6. Worksheet: Among Site (Beta) Diversity – Part 1

Elaine Hoffman; Z620: Quantitative Biodiversity, Indiana University

05 February, 2025

OVERVIEW

In this worksheet, we move beyond the investigation of within-site α-diversity. We will explore β-diversity, which is defined as the diversity that occurs among sites. This requires that we examine the compositional similarity of assemblages that vary in space or time.

After completing this exercise you will know how to:

1. formally quantify β-diversity
2. visualize β-diversity with heatmaps, cluster analysis, and ordination
3. test hypotheses about β-diversity using multivariate statistics

Directions:

1. In the Markdown version of this document in your cloned repo, change "Student Name" on line 3 (above) with your name.
2. Complete as much of the worksheet as possible during class.
3. Use the handout as a guide; it contains a more complete description of data sets along with examples of proper scripting needed to carry out the exercises.
4. Answer questions in the worksheet. Space for your answers is provided in this document and is indicated by the ">" character. If you need a second paragraph be sure to start the first line with ">". You should notice that the answer is highlighted in green by RStudio (color may vary if you changed the editor theme).
5. Before you leave the classroom, **push** this file to your GitHub repo.
6. For the assignment portion of the worksheet, follow the directions at the bottom of this file.
7. When you are done, **Knit** the text and code into a PDF file.
8. After Knitting, submit the completed exercise by creating a **pull request** via GitHub. Your pull request should include this file (**6.BetaDiversity_1_Worksheet.Rmd**) with all code blocks filled out and questions answered) and the PDF output of `Knitr` (**6.BetaDiversity_1_Worksheet.pdf**).

The completed exercise is due on **Wednesday, February 5th, 2025 before 12:00 PM (noon)**.

1) R SETUP

Typically, the first thing you will do in either an R script or an RMarkdown file is setup your environment. This includes things such as setting the working directory and loading any packages that you will need.

In the R code chunk below, please provide the code to:

1) Clear your R environment,
2) Print your current working directory,
3) Set your working directory to your `Week3-Beta/` folder folder, and
4) Load the `vegan` R package (be sure to install first if you have not already).

```
rm(list = ls())
getwd()
```

```
## [1] "/cloud/project/Week3-Beta"
```

```
setwd("/cloud/project/Week3-Beta")

package.list <- c('vegan', 'ade4', 'viridis', 'gplots', 'indicspecies')
for (package in package.list){
  if (!require(package, character.only = TRUE, quietly = TRUE)) {
    install.packages(package)
    library(package, character.only = TRUE)
  }
}
```

```
## This is vegan 2.6-8
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
#install.packages("Rcmdr", dependencies = TRUE)
#install.packages("BiodiversityR", dependencies = TRUE)

library("BiodiversityR")
```

```
## Loading required package: tcltk
```

```
## Warning in fun(libname, pkgname): couldn't connect to display ":0"
```

```
## BiodiversityR 2.17-1.1: Use command BiodiversityRGUI() to launch the Graphical User Interface;
## to see changes use BiodiversityRGUI(changeLog=TRUE, backward.compatibility.messages=TRUE)
```

```
#BiodiversityR was loaded separately using the Tools tab.
```

2) LOADING DATA

Load dataset

In the R code chunk below, do the following:

1. load the `doubs` dataset from the `ade4` package, and
2. explore the structure of the dataset.

```
# note, please do not print the dataset when submitting
data(doubs)
#str(doubs)
#length(doubs)
#str(doubs, max.level = 1)
#head(doubs$env)
```

Question 1: Describe some of the attributes of the `doubs` dataset.

a. How many objects are in `doubs`?
b. How many fish species are there in the `doubs` dataset?
c. How many sites are in the `doubs` dataset?

Answer 1a: 4 ***Answer 1b***: 27 ***Answer 1c***: 30

Visualizing the Doubs River Dataset

Question 2: Answer the following questions based on the spatial patterns of richness (i.e., α-diversity) and Brown Trout (*Salmo trutta*) abundance in the Doubs River.

a. How does fish richness vary along the sampled reach of the Doubs River?
b. How does Brown Trout (*Salmo trutta*) abundance vary along the sampled reach of the Doubs River?
c. What do these patterns say about the limitations of using richness when examining patterns of biodiversity?

Answer 2a: Species richness appears to be higher downstream than upstream based on the plot of fish richness at each site provided in the packet for this assignment. ***Answer 2b***: Brown trout abundance is higher upstream than downstream with almost no individuals being dound downstream. ***Answer 2c***: This shows that richness alone will not tell you about the distribution of individual species in an environment. With the richness graph, it appears that the distribution of species throughout the stream is equal, but after looking at the brown trout abundance graph, it is clear that this is not the case.

3) QUANTIFYING BETA-DIVERSITY

In the R code chunk below, do the following:

1. write a function (`beta.w()`) to calculate Whittaker's β-diversity (i.e., β_w) that accepts a site-by-species matrix with optional arguments to specify pairwise turnover between two sites, and
2. use this function to analyze various aspects of β-diversity in the Doubs River.

```
fish <- doubs$fish


beta.w <- function(site.by.species = ""){
  SbyS.pa <- decostand(site.by.species, method = "pa")
  #convert to presence-absence
  S <- ncol(SbyS.pa[, which(colSums(SbyS.pa) > 0)])
  #number of species in the region
  a.bar <- mean(specnumber(SbyS.pa))
  #average richness at each site
  b.w <- round(S/a.bar, 3)
  #round to 3 decimal places
  return(b.w)
}

beta.w <- function(site.by.species = "", sitenum1 = "", sitenum2 = "", pairwise = FALSE){
  #only if we specify pairwise as TRUE, do this:
  if(pairwise == TRUE){
    #As a check, let's print an error if we do not provide needed arguments
    if(sitenum1 == "" | sitenum2 == "") {
      print("Error: please specify sites to compare")
      return(NA)}
    #If our function made it this far, let us calculate pairwise beta diversity
    site1 = site.by.species[sitenum1,]
    #select site 1
    site2 = site.by.species[sitenum2,]
    #select site 2
    site1 = subset(site1, select = site1 > 0)
    #Removes absences
    site2 = subset(site2, select = site2 > 0)
```

```
    #Removes absences
    gamma = union(colnames(site1), colnames(site2))
    #Gamma species pool
    s = length(gamma)
    #Gamma richness
    a.bar = mean(c(specnumber(site1), specnumber(site2)))
    #Mean sample richness
    b.w = round(s/a.bar -1, 3)
    return(b.w)
  }
  #otherwise pairwise defaults to FALSE, so do this, like before:
  else{
    SbyS.pa <- decostand(site.by.species, method = "pa")
    #convert to presence-absence
    S <- ncol(SbyS.pa[, which(colSums(SbyS.pa) > 0)])
    #number of species in region
    a.bar <- mean(specnumber(SbyS.pa))
    #average richness at each site
    b.w <- round(S/a.bar, 3)
    return(b.w)
  }
}


beta.w(fish)
```

```
## [1] 2.16
```

```
# Compute beta diversity between site 1 and site 2
beta_1_2 <- beta.w(site.by.species = fish, sitenum1 = 1, sitenum2 = 2, pairwise = TRUE)
cat(" w between site 1 and site 2:", beta_1_2, "\n")
```

```
##  w between site 1 and site 2: 0.5
```

```
# Compute beta diversity between site 1 and site 10
beta_1_10 <- beta.w(site.by.species = fish, sitenum1 = 1, sitenum2 = 10, pairwise = TRUE)
cat(" w between site 1 and site 10:", beta_1_10, "\n")
```

```
##  w between site 1 and site 10: 0.714
```

```
# Compare results
if (beta_1_2 < beta_1_10) {
  cat("Site 1 is more similar to site 2 than to site 10.\n")
} else if (beta_1_2 > beta_1_10) {
  cat("Site 1 is more similar to site 10 than to site 2.\n")
} else {
  cat("Site 1 is equally similar to both site 2 and site 10.\n")
}
```

```
## Site 1 is more similar to site 2 than to site 10.
```

Question 3: Using your `beta.w()` function above, answer the following questions:

a. Describe how local richness (α) and turnover (β) contribute to regional (γ) fish diversity in the Doubs.
b. Is the fish assemblage at site 1 more similar to the one at site 2 or site 10?
c. Using your understanding of the equation $\beta_w = \gamma/\alpha$, how would your interpretation of β change if we instead defined beta additively (i.e., $\beta = \gamma - \alpha$)?

Answer 3a: According to the information in the appendix, the relationship between gamma (g) diversity and alpha (a) and beta (b) diversity is more recently regarded as g = a + b which would indicate that alpha diversity and beta diversity exhibit additive partitioning. Whitaker, however, believed in a multiplicative approach (g = a * b), so given the b.w value of 2.16 it appears that there is a lot of turnover (beta) or variation between species at the individual sites (alpha) and the total species pool (gamma). Another way to explain this would be to say that each site proportionally has few species in comparison to the total species pool. ***Answer 3b***: Site 1 is more similar to site 2 than to site 10. ***Answer 3c***: Defining beta diversity as additively changes the interpretation of beta diversity to an absolute value of species turnover rather than a viewing turnover as a ratio as we were doing before. This indicates that the total species pool is much larger than the species pool at any one site. This is basically the same overall finding as the interpretation in 3a, we just reached the conclusion through different paths.

The Resemblance Matrix

In order to quantify β-diversity for more than two samples, we need to introduce a new primary ecological data structure: the **Resemblance Matrix**.

Question 4: How do incidence- and abundance-based metrics differ in their treatment of rare species?

Answer 4:

In the R code chunk below, do the following:

1. make a new object, `fish`, containing the fish abundance data for the Doubs River,
2. remove any sites where no fish were observed (i.e., rows with sum of zero),
3. construct a resemblance matrix based on Sørensen's Similarity ("fish.ds"), and
4. construct a resemblance matrix based on Bray-Curtis Distance ("fish.db").

```
fish <- doubs$fish
fish <- fish[-8, ] #remove site 8 from data

#calculate Jaccard
fish.dj <- vegdist(fish, method = "jaccard", binary = TRUE)

#calculate bray-curtis
fish.db <- vegdist(fish, method = "bray")

#calculate sorensen
fish.ds <- vegdist(fish, method = "bray", binary = TRUE)
```

Question 5: Using the distance matrices from above, answer the following questions:

a. Does the resemblance matrix (`fish.db`) represent similarity or dissimilarity? What information in the resemblance matrix led you to arrive at your answer?
b. Compare the resemblance matrices (`fish.db` or `fish.ds`) you just created. How does the choice of the Sørensen or Bray-Curtis distance influence your interpretation of site (dis)similarity?

Answer 5a: The resemblance matrix fish.db represents dissimilarity because the values are closer to 1 than they are to 0. This interpretation can be made because the matrix was constructed using Bray-Curtis values which are a measure of percentage of difference.

Answer 5b:Using the Sorensen index generates a matrix where values closer to 1 are similar and values closer to 0 are not which is the opposite of Bray-Curtis. When looking at the fish.ds matrix, the majority of the values are close to 0 which agrees with the findings of the Bray-Curtis matrix that most of the sites are dissimilar.

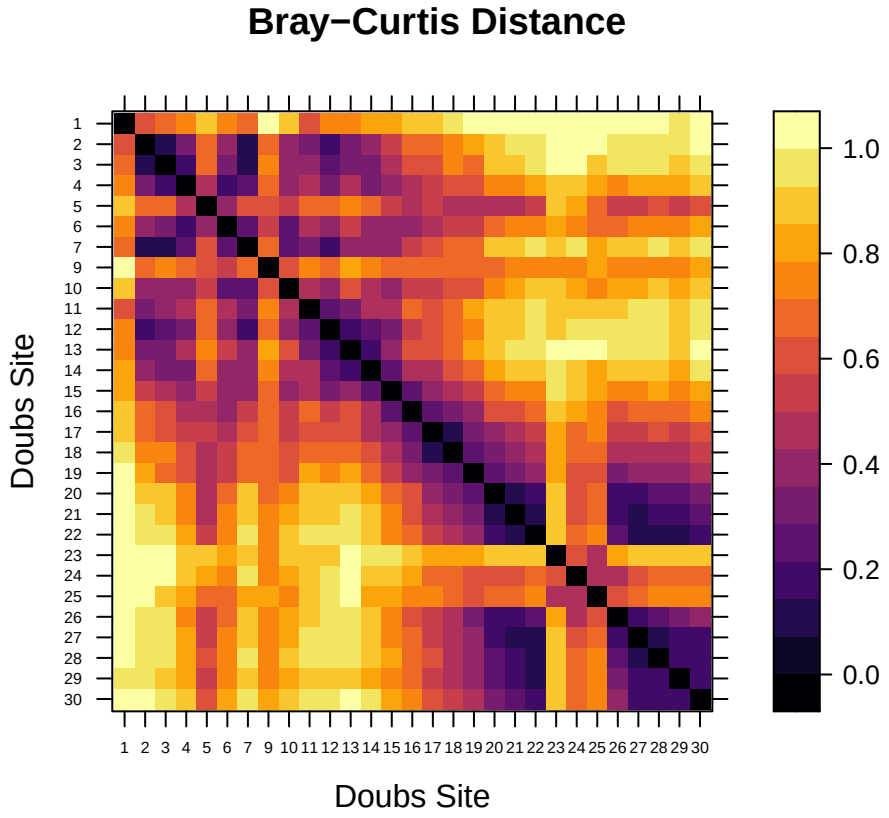
4) VISUALIZING BETA-DIVERSITY

A. Heatmaps

In the R code chunk below, do the following:

1. define a color palette,
2. define the order of sites in the Doubs River, and
3. use the `levelplot()` function to create a heatmap of fish abundances in the Doubs River.

```
#Define order of sites
order <- rev(attr(fish.db, "Labels"))

#Plot Heatmap
levelplot(as.matrix(fish.db)[, order], aspect = "iso", col.regions = inferno,
          xlab = "Doubs Site", ylab = "Doubs Site", scales = list(cex = 0.5),
          main = "Bray-Curtis Distance")
```

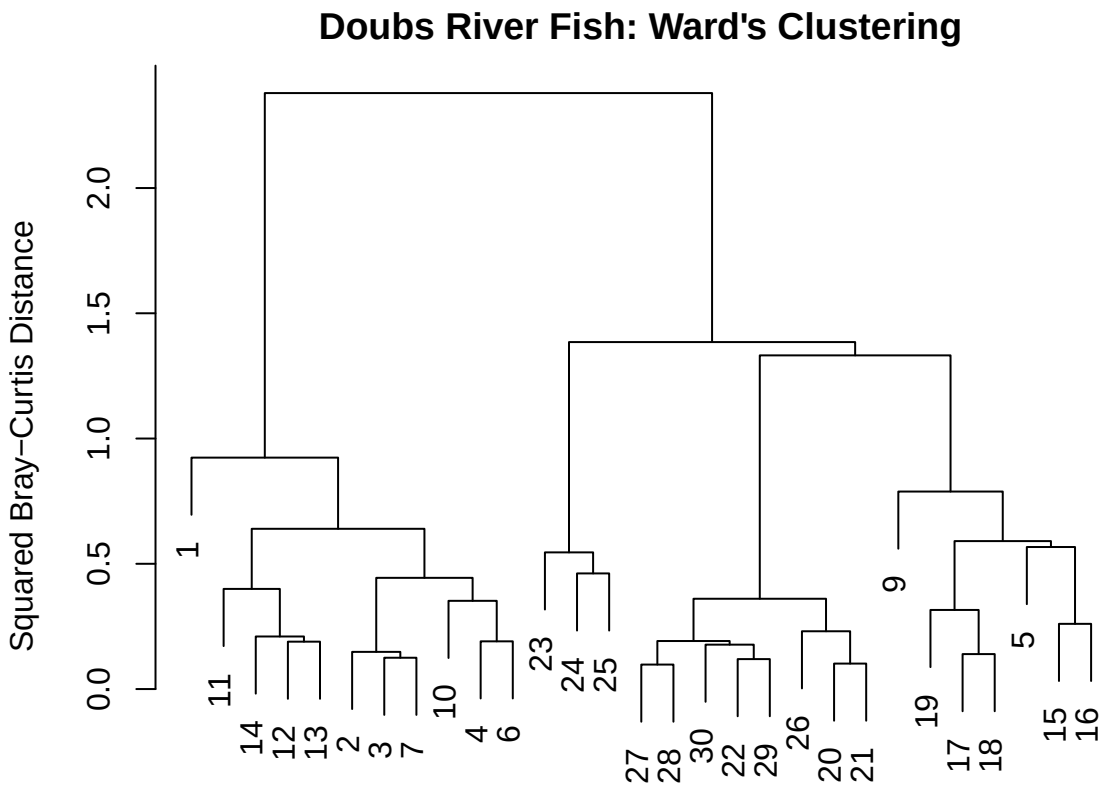


B. Cluster Analysis

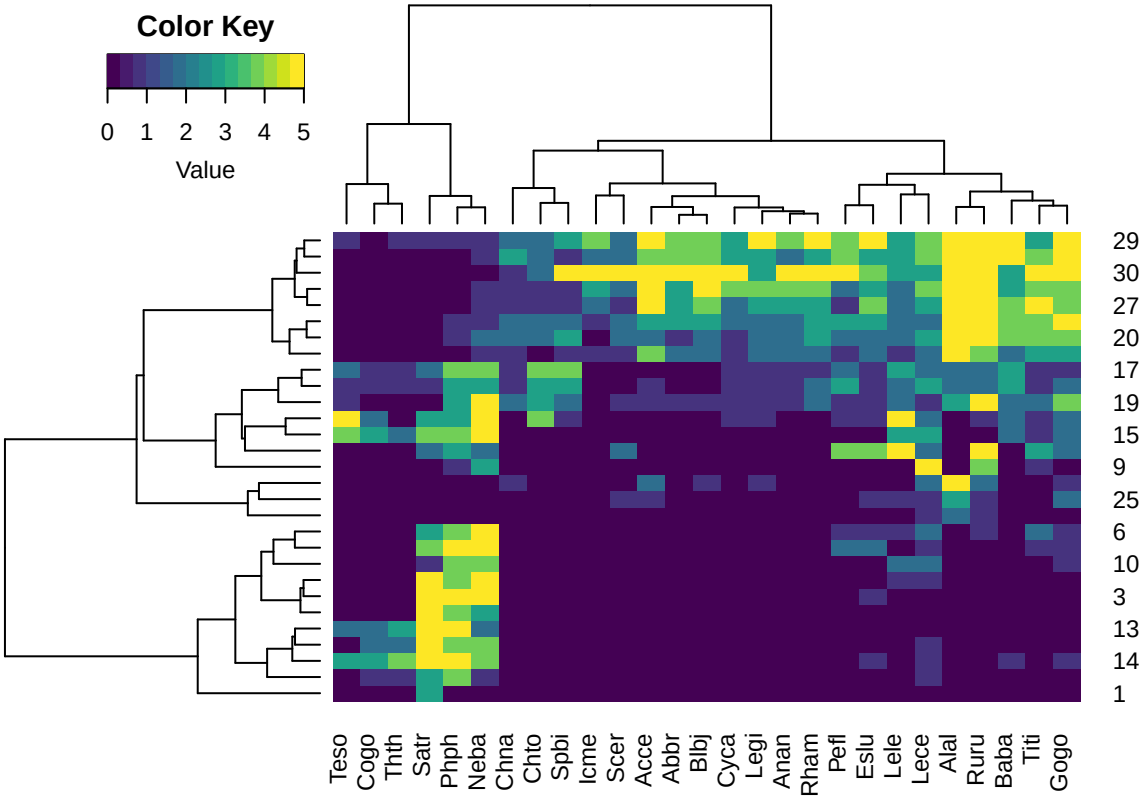
In the R code chunk below, do the following:

1. perform a cluster analysis using Ward's Clustering, and
2. plot your cluster analysis (use either `hclust` or `heatmap.2`).

```
#Perform cluster analysis
fish.ward <- hclust(fish.db, method = "ward.D2")

#Plot cluster
par(mar = c(1, 5, 2, 2) + 0.1)
plot(fish.ward, main = "Doubs River Fish: Ward's Clustering", ylab = "Squared Bray-Curtis Distance")
```




```
gplots::heatmap.2(as.matrix(fish),
                  distfun = function(x) vegdist(x, method = "bray"),
                  hclustfun = function(x) hclust(x, method = "ward.D2"),
                  col = viridis, trace = "none", density.info = "none")
```



Question 6: Based on cluster analyses and the introductory plots that we generated after loading the data, develop an ecological hypothesis for fish diversity the `doubs` data set?

> ***Answer 6***: Fish diversity in the doubs river exhibits spatial variation with a gradient of fish diversity starting with a higher fish diversity upstream and lower fish diversity downstream. This is likely due to a gradient of environmental conditions such as temperature, flow rate, salinity, etc.

C. Ordination

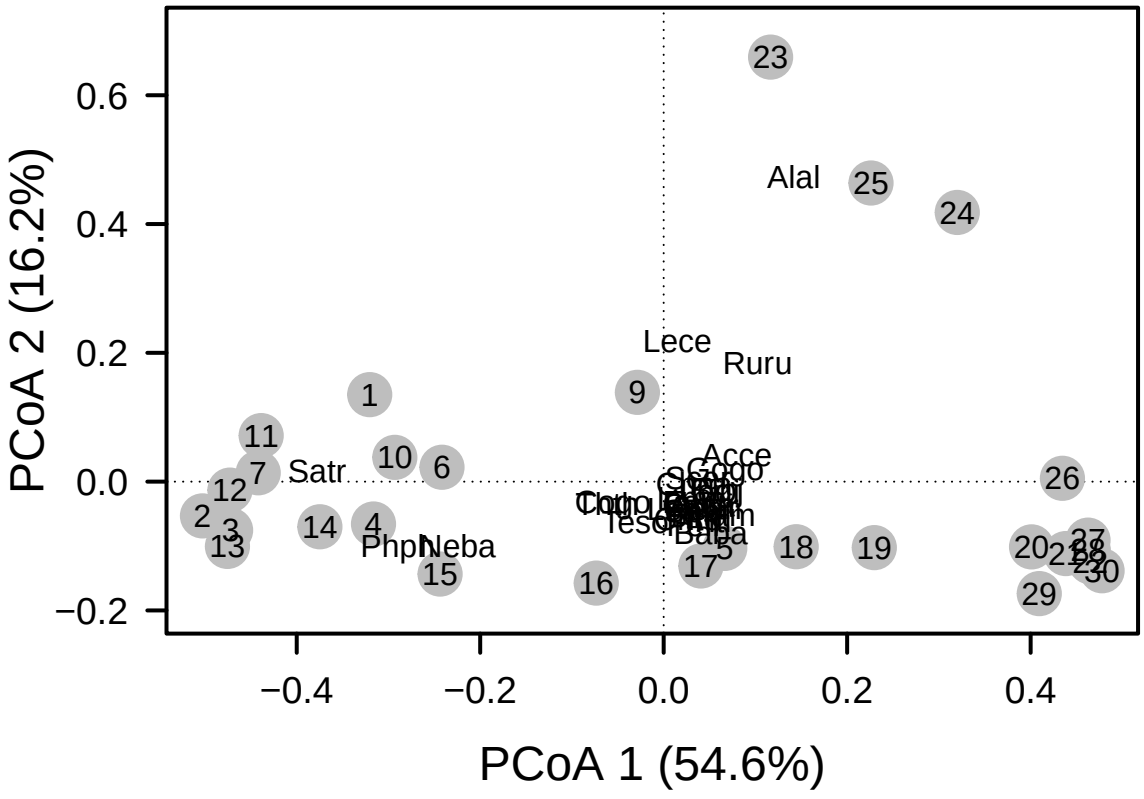
Principal Coordinates Analysis (PCoA)

In the R code chunk below, do the following:

1. perform a Principal Coordinates Analysis to visualize beta-diversity
2. calculate the variation explained by the first three axes in your ordination
3. plot the PCoA ordination,
4. label the sites as points using the Doubs River site number, and
5. identify influential species and add species coordinates to PCoA plot.

```
fish.pcoa <- cmdscale(fish.db, eig = TRUE, k = 3)

explainvar1 <- round(fish.pcoa$eig[1] / sum(fish.pcoa$eig), 3) * 100
explainvar2 <- round(fish.pcoa$eig[2] / sum(fish.pcoa$eig), 3) * 100
explainvar3 <- round(fish.pcoa$eig[3] / sum(fish.pcoa$eig), 3) * 100
sum.eig <- sum(explainvar1, explainvar2, explainvar3)

#Define plot parameters
par(mar = c(5, 5, 1, 2) + 0.1)

#Initiate plot
```

```r
plot(fish.pcoa$points[ ,1], fish.pcoa$points[ ,2], ylim = c(-0.2, 0.7),
     xlab = paste("PCoA 1 (", explainvar1, "%)", sep = ""),
     ylab = paste("PCoA 2 (", explainvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5,
     cex.axis = 1.2, axes = FALSE)

# Add axes
axis(side = 1, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
axis(side = 2, labels = T, lwd.ticks = 2, cex.axis = 1.2, las = 1)
abline(h = 0, v = 0, lty = 3)
box(lwd = 2)

#Add points and labels
points(fish.pcoa$points[ ,1], fish.pcoa$points[ ,2],
       pch = 19, cex = 3, bg = "gray", col = "gray")
text(fish.pcoa$points[ ,1], fish.pcoa$points[ ,2],
     labels = row.names(fish.pcoa$points))

#First we calculate the relative abundances of each species at each site
fishREL <- fish
for(i in 1:nrow(fish)){
  fishREL[i, ] = fish[i, ] / sum(fish[i, ])
}

#Now we use this information to calculate and add species scores
fish.pcoa <- add.spec.scores(fish.pcoa, fishREL, method = "pcoa.scores")
text(fish.pcoa$cproj[ ,1], fish.pcoa$cproj[ ,2],
     labels = row.names(fish.pcoa$cproj), col = "black")
```



In the R code chunk below, do the following:

1. identify influential species based on correlations along each PCoA axis (use a cutoff of 0.70), and
2. use a permutation test (999 permutations) to test the correlations of each species along each axis.

```
# Compute species correlations with PCoA axes
spe.corr <- add.spec.scores(fish.pcoa, fishREL, method = "cor.scores")$cproj

# User-defined cutoff for influential species
corrcut <- 0.7

# Select influential species based on correlation cutoff
imp.spp <- spe.corr[abs(spe.corr[, 1]) >= corrcut | abs(spe.corr[, 2]) >= corrcut, ]

# Extract species names
influential_species <- rownames(imp.spp)

# Print list of influential species
cat("Influential species (correlation  ", corrcut, "):\n")
```

```
## Influential species (correlation  0.7 ):
```

```
print(influential_species)
```

```
##  [1] "Phph" "Neba" "Rham" "Legi" "Cyca" "Abbr" "Acce" "Blbj" "Alal" "Anan"
```

```
#Permutation test for species abundances across axes
fit <- envfit(fish.pcoa, fishREL, perm = 999)
print(fit)
```

```
##
## ***VECTORS
##
##          Dim1     Dim2     r2 Pr(>r)
## Cogo -0.83884 -0.54438 0.2982  0.007 **
## Satr -0.99904  0.04371 0.4326  0.005 **
## Phph -0.94110 -0.33813 0.7814  0.001 ***
## Neba -0.91413 -0.40543 0.6234  0.001 ***
## Thth -0.87692 -0.48063 0.2634  0.029 *
## Teso -0.44704 -0.89452 0.1700  0.087 .
## Chna  0.99707 -0.07644 0.4612  0.001 ***
## Chto  0.42032 -0.90738 0.2579  0.026 *
## Lele  0.33041 -0.94384 0.0495  0.528
## Lece  0.06856  0.99765 0.3399  0.008 **
## Baba  0.54118 -0.84091 0.6752  0.001 ***
## Spbi  0.57341 -0.81927 0.4138  0.004 **
## Gogo  0.97507  0.22188 0.3753  0.002 **
## Eslu  0.72044 -0.69352 0.1673  0.082 .
## Pefl  0.43762 -0.89916 0.3048  0.007 **
## Rham  0.72476 -0.68901 0.8301  0.001 ***
## Legi  0.93461 -0.35568 0.7016  0.001 ***
## Scer  0.98569  0.16858 0.3533  0.005 **
## Cyca  0.68181 -0.73153 0.7743  0.001 ***
## Titi  0.64378 -0.76521 0.4586  0.002 **
## Abbr  0.77254 -0.63497 0.7128  0.001 ***
## Icme  0.75626 -0.65427 0.5270  0.002 **
## Acce  0.88799  0.45986 0.6294  0.002 **
```

```
## Ruru  0.48379  0.87518 0.5177  0.002 **
## Blbj  0.95802 -0.28671 0.6766  0.001 ***
## Alal  0.28755  0.95777 0.8592  0.001 ***
## Anan  0.74277 -0.66954 0.7894  0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

Question 7: Address the following questions about the ordination results of the `doubs` data set:

a. Describe the grouping of sites in the Doubs River based on fish community composition.
b. Generate a hypothesis about which fish species are potential indicators of river quality.

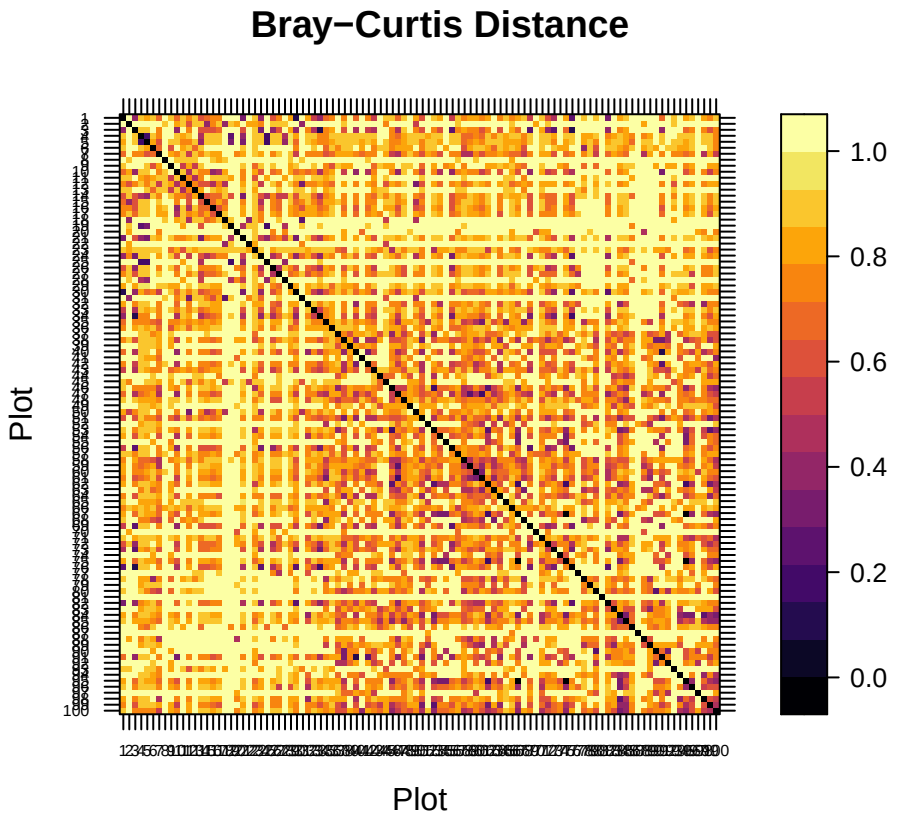
> ***Answer 7a***: The sites can be grouped into roughly three categories based on the analysis that was performed. These groups include upstream sites characterized by low Dim 1 values in the permutation test, midstream sites characterized by intermediate Dim 1 values, and downstream sites characterized by high Dim 1 values. ***Answer 7b***: The presence of fish species grouped into the upstream sites are potential indicators of river quality because there is some environmental factor (potentially contamination) preventing them from being able to inhabit the lower portion of the river. If they are present in a site, then that site is likely optimal for supporting fish and free from significant sources of polution.

SYNTHESIS

Load the dataset from that you and your partner are using for the team project. Use one of the tools introduced in the beta diversity module to visualize your data. Describe any interesting patterns and identify a hypothesis is relevant to the principles of biodiversity.

```
#I tried to run a PCoA with this data set but it maxed out the RAM on posit and eventually crashed.
#Anna had the same issue so we filtered to the first 100 plots and made a heat map so we could still in

library(vegan)
library(tibble)
library(lattice)

tree <-read.csv("https://raw.githubusercontent.com/anna-l-2/QB_biodiversity_project_EH/d5b1465aaa135077

tree.species.df <- data.frame(Plot_ID = tree$PLOT, Species_ID = tree$SPCD)
#print(tree.species.df)

tree.ss.df <- as.data.frame.matrix(table(tree.species.df$Plot_ID, tree.species.df$Species_ID))
tree.ss.df <- rownames_to_column(tree.ss.df, var = "Plot_ID")
tree.ss.df <-tree.ss.df[1:100,1:100]
#print(tree.ss.df)

tree.species.only.df <- tree.ss.df[, -1]
#print(tree.species.only.df)

tree.species.only.df.L <-tree.species.df

#Bray-Curtis

tree.db <-vegdist(tree.species.only.df, method = "bray")

order <-rev(attr(tree.db, "Labels"))
```

```
levelplot(as.matrix(tree.db)[,order], aspect = "iso", col.regions =inferno,
          xlab = "Plot", ylab = "Plot", scales =list(cex =.5),
          main = "Bray-Curtis Distance")
```



>By looking at this heat map it is difficult to make any conclusions about the dataset as a whole because the sample size we were able to pull is so small in comparison to the entire dataset as a whole. I do notice what seems to be some correllation between the plots represented by small numbers and those represented by large numbers. Further analysis will be needed to determine what is causing that differece. > We have developed an overarching question for this project, so I will be offering our hypothesis for that question. Question: What is the influence of the type of mycorrhizal association on tree species abundance when exposed to the pressure of invasive species? Hypothesis: Mycorrhizal association will influence tree species abundance in plots containing invasive species.