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

Jocelyn Huang; 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 5$^{\text{th}}$, 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()
#setwd("/cloud/project/QB2025_Huang/Week3-Beta)
```

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
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
```

```
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:
```

```
#number of fish species
#nrow(doubs$fish)
```

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***: There are 4 objects. ***Answer 1b***: There are 27 fish species. ***Answer 1c***: There are 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 is relatively higher at the curve (x-coordinate from around 200 to 250 km, y corrdinate around 200 km) and around the downstream, while it's relatively lower at upstream and until it changes direction.

Answer 2b:
The Brown trout is more abundant before the stream changes direction (around (x,y) = (200~225, 150~200)) and relatively abundant around the upstream area. Brown trout abundance is low downstream where it's almost zero.

Answer 2c:
If we are only using the richness, we cannot get the information for how abundant each species is. It only takes into accoount the total number of species that occurred. For example, Brown trout is not abunddant around downstream, but the total species richness downstream is very high.

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.tem <- 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) {
  if (pairwise == TRUE){

    # As a check, 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)
  }

  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(doubs$fish) #For 3a
```

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

```
beta.w(doubs$fish, "1", "2", TRUE)
```

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

```
beta.w(doubs$fish, "1", "10", 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:
The regional fish diversity γ in the Doubs is 2.16 times more diverse than the local sites on average.

Answer 3b:
Since the value of $\beta_w - 1$ ranges from 0 to 1, with 0 being the minimum β -diversity (i.e. the two sites have the highest similarlities in terms of species assemblage) and 1 the opposite, site 1 and 2 ($\beta_w - 1 = 0.5$) are more similar than site 1 and 10 ($\beta_w - 1 = 0.714$).

Answer 3c:
If we define β additively as $\beta = \gamma - \alpha$, then that means the regional diversity γ is going to be how many more species are there in the region comparing to that in local sites on avergae.

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 metrices like Jaccard and Sorensen emphasize the importance of rare species because b and c (both are variable in the equation) represent the unique (rare) species in each site, while abundance-based ones do not emphasize rare 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
rowSums(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 
##  3 12 16 21 34 21 16  0 14 14 11 18 19 28 33 40 44 42 46 56 62 72  4 15 11 43 
## 27 28 29 30 
## 63 70 87 89
```

```
fish <- fish[-8,]

fish.ds <- vegdist(fish, method = "bray", binary = TRUE, upper = TRUE, diag = TRUE) #Sorensen's Similar
fish.db <- vegdist(fish, method = "bray", upper = TRUE, diag = TRUE) #Bray-Curtis Distance
```

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***:
> I think `fish.db` represents dissimilarity, because the diagonals (the same sites) are all 0, which will not make sense if the matrix represents similarities.
>
> ***Answer 5b***: While both matrices represents pairwise dissimilarities between sites, the values in the matrix generated by Sorensen's (binary) are generally smaller than that generated by Bray-Curtis, which means Sorensen's matrix provides values that shows more pairwise similarities between sites. If using Sorensen's, we might neglect many of the variations between sites, resulting in a lower beta diversity, potentially.

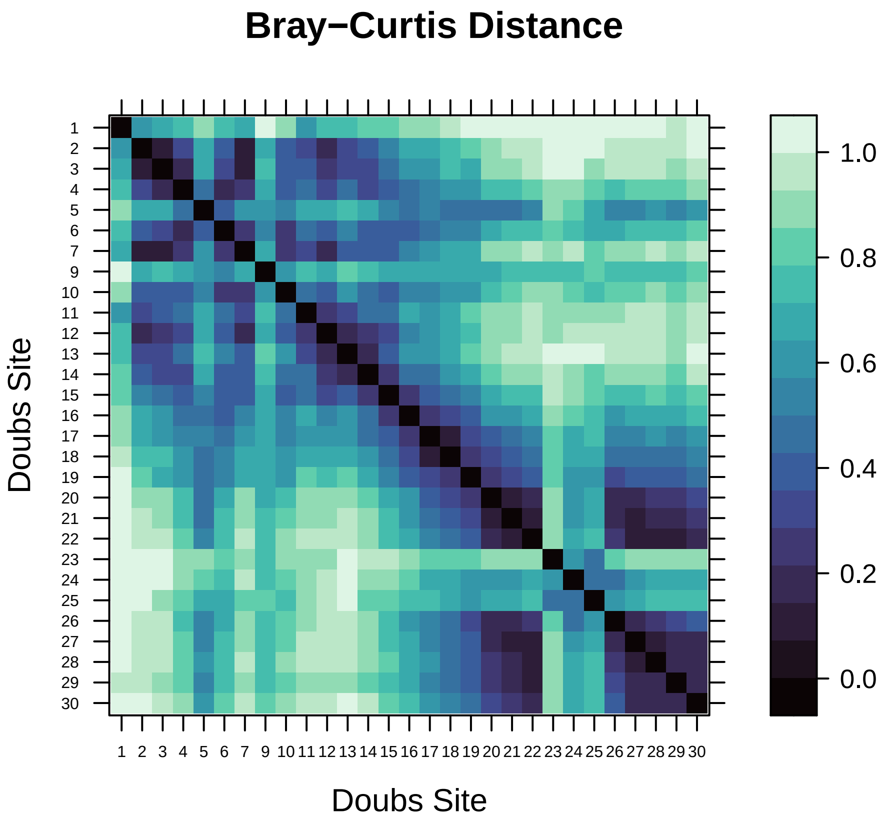
4) VISUALIZING BETA-DIVERSITY

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.

```
library(viridis)
order <- rev(attr(fish.db, "Labels")) #Define Order of Sites
levelplot(as.matrix(fish.db)[, order], aspect = "iso", col.regions = mako,
          xlab = "Doubs Site", ylab = "Doubs Site", scales = list(cex = 0.5),
          main = "Bray-Curtis Distance")
```

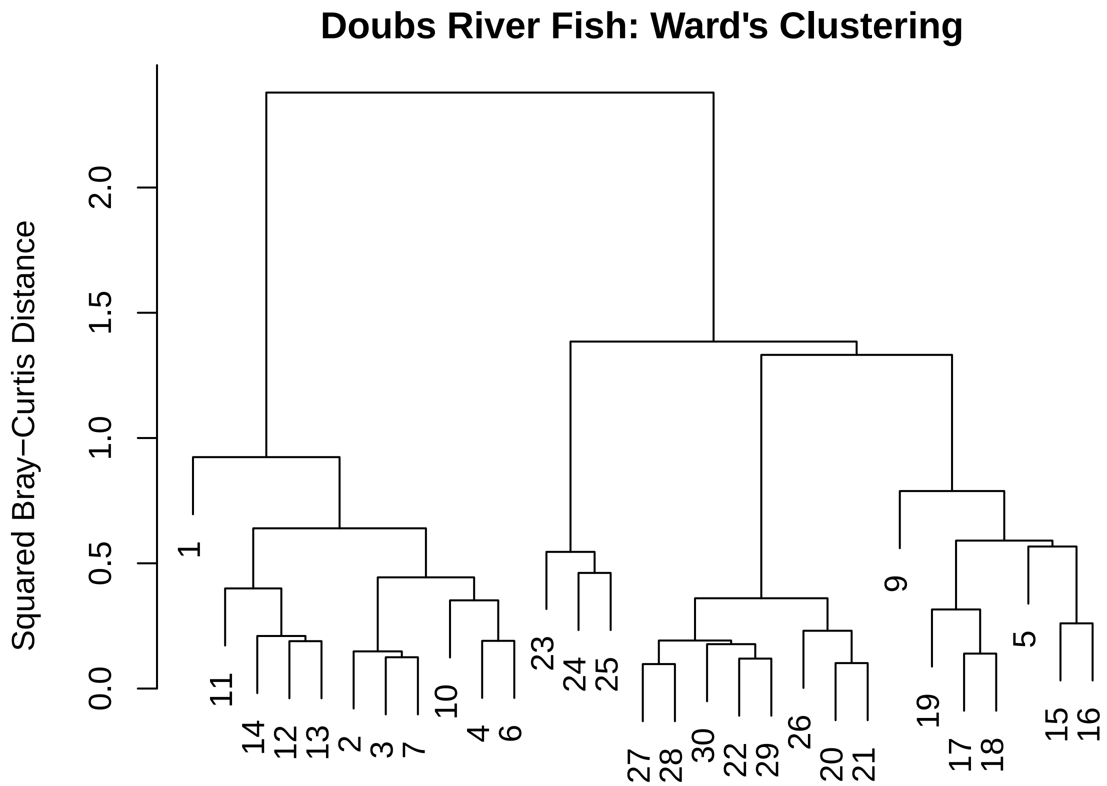


B. Cluster Analysis

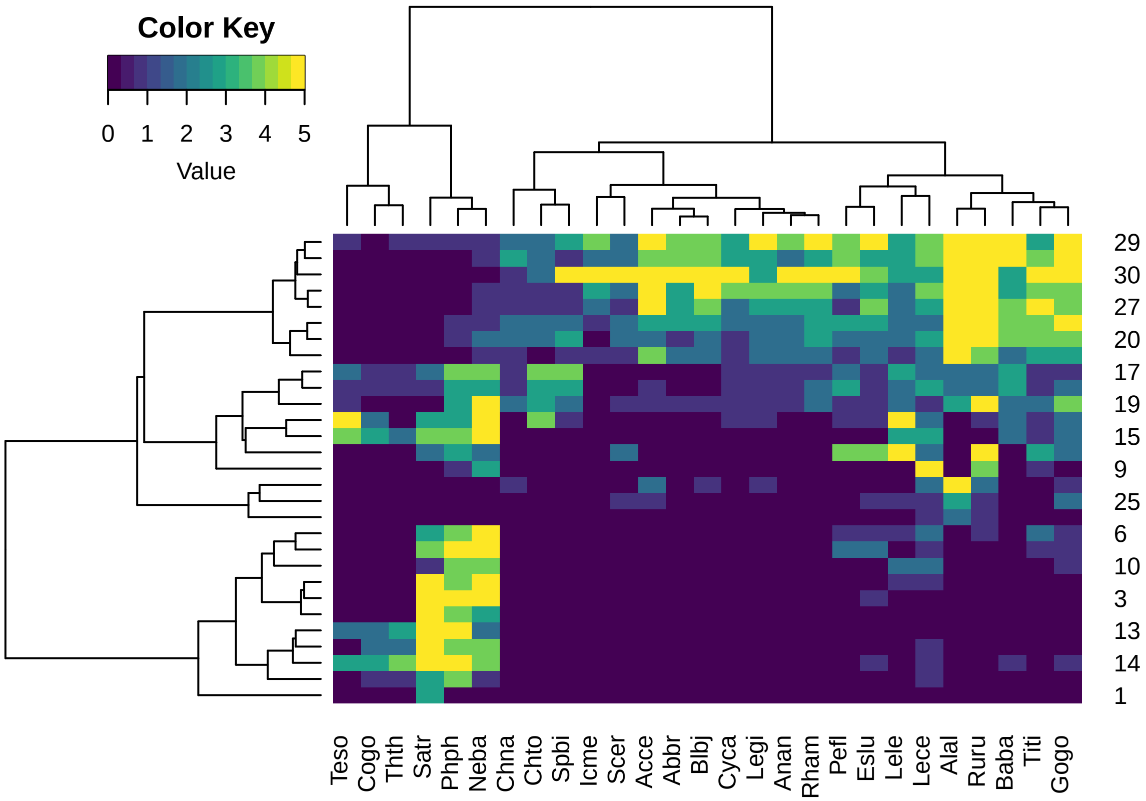
In the R code chunk below, do the following:

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")

#plot with `hclust`:
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
#plot with `heatmap.2`:
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***: High deposite of resources at the downstream region (sites with higher number) contributes to higher diversity at these sites, and the fish species that are relatively more abundant at these sites are different from the ones dominated upstream sites.

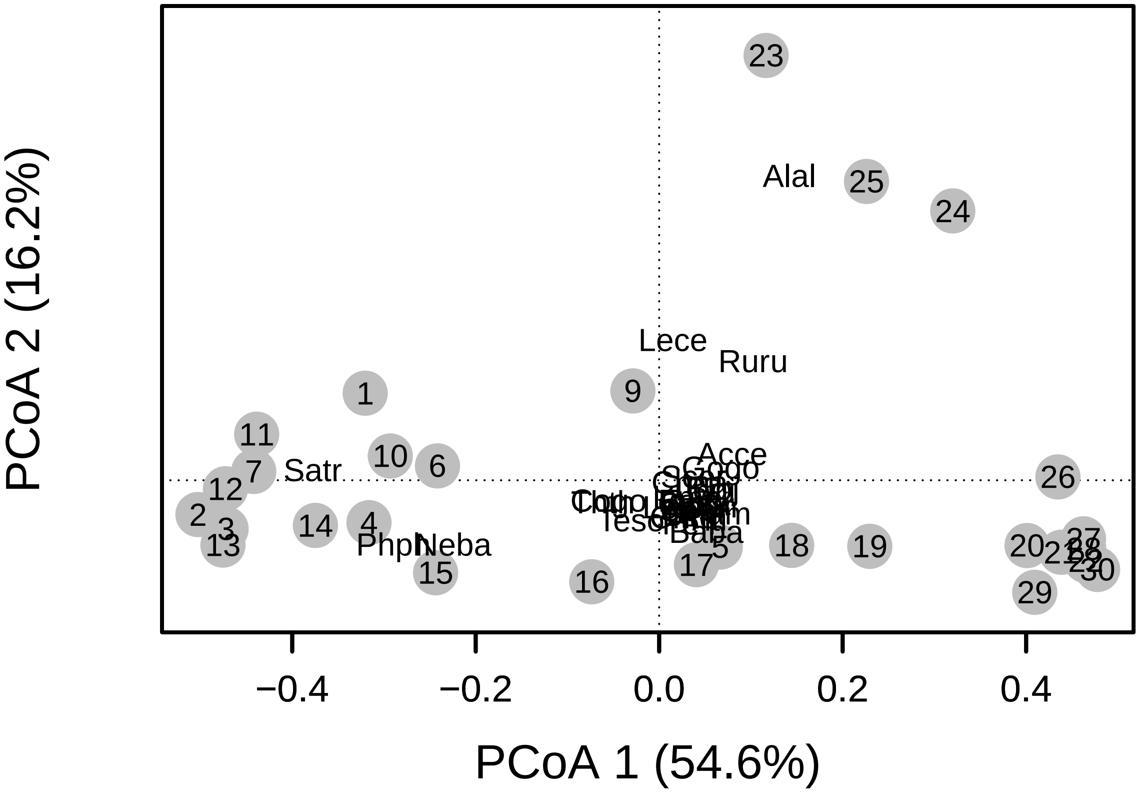
C. Ordination

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.

```r
fish.pcoa <- cmdscale(fish.db, eig = TRUE, k = 3) #performed a PCoA

# Variation explained by the first three axes:
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)

# Identify the influential species
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)
  }

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

# Plot the PCoA ordination:
par(mar = c(5, 5, 1, 2) + 0.1)
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 = 1, 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")   #add species scores
```



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.

```
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, ]
fit <- envfit(fish.pcoa, fishREL, perm = 999)
```

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

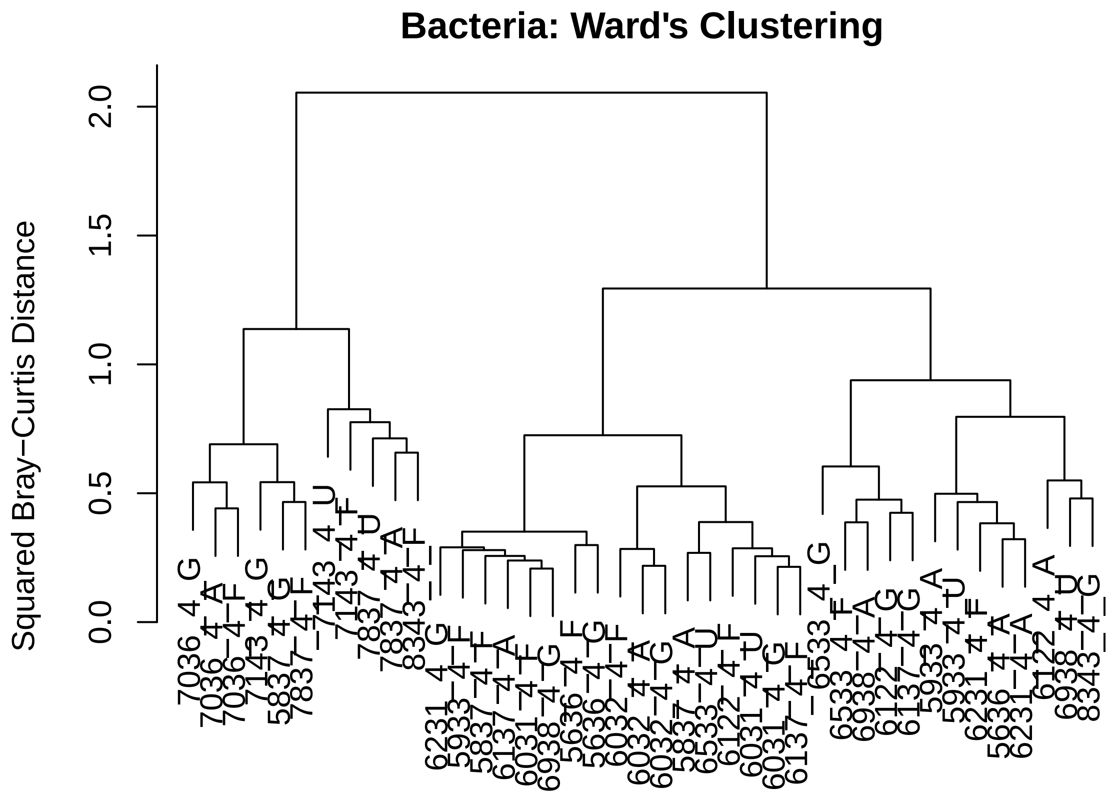
a. Describe the grouping of sites in the Doubs River based on fish community composition.
b. Generate a hypothesis about which fish species are potential indicators of river quality.

> ***Answer 7a***:
> From the PCoA ordination plot, sites that have similar fish assemblage composition are close to each other on the plot, and can be roughly grouped into four groups. Sites 23, 24, and 25 are one group that is contributed to the relatively high abundance in the Common bleaks. Abundance of the Brown Trouts, the Minnows, and the Stone Loaches contribute to the similarities in sites that are distributing at the lower left corner of the plot. Sites 5, 9, and 16-19 are grouped, and the influential fish species whose abundance formed such similarity is very diverse, which means these sites are highly diverse and are tending to not be dominated by a single species.
>
> ***Answer 7b***:
> Minnows, brown trouts, Bitterling, Pumpkinseed, Carp, Freshwater bream, Ruffe, Silver bream, Bleak, Eel are considered important species since they have relatively stronger correlation. The abundance of these species contribute relatively more to the similarities between sites, comparing to the other species, which may make these species suitable indicators of environmental resources in different sites.

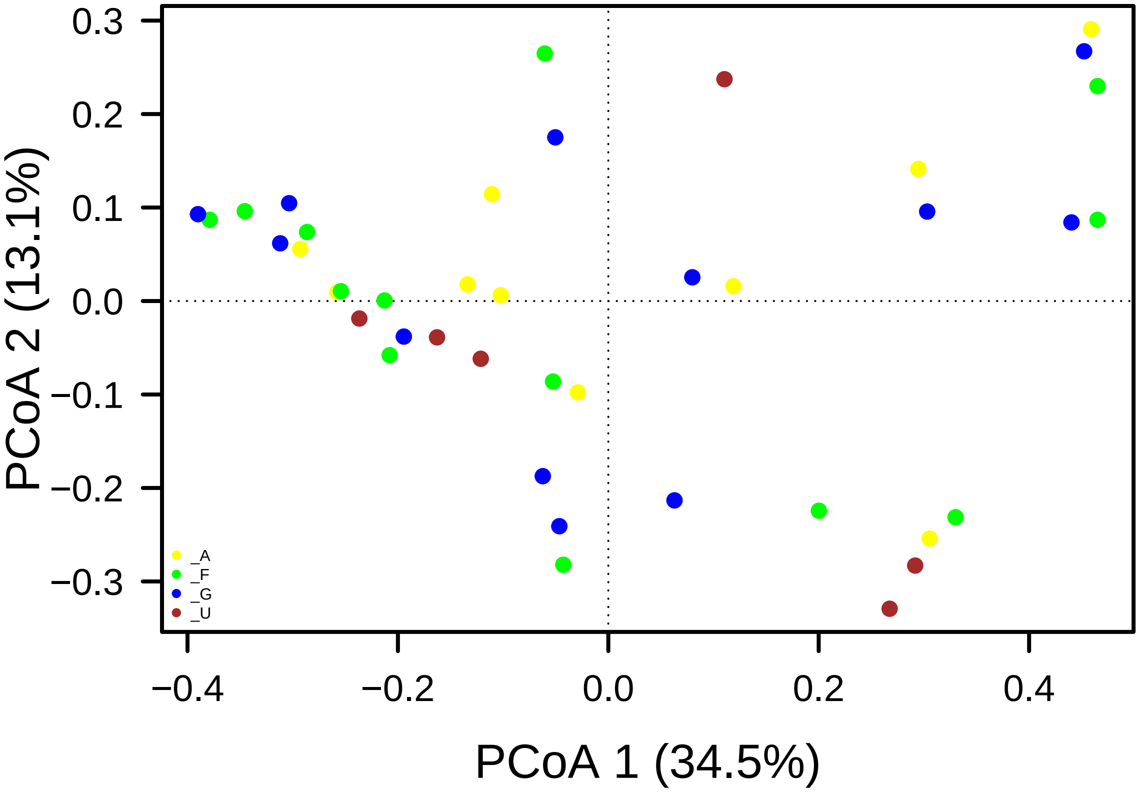
SYNTHESIS

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.

```
load("/cloud/project/QB2025_Huang/Group-project/longdataBac_objects2_datadryad.rda")
Bacteria <- longdataBac_datadryad #rename
rm(longdataBac_datadryad)
bac_by_site <- with(Bacteria, tapply(Counts, list(PlotID, Sender), sum, default = 0))
#write.table(bac_by_site, file = "bacteria_div.txt", sep = "\t", row.names = TRUE, col.names = NA, quot

#drop some:
bac_fi <- bac_by_site[!grepl("_3|_2|_1", rownames(bac_by_site)), ]

# Make a resemblance matrix based on Bray-Curtis
bac.db <- vegdist(bac_fi, method = "bray", upper = TRUE, diag = TRUE)

#Cluster analysis
bac.ward <- hclust(bac.db, method = "ward.D2")

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



```
#PCoA
bac.pcoa <- cmdscale(bac.db, eig = TRUE, k = 3) #performed a PCoA

# Variation explained by the first three axes:
explainvar1_b <- round(bac.pcoa$eig[1]/sum(bac.pcoa$eig), 3)*100
explainvar2_b <- round(bac.pcoa$eig[2]/sum(bac.pcoa$eig), 3)*100
explainvar3_b <- round(bac.pcoa$eig[3]/sum(bac.pcoa$eig), 3)*100
sum.eig <- sum(explainvar1_b, explainvar2_b, explainvar3_b)

#identify influential species
#bacREL <- bac_fi
#  for(i in 1:nrow(bac_fi)){
#    bacREL[i, ] = bac_fi[i, ]/sum(bac_fi[i, ])
#  }

#bac.pcoa <- add.spec.scores.class(bac.pcoa, bacREL, method = "pcoa.scores")

# Plot the PCoA ordination:

habitat_colors <- c("_A" = "yellow", "_F" = "green", "_G" = "blue", "_U" = "brown")

row_names_b <- rownames(bac.pcoa$points)

point_colors <- sapply(row_names_b, function(name) {
  match <- grep("_A|_F|_G|_U", name, value = TRUE)
  if (length(match) > 0) {
    return(habitat_colors[substr(match, nchar(match) - 1, nchar(match))])
  }
  else {
    return("black")  # Default color for other rows
  }
})

par(mar = c(5, 5, 1, 2) + 0.1)
plot(bac.pcoa$points[, 1], bac.pcoa$points[, 2],
     xlim = range(bac.pcoa$points[, 1]),
     ylim = range(bac.pcoa$points[, 2]),
     xlab = paste("PCoA 1 (", explainvar1_b, "%)", sep = ""),
     ylab = paste("PCoA 2 (", explainvar2_b, "%)", sep = ""),
     pch = 16, cex = 2.0, col = point_colors,
     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(bac.pcoa$points[, 1], bac.pcoa$points[, 2],
       pch = 19, cex = 1, bg = "gray", col = point_colors)
legend("bottomleft", legend = names(habitat_colors), col = habitat_colors, pch = 19,
       cex = 0.5, pt.cex = 0.5, bty = "n")
```



Answer: The PCoA plot shows that the bacterial community composition in the four habitats (agriculture”_A”; forest”_F”; grassland”_G”; and urban”_U”) are all very diverse. Sites that are on the left side of the panels seem to have an assemblage of bacterial community that are more similar to each other. This might suggests, and we hypothesizes that habitat type does not have an effect on the bacterial diversity.