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

Lauren Albert; Z620: Quantitative Biodiversity, Indiana University

01 February, 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).

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

```
## [1] "/Users/laurenalbert/GitHub/QB2023_Albert/2.Worksheets/6.BetaDiversity"
```

```r
package.list <-c('vegan','ade4','viridis','gplots','BiodiversityR','indicspecies')
options(repos = list(CRAN="http://cran.rstudio.com/"))
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

## BiodiversityR 2.15-1: Use command BiodiversityRGUI() to launch the Graphical User Interface;
## to see changes use BiodiversityRGUI(changeLog=TRUE, backward.compatibility.messages=TRUE)
```

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.

```r
# note, please do not print the dataset when submitting
data(doubs)
#objects(doubs)
#str(doubs)
#str(doubs, max.level = 1)
```

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 objects
> ***Answer 1b***: 27 fish 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?

> ***Answer 2a***: Fish richness tends to increase moving downstream in the Doubs River, specifically, richness increases at coordinate (200,200).
> ***Answer 2b***: Brown trouch abundance decreases moving downstream in the Doubs River. Adundance of Brown Trout is greatest at x-coordinates 100 to 200. ***Answer 2c***: These patterns suggest that richness could overestimate the abundance of certain species at certain sites. By providing an estimate of overall species richness, the specific abundance of different species is missed, in this case the representation of Brown Trout is missed in the richness measure.

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.

```r
beta.w <- function(site.by.species = ""){
  SbyS.pa <- decostand(site.by.species, method = "pa")
  S <- ncol(SbyS.pa[,which(colSums(SbyS.pa)>0)])
  a.bar <- mean(specnumber(SbyS.pa))
  b.w <- round(S/a.bar, 3)
  return(b.w)
}

beta.w <- function(site.by.species = "", sitenum1 = "", sitenum2 = "",
                   pairwise = FALSE){
  if (pairwise == TRUE){
    if (sitenum1 == "" | sitenum ==""){
      print("Error : please specify site to compare")
      return(NA)}
    site1 = site.by.species[sitenum1,]
    site2 = site.by.species[sitenum2,]
    site1 = subset(site1, select = site1 >0)
    site2 = subset(site2, select = site2 >0)
    gamma = union(colnames(site1), colnames(site2))
    s = length(gamma)
    a.bar = mean(c(specnumber(site1), specnumber(site2)))
    b.w = round(s/a.bar - 1, 3)
    return(b.w)
  }
  else{
    SbyS.pa <- decostand(site.by.species, method = "pa")
    S <- ncol(SbyS.pa[,which(colSums(SbyS.pa) > 0)])
    a.bar <- mean(specnumber(SbyS.pa))
    b.w <- round(S/a.bar, 3)
    return(b.w)
  }
}

beta.w(doubs$fish)
```

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

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***: ***Answer 3b***:
> ***Answer 3c***: An additive definition of beta diversity would instead measure how many more species exist in a regional pool than in local sites, whereas a multiplicative approach demonstrates how many times more diverse the region is than local sites.

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 metrics place more weight on shared species, whereas abundance-based metrics are more influenced by more abundant species.

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, ]

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 resemblance matrix created using the Bray-Curtis Distance represents dissimilarity. In the resemblance matrix, values range from 0 to 1, where 1 indicates two sites have complete dissimilarity. Because there are values in the matrix that are at 1, the dissimilarity is calculated by 1 - similarity. ***Answer 5b***: Using Sorenson with this data ignores the fish abundance and instead creates a matrix of similarity based on presense or absense. In the matrix, the values are slightly different then where sites that are dissimilar based on abundance may be more or less dissimilar in the incidence of the fish species.

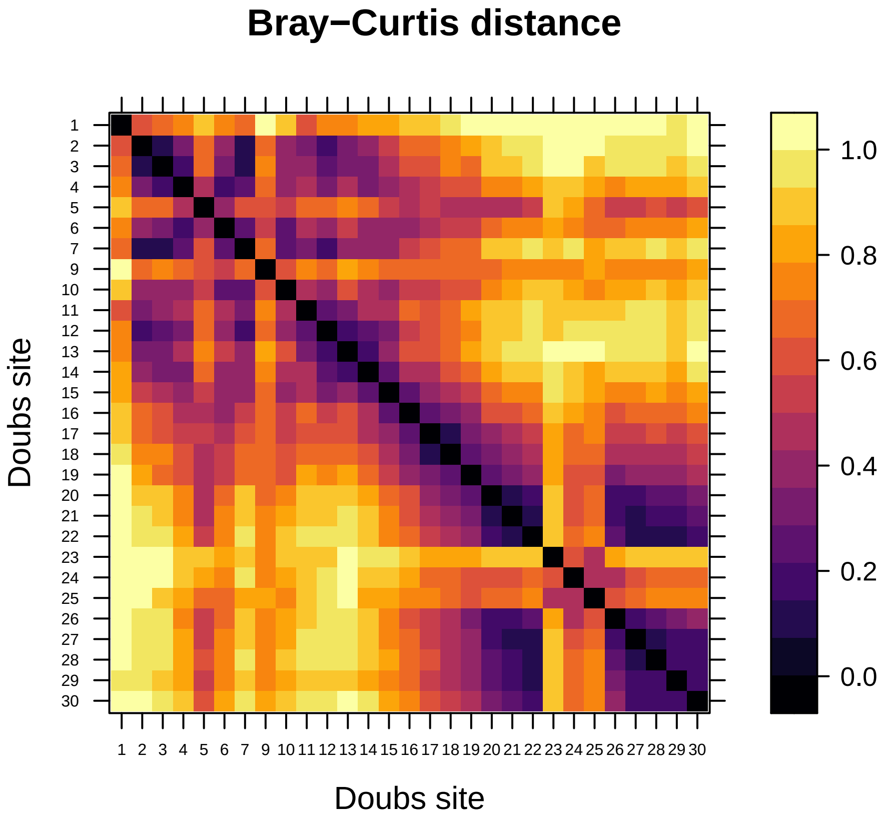
4) VISUALIZING BETA-DIVERSITY

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
```

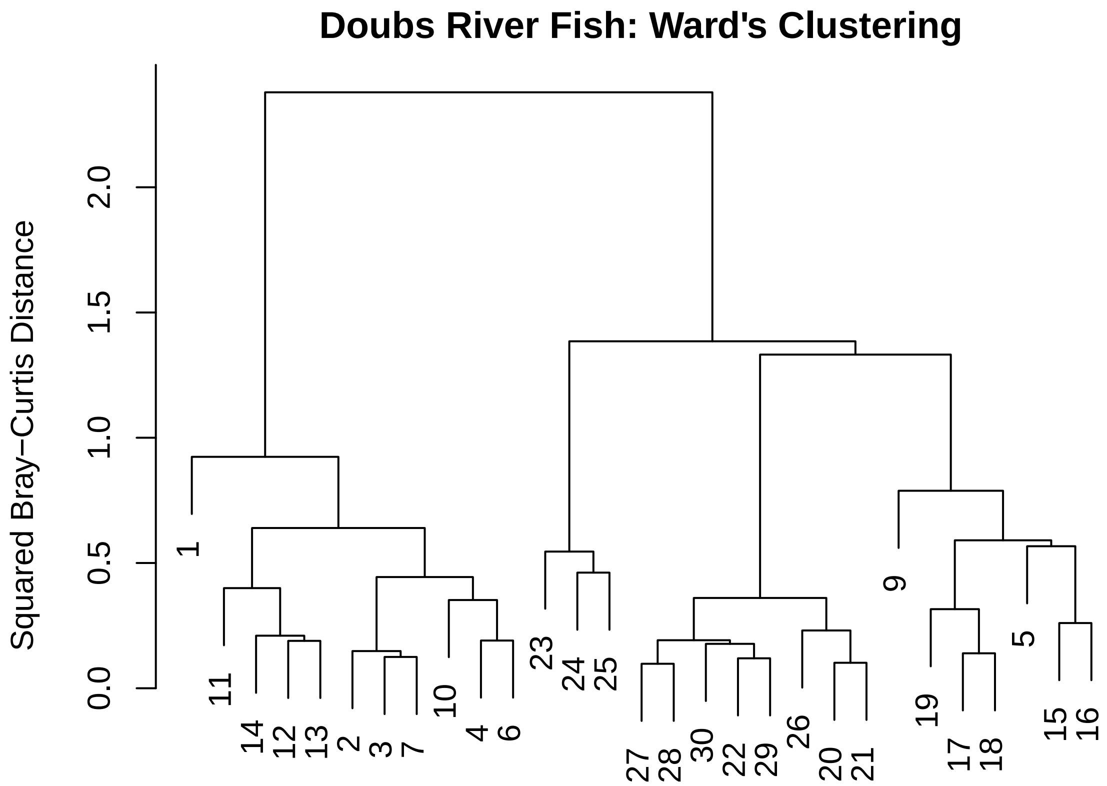


B. Cluster Analysis

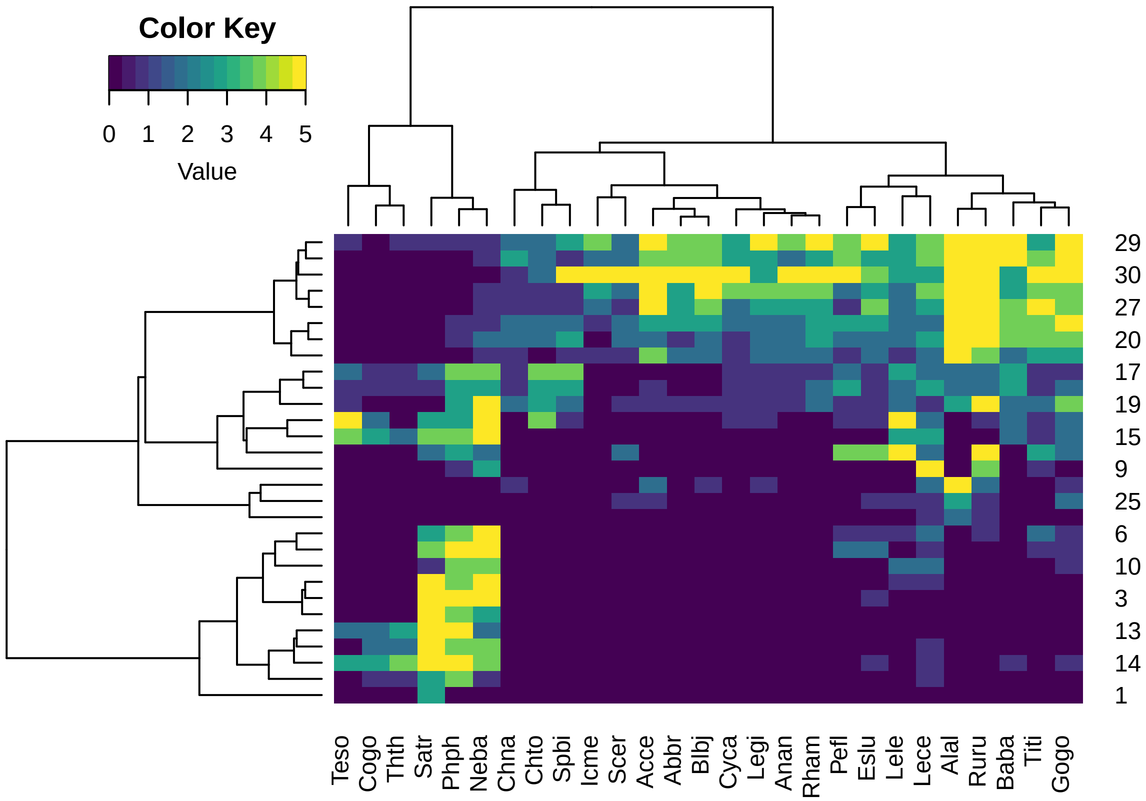
In the R code chunk below, do the following:

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

```r
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")
```



```r
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***: The cluster analysis and introductory plots suggest most sites in the Doubs River dataset are similiar and a few sites and species contribute to the dissimilarity.

C. Ordination

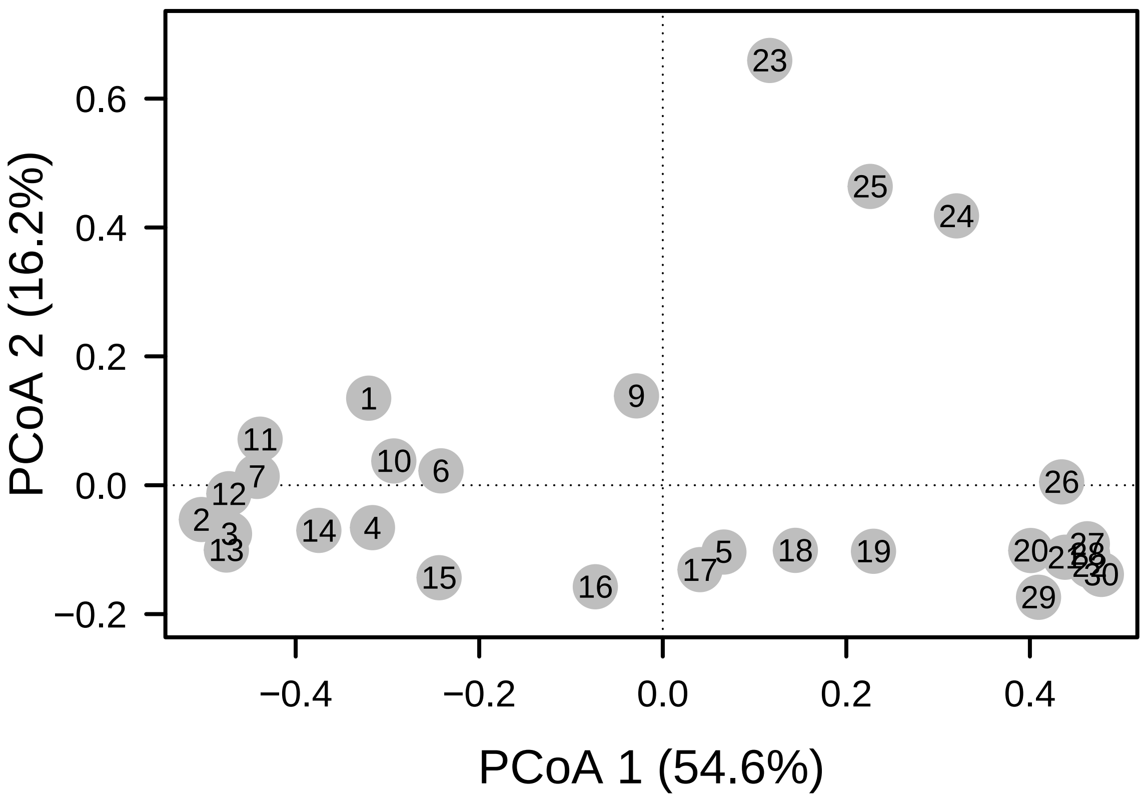
Principal Coordinates Analysis (PCoA)

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)
```

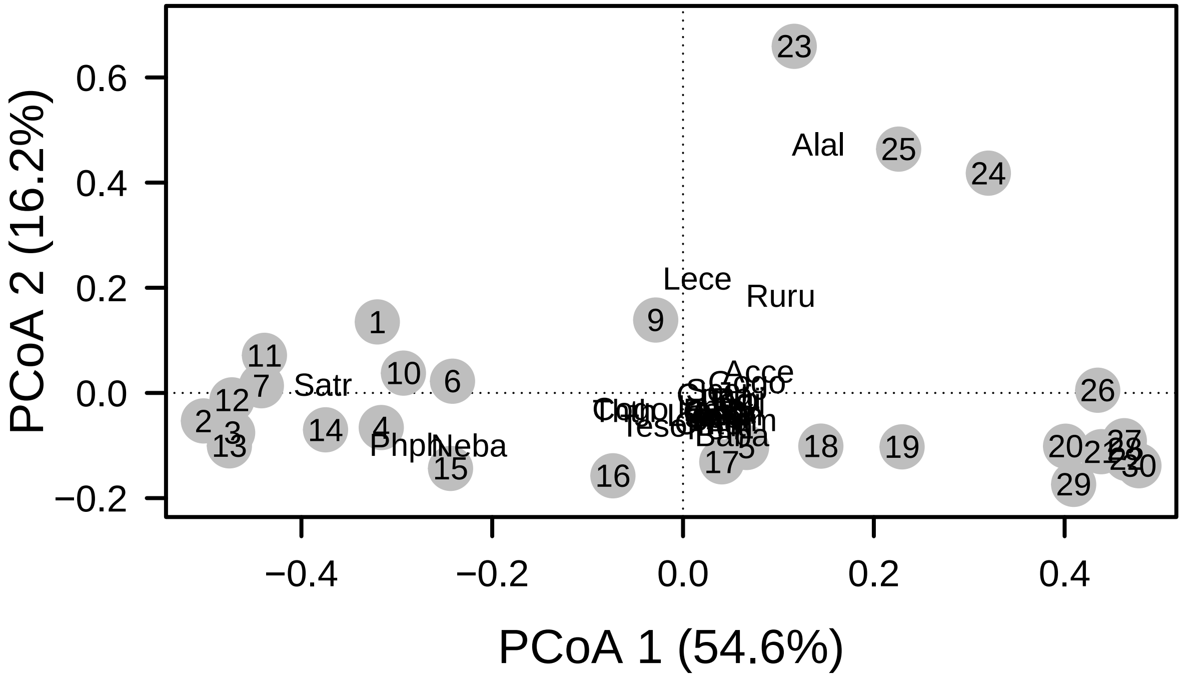
```r
par(mar = c(5,5,1,2) + 0.1)

plot(fish.pcoa$point[ ,1], fish.pcoa$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.

```r
plot.new()
fishREL <- fish
for(i in 1:nrow(fish)){
  fishREL[i, ] = fish[i, ] / sum(fish[i, ])
}

`add.spec.scores.class` <-
  function(ordi,comm,method="cor.scores",multi=1,Rscale=F,scaling="1") {
    ordiscores <- scores(ordi,display="sites")
    n <- ncol(comm)
    p <- ncol(ordiscores)
    specscores <- array(NA,dim=c(n,p))
    rownames(specscores) <- colnames(comm)
    colnames(specscores) <- colnames(ordiscores)
    if (method == "cor.scores") {
      for (i in 1:n) {
        for (j in 1:p) {specscores[i,j] <- cor(comm[,i],ordiscores[,j],method="pearson")}
      }
    }
    if (method == "wa.scores") {specscores <- wascores(ordiscores,comm)}
    if (method == "pcoa.scores") {
      rownames(ordiscores) <- rownames(comm)
      eigenv <- ordi$eig
      accounted <- sum(eigenv)
      tot <- 2*(accounted/ordi$GOF[2])-(accounted/ordi$GOF[1])
      eigen.var <- eigenv/(nrow(comm)-1)
      neg <- length(eigenv[eigenv<0])
      pos <- length(eigenv[eigenv>0])
      tot <- tot/(nrow(comm)-1)
      eigen.percen <- 100*eigen.var/tot
      eigen.cumpercen <- cumsum(eigen.percen)
      constant <- ((nrow(comm)-1)*tot)^0.25
      ordiscores <- ordiscores * (nrow(comm)-1)^-0.5 * tot^-0.5 * constant
      p1 <- min(p, pos)
      for (i in 1:n) {
        for (j in 1:p1) {
          specscores[i,j] <- cor(comm[,i],ordiscores[,j])*sd(comm[,i])/sd(ordiscores[,j])
          if(is.na(specscores[i,j])) {specscores[i,j]<-0}
        }
      }
      if (Rscale==T && scaling=="2") {
        percen <- eigen.var/tot
        percen <- percen^0.5
        ordiscores <- sweep(ordiscores,2,percen,"/")
        specscores <- sweep(specscores,2,percen,"*")
      }
      if (Rscale==F) {
        specscores <- specscores / constant
        ordiscores <- ordi$points
      }
      ordi$points <- ordiscores
      ordi$eig <- eigen.var
      ordi$eig.percen <- eigen.percen
      ordi$eig.cumpercen <- eigen.cumpercen
      ordi$eigen.total <- tot
      ordi$R.constant <- constant
      ordi$Rscale <- Rscale
      ordi$scaling <- scaling
    }
    specscores <- specscores * multi
    ordi$cproj <- specscores
    return(ordi)
  }

plot(fish.pcoa$point[ ,1], fish.pcoa$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$points))
fish.pcoa <- add.spec.scores.class(fish.pcoa, fishREL, method = "pcoa.scores")
text(fish.pcoa$cproj[ ,1], fish.pcoa$cproj[ ,2],
     labels = row.names(fish.pcoa$cproj), col = "black")
```



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

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 principal coordinate on the x-axis explains 54.6% of variation while the PC on the y-axis explain 16.2%. The grouping of sites in the Doubs River fall within two regions of the ordination plot. ***Answer 7b***: The species that are aggregated in the center of the ordination plot may suggest that those sites have greater river quality.

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.

```r
getwd()
```

```
## [1] "/Users/laurenalbert/GitHub/QB2023_Albert/2.Worksheets/6.BetaDiversity"
```

```
require(plyr)
```

```
## Loading required package: plyr
```

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

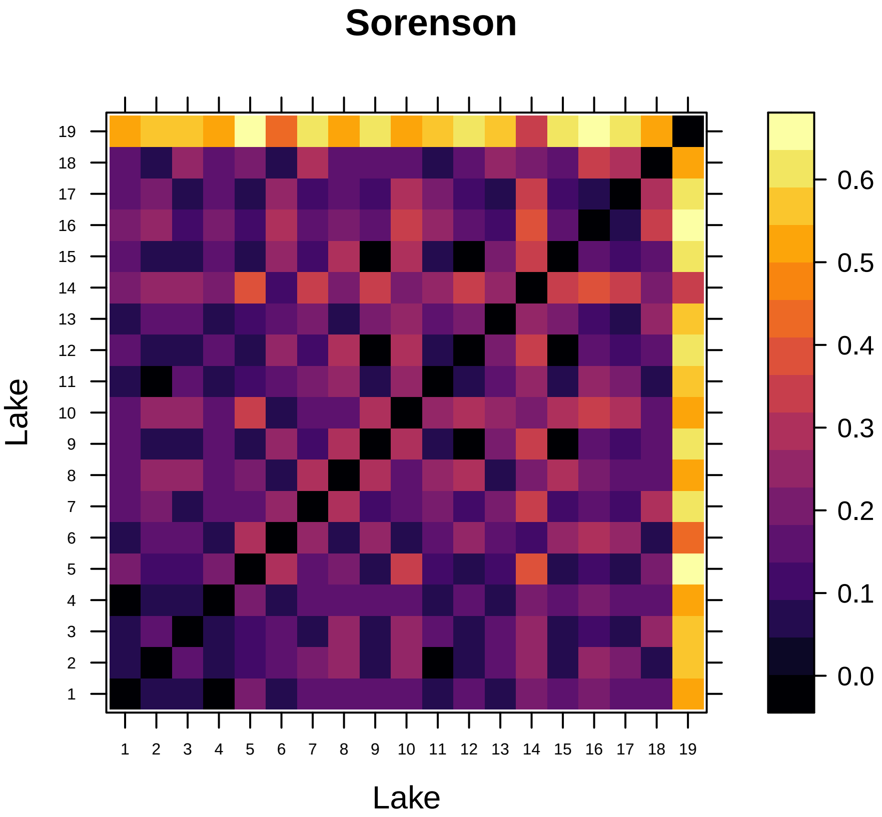
```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::arrange()   masks plyr::arrange()
## x purrr::compact()   masks plyr::compact()
## x dplyr::count()     masks plyr::count()
## x dplyr::failwith()  masks plyr::failwith()
## x dplyr::filter()    masks stats::filter()
## x dplyr::id()        masks plyr::id()
## x dplyr::lag()       masks stats::lag()
## x dplyr::mutate()    masks plyr::mutate()
## x dplyr::rename()    masks plyr::rename()
## x dplyr::summarise() masks plyr::summarise()
## x dplyr::summarize() masks plyr::summarize()
```

```
library(readr)
zoopfield_2009 <- read_csv("zoopfield_2009.csv")
```

```
## New names:
## Rows: 282 Columns: 88
## -- Column specification
## -------------------------------------------------------- Delimiter: "," chr
## (2): Lake_Name, date dbl (55): Round, JD, M_m_l, M_s_l, M_n_l, M_m_e, M_s_e,
## M_n_e, U_m_l, U_s_l,... num (11): dent, cerio, pulic, Bosmina, parv, ambig,
## diaph, scaph, alona, pre... lgl (20): M_m_l2, M_s_l2, U_m_l2, U_s_l2, O_m_l2,
## O_s_l2, E_m_l2, E_s_l2, os...
## i Use ‘spec()‘ to retrieve the full column specification for this data. i
## Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
## * ‘P_ug‘ -> ‘P_ug...66‘
## * ‘P_ug‘ -> ‘P_ug...67‘
```

```
#looking at structure imported data -- only looking at data collected in 2009
#str(zoopfield_2009)

#subset into site-by-species matrix (site = lake name)
zoopSbyS <- zoopfield_2009[, c(1, 37:49)]
#str(zoopSbyS)

#have to change lake names to numbered sites
zoopSbyS$Lake_Name <- revalue(zoopSbyS$Lake_Name, c("Airline" ="1", "Beaver Dam" ="2", "Beaver dam" ="2
## The following ‘from‘ values were not present in ‘x‘: Beaver dam, T-lake, university, Willow, Wampler
```

```
zoopSbyS$Lake_Name<- as.numeric(as.character(zoopSbyS$Lake_Name))
#str(zoopSbyS)

#aggregating sampling rounds at each lake
zoopagg <- zoopSbyS %>%
  group_by(Lake_Name)%>%
  summarize_all(sum, na.rm = TRUE)

zoopagg$Lake_Name<-as.numeric(as.character(zoopagg$Lake_Name))
#resemblance matrix of Sorenson and Bray-Curtis
zoop.ds <- vegdist(zoopagg, method = "bray", binary = TRUE, na.rm = TRUE)
zoop.db <- vegdist(zoopagg, method = "bray", na.rm = TRUE) #species abundance paradox

#heat-map
order <- rev(attr(zoop.ds, "Labels"))
levelplot(as.matrix(zoop.ds), col.regions = inferno, xlab = "Lake", ylab = "Lake", scales = list(cex = 
```



> Though not super descriptive, a look at beta-diversity for the year 2009 shows that most lakes are similar in the incidence-based metric. I would be interested to look more closely at site 19 where there seems to be most dissimilarity to other sites. I would like to include an environment by species matrix to consider environmental characteristics. Importantly, the data are density measures of zooplankton species, so I’d like to further look at abundance-based metrics once I better understand how to work with zero values in the data set.