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

Erica Nadolski; Z620: Quantitative Biodiversity, Indiana University

31 January, 2023

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 today, it is *imperative* that you **push** this file to your GitHub repo, at whatever stage you are. Ths will enable you to pull your work onto your own computer.
6. When you have completed the worksheet, **Knit** the text and code into a single PDF file by pressing the `Knit` button in the RStudio scripting panel. This will save the PDF output in your '6.BetaDiversity' folder.
7. After Knitting, please submit the worksheet by making a **push** to your GitHub repo and then create 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 1st, 2023 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, provide the code to:

1. clear your R environment,
2. print your current working directory,
3. set your working directory to your "*/6.BetaDiversity*" folder, and
4. load the `vegan` R package (be sure to install if needed).

```
getwd()
```

```
## [1] "/Users/ericanadolski/GitHub/QB2023_Nadolski/2.Worksheets/6.BetaDiversity"
```

```
setwd("/Users/ericanadolski/GitHub/QB2023_Nadolski/2.Worksheets/6.BetaDiversity")

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 2.6-4
```

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

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

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, max.level=1)
```

```
## List of 4
##  $ env    :'data.frame': 30 obs. of  11 variables:
##  $ fish   :'data.frame': 30 obs. of  27 variables:
##  $ xy     :'data.frame': 30 obs. of  2 variables:
##  $ species:'data.frame': 27 obs. of  4 variables:
```

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 data frames ***Answer 1b***: 27 species ***Answer 1c***: 30 sites

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?

```
fish <- doubs$fish
env <- doubs$env
specnumber(fish)
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  1  3  4  8 11 10  5  0  5  6  6  6  6 10 11 17 22 23 23 22 23 22  3  8  8 21
## 27 28 29 30
## 22 22 26 21
```

> ***Answer 2a***: Richness increases on average from sites 1-30. ***Answer 2b***: Brown trout abudance shows an opposite relationship to total richness, its abundance is higher at the first sites and decreases at later sites. ***Answer 2c***: This indicates that overall richness measures can mask interesting aspects of the data including patterns of individual species.

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.

```
beta.w <- function(site.by.species="", sitenum1="", sitenum2="", pairwise=FALSE){
  # ONLY if pairwise = TRUE do this
  if (pairwise == TRUE){
    # as a check, print error if needed arguments are missing
    if (sitenum1 == "" | sitenum2 == ""){
      print("Error: please specify sites to compare")
      return(NA)}
    # if our function made it this far, calculate pairwise Beta diversity
    site1 = site.by.species[sitenum1,]
    site2 = site.by.species[sitenum2,]
    # remove absences
    site1 = subset(site1, select = site1 > 0)
    site2 = subset(site2, select = site2 > 0)
    # gamma species pool
    gamma = union(colnames(site1), colnames(site2))
    # gamma richness
    s = length(gamma)
    # mean sample richness
    a.bar = mean(c(specnumber(site1), specnumber(site2)))
```

```
    b.w = round(s/a.bar - 1, 3)
    return(b.w)
  }
  # otherwise, pairwise defaults to false, so do this like before:
  else{
    # convert to presence absence
    SbyS.pa <- decostand(site.by.species, method="pa")
    # number of sp in the region
    S <- ncol(SbyS.pa[,which(colSums(SbyS.pa)>0)])
    # avg richness at each site
    a.bar <- mean(specnumber(SbyS.pa))
    # convert to 3 decimal places
    b.w <- round(S/a.bar, 3)
    return(b.w)
  }
}

beta.w(fish)
```

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

```
beta.w(fish,1,2,pairwise=TRUE)
```

```
## [1] 0.5
```

```
beta.w(fish,1,10,pairwise=TRUE)
```

```
## [1] 0.714
```

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***: Beta diversity explains many more nuances of the regional gamma diversity of fish in the Doubs river. ***Answer 3b***: Site 1 is more similar to site 2 than site 10 ($0.5 < 0.714$) ***Answer 3c***: Whittaker originally proposed that the relationship between alpha and gamma was multiplicative. If it was additive, then it would be interpreted as quantifying how much more dissimilarity (or species diversity) the whole dataset gamma contains than the average site alpha within the dataset.

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: Incidence-based measures weigh rare species equally to abundant species in their contribution to diversity; abundance-based metrics weigh species based on their abundance so rare species get less weight overall.

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 <- fish[-8,]
fish.ds <- vegdist(fish, method="bray", binary=TRUE)
fish.db <- vegdist(fish, method="bray")
```

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 values in fish.db represent dissimilarity, so closer to 0 means more similar and closer to 1 means more dissimilar. I deduced this because the earlier sites has higher values pairwise to the later sites, and the later sites had lower values pairwise with other later sites.

Answer 5b: The two methods generate overall similar sets of values.

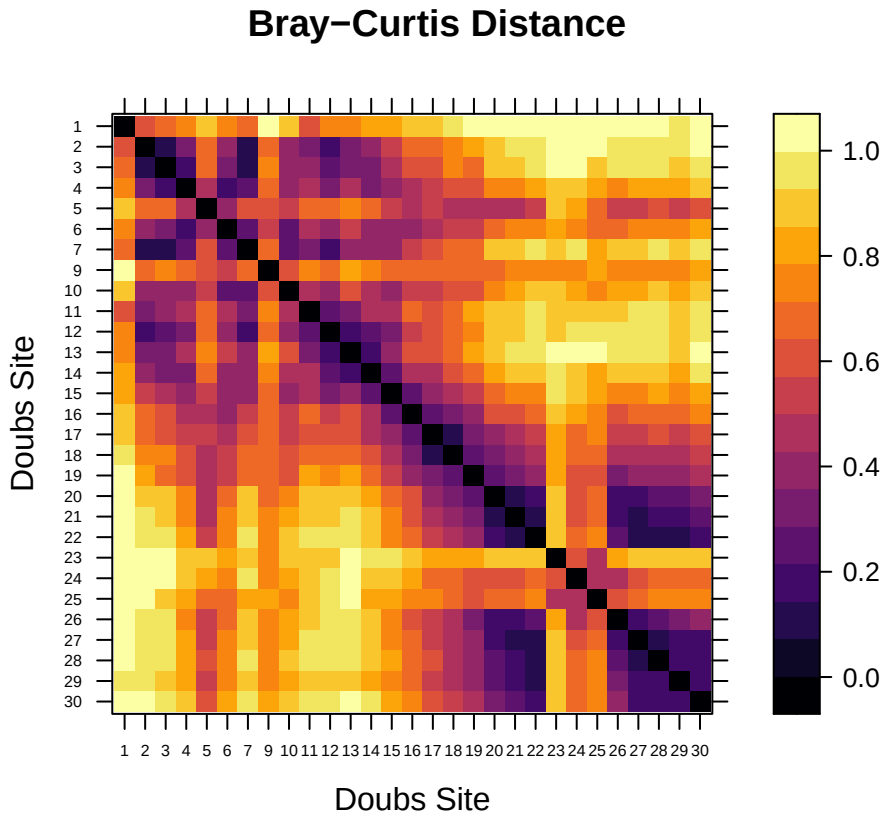
4) VISUALIZING BETA-DIVERSITY

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.

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

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")
```

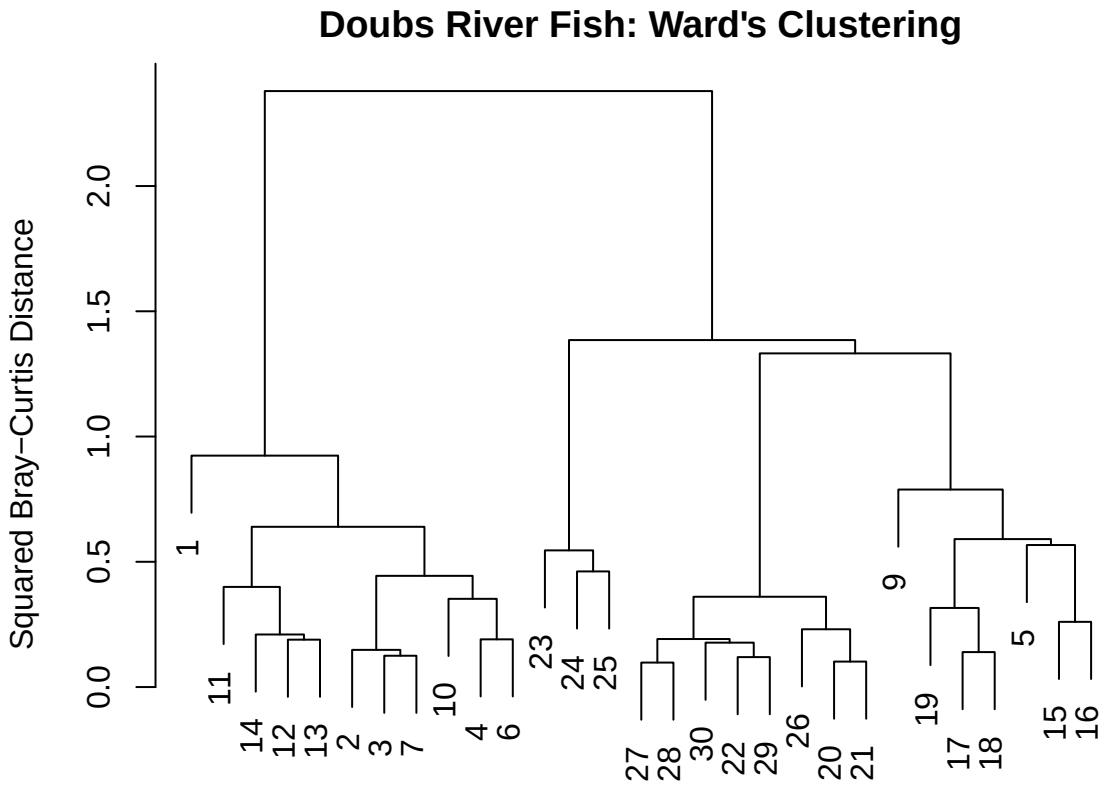


B. Cluster Analysis

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`).

```
fish.ward <- hclust(fish.db, method= "ward.D2")

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



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***: Ecological hypothesis: later sites that are lower down the river are more similar in these environmental qualities: _____ and this explains the fish diversity.

C. Ordination

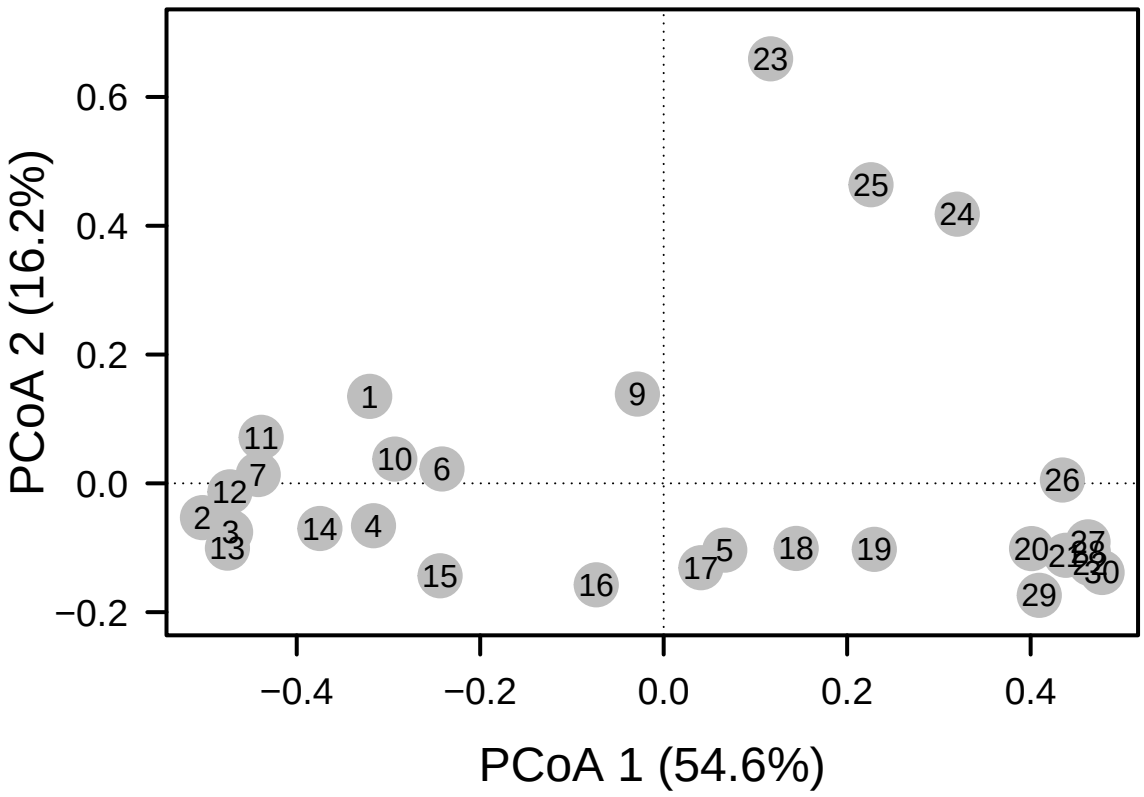
Principal Coordinates Analysis (PCoA)

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)

par(mar = c(5, 5, 1, 2) + 0.1)
pcoa <- 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);
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);
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))
```

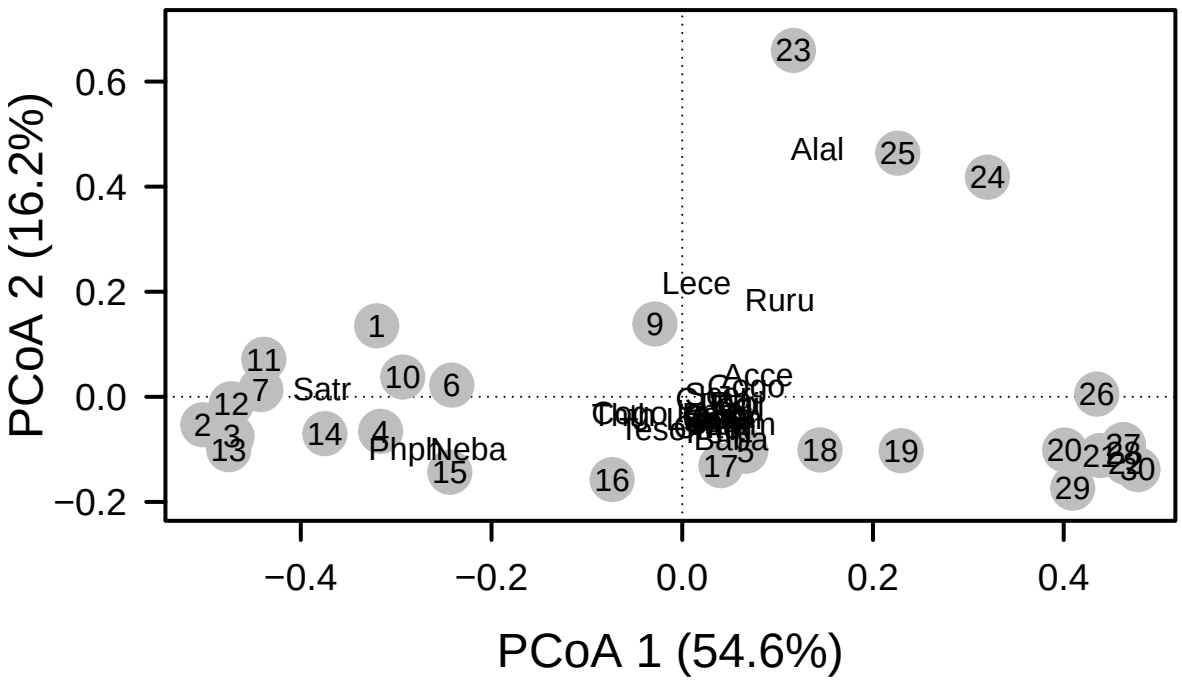


In the R code chunk below, do the following:

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.

```
# now calculating relative abundance of each fish at each site to calculate and add spec scores
fishREL <- fish
for(i in 1:nrow(fish)){
  fishREL[i, ] = fish[i, ] / sum(fish[i, ])
}

fish.pcoa <- add.spec.scores.class(fish.pcoa,fishREL, method = "pcoa.scores")

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);
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);
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$points));
text(fish.pcoa$cproj[ ,1], fish.pcoa$cproj[ ,2],
     labels = row.names(fish.pcoa$cproj), col = "black")
```




```
spe.corr <- add.spec.scores.class(fish.pcoa, fishREL, method = "cor.scores")$cproj
corrcut <- 0.7 #user defined cutoff
imp.spp <- spe.corr[abs(spe.corr[ ,1]) >= corrcut | abs(spe.corr[ ,2]) >= corrcut]

# permutation test for species abundances across axes
fit <- envfit(fish.pcoa, fishREL, perm=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.

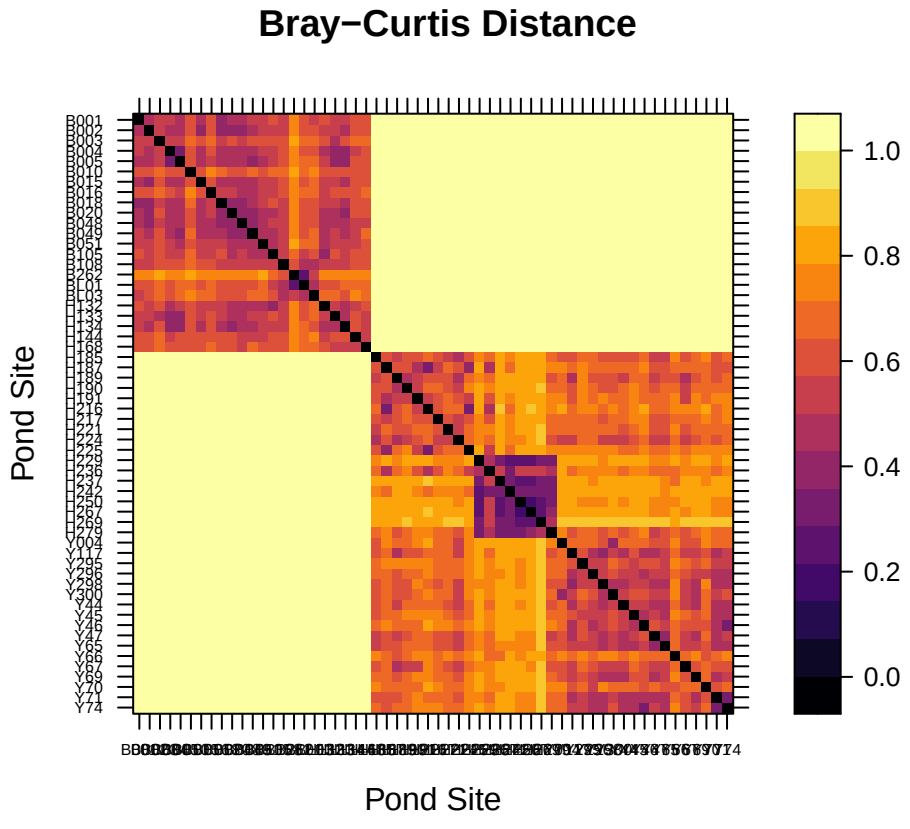
Answer 7a: The 1st principal component separates out clusters of sites that contain Satr, and secondarily that contain Phph and Neba. The 2nd principal component separates out clusters of sites that contain Alal, and secondarily that contain Lece and Ruru. The rest of the fish species cluster tightly on the PCoA plot near the origin, meaning that they do not contribute heavily

to the 1st or 2nd PCs. ***Answer 7b***: Satr, Phph, and Neba are more abundant at sites near the river source, where the water temperature is cooler, and the water is clearer and has higher oxygen levels; these fish can potential indicators of this type of river. In contrast, Alal, Lece, and Ruru are abundant downstream the river, where the water contains higher levels of sediments and therefore nutrients, there will be less light through the murkier water, and lower oxygen levels; these fish are potential indicators of this type of river environment.

SYNTHESIS

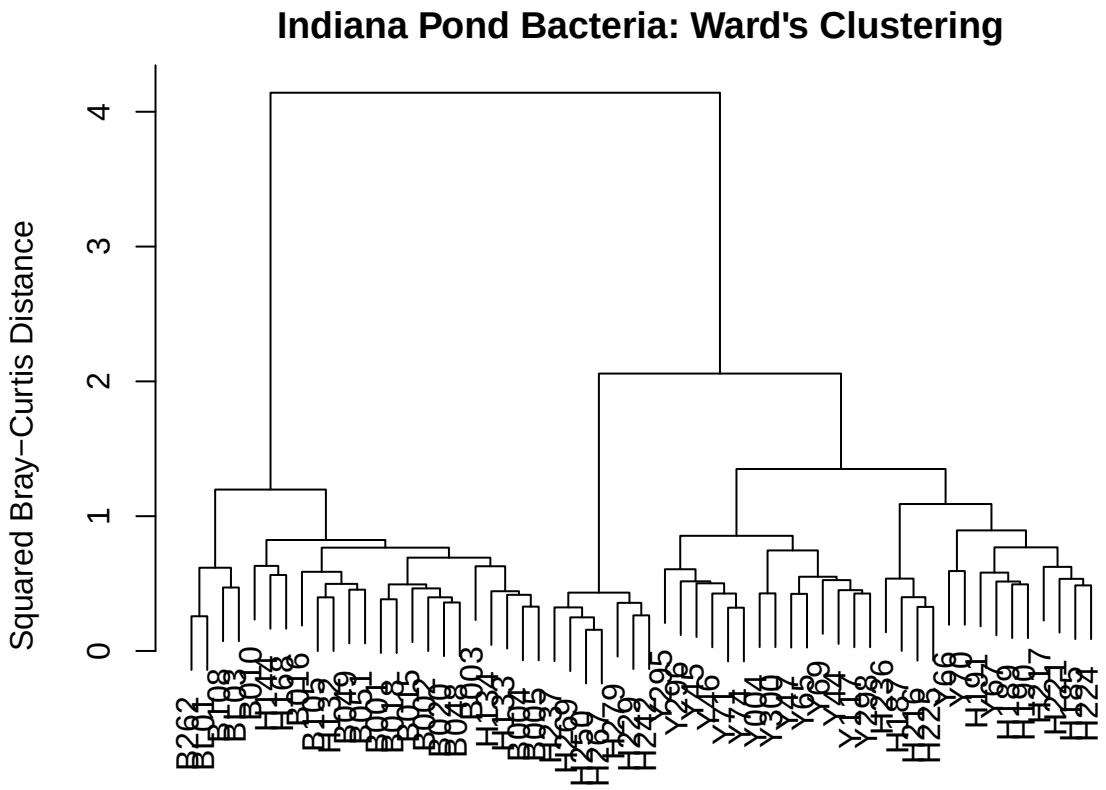
Load the dataset from that you and your partner are using for the team project. Use one of the tools introduced in the beta diversity module to visualize your data. Describe any interesting patterns and identify a hypothesis is relevant to the principles of biodiversity.

```
# Bray Curtis resemblance matrix
total.db <- vegdist(total, method="bray")

# Heatmap
order.p <- rev(attr(total.db, "Labels"))

levelplot(as.matrix(total.db)[,order.p], aspect="iso", col.regions=inferno,
          xlab="Pond Site", ylab= "Pond Site", scales=list(cex=0.5),
          main= "Bray-Curtis Distance")
```

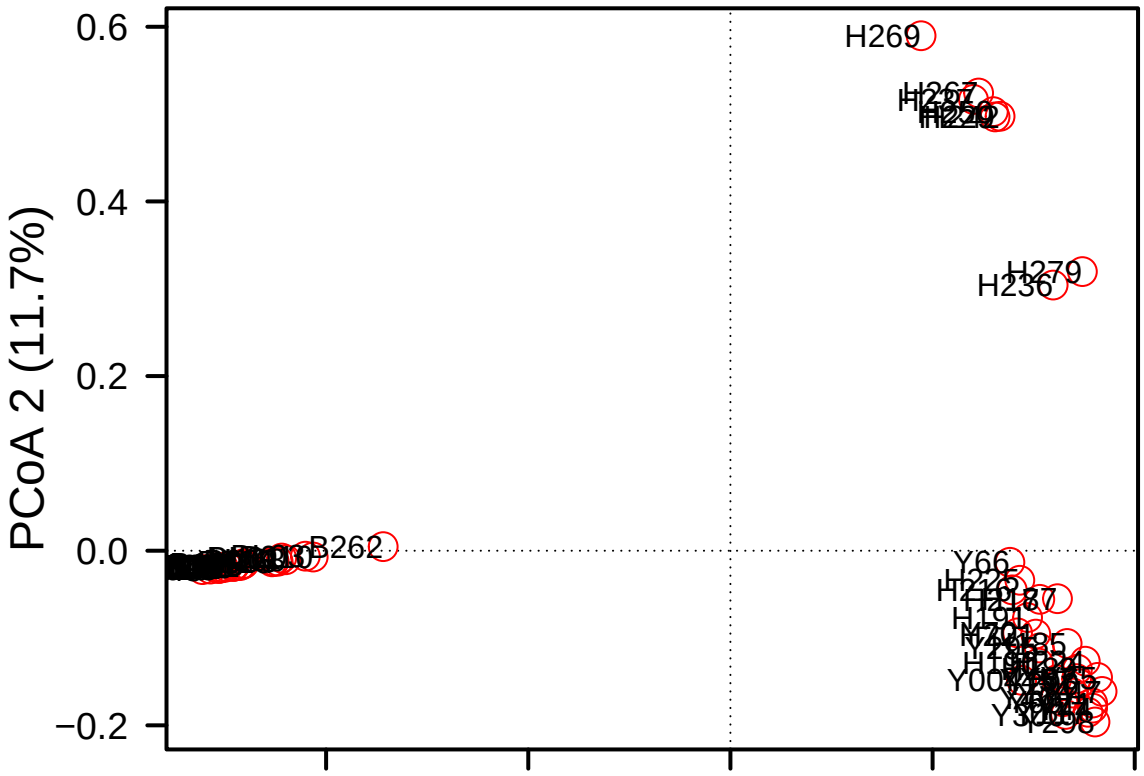



```
# Wards cluster analysis
total.ward <- hclust(total.db, method= "ward.D2")
```

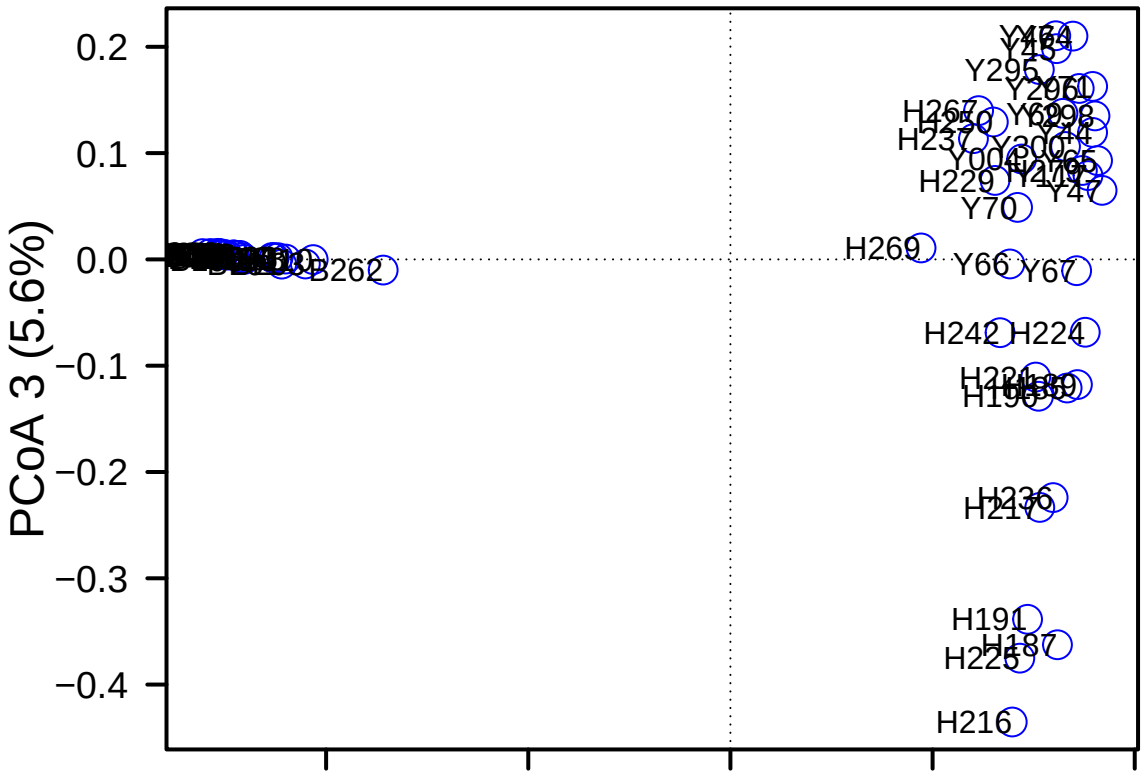
```
par(mar = c(1, 5, 2, 2) + 0.1)
plot(total.ward, main="Indiana Pond Bacteria: Ward's Clustering", ylab= "Squared Bray-Curtis Distance")
```



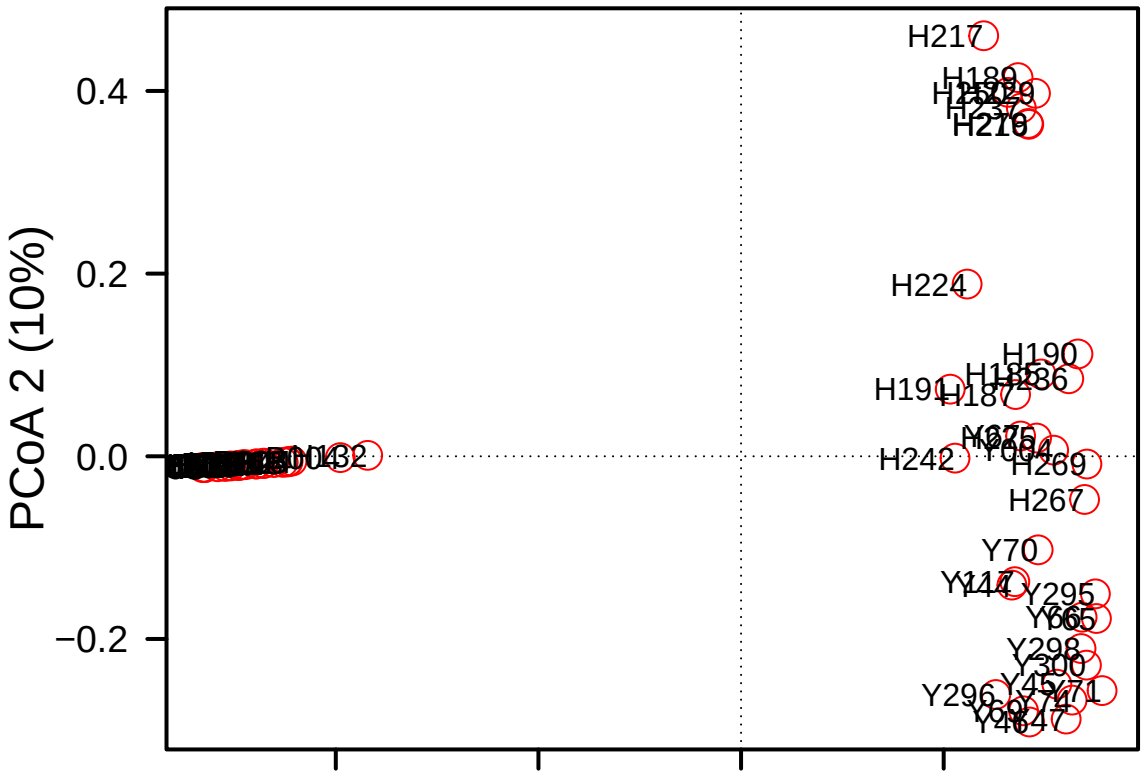

```
# Principal Component Analysis - TOTAL
total.pcoa <- cmdscale(total.db, eig=TRUE, k=3)

exvar1 <- round(total.pcoa$eig[1] / sum(total.pcoa$eig), 3) * 100
exvar2 <- round(total.pcoa$eig[2] / sum(total.pcoa$eig), 3) * 100
exvar3 <- round(total.pcoa$eig[3] / sum(total.pcoa$eig), 3) * 100
total.sum.eig <- sum(exvar1, exvar2, exvar3)

# PCoA Plot PC1 x PC2
plot(total.pcoa$points[ ,1], total.pcoa$points[ ,2], #ylim = c(-0.2, 0.7),
     xlab= paste("PCoA 1 (", exvar1, "%)", sep = ""),
     ylab= paste("PCoA 2 (", exvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5,
     cex.axis=1.2, axes=FALSE);
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);
points(total.pcoa$points[ ,1], total.pcoa$points[ ,2],
       pch = 1, cex = 2, bg = "red", col = "red");
text(total.pcoa$points[ ,1], total.pcoa$points[ ,2],
     labels = row.names(total.pcoa$points), adj=1)
```




```
# PCoA Plot PC1 x PC 3
plot(total.pcoa$points[ ,1], total.pcoa$points[ ,3], #ylim = c(-0.2, 0.7),
     xlab= paste("PCoA 1 (", exvar1, "%)", sep = ""),
     ylab= paste("PCoA 3 (", exvar3, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5,
     cex.axis=1.2, axes=FALSE);
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);
points(total.pcoa$points[ ,1], total.pcoa$points[ ,3],
       pch = 1, cex = 2, bg = "blue", col = "blue");
text(total.pcoa$points[ ,1], total.pcoa$points[ ,3],
     labels = row.names(total.pcoa$points), adj=1)
```




```
## would be great to get the points colored by state park

# Principal Component Analysis - ACTIVE
# bray curtis
active.db <- vegdist(active, method="bray")
active.pcoa <- cmdscale(active.db, eig=TRUE, k=3)

a.exvar1 <- round(active.pcoa$eig[1] / sum(active.pcoa$eig), 3) * 100
a.exvar2 <- round(active.pcoa$eig[2] / sum(active.pcoa$eig), 3) * 100
a.exvar3 <- round(active.pcoa$eig[3] / sum(active.pcoa$eig), 3) * 100
active.sum.eig <- sum(a.exvar1, a.exvar2, a.exvar3)

# PCoA Plot PC1 x PC2
plot(active.pcoa$points[ ,1], active.pcoa$points[ ,2], #ylim = c(-0.2, 0.7),
     xlab= paste("PCoA 1 (", a.exvar1, "%)", sep = ""),
     ylab= paste("PCoA 2 (", a.exvar2, "%)", sep = ""),
     pch = 16, cex = 2.0, type = "n", cex.lab = 1.5,
     cex.axis=1.2, axes=FALSE);
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);
points(active.pcoa$points[ ,1], active.pcoa$points[ ,2],
       pch = 1, cex = 2, bg = "red", col = "red");
text(active.pcoa$points[ ,1], active.pcoa$points[ ,2],
     labels = row.names(active.pcoa$points), adj=1)
```




```
## sorry that the plot axes don't show up on the knitted PDF, I don't know why because they show up in
```

Synthesis Answer: The sites show large variation overall, notably with two main clusters as shown by the heatmap and Ward's clustering. There were three grouped site locations (Brown County State Park, Hoosier National Forest, and Yellowwood State Forest), and sites within each area generally clustered closer together than with sites from another area, meaning sites within an area generally were more similar in diversity than they were similar sites from a different area. B sites and Y sites show this trend the most strongly, while H sites actually showed split clusters, with some sites clustering with the B cluster and others with the Y cluster. The 1st principal component of the PCoA on the TOTAL samples captures 43% of the diversity among sites, and plotting it shows that it split the B cluster away from the Y cluster. Plotting the 2nd and 3rd PCs did not separate the tight B cluster but did capture variation among the Y and H sites, 11% and 5% of the overall variation, respectively. An initial hypothesis that we were considering was that dormancy (including total DNA versus only active RNA) may be an interesting driver of Beta diversity patterns in this pond dataset; however, comparison of these initial visualizations (first two PCoA plots of TOTAL vs. third plot of ACTIVE) indicate that the patterns do not differ much at all when dormancy is factored in. I will start considering additional hypotheses regarding environmental variables that might explain the stark difference in diversity of B from Y and most of H (shown in PC 1), and others that could explain the additional variation among the H and Y sites (shown in PC 2 and 3).