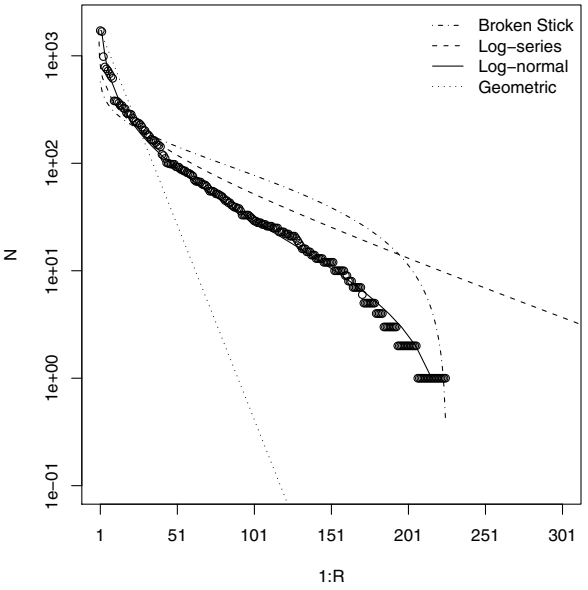


Fig. 10.8: A few common rank–abundance distributions, along with the BCI data [36]. The log-normal curve is fit to the data, and the broken stick distribution is always determined by the number of species. Here we let the geometric distribution be determined by the abundance of the most common species. The log-series was plotted so that it matched qualitatively the most abundant species.

Generating other rank–abundance distributions)

We can illustrate the above rank–abundance distributions as they might relate to the BCI tree data (see previous code). We start with MacArthur's broken stick model. We use cumulative summation backwards to add all $1/n_i$, and then re-sort it by rank (cf. eq. 10.9).

```
> N <- sort(colSums(BCI), decr = TRUE)
> f1 <- sort(cumsum(1/(R:1)), decr = TRUE)
> Nt <- sum(N)
> NMac <- Nt * f1/sum(f1)
```

Next, we create the geomtric rank–abundance distribution, where we let the BCI data tell us d, the density of the most abundant species; therefore we can multiply these by N_T to get expected abundances.

```
> d <- N[1]/Nt
> Ngeo.f <- d * (1 - d)^(0:(R - 1))
> Ngeo <- Nt * Ngeo.f
```

Last, we generate a log-series relevant to the BCI data. First, we use the `optimal.theta` function in the `untb` package to find a relevant value for Fisher's α. (See more and θ and α below under neutral theory).

```
> library(untb)
> alpha <- optimal.theta(N)
```

To calculate the rank abundance distribution for the log-series, we first need a function for the "standard exponential integral" which we then integrate for each population size.

```
> sei <- function(t = 1) exp(-t)/t
> alpha <- optimal.theta(N)
> ranks.logseries <- sapply(N, function(x) {
+     n <- x * log(1 + alpha/Nt)
+     f <- integrate(sei, n, Inf)
+     fv <- f[["value"]]
+     alpha * fv
+ })
```

Plotting other rank–abundance distributions (Fig. 10.8)

Now we can plot the BCI data, and all the distributions, which we generated above. Note that for the log-normal and the log-series, we calculated ranks, based on the species–abundance distributions, whereas in the standard form of the geometric and broken stick distributions, the expected abundances are calculated, in part, from the ranks.

```
> plot(1:R, N, ylim = c(0.1, 2000), xlim = c(1,
+     301), axes = FALSE, log = "y")
> axis(2)
> axis(1, 1 + seq(0, 300, by = 50))
> box()
> lines(1:R, NMac, lty = 4)
> lines(1:R, Ngeo, lty = 3)
> lines(ranks.logseries, N, lty = 2)
> lines(ranks.lognormal, N, lty = 1)
> legend("topright", c("Broken Stick", "Log-series",
+     "Log-normal", "Geometric"), lty = c(4, 2,
+     1, 3), bty = "n")
```

Note that we have not *fit* the the log-series or geometric distributions to the data, but rather, placed them in for comparison. Properly fitting curves to distributions is a picky business [124, 151], especially when it comes to species abundance distributions.

10.3.3 Pattern *vs.* process

Note the resemblance between stick-breaking and niche evolution — if we envision the whole stick as all of the available niche space, or energy, or limiting resources, then each fragment represents a portion of the total occupied by each species. Thus, various patterns of breakage act as models for niche partitioning and relative abundance patterns. Other biological and stochastic processes create specific distributions. For instance, completely random births, deaths, migration, and speciation will create the log-series distribution and the log-normal-like distributions (see *neutral theory* below). We noted above that independent, multiplicatively interacting factors can create the log-normal distribution.

On the other hand, all of these abundance distributions should probably be used primarily to describe *patterns* of commonness, rarity, and not to infer anything about the processes creating the patterns. These graphical descriptions are merely attractive and transparent ways to illustrate the abundances of the species in your sample/site/experiment.

The crux of the issue is that different *processes* can cause the same abundance distribution, and so, sadly, *we cannot usually infer the underlying processes from the patterns we observe.* That is, correlation is different than causation. Abundances of real species, in nature and in the lab, are the result of *mechanistic processes*, including as those described in models of abundance

distributions. However, we cannot say that a particular pattern was the result of a particular process, based on the pattern alone. Nonetheless, they are good summaries of ecological communities, and they show us, in a glance, a lot about *diversity*.

Nonetheless, interesting questions remain about the relations between patterns and processes. Many ecologists would argue that describing patterns and predicting them are the most important goals of ecology [156], while others argue that process is all important. However, both of these camps would agree about the fallacy of drawing conclusions about processes based on pattern — nowhere in ecology has this fallacy been more prevalent than with abundance distributions [66].

10.4 Neutral Theory of Biodiversity and Biogeography

One model of abundance distributions is particularly important, and we elaborate on it here. It is referred to variously as the *unified neutral theory of biodiversity and biogeography* [81], or often "neutral theory" for short. Neutral theory is important because it does much more than most other models of abundance distributions. It is a testable theory that makes quantitative predictions across several levels of organizations, for both evolutionary and ecological processes.

Just as evolutionary biology has neutral theories of gene and allele frequencies, ecology has neutral theories of population and community structure and dynamics [9, 22, 81]. Neutral communities of species are computationally and mathematically related and often identical to models of genes and alleles [53] (Table 10.2). Thus, lessons you have learned about genetic drift often apply to neutral models of communities. Indeed, neutral ecological dynamics at the local scale are often referred to as *ecological drift*, and populations change via demographic stochasticity.[9]

Stephen Hubbell proposed his "unified neutral theory of biodiversity and biogeography" (hereafter NTBB, [81]) as a null model of the dynamics of individuals, to help explain relative abundances of tropical trees. Hubbell describes it as a direct descendant of MacArthur and Wilson's theory of island biogeography [121, 122] (see below, species–area relations). Hubbell proposed it both as a null hypothesis and also — and this is the controversial part — as a model of community dynamics that closely approximates reality.

The relevant "world" of the NTBB is a metacommunity (Fig. 10.9), that is, a collection of similar local communities connected by dispersal [102].[10] The metacommunity is populated entirely by individuals that are functionally identical. The NTBB is a theory of the *dynamics of individuals*, modeling individual

[9] Some have called neutral theory a *null model*, but others disagree [61], describing distinctions between dynamical, process-based neutral models with fitted parameters, *vs.* static data-based null models [60]. Both can be used as benchmarks against which to measure other phenomena. Under those circumstances, I suppose they might both be null hypotheses.

[10] The metacommunity concept is quite general, and a neutral metacommunity is but one caricature [102].

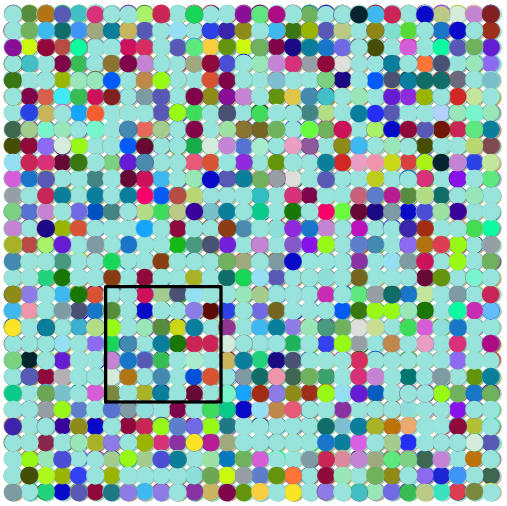


Fig. 10.9: A cartoon of a local community of forest canopy trees (small box) nested inside part of the metacommunity of a tropical forest. The true metaccommunity would extend far beyond the boundaries of this figure to include the true potential source of propagules. Shades of grey indicate different species. The local community is a sample of the larger community (such as the 50 ha forest dynamics plot on BCI) and receives migrants from the metacommunity. Mutation gives rise to new species in the metacommunity. For a particular local community, such as a 50 ha plot on an island in the Panama canal, the metacommunity will include not only the surrounding forest on the island, but also Panama, and perhaps much of the neotropics [86].

births, deaths, migration and mutation. It assumes that within a guild, such as late successional tropical trees, species are essentially neutral with respect to their fitness, that is, they exhibit *fitness equivalence*. This means that the probabilities of birth, death, mutation and migration are *identical for all individuals*. Therefore, changes in population sizes occur via *random walks*, that is, via stochastic increases and decreases with time step (Fig. 10.10). Random walks do not imply an absence of competition or other indirect enemy mediated negative density dependence. Rather, competition is thought to be diffuse, and equal among individuals. We discuss details of this in the context of simulation. Negative density dependence arises either through a specific constraint on the total number of individuals in a community [81], or as traits of individuals related to the probabilities of births, deaths, and speciation [214].

A basic paradox of the NTBB, is that in the absence of migration or mutation, diversity gradually declines to zero, or monodominance. A random walk due to fitness equivalence will eventually result in the loss of all species except one.[11] However, the loss of diversity in any single local community is predicted to be very, very slow, and is countered by immigration and speciation (we dis-

[11] This is identical to allele frequency in genetic drift.

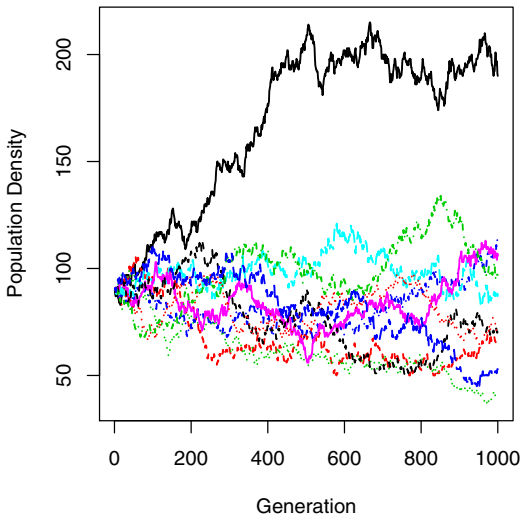


Fig. 10.10: Neutral ecological drift. Here we start with 10 species, each with 90 individuals, and let their abundances undergo random walks within a finite local community, with no immigration. Here, one generation is equal to nine deaths and nine births. Note the slow decline in unevennes — after 1000 deaths, no species has become extinct.

cuss more details below). Thus, species do not increase deterministically when rare — this makes the concept of coexistence different than the stable coexistance criteria discussed in previous chapters. Coexistance here *is not stable* but rather only a stochastic event with a limited time horizon which is balanced by the emergence of new species.

If all communities are thought to undergo random walks toward monodominance, how is diversity maintained in any particular region? Two factors maintain species in any local community. First, immigration into the local community from the metacommunity can bring in new species. Even though each local community is undergoing a random walk toward monodominance, each local community may become dominated by any one of the species in the pool because all species have equal fitness. Thus separate local communities are predicted to become dominated by different species, and these differences among local communities help maintain diversity in the metacommunity landscape.[12] Second, at longer time scales and larger spatial scales, speciation (i.e., mutation and lineage-splitting) within the entire metacommunity maintains biodiversity. Mutation and the consequent speciation provide the ultimate source of variation. Random walks toward extinction in large communities are *so* lengthy that the extinctions are balanced by speciation.

[12] This is the same as how genetic drift operates across subpopulations connected by low rates of gene exchange.

Introducing new symbols and jargon for ecological neutral theory, we state that the diversity of the metacommunity, θ, is a function of the number of individuals in the metacommunity, J_M, and the per capita rate at which new species arise (via mutation) ν ($\theta = 2J_M\nu$; Table 10.2). A local community undergoes ecological drift; drift causes the slow loss of diversity, which is balanced by a per capita (J_L) immigration rate m.

Table 10.2: Comparison of properties and jargon used in ecological and population genetic neutral theory (after Alonso et al. [1]). Here x is a continuous variable for relative abundance of a species or allele ($0 \leq x \geq 1$, $x = n/J$). Note this is potentially confused with Fisher's log-series $\langle \phi_n \rangle = \theta x^n/n$, which is a discrete species abundance distribution in terms of numbers of individuals, n (not relative abundance), and where $x = b/d$.

Property	**Ecology**	**Population Genetics**
Entire System (size)	Metacommunity (J_M)	Population (N)
Nested subsystem (size)	Local community (J_L)	Subpop. or Deme (N)
Smallest neutral system unit	Individual organism	Individual gene
Diversity unit	Species	Allele
Stochastic process	Ecological drift	Genetic drift
Generator of diversity (rate symbol)	Speciation (ν)	Mutation (μ)
Fundamental diversity number	$\theta \approx 2J_M\nu$	$\theta \approx 4N\mu$
Fundamental dispersal number	$I \approx 2J_Lm$	$\theta \approx 4Nm$
Relative abundance distribution ($\Phi(x)$)	$\frac{\theta}{x}(1-x)^{\theta-1}$	$\frac{\theta}{x}(1-x)^{\theta-1}$
Time to common ancestor	$\frac{-J_Mx}{1-x}\log x$	$\frac{-Nx}{1-x}\log x$

It turns out that the abundance distribution of an infinitely large metacommunity is Fisher's log-series distribution, and that θ of neutral theory is α of Fisher's log-series [2, 105, 215]. However, in any one local community, random walks of rare species are likely to include zero, and thus become extinct in the local community by chance alone. This causes a deviation from Fisher's log-series in any *local community* by reducing the number of the the rarest species below that predicted by Fisher's log-series. In particular, it tends to create a log-normal–like distribution, much as we often see in real data (Fig. 10.8). These theoretical findings and their match with observed data are thus consistent with the hypothesis that communities may be governed, in some substantive part, by neutral drift and migration.

Both theory and empirical data show that species which coexist may be more similar than predicted by chance alone [103], and that similarity (i.e., fitness equality) among species helps maintain higher diversity than would otherwise be possible [27]. Chesson makes an important distinction between *stabilizing mechanisms*, which create attractors, and *equalizing mechanisms*, which reduce differences among species, slow competitive exclusion and facilitate stabilization [27].

The NTBB spurred tremendous debate about the roles of chance and determinism, of dispersal-assembly and niche-assembly, of evolutionary processes in ecology, and how we go about "doing community ecology" (see, e.g., articles in *Functional Ecology*, 19(1), 2005; *Ecology*, 87(6), 2006). This theory in its narrowest sense has been falsified with a few field experiments and observation studies [32,223]. However, the *degree* to which stochasticity and dispersal *versus* niche partitioning structure communities remains generally unknown. Continued refinements and elaboration (e.g., [2, 52, 64, 86, 164, 216]) seem promising, continuing to intrigue scientists with different perspectives on natural communities [86,99]. Even if communities turn out to be completely non-neutral, NTBB provides a baseline for community dynamics and patterns that has increased the rigor of evidence required to demonstrate mechanisms controlling coexistence and diversity. As Alonso *et al.* state, "... good theory has more predictions per free parameter than bad theory. By this yeardstick, neutral theory fares fairly well" [1].

10.4.1 Different flavors of neutral communities

Neutral dynamics in a local community can be simulated in slightly different ways, but they are all envisioned some type of *random walk*. A random walk occurs when individuals reproduce or die at random, with the consequence that each population increases or decreases by chance.

The simplest version of a random walk assumes that births and deaths and consequent increases and decreases in population size are equally likely and equal in magnitude. A problem with this type of random walk is that a community can increase in size (number of individuals) without upper bound, or can disappear entirely, by chance. We know this doesn't happen, but it is a critically important first step in conceptualizing a neutral dynamic [22].

Hubbell added another level of biological reality by fixing the total number of individuals in a local community, J_L, as constant. When an individual dies, it is replaced with another individual, thus keeping the population size constant. Replacements come from within the local community with probability $1-m$, and replacements come from the greater metacommunity with probability m. The dynamics of the metacommunity are so slow compared to the local community that we can safely pretend that it is fixed, unchanging.

Volkov *et al.* [215] took a different approach by assuming that each species undergoes independent *biased random walks*. We imagine that each species undergoes its own completely independent random walk, *as if* it is not interacting with any other species. The key is that the birth rate, b, is slightly *less* than the death rate, d — this bias toward death gives us the name *biased random walk*. In a deterministic model with no immigration or speciation, this would result in a slow population decline to zero. In a stochastic model, however, some populations will increase in abundance by chance alone. Slow random walks toward extinctions are balanced by speciation in the metacommunity (with probability ν).

If species all undergo independent biased random walks, does this mean species don't compete and otherwise struggle for existance? No. The *reason* that

$b < d$ is precisely because all species struggle for existance, and only those with sufficiently high fitness, *and which are lucky*, survive. Neutral theory predicts that it is these species that we observe in nature — those which are lucky, and also have sufficiently high fitness.

In the metacommunity, the average number of species, $\langle \phi_n^M \rangle$, with population size n is

$$\langle \phi_n^M \rangle = \theta \frac{x^n}{x} \tag{10.10}$$

where $x = b/d$, and $\theta = 2J_M\nu$ [215]. The M superscript refers to the metacommunity, and the angle brackets indicate merely that this is the average. Here b/d is barely less than one, because it a biased random walk which is then offset by speciation, ν. Now we see that this is exactly Fisher's log-series distribution (eq. 10.7), where that $x = b/d$ and $\theta = \alpha$. Volkov *et al.* thus show that in a neutral metacommunity, x has a biological interpretation.

The expected size of the entire metacommunity is simply the sum of all of the average species' n [215].

$$J_M = \sum_{n=1}^{\infty} n\langle \phi_n^M \rangle = \theta \frac{x}{1-x} \tag{10.11}$$

Thus the size of the metacommunity is an emergent property of the dynamics, rather than an external constraint. To my mind, it seems that the number of individuals in the metacommunity must result from responses of individuals *to* their external environment.

Volkov *et al.* went on to derive expressions for births, deaths, and average relative abundances in the local community [139,215]. Given that each individual has an equal probability of dying and reproducing, and that replacements can also come from the metacommunity with a probability proportional to their abundance in the metacommunity, one can specify rules for populations in local communities of a fixed size. These are the probability of increase, $b_{n,k}$, or decrease, $d_{n,k}$, in a species, k, of population size n.

$$b_{n,k} = (1-m)\frac{n}{J_L}\frac{J_L - n}{J_L - 1} + m\frac{\mu_k}{J_M}\left(1 - \frac{n}{J_L}\right) \tag{10.12}$$

$$d_{n,k} = (1-m)\frac{n}{J_L}\frac{J_L - n}{J_L - 1} + m\left(1 - \frac{mu_k}{J_M}\right)\frac{n}{J_L} \tag{10.13}$$

These expressions are the sum of two joint probabilities, each of which is comprised of several independent events. These events are immigration, and birth and death of individuals of different species. Here we describe these probabilities. For a population of size n of species k, we can indicate per capita, per death probabilities including

- m, the probability that a replacement is immigrant from the metacommunity, and $1-m$, the probability that the replacement is from the local community.

- n/J_L, the probability that an individual selected randomly from the local community belongs to species k, and $1-n/J_L$ or $(J-n)/(J_L)$, the probability that an individual selected randomly from the local community belongs to any species other than k.
- $(J-n)/(J_L-1)$, the conditional probability that, given that an individual of species k has already been drawn from the population, an individual selected randomly from the local community belongs to any species other than to species k.
- μ_k/J_M, the probability that an individual randomly selected from the metacommunity belongs to species k, and $1-n/J_M$, the probability that an individual randomly selected from the metacommunity belongs to any species other than k.

Each of these probabilities is the probability of some unspecified event — that event might be birth, death, or immigration.

Before we translate eqs. 10.12, 10.13 literally, we note that b and d each have two terms. The first term is for dynamics related to the local community, which happen with probability $1-m$. The second is related to immigration from the metacommunity which occurs with probability m. Consider also that if a death is replaced by a birth of the same species, or a birth is countered by a death of the same species, they cancel each other out, as if nothing ever happened. Therefore each term requires a probability related to species k and to non-k.

Eq. 10.12, $b_{n,k}$, is the probability that an individual will be added to the population of species k. The first term is the joint probability that an addition to the population comes from within the local community $(1-m)$ *and* the birth comes from species k (n/J_L) *and* there is a death of an individual of any other species $((J_L-n)/(J_L-1))$.[13] The second term is the joint probability that the addition to the population comes from the metacommunity via immigration (m) *and* that the immigrant is of species k (μ_k/J_M) *and* is not accompanied by a death of an individual of its species (n/J_L).

An individual may be subtracted from the population following similar logic. Eq. 10.13, $d_{n,k}$, is the probability that a death will remove an individual from the population of species k. The first term is the joint probability that the death occurs in species k (n/J_L) and the replacement comes from the local community $(1-m)$ and is some species other than k $((J_L-n)/(J_L-1))$. The second term is the joint probability that the death occurs in species k (n/J_L), and that it is replaced by an immigrant (m) *and* the immigrant is any species in the metacommunity other than k $(1-\mu_k/J_M)$.

10.4.2 Investigating neutral communities

Here we exxplore netural communities using the `untb` package, which contains a variety of functions for teaching and research on neutral theory.

[13] The denominator of the death probability is J_L-1 instead of J_L because we have already selected the first individual who will do the reproduction, so the total number of remaining individuals is J_L-1 rather than J_L; the number of non-k individuals remains J_L-n

Pure drift

After loading the package and setting graphical parameters, let's run a simulation of drift. Recall that drift results in the slow extinction of all but one species (Fig. 10.10). We start with a local community with 20 species, each with 25 individuals[14] The simulation then runs for 1000 generations (where 9/900 individuals die per generation).[15]

```
> library(untb)
> a <- untb(start = rep(1:20, 25), prob = 0, gens = 2500,
+     D = 9, keep = TRUE)
```

We `keep`, in a matrix, all 450 trees from each of the 1000 time steps so that we can investigate the properties of the community. The output is a matrix where each element is an integer whose value represents a species ID. Rows are time steps and we can think of columns as spots on the ground occupied by trees. Thus a particular spot of ground may be occupied by an individual of one species for a long time, and suddenly switch identity, because it dies and is replaced by an individual of another species. Thus the community always has 450 individuals (columns), but the identities of those 450 change through time, according to the rules laid out in eqs. 10.12, 10.13. Each different species is represented by a different integer; here we show the identitites of ten individuals (columns) for generations 901–3.

```
> (a2 <- a[901:903, 1:10])
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
[1,]    7   15    4   15   11    7   15   14   20    20
[2,]    7   15    4   15   11    7   15   14   20    20
[3,]    7   15    4   15   11    7   15   14   20    20
```

Thus, in generation 901, tree no. 1 is species 7 and in generation 902, tree no. 3 is species 4.

We can make pretty pictures of the communities at time steps 1, 100, and 2000, by having a single point for each tree, and coding species identity by shades of grey.[16]

```
> times <- c(1, 50, 2500)
> sppcolors <- sapply(times, function(i) grey((a[i,
+     ] - 1)/20))
```

This function applies to the community, at each time step, the `grey` function. Recall that species identity is an integer; we use that integer to characterize each species' shade of grey.

Next we create the three graphs at three time points, with the appropriate data, colors, and titles.

[14] This happens to be the approximate tree density (450 trees ha^{-1}, for trees > 10 cm DBH) on BCI.

[15] Note that `display.untb` is great for pretty pictures, whereas `untb` is better for more serious simulations.

[16] We could use a nice color palette, `hcl`, based on hue, chroma, and luminance, for instance `hcl(a[i,]*30+50)`

```
> layout(matrix(1:3, nr = 1))
> par(mar = c(1, 1, 3, 1))
> for (j in 1:3) {
+     plot(c(1, 20), c(1, 25), type = "n", axes = FALSE)
+     points(rep(1:20, 25), rep(1:25, each = 20),
+         pch = 19, cex = 2, col = sppcolors[, j])
+     title(paste("Time = ", times[j], sep = ""))
+ }
```

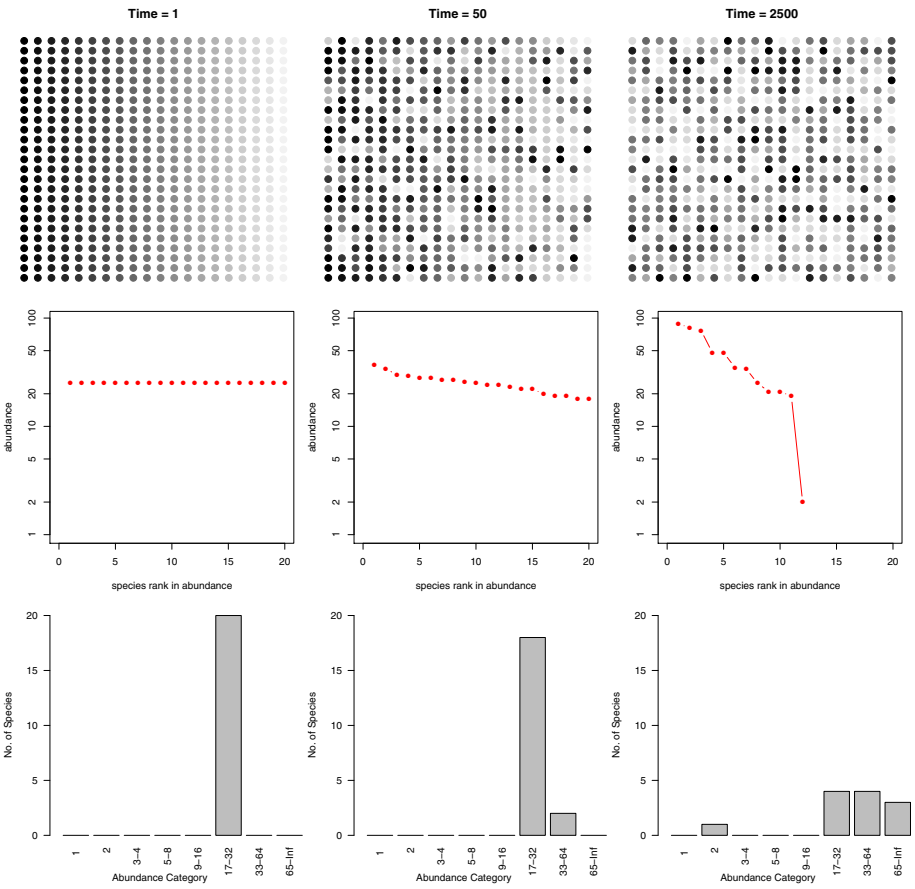


Fig. 10.11: Three snapshots of one community, drifting through time. Shades of grey represent different species. Second row contains rank abundance distributions; third row contains species abundance distributions. Drift results in the slow loss of diversity.

From these graphs (Fig. 10.11), we see that, indeed, the species identity of the 450 trees changes through time. Remember that this is *not spatially explicit* — we are not stating that individual *A* is next, or far away from, individual *B*.

Rather, this is a spatially implicit representation — all individuals are characterized by the same probaibllities.

Next let's graph the rank abundance distribution of these communities (Fig. 10.11). For each time point of interest, we first coerce each "ecosystem" into a `count` object, and then plot it. Note that we plot them all on a common scale to facilitate comparison.

```
> layout(matrix(1:3, nr = 1))
> for (i in times) {
+     plot(as.count(a[i, ]), ylim = c(1, max(as.count(a[times[3],
+         ]))), xlim = c(0, 20))
+     title(paste("Time = ", i, sep = ""))
+ }
```

Next we create species abundance distributions, which are histograms of species' abundances (Fig. 10.11). If we want to plot them on a common scale, it takes a tad bit more effort. We first create a matrix of zeroes, with enough columns for the last community, use `preston` to create the counts of species whose abundances fall into logarithmic bins, plug those into the matrix, and label the matrix columns with the logarithmic bins.

```
> out <- lapply(times, function(i) preston(a[i,
+     ]))
> bins <- matrix(0, nrow = 3, ncol = length(out[[3]]))
> for (i in 1:3) bins[i, 1:length(out[[i]])] <- out[[i]]
> bins
```

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    0    0    0    0    0   20    0    0
[2,]    0    0    0    0    0   18    2    0
[3,]    0    1    0    0    0    4    4    3
```

```
> colnames(bins) <- names(preston(a[times[3], ]))
```

Finally, we plot the species–abundance distributions.

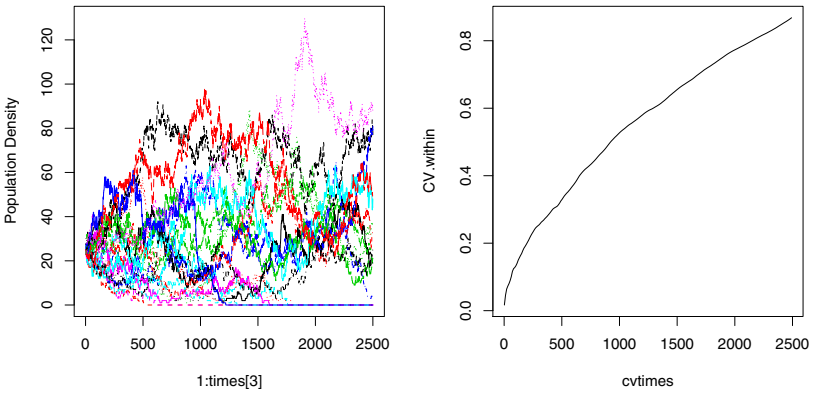
```
> layout(matrix(1:3, nr = 1))
> for (i in 1:3) {
+     par(las = 2)
+     barplot(bins[i, ], ylim = c(0, 20), xlim = c(0,
+         8), ylab = "No. of Species", xlab = "Abundance Category")
+ }
```

Bottom line: *drift causes the slow loss of species from local communities* (Fig. 10.11). What is not illustrated here is that without dispersal, drift will cause different species to become abundant in different places because the variation is random. In that way, drift maintains diversity at large scales, in the metacommunity. Last, low rates of dispersal among local communities maintains some diversity in local communities without changing the entire metacommunity into a single large local community. Thus dispersal limitiation, but not its absence, maintains diversity.

Next, let's examine the dynamics through time. We will plot individual species trajectories (Fig. 10.12a).

```
> sppmat <- species.table(a)
> matplot(1:times[3], sppmat[1:times[3], ], type = "l",
+     ylab = "Population Density")
```

The trajectories all start at the same abundance, but they need not have. The trajectories would still have the same drifting quality.



(a) 20 populations (b) Average variability

Fig. 10.12: Dynamics and average variation within populations. In random walks, average variation (measured with the coefficient of variation) increases with time.

For random walks in general, the observed variance and coefficient of variation ($CV = \hat{\sigma}/\bar{x}$) of a population will grow over time [32]. Here we calculate the average population CV of cumulative observations (note the convenient use of nested `(s)apply` functions). Let's calculate the CV's for every tenth time step.

```
> cvtimes <- seq(2, 2500, by = 10)
> CV.within <- sapply(cvtimes, function(i) {
+     cvs <- apply(sppmat[1:i, ], 2, function(x) sd(x)/mean(x))
+     mean(cvs)
+ })
```

Now plot the average CV through time. The average observed CV should increase (Fig. 10.12b).

```
> plot(cvtimes, CV.within, type = "l")
```

This shows us that the populations never approach an equilibrium, but wander aimlessly.

Real data

Last, we examine a BCI data set [36]. We load the data (data from 50 1 ha plots × 225 species, from the `vegan` package), and sum species abundances to get each species total for the entire 50 h plot (Fig. 10.13a).

```
> library(vegan)
> data(BCI)
> n <- colSums(BCI)
> par(las = 2, mar = c(7, 4, 1, 1))
> xs <- plot(preston(rep(1:225, n), n = 12, original = TRUE))
```

We would like to estimate θ and m from these data, but that requires specialized software for any data set with a realistic number of individuals. Specialized software would provide maximum likelihood estimates (in a reasonable amount of computing time) for m and θ for large communities [51,69,86]. The BCI data have been used repeatedly, so we rely on estimates from the literature ($\theta \approx 48$, $m \approx 0.1$) [51,215]. We use the approach of Volkov et al. [215] to generate expected species abundances (Fig. 10.13a).

```
> v1 <- volkov(sum(n), c(48, 0.1), bins = TRUE)
> points(xs, v1[1:12], type = "b")
```

More recently, Jabot and Chave [86] arrived at estimates that differed from previous estimates by orders of magnitude (Fig. 10.13b). Their novel approach estimated θ and m were derived from *both* species abundance data and from phylogenetic data ($\theta \approx 571, 724$, $m \approx 0.002$). This is possible because neutral theory makes a rich array of predictions, based on both ecological and evolutionary processes. Their data were only a little bit different (due to a later census), but their analyses revealed radically different estimates, with a much greater diversity and larger metacommunity (greater θ), and much lower immigration rates (smaller m).

Here we derive expected species abundance distributions. The first is based soley on census data, and is similar to previous expections (Fig. 10.13b).

```
> v2 <- volkov(sum(n), c(48, 0.14), bins = TRUE)
> xs <- plot(preston(rep(1:225, n), n = 12, original = TRUE),
+     col = 0, border = 0)
> axis(1, at = xs, labels = FALSE)
> points(xs, v2[1:12], type = "b", lty = 2, col = 2)
```

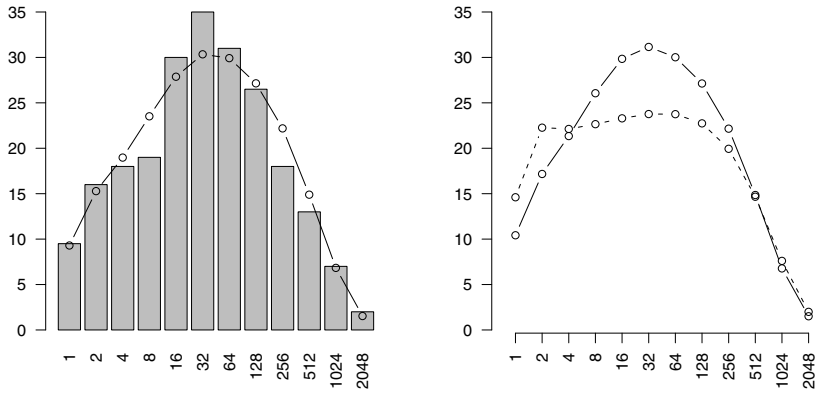
However, when they also included phylogenetic data, they found very different expected species abundance distributions (Fig. 10.13b).

```
> v4 <- volkov(sum(n), c(571, 0.002), bins = TRUE)
> points(xs, v4[1:12], type = "b", lty = 4, col = 2)
```

If nothing else, these illustrate the effects of increasing θ and reducing m.

10.4.3 Symmetry and the rare species advantage

An important advance in neutral theory is the quantitfication of a *symmetric rare species advantage*. The symmetry hypothesis posits that all species have a *symmetrical rare species advantage* [214, 216]. That is, all species increase when rare to the same degree (equal negative density dependence). In a strict sense, all individuals remain the same in that their birth and death probabilities change with population size in the same manner for all species. This obviously



(a) Original estimates (b) Jabot and Chave estimates

Fig. 10.13: Species abundance distributions for BCI trees. (a) Data for histogram from [36], overlain with expected abundances with θ and m values fitted to the data [51,215]. (b) Jabot and Chave found that when they used only species abundances (as did previous investigators) their pattern was similar to previous findings (solid line). However, adding phylogenetic information led to very different expectations (dashed line).

reduces the chance of random walks to extinction, but is nonetheless the same among all species. Estimation of the magnitude of the rare species advantage is interesting addition to stepwise increasing complexity.

To sum up: it is safe to say that neutral theory has already made our thinking about community structure and dynamics more sophisticated and subtle, by extending island biogeography to individuals. The theory is providing quantitative, pattern-generating models, that are analogous to null hypotheses. With the advent of specialized software, theory is now becoming more useful in our analysis of data [51, 69, 86].

10.5 Diversity Partitioning

We frequently refer to biodiversity (i.e., richness, Simpson's, and Shannon-Wiener diversity) at different spatial scales as α, β, and γ diversity (Fig. 10.14).

- Alpha diversity, α, is the diversity of a point location or of a single sample.
- Beta diversity, β, is the diversity due to multiple localities; β diversity is sometimes thought of as *turnover* in species composition among sites, or alternatively as the number of species in a region that are *not* observed in a sample.
- Gamma diversity, γ, is the diversity of a region, or at least the diversity of all the species in a set of samples collected over a large area (with large extent relatve to a single sample).

Diversity across spatial scales can be further be *partitioned* in one of two ways, either using *additive* or *multiplicative* partitioning.

Additive partitioning [42, 43, 97] is represented as

$$\bar{\alpha} + \beta = \gamma \tag{10.14}$$

where $\bar{\alpha}$ is the average diversity of a sample, γ is typically the diversity of the pooled samples, and β is found by difference ($\beta = \gamma - \bar{\alpha}$). We can think of β as the *average* number of species *not* found in a sample, but which we know to be in the region. Additive partitioning allows direct comparison of average richness among samples at any hierarchical level of organization because all three measures of diversity (α, β, and γ) are *expressed in the same units*. This makes it analogous to partitioning variance in ANOVA. This is not the case for multiplicative partitioning diversity.

Partitioning can also be *multiplicative* [219],

$$\bar{\alpha}\beta = \gamma \tag{10.15}$$

where β is a conversion factor that describes the relative change in species composition among samples. Sometimes this type of β diversity is thought of as the number of different community types in a set of samples. However, one must use this interpretation with great caution, as either meaning of β diversity depends completely on the sizes or extent of the samples used for α diversity.

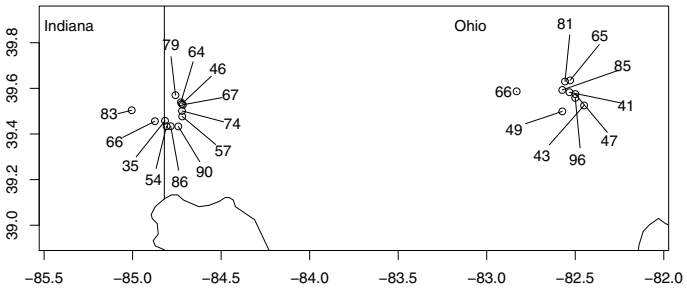


Fig. 10.14: Hierarchical sampling of moth species richness in forest patches in Indiana and Ohio, USA [198]. α-diversity is the diversity of a single site (richness indicated by numbers). γ-diversity is the total number of species found in any of the samples (here $\gamma = 230$ spp.). Additive β-diversity is the difference, $\gamma - \bar{\alpha}$, or the average number of species *not* observed in a single sample. Diversity partitioning can be done at two levels, sites within ecoregions and ecoregions within the geographic region (see example in text for details).

Let us examine the the limits of β diversity in extremely small and extremely large samples. Imagine that our sample units each contain, on average, one individual (and therefore one species) and we have 10^6 samples. If richness is our measure of diversity, then $\bar{\alpha} = 1$. Now imagine that in all of our samples we find a total of 100 species, or $\gamma = 100$. Our additive and multiplicative partitions would then be $\beta_A = 99$, and $\beta_M = 100$, respectively. If the size of the sample unit increases, each sample will include more and more individuals and therefore more of the diversity, and by definition, β will decline. If each sample gets large enough, then each sample will capture more and more of the species until a sample gets so large that it includes all of the species (i.e., $\bar{\alpha} \to \gamma$). At this point, $\beta_A \to 0$ and $\beta_M \to 1$.

Note that β_A and β_M do not change at the same rates (Fig. 10.15). When we increase sample size so that each sample includes an average of two species ($\bar{\alpha} = 2$), then $\beta_A = 98$ and $\beta_M = 50$. If each sample were big enough to have on average 50 species ($\bar{\alpha} = 50$), then $\beta_A = 50$ and $\beta_M = 2$. So, the β's do not change at the same rate (Fig. 10.15).

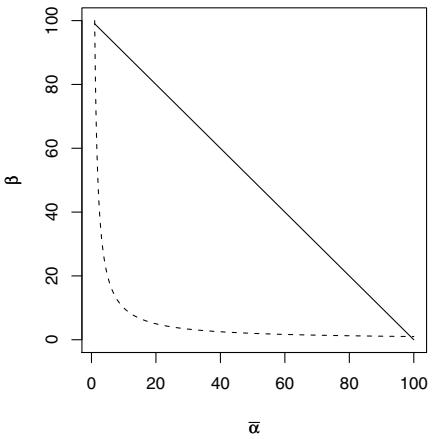


Fig. 10.15: Relations of β_A (with additive partitioning) and β_M (with multiplicative partitioning) to $\bar{\alpha}$, for a fixed $\gamma = 500$ species. In our example, we defined diversity as species richness, so the units of β_A and α are number of species per sample, and $\bar{\alpha}$ is the mean number of species in a sample.

Multiplicative β_M is sometimes thought of as the number of independent "communities" in a set of samples. This would make sense if our sampling regime were designed to capture representative parts of different communities. For example, if we sampled an elevational gradient, or a productivity gradient, and our smallest sample was sufficiently large so as to be representative of that point along the gradient[17] then β_M could provide a measure of the relative turnover in

[17] One type of sample that attempts to do this is a relevé.

composition or "number of different communities." However, we know that composition is also predicted to vary randomly across the landscape [36]. Therefore, if each sample is small, and not really representative of a "community," then the small size of the samples will inflate β_M and change the interpretation.

As an example, consider the BCI data, which consists of 50 contiguous 1 ha plots. First, we find γ (all species in the data set, or the number of columns), and $\bar{\alpha}$ (mean species number per 1 ha plot).

```
> (gamma <- dim(BCI)[2])

[1] 225

> (alpha.bar <- mean(specnumber(BCI)))

[1] 90.78
```

Next we find additive β-diversity and multiplicative β-diversity.

```
> (beta.A <- gamma - alpha.bar)

[1] 134.2

> (beta.M <- gamma/alpha.bar)

[1] 2.479
```

Now we interpret them. These plots are located in a relatively uniform tropical rainforest. Therefore, they each are samples drawn from a single community type. However, the samples are small. Therefore, each 1 ha plot (10^4 m^2 in size) misses more species than it finds, on average ($\beta_A > \bar{\alpha}$). In addition, $\beta_M = 2.48$, indicating a great deal of turnover in species composition. We could mistakenly interpret this as indicating something like ~ 2.5 independent community types in our samples. Here, however, we have a single community type — additive partitioning is a little simpler and transparent in its interpretation.

For other meanings of β-diversity, linked to neutral theory, see [36, 144].

10.5.1 An example of diversity partitioning

Let us consider a study of moth diversity by Keith Summerville and Thomas Crist [197,198]. The subset of their data presented here consists of woody plant feeding moths collected in southwest Ohio, USA. Thousands of individuals were trapped in 21 forest patches, distributed in two adjacent ecoregions (12 sites - North Central Tillplain [NCT], and 9 sites - Western Allegheny Plateau [WAP], Fig. 10.14). This data set includes a total of 230 species, with 179 species present in the NCT ecoregion and 173 species present in the WAP ecoregion. From these subtotals, we can already see that each ecoregion had most of the combined total species (γ).

We will partition richness at three spatial scales: sites within ecoregions ($\bar{\alpha}_1$), ecoregions ($\bar{\alpha}_2$), and overall (γ). This will result in two β-diversities: β_1 among sites within each ecoregion, and β_2 between ecoregions. The relations among these are straightforward.

$$\bar{\alpha}_2 = \bar{\alpha}_1 + \beta_1 \tag{10.16}$$

$$\gamma = \bar{\alpha}_2 + \beta_2 \tag{10.17}$$

$$\gamma = \bar{\alpha}_1 + \beta_1 + \beta_2 \tag{10.18}$$

To do this in R, we merely implement the above equations using the data in Fig. 10.14 [198]. First, we get the average site richness, $\bar{\alpha}_1$. Because we have different numbers of individuals from different site, and richness depends strongly on the number of individuals in our sample, we may want to weight the sites by the number of individuals. However, I will make the perhaps questionable argument for the moment that because trapping effort was similar at all sites, we will not adjust for numbers of individuals. We will assume that different numbers of individuals reflect different population sizes, and let number of individuals be one of the local determinants of richness.

```
> data(moths)

> a1 <- mean(moths[["spp"]])
```

Next we calculate average richness richness for the ecoregions. Because we had 12 sites in NCT, and only nine sites in WAP for what might be argued are landscape constraints, we will use the weighted average richness, adjusted for the number of sites.[18] We also create an object for $\gamma = 230$.

```
> a2 <- sum(c(NCT = 179, WAP = 173) * c(12, 9)/21)
> g <- 230
```

Next, we get the remaining quantities of interest, and show that the partition is consistent.

```
> b1 <- a2 - a1
> b2 <- g - a2
> abg <- c(a1 = a1, b1 = b1, a2 = a2, b2 = b2, g = g)
> abg

    a1     b1     a2     b2      g
 65.43 111.00 176.43  53.57 230.00

> a1 + b1 + b2 == g

[1] TRUE
```

The partitioning reveals that β_1 is the largest fraction of overall γ-richness (Fig. 10.16). This indicates that in spite of the large distance between sampling areas located in different ecoregions, and the different soil types and associated flora, most of the variation occurs *among sites within regions*. If there had been a greater difference in moth community composition among ecoregions, then β_2-richness would have made up a greater proportion of the total.

[18] The arithmetic mean is $\sum a_i Y_i$, where all $a_i = 1/n$, and n is the total number of observations. A weighted average is the case where the a_i represent unequal *weights*, often the fraction of n on which each Y_i is based. In both cases, $\sum a = 1$.

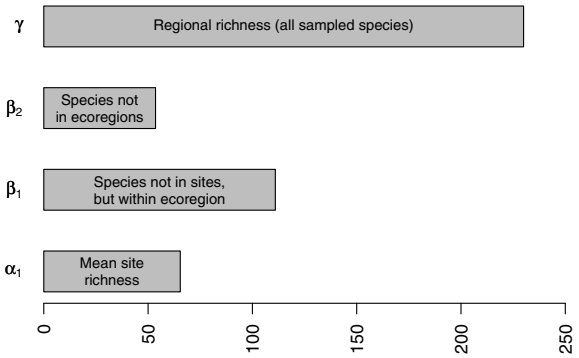


Fig. 10.16: Hierarchical partitioning of moth species richness in forest patches [198]. See Fig. 10.14 for geographical locations.

These calculations show us how simple this additive partition can be, although more complicated analyses are certainly possible. It can be very important to weight appropriately the various measures of diversity (e.g., the number of individuals in each sample, or number of samples per hierarchical level). The number of individuals in particular has a tremendous influence on richness, but has less influence on Simpson's diversity partitioning. The freely available PARTITION software will perform this additive partitioning (with sample sizes weights) and perform statistical tests [212].

10.5.2 Species–area relations

The relation between the number of species found in samples of different area has a long tradition [5, 10, 120–122, 169, 178], and is now an important part of the metastasizing subdiscipline of *macroecology* [15, 71].

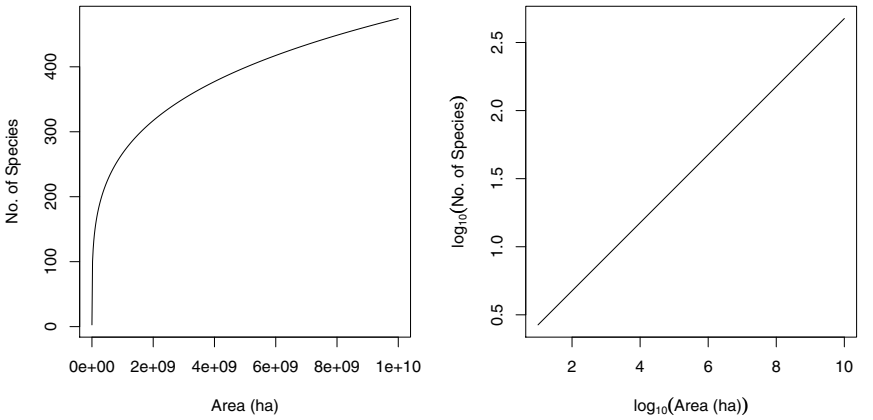
Most generally, the species–area relation (SAR) is simply an empirical pattern of the number of species found in patches of different size, plotted as a function of the sizes of the respective patches (Fig. 10.17). These patches may be isolated from each other, as islands in the South Pacific [121], or mountaintops covered in coniferous forest surrounded by a sea of desert [16], or calcareous grasslands in an agricultural landscape [68]. On the other hand, these patches might be nested sets, where each larger patch contains all others [41, 163].

Quantitatively, the relation is most often proposed as a simple power law,

$$R = cA^z \tag{10.19}$$

where R is the number of species in a patch of size A, and c and z are fitted constants. This is most often plotted as a log–log relation, which makes it linear.

$$\log(R) = b + zA \tag{10.20}$$



(a) Arithmetic scale (b) Logarithmic scale

Fig. 10.17: Power law species–area relations.

where b is the intercept (equal to $\log c$) and z is the slope.

Drawing power law species–area relations (Fig. 10.17)

Here we simply draw some species area curves.

```
> A <- 10^1:10
> c <- 1.5
> z <- 0.25
> curve(c * x^z, 10, 10^10, n = 500, ylab = "No. of Species",
+     xlab = "Area (ha)")

> A <- 10^1:10
> c <- 1.5
> z <- 0.25
> curve(log(c, 10) + z * x, 1, 10,
+     ylab = quote(log[10]("No. of Species")),
+     xlab = quote(log[10]("Area (ha)")))
```

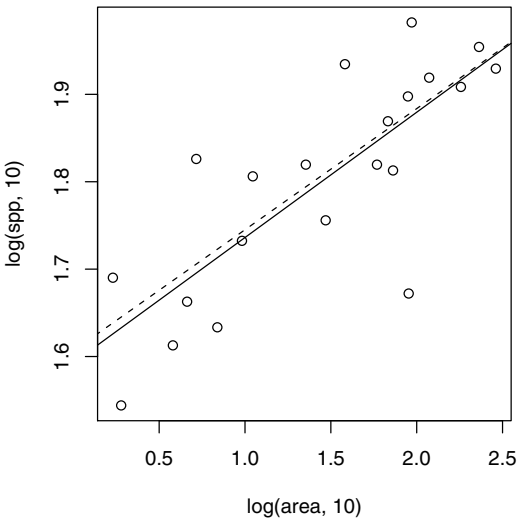


Fig. 10.18: Fitted power law species–area relations.

Fitting a species–area relation (Fig. 10.18)

Here we fit a species–area curve to data, and examine the slope. We could fit a nonlinear power relation ($y = cA^z$); this would be appropriate, for instance, if the residual noise around the line were of the same magnitude for all patch areas. We could use reduced major axis regression, which is appropriate when there is equivalent uncertainty or error on both x and y. Last (and most often), we could use a simple linear regression on the log-transformed data, which is appropriate when we know x to a high degree of accuracy, but measure y with some error, and the transformation causes the residual errors to be of similar magnitude at all areas. We start with the last (log-transformed). Here we plot the data, and fit a linear model to the common log-transformed data.

```
> plot(log(spp, 10) ~ log(area, 10), moths)
> mod <- lm(log(spp, 10) ~ log(area, 10), data = moths)
> abline(mod)
```

Next we fit the nonlinear curve to the raw data, and overlay that fit (on the log scale).

```
> mod.nonlin <- nls(spp ~ a * area^z, start = list(a = 1,
+     z = 0.2), data = moths)
> curve(log(coef(mod.nonlin)[1], 10) + x * coef(mod.nonlin)[2],
+     0, 3, add = TRUE, lty = 2)
```

Assessing species–area relations

Note that in Figure 10.18, the fits differ slightly between the two methods. Let's compare the estimates of the slope — we certainly expect them to be similar, given the picture we just drew.

```
> confint(mod)

                2.5 % 97.5 %
(Intercept)   1.50843 1.6773
log(area, 10) 0.09026 0.1964

> confint(mod.nonlin)

      2.5%   97.5%
a 31.61609 50.1918
z  0.08492  0.1958
```

We note that the estimates of the slopes are quite similar. Determining the better of the two methods (or others) is beyond the scope of this book, but be aware that methods can matter.

The major impetus for the species–area relation came from (i) Preston's work on connections between the species–area relation and the log-normal species abundance distribution [169,170], and (ii) MacArthur and Wilson's theory of island biogeography[19] [122].

Preston posited that, given the log-normal species abundance distributions (see above), then increasingly large samples should accumulate species at particular rates. Direct extensions of this work, linking neutral theory and maximum entropy theory to species abundances and species–area relations continues today [10,71,81]

Island biogeography

MacArthur and Wilson proposed a simple theory wherein the number of species on an oceanic island was a function of the immigration rate of new species, and extinction rate of existing species (Fig. 10.19). The number of species at any one time was *a dynamic equilibrium*, resulting from both slow inevitable extinction and slow continual arrival of replacements. Thus species composition on the island was predicted to change over time, that is, to undergo turnover.

Let us imagine immigration rate, y, as a function of the number of species already on an island, x (Fig. 10.19). This relation will have a negative slope, because as the number of species rises, that chance that a new individual actually represents a new species will decline. The immigration rate will be highest when there are no species on the island, $x = 0$, and will fall to zero when every conceivable species is already there. In addition, the slope should be decelerating

[19] (Originally proposed in a paper entitled "An Equilibrium Theory of Insular Zoogeography" [121])

(concave up) because some species will be much more likely than others to immigrate. This means that the immigration rate drops steeply as the most likely immigrants arrive, and only the unlikely immigrants are missing. Immigrants may colonize quickly for two reasons. First, as Preston noted, some species are simply much more common than others. Second, some species are much better dispersers than others.

Now let us imagine extinction rate, y, as a function of the number of species, x (Fig. 10.19). This relation will have a positive slope, such that the probability of extinction increases with the number of species. This is predicted to have an accelerating slope (concave-up), for essentially the same sorts of reasons governing the shape of the immigration curve: Some species are less common than others, and therefore more likely to become extinct do to demographic and environmental stochasticity, and second, some species will have lower fitness for any number of reasons. As the number of species accumulates, the more likely it will become that these extinction-prone species (rare and/or lower fitness) will be present, and therefore able to become extinct.

The *rate of change* of the number of species on the island, ΔR, will be the difference between immimigration, I, and extinction, E, or

$$\Delta R = I - E. \tag{10.21}$$

When $\Delta R = 0$, we have an equilibrium. If we knew the quantitative form of immigration and extinction, we could solve for the equilibrium. That equilibrium would be the point on the x axis, R, where the two rates cross (Fig. 10.19).

In MacArthur and Wilson's theory of island biogeography, these rates could be driven by the *sizes* of the islands, where

- larger islands had lower extinction rates because of larger average population sizes.
- larger islands had higher colonization rates because they were larger targets for dispersing species.

The distance between an island and sources of propagules was also predicted to influence these rates.

- Islands closer to mainlands had higher colonization rates of new species because more propagules would be more likely to arrive there,
- The effect of area would be more important for islands far from mainlands than for islands close to mainlands.

Like much good theory, these were simple ideas, but had profound effects on the way ecologists thought about communities. Now these ideas, of dispersal mediated coexistence and landscape structure, continue to influence community ecologists [15, 81, 102].

Drawing immigration and extinction curves

It would be fun to derive a model of immigration and extinction rates from first principles [121], but here we can illustrate these relations with some simple

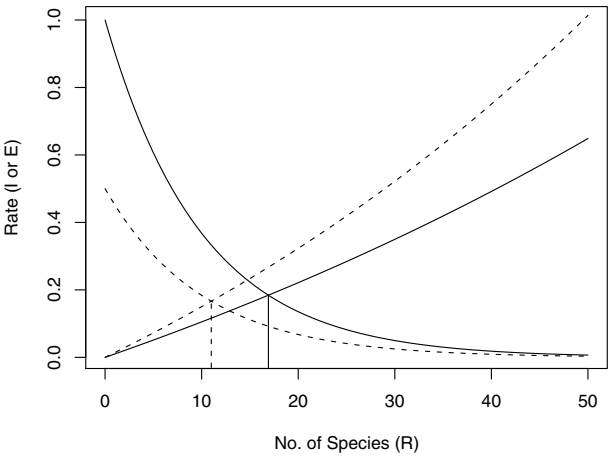


Fig. 10.19: Immigration and extinction curves for the theory of island biogeography. The declining curves represent immigration rates as functions of the number of species present on an island. The increasing curves represent extinction rates, also as functions of island richness. See text for discussion of the heights of these curves, i.e., controls on these rates. Here the dashed lines represent an island that is shrinking in size.

phenomenological graphs. We will assume that immmigration rate, I, can be represented as a simple negative exponential function $\exp(I_0 - iR)$, where I_0 is the rate of immigration to an empty island, and i is the per species negative effect on immigration rate.

```
> I0 <- log(1)
> b <- 0.1
> curve(exp(I0 - b * x), 0, 50, xlab = "No. of Species (R)",
+     ylab = "Rate (I or E)")
```

Note that extinction rate, E, must be zero if there are no species present. Imagine that extinction rate is a function of density and that average density declines as the number of species increases, or $\bar{N} = 1/R$.[20].

```
> d <- 0.01
> curve(exp(d * x) - 1, 0, 50, add = TRUE)
```

We subtract 1 merely to make sure that $E = 0$ when $R = 0$.

The number of species, R, will result from $\Delta R = 0 = I - E$, the point at which the lines cross.

$$I = e^{I_0 - bR} \tag{10.22}$$

$$E = e^{dR} - 1 \tag{10.23}$$

$$\delta R = 0 = I - E \tag{10.24}$$

[20] Why would this make sense ecologically?

Here we find this empricially by creating a function of R to minimize — we will minimize $(I - E)^2$; squaring the difference gives the quantity the convenient property that the minimum will be approached from either positive or negative values.

```
> deltaR <- function(R) {
+     (exp(I0 - b * R) - (exp(d * R) - 1))^2
+ }
```

We feed this into an optimizer for one-parameter functions, and specify that we know the optimum will be achieved somewhere in the interval between 1 and 50.

```
> optimize(f = deltaR, interval = c(1, 50))[["minimum"]]

[1] 16.91
```

The output tells us that the minimum was achieved when $R \approx 16.9$.

Now imagine that rising sea level causes island area to shrink. What is this predicted to do? It could

1. reduce the base immigration rate because the island is a smaller target,
2. increase extinction rate because of reduced densities.[21]

Let us represent reduced immigration rate by reducing I_0.

```
> I0 <- log(1/2)
> curve(exp(I0 - b * x), 0, 50, add = TRUE, lty = 2)
```

Next we increase extinction rate by increasing the per species rate.

```
> d <- 0.014
> curve(exp(d * x) - 1, 0, 50, add = TRUE, lty = 2)
```

If we note where the rates cross, we find that the number of predicted species has declined. With these new immigration and extinciton rates the predicted number of species is

```
> optimize(f = deltaR, interval = c(1, 50))[["minimum"]]

[1] 11.00
```

or 11 species, roughly a 35% decline ($(17 - 11)/17 = 0.35$).

The beauty of this theory is that it focuses our attention on *landscape level processes*, often outside the spatial and temporal limits of our sampling regimes. It specifies that any factor which helps determine the immigration rate or extinction rate, including island area or proximity to a source of propagules, is predicted to alter the equilibrium number of species at any point in time. We should further emphasize that the *identity* of species should change over time, that is, undergo turnover, because new species arrive and old species become extinct. The rate of turnover, however, is likely to be slow, because the species that are most likely to immigrate and least likely to become extinct will be the same species from year to year.

[21] We might also predict an increased extinction rate because of reduced rescue effect (Chapter 4).

10.5.3 Partitioning species–area relations

You may already be wondering if there is a link island biogeography and β-diversity. After all, as we move from island to island, and as we move from small islands to large islands, we typically encounter additional species, and that is what we mean by β-diversity. Sure enough, there are connections [42].

Let us consider the moth data we used above (Fig. 10.14). The total number of species in all of the patches is, as before, γ. The average richness of these patches is $\bar{\alpha}$, and also note that *part of what determines that average is the area of the patch.* That is, when a species is missing from a patch, part of the reason might be that the patch is smaller than it could be. We will therefore partition β into yet one more category: species missing due to patch size, β_{area}. This new quantity is the average difference between $\bar{\alpha}$ and the diversity *predicted* for the largest patch (Fig. 10.20). In general then,

$$\beta = \beta_{area} + \beta_{replace} \tag{10.25}$$

where $\beta_{replace}$ is the average number of species missing that are *not* explained by patch size.

In the context of these data (Fig. 10.20), we now realize that $\beta_1 = \beta_{area} + \beta_{ecoregion}$, so the full partition becomes

$$\gamma = \bar{\alpha}_1 + \beta_{area} + \beta_{ecoregion} + \beta_{geogr.region} \tag{10.26}$$

where $\beta_{replace} = \beta_{ecoregion} + \beta_{geogr.region}$. Note that earlier in the chapter, we did not explore the effect of area. In that case, $\beta_{ecoregion}$ included both the effect of area and the effect of ecoregion; here we have further partitioned this variation into variation due to patch size, as well as variation due to ecoregion. This reduces the amount of unexplained variation among sites within each ecoregion.

Let's calculate those differences now. We will use quantities we calculated above for $\bar{\alpha}_1$, $\bar{\alpha}_2$, γ, and a nonlinear species–area model from above. We can start to create a graph similar to Fig. 10.20.

```
> plot(spp ~ area, data = moths, ylim = c(30, 230),
+     xlab = "Area (ha)", ylab = "No. of Species (R)")
> curve(coef(mod.nonlin)[1] * x^coef(mod.nonlin)[2],
+     0, max(moths[["area"]]), add = TRUE, lty = 2,
+     lwd = 2)
> abline(h = g, lty = 3)
> text(275, g, quote(gamma), adj = c(0.5, 1.5),
+     cex = 1.5)
```

Next we need to find the *predicted richness* for the maximum area. We use our statistical model to find that.

```
> (MaxR <- predict(mod.nonlin, list(area = max(moths[["area"]]))))

[1] 88.62
```

We can now find β_{area}, β_{eco} and β_{geo}.

```
> b.area <- MaxR - a1
> b.eco <- a2 - (b.area + a1)
> b.geo <- g - a2
```

Now we have partitioned γ a little bit more finely with a beastiary of β's, where

- $\bar{\alpha}_1$ is the average site richness.
- β_{area} is the average number of species not observed, due to different patch sizes.
- β_{eco} is the average number of species not observed at a site, is not missing due to patch size, but is in the ecoregion.
- β_{geo} is the average number of species not found in the samples from different ecoregions.

Finally, we add lines to our graph to show the partitions.

```
> abline(h = g, lty = 3)
> abline(h = b.eco + b.area + a1, lty = 3)
> abline(h = b.area + a1, lty = 3)
> abline(h = a1, lty = 3)
```

Now we have further quantified how forest fragment area explains moth species richness. Such understanding of the spatial distribution of biodiversity provides a way to better quantify patterns governed by both dispersal and habitat preference, and allows us to better describe and manage biodiversity in human-dominated landscapes.

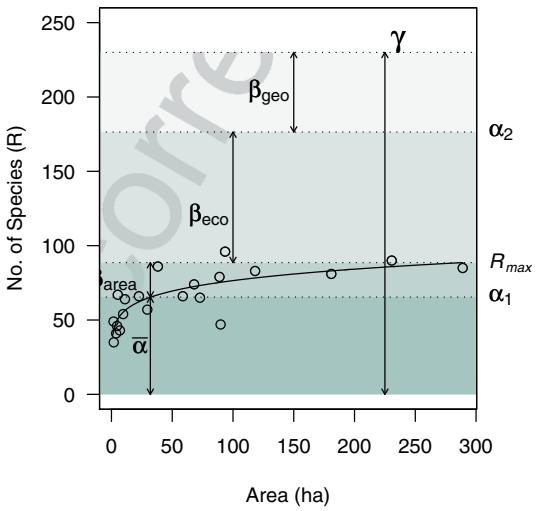


Fig. 10.20: Combining species–area relations with additive diversity partitioning. Forest fragment area explains relatively little of the diversity which accumulates in isolated patches distributed in space. However, it is likely that area associated with the collection of samples (i.e., the distances among fragments) contributes to β_{eco} and β_{geo}.

10.6 Summary

We have examined communities as multivariate entities which we can describe and compare in a variety of ways.

- Composition includes all species (multivariate data), whereas species diversity is a univariate description of the variety of species present.
- There are many ways to quantify species diversity, and they tend to be correlated. The simplest of these is richness (the number of species present), whereas other statistics take species' relative abundances into account.
- Species abundance distributions and rank abundance distributions are analogous to probability distributions, and provide more thorough ways to describe the patterns of abundance and variety in communities. These all illustrate a basic law of community ecology: most species are rare. Null models of community structure and processes make predictions about the shape of these distributions.
- Ecological neutral theory provides a dynamical model, not unlike a null model, which allows quantitative predictions relating demographic, immigration, and speciation rates, species abundance distributions, and patterns of variation in space and time.
- Another law of community ecology is that the number of species increases with sample area and appears to be influenced by immigration and extinction rates.
- We can partition diversity at different spatial scales to understand the structure of communities in landscapes.

Problems

Table 10.3: Hypothetical data for Problem 1.

Site	Sp. A	Sp. B	Sp. C
Site 1	0	1	10
Site 2	5	9	10
Site 3	25	20	10

10.1. How different are the communities in Table 10.3?
(a) Convert all data to relative abundance, where the relative abundance of each site sum to 1.
(b) Calculate the Euclidean and Bray-Curtis (Sørensen) distances between each pair of sites for both relative and absolute abundances.
(c) Calculate richness, Simpson's and Shannon-Wiener diversity for each site.

10.2. Use rarefaction to compare the tree richness in two 1 ha plots from the BCI data in the `vegan` package. Provide code, and a single graph of the expectations

for different numbers of individuals; include in the graph some indication of the uncertainty.

10.3. Select one of the 1 ha BCI plots (from the `vegan` package), and fit three different rank abundance distributions to the data. Compare and contrast their fits.

10.4. Simulate a neutral community of 1000 individuals, selecting the various criteria on yur own. Describe the change through time. Relate the species abundance distributions that you observe through time to the parameters you choose for the simulation.

10.5. Using the `dune` species data (`vegan` package), partition species richness into $\bar{\alpha}$, β, and γ richness, where rows are separate sites. Do the same thing using Simpson's diversity.