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

Pat Wall; Z620: Quantitative Biodiversity, Indiana University

27 April, 2021

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 '8.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 (**8.BetaDiversity_1_Worksheet.Rmd**) with all code blocks filled out and questions answered) and the PDF output of `Knitr` (**8.BetaDiversity_1_Worksheet.pdf**).

The completed exercise is due on **Friday, April 16$^{\text{th}}$, 2021 before 09:00 AM**.

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 "*/8.BetaDiversity*" folder, and

4. load the `vegan` R package (be sure to install if needed).

```
rm(list = ls())
setwd('~/Classwork/Current/QB/2.Worksheets/8.BetaDiversity/')
require(vegan)
```

```
## Loading required package: vegan

## Loading required package: permute

## Loading required package: lattice

## This is vegan 2.5-7
```

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.

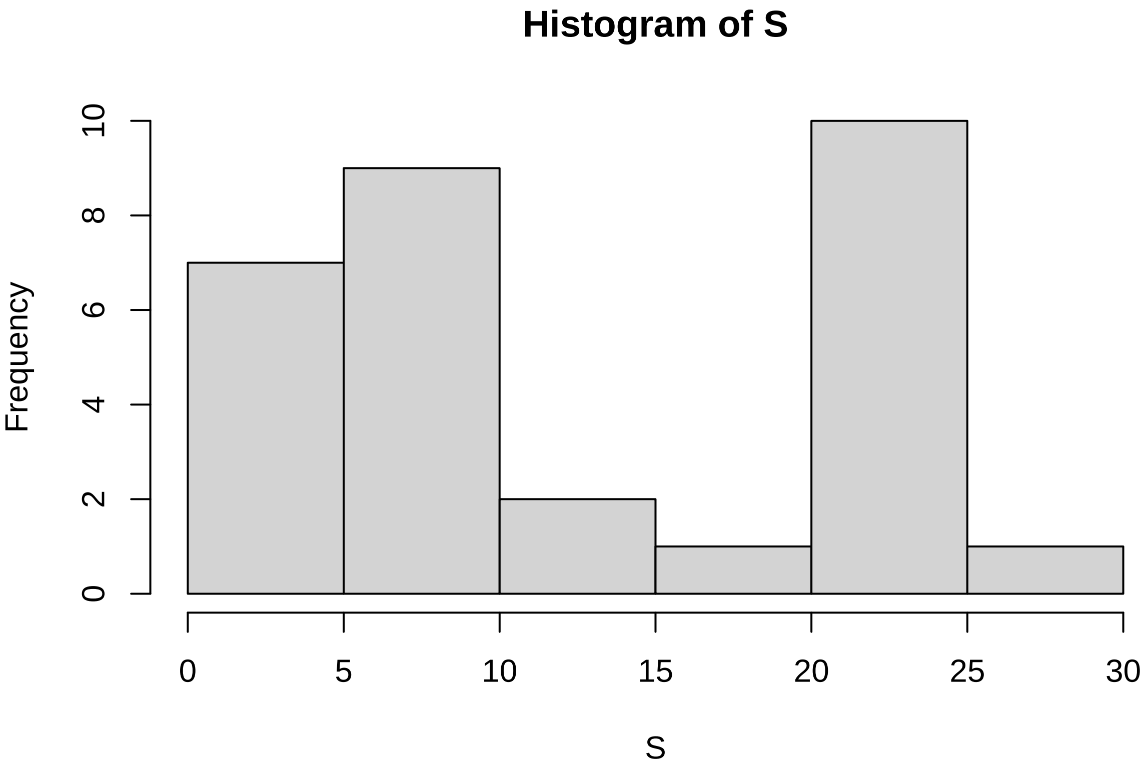
```
# note, please do not print the dataset when submitting
require(ade4)
```

```
## Loading required package: ade4
```

```
data(doubs)
# str(doubs[['fish']])
fish <- doubs[['fish']]
print(fish$Satr)
```

```
##  [1] 3 5 5 4 2 3 5 0 0 1 3 5 5 5 4 3 2 1 0 0 0 0 0 0 0 0 0 0 1 0
```

```
S <- rowSums(fish > 0)
hist(S)
```



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 things in `doubs`. ***Answer 1b***: doubs[["fish"]] is a dataframe with 27 species. ***Answer 1c***: There are 30 sites in the dataframe.

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***: Richness spans almost the entire allowable range of richnesses with a minimum observed richness of 0 and a maximum observed richness of 26. ***Answer 2b***: *Salmo trutta* spans a range of 0 to 5 individuals. in many of them the fish is not observed at all. ***Answer 2c***: Richness from presensce absence data contains much less information that abundance data

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.

```
require(vegan)
beta_w <- function(sbs, site1=FALSE, site2=FALSE, pairwise=FALSE) {
  if (pairwise) {
    if (site1 == FALSE | site2 == FALSE) {
      stop('You must provide sites for pairwise comparison')
    }
    s1 <- subset(sbs[site1,], select = sbs[site1,] > 0)
    s2 <- subset(sbs[site2,], select = sbs[site2,] > 0)
    gamma <- union(colnames(s1), colnames(s2))
    s <-  length(gamma)
    a_bar <-  mean(c(vegan::specnumber(s1), vegan::specnumber(s2)))
    b_w <- round((s / a_bar) - 1, 3)
    return(b_w)
  }
  else { # not pairwise
    sbs_pa <- vegan::decostand(sbs, method = "pa")
    S <- ncol(sbs_pa[,which(colSums(sbs_pa) > 0)])
    a_bar <- mean(vegan::specnumber(sbs_pa))
    b_w <- round(S/a_bar, 3)
    return(b_w)
  }
}
print(paste("Overall beta_w:", beta_w(fish)))
```

```
## [1] "Overall beta_w: 2.16"
```

```
print(paste("b(1,2):", beta_w(fish, 1, 2, TRUE)))
```

```
## [1] "b(1,2): 0.5"
```

```
print(paste("b(1,10):", beta_w(fish, 1, 10, TRUE)))
```

```
## [1] "b(1,10): 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 β diversity above 1 which means that $\gamma > \alpha$. This implies that much of the γ diversity is not captured in within site diversity. This value tells us that there are about twice the number of species regionally than are found in the average site. About half of the diversity then must be β diversity. ***Answer 3b***: $\beta(1,2) < \beta(1,10)$ so sites 1 and 2 are more similar than site 1 and 10 ***Answer 3c***: I dont think the interpretation actually changes much qualitatively. This change redifines β diversity in absolute number of species which likely makes its range much larger in practice and in general confuses interpretation a little bit.

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 consider rare species as equally important to common 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_nonzero <- fish[-8,]
fish_ds <- vegan::vegdist(fish_nonzero, method = 'bray', binary = TRUE)
fish_db <- vegan::vegdist(fish_nonzero, method = 'bray', upper = TRUE, diag = TRUE)

# to see whether similarity or dissimilarity
# print(vegan::vegdist(fish_nonzero, method = 'bray', upper = TRUE, diag = TRUE, binary = TRUE))

# the whole things
#print(fish_ds)
#print(fish_db)

print(sum(fish_ds >= fish_db) / length(fish_ds))
```

```
## [1] 0.1847291
```

```
idx <- which(fish_ds > fish_db)[1] # corresponds to site 4 compared to site 1
print(fish[c(4, 1),])
```

```
##   Cogo Satr Phph Neba Thth Teso Chna Chto Lele Lece Baba Spbi Gogo Eslu Pefl
## 4    0    4    5    5    0    0    0    0    0    1    0    0    1    2    2
## 1    0    3    0    0    0    0    0    0    0    0    0    0    0    0    0
##   Rham Legi Scer Cyca Titi Abbr Icme Acce Ruru Blbj Alal Anan
## 4    0    0    0    0    1    0    0    0    0    0    0    0
## 1    0    0    0    0    0    0    0    0    0    0    0    0
```

```
print(fish[c(3, 1),])
```

```
##   Cogo Satr Phph Neba Thth Teso Chna Chto Lele Lece Baba Spbi Gogo Eslu Pefl
## 3    0    5    5    5    0    0    0    0    0    0    0    0    0    1    0
## 1    0    3    0    0    0    0    0    0    0    0    0    0    0    0    0
##   Rham Legi Scer Cyca Titi Abbr Icme Acce Ruru Blbj Alal Anan
## 3    0    0    0    0    0    0    0    0    0    0    0    0
## 1    0    0    0    0    0    0    0    0    0    0    0    0
```

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***: Both of these resemblances capture measures of dissimilarity. We can see this in the fact that the diagonals, self comparisons, show all zero values. ***Answer 5b***: Bray-Curtis uses more information than Sorenson and almost always reports a greater value than Sorenson. Looking at specific cases, we see that for a site with only a single species [1] when compared to another site with many rare species [4] is more dissimilar according to Sorenson than Bray-Curtis while the comparisoon of site [3] with a few intermediate species to site [1].

4) VISUALIZING BETA-DIVERSITY

A. Heatmaps

In the R code chunk below, do the following:

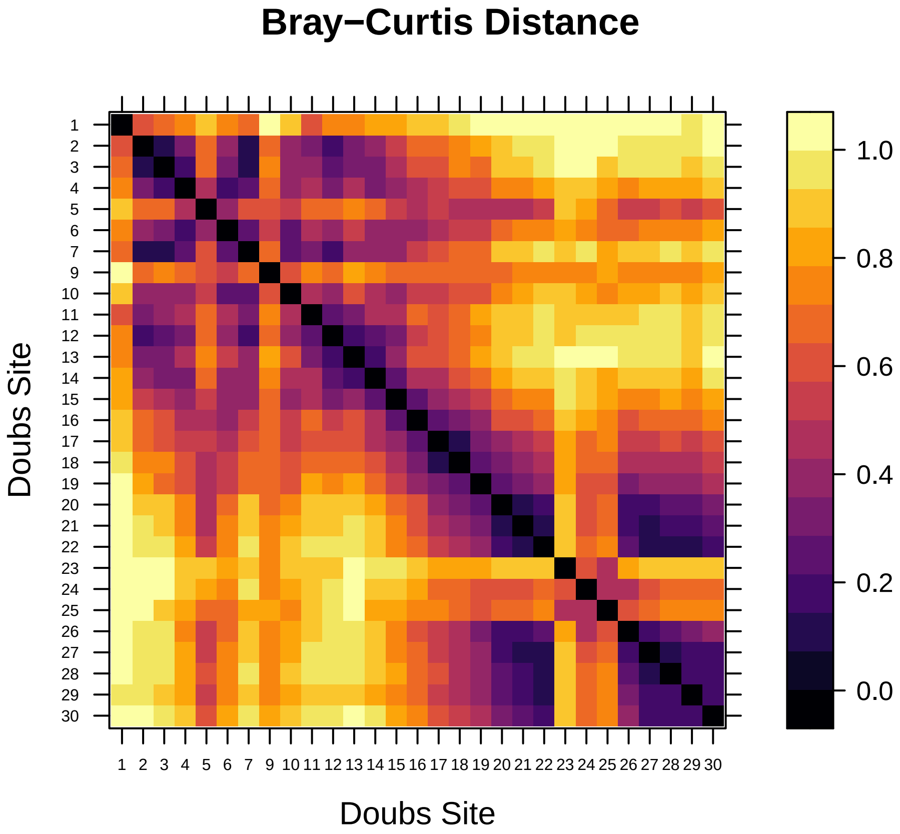
1. define a color palette,
2. define the order of sites in the Doubs River, and
3. use the `levelplot()` function to create a heatmap of fish abundances in the Doubs River.

```
require(lattice)
require(viridis)
```

```
## Loading required package: viridis
```

```
## Loading required package: viridisLite
```

```
# order 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")
```

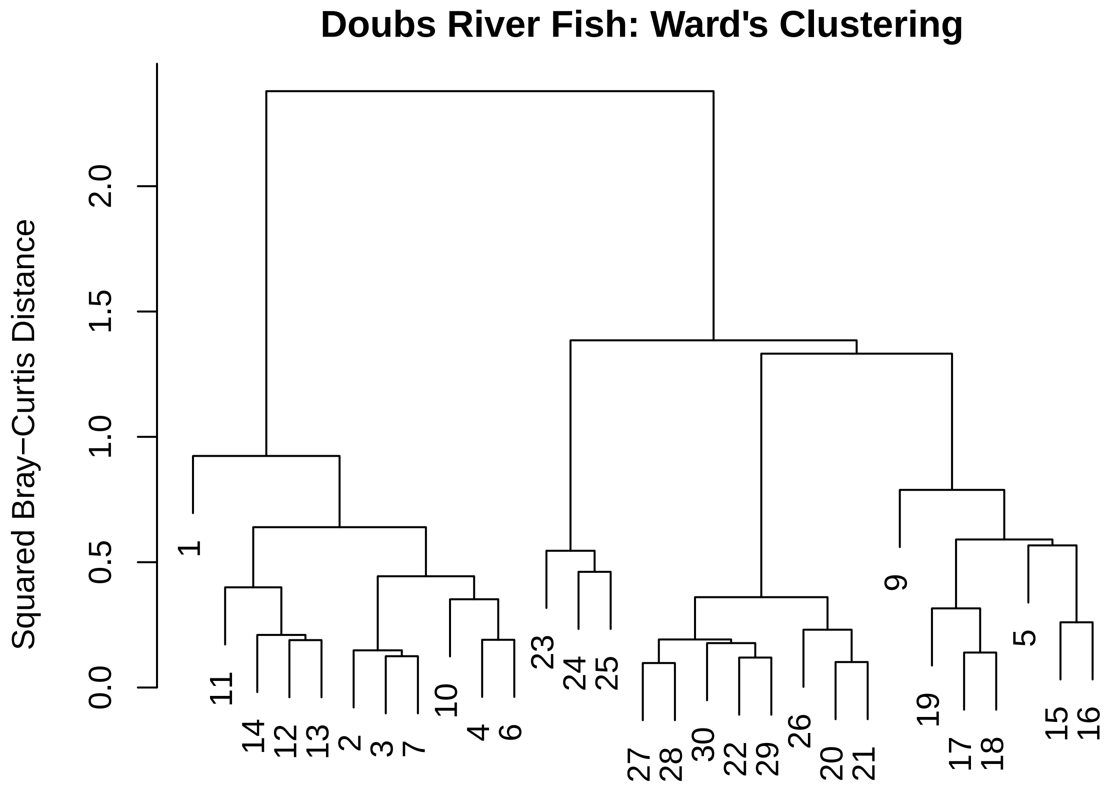


B. Cluster Analysis

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`).

```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")
```



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***: There are effectively three major different types of sites and one less frequent type. These correspond roughly to where they are on the river e.g. upstream or downstream.

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
require(dplyr)
```

```
## Loading required package: dplyr
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
```

```
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

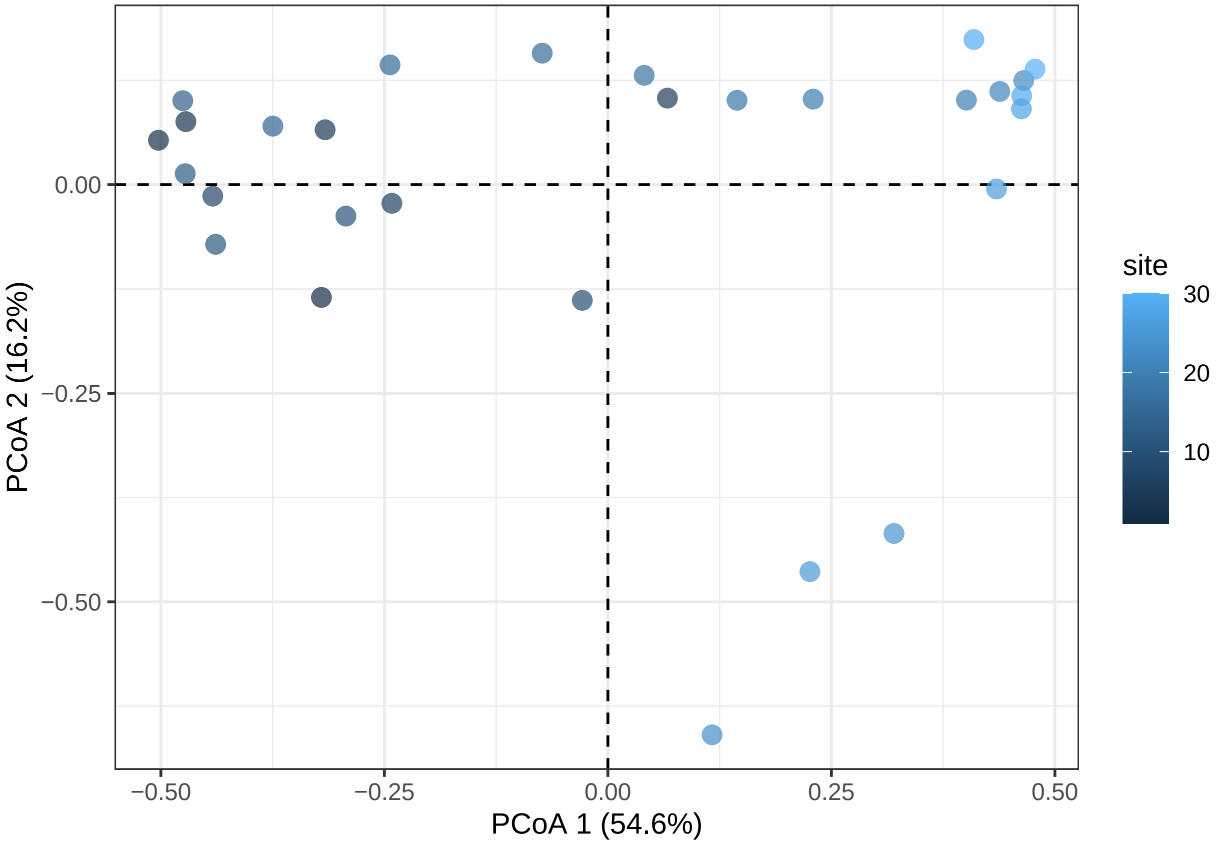
```
fish_pcoa <- cmdscale(fish_db, eig = TRUE, k = 3)
expvar <- round(sapply(fish_pcoa$eig, function(x) x / sum(fish_pcoa$eig)), 3) * 100
pcoa_df <- tibble::as_tibble(fish_pcoa$points)
```

```
## Warning: The `x` argument of `as_tibble.matrix()` must have unique column names if `.name_repair` is
## Using compatibility `.name_repair`.
```

```
ord_plot_df <- pcoa_df %>% mutate(site = as.numeric(rownames(fish_nonzero)))
print(ord_plot_df)
```

```
## # A tibble: 29 x 4
##         V1      V2      V3  site
##      <dbl>   <dbl>   <dbl> <dbl>
##  1 -0.320  -0.135   0.610      1
##  2 -0.503   0.0532  0.115      2
##  3 -0.472   0.0755  0.0252     3
##  4 -0.316   0.0658 -0.0547     4
##  5  0.0666  0.104  -0.0616     5
##  6 -0.242  -0.0223 -0.118      6
##  7 -0.442  -0.0137 -0.0196     7
##  8 -0.0286 -0.139  -0.190      9
##  9 -0.293  -0.0377 -0.202     10
## 10 -0.439  -0.0715  0.154     11
## # ... with 19 more rows
```

```
# the plot
g <- ggplot(ord_plot_df, aes(x=V1, y=V2, color=site)) +
  geom_hline(yintercept = 0, linetype=2) +
  geom_vline(xintercept = 0, linetype=2) +
  geom_point(size=3, alpha=0.7) +
  xlab(paste("PCoA 1 (", expvar[1], "%)", sep="")) +
  ylab(paste("PCoA 2 (", expvar[2], "%)", sep="")) +
  theme_bw()
g
```



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.

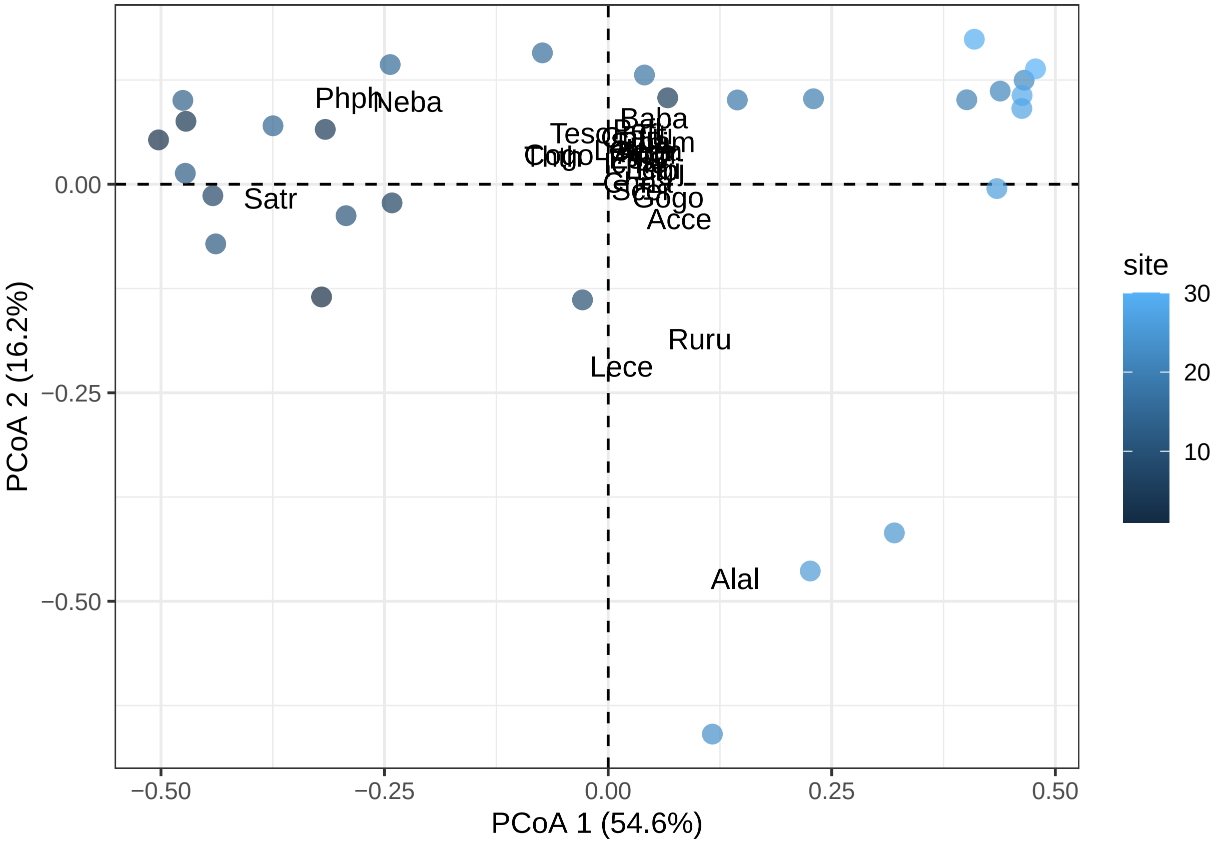
```
fishREL <- fish_nonzero
for (i in 1:nrow(fish_nonzero)) {
  fishREL[i, ] = fish_nonzero[i, ] / sum(fish_nonzero[i, ])
}

fish_pcoa2 <- BiodiversityR::add.spec.scores(fish_pcoa, fishREL, method = "pcoa.scores")
```

```
## Registered S3 methods overwritten by 'lme4':
##   method                          from
##   cooks.distance.influence.merMod car
##   influence.merMod                car
##   dfbeta.influence.merMod         car
##   dfbetas.influence.merMod        car
```

```
s_proj <- tibble::as_tibble(fish_pcoa2$cproj) %>% mutate(names = colnames(fish))
print(s_proj)
```

```
## # A tibble: 27 x 4
##       Dim1     Dim2     Dim3 names
##      <dbl>    <dbl>    <dbl> <chr>
##  1 -0.0552  0.0358  -0.0177  Cogo
##  2 -0.378  -0.0165   0.882   Satr
##  3 -0.290   0.104   -0.0949  Phph
##  4 -0.224   0.0995  -0.246   Neba
##  5 -0.0614  0.0336   0.00241 Thth
##  6 -0.0308  0.0617  -0.0648  Teso
##  7  0.0331  0.00254 -0.00406 Chna
##  8  0.0263  0.0567  -0.0492  Chto
##  9  0.0145  0.0415  -0.124   Lele
## 10  0.0150 -0.218   -0.186   Lece
## # ... with 17 more rows
```

```
g <- ggplot() +
  geom_hline(yintercept = 0, linetype=2) +
  geom_vline(xintercept = 0, linetype=2) +
  geom_point(data=ord_plot_df, aes(x=V1, y=V2, color=site), size=3, alpha=0.7) +
  geom_text(data=s_proj, aes(x=Dim1, y=Dim2, label=names)) +
  xlab(paste("PCoA 1 (", expvar[1], "%)", sep="")) +
  ylab(paste("PCoA 2 (", expvar[2], "%)", sep="")) +
  theme_bw()
g
```



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***: There are four major groups three of them are along $y = 0$ and one, the smallest group, is offset from it. ***Answer 7b***: Assuming lower site number is more upstream, the cluster far to the right is the most downstream which I would expect to be the most polluted (note: is this how rivers work?). There does not seem to be a good marker fish for this although absence of Satr is a candidate marker. Alternately we might think that the second eiganvector of the BC matrix (PCoA 2) is capturing quality somehow in which case the best marker fish would be Alal.

SYNTHESIS

Using the `mobsim` package from the DataWrangling module last week, simulate two local communities each containing 1000 individuals (N) and 25 species (S), but with one having a random spatial distribution and the other having a patchy spatial distribution. Take ten (10) subsamples from each site using the quadrat function and answer the following questions:

```
require(mobsim)
```

```
## Loading required package: mobsim
```

```
require(dplyr)
pois_comm <- mobsim::sim_poisson_community(25, 1000)
pois_quads <- mobsim::sample_quadrats(pois_comm, 10, avoid_overlap = TRUE, plot = FALSE)
thom_comm <- mobsim::sim_thomas_community(25, 1000)
thom_quads <- mobsim::sample_quadrats(thom_comm, 10, avoid_overlap = TRUE, plot = FALSE)
```

1) Compare the average pairwise similarity among subsamples in site 1 (random spatial distribution) to the average pairwise similarity among subsamples in site 2 (patchy spatial distribution). Use a t-test to determine whether compositional similarity was affected by the spatial distribution. Finally, compare the compositional similarity of site 1 and site 2 to the source community?

```
# default values are nonbinary bray-curtis
pois_beta <- vegan::vegdist(pois_quads$spec_dat)
thom_beta <- vegan::vegdist(thom_quads$spec_dat)
```

```
## Warning in vegan::vegdist(thom_quads$spec_dat): you have empty rows: their
## dissimilarities may be meaningless in method "bray"
```

```
t_res <- t.test(pois_beta, y = thom_beta)
t_res
```

```
##
##  Welch Two Sample t-test
##
## data:  pois_beta and thom_beta
## t = -9.7863, df = 66.809, p-value = 1.573e-14
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3611540 -0.2387842
## sample estimates:
## mean of x mean of y
## 0.6416121 0.9415812
```

β diversity is significantly higher in the thomas community than in the poisson community which is shown in the results of the t-test. I'm not entirely sure how we can compare them to the full community, though I'm sue that given enough time I could figure it out.

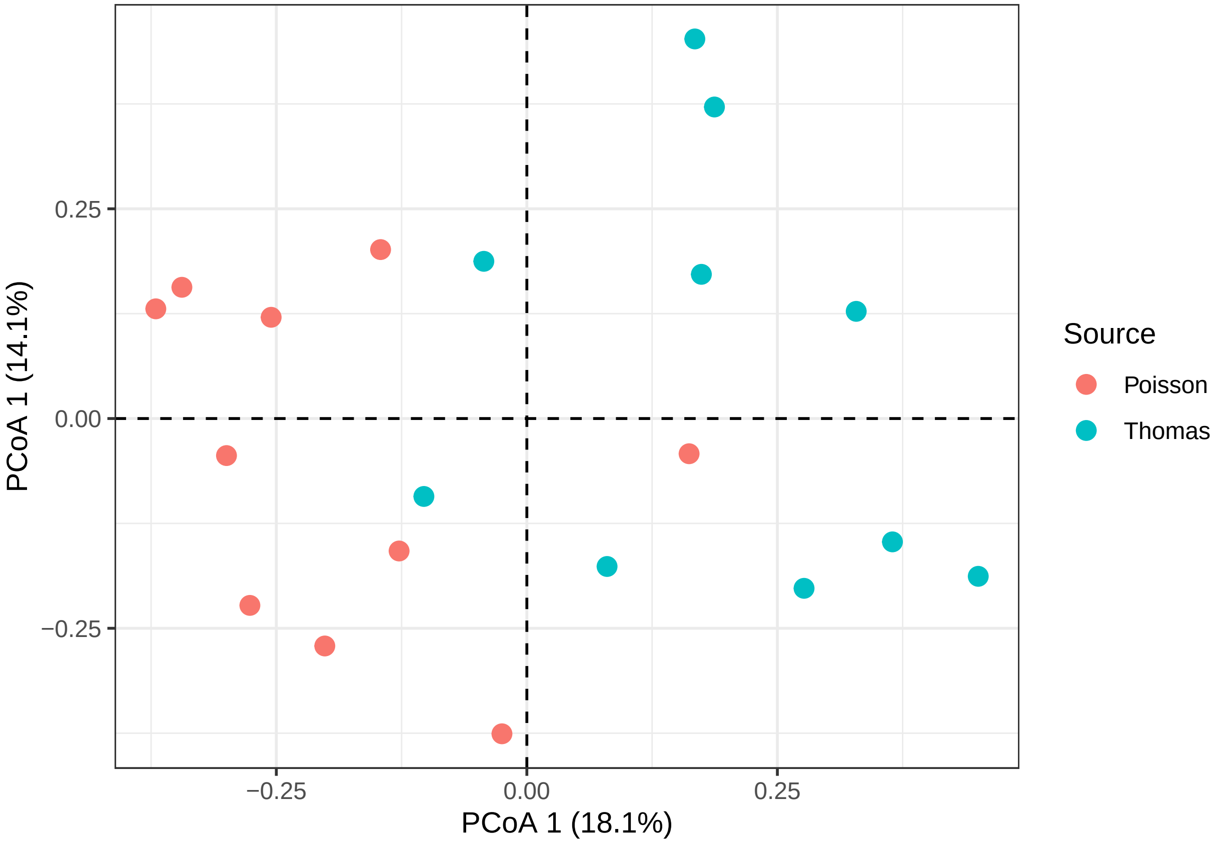
2) Create a cluster diagram or ordination using your simulated data. Are there any visual trends that would suggest a difference in composition between site 1 and site 2? Describe.

```
# construct and decompose resemblance matrix
all_sites <- bind_rows(pois_quads$spec_dat, thom_quads$spec_dat)
all_bray <- vegan::vegdist(all_sites, diag = TRUE, upper= TRUE)
```

```
## Warning in vegan::vegdist(all_sites, diag = TRUE, upper = TRUE): you have empty
## rows: their dissimilarities may be meaningless in method "bray"
```

```
pcoa_list <- cmdscale(all_bray, eig = TRUE, k = 3)
all_pcoa <- tibble::as_tibble(pcoa_list$points)
```

```r
# add source identities
all_pcoa <- mutate(all_pcoa, Source = c(rep('Poisson', 10), rep('Thomas', 10)))

# get the explained variance values
expvar <- round(sapply(pcoa_list$eig, function(x) x / sum(fish_pcoa$eig)), 3) * 100

g <- ggplot(all_pcoa, aes(x=V1, y=V2, color=Source)) +
  geom_vline(xintercept = 0, linetype=2) +
  geom_hline(yintercept = 0, linetype=2) +
  geom_point(size=3) +
  xlab(paste("PCoA 1 (", expvar[1], "%)", sep="")) +
  ylab(paste("PCoA 1 (", expvar[2], "%)", sep="")) +
  theme_bw()
g
```



The poisson and thomas communities do mostly form different clusters. Without coloring the points according to the source, however, it might be hard to identify the clusters in many cases.