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

Caroline Edwards; Z620: Quantitative Biodiversity, Indiana University

16 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 16th, 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())
getwd()
```

```
## [1] "/Users/carolineedwards/quant_bio/GitHub/QB2021_Edwards/2.Worksheets/8.BetaDiversity"
```

```
setwd("~/quant_bio/GitHub/QB2021_Edwards/2.Worksheets/8.BetaDiversity/")
package.list<-c('vegan','ade4','viridis','gplots','BiodiversityR','indicspecies','mobsim')
for (package in package.list){
  if (!require(package, character.only = TRUE, quietly =TRUE)){
    install.packages(package)
    library(package, character.only = TRUE)
  }
}
```

```
## This is vegan 2.5-6
```

```
## Warning: package 'ade4' was built under R version 3.6.2
```

```
## Warning: package 'gplots' was built under R version 3.6.2
```

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

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

```
## Warning: package 'BiodiversityR' was built under R version 3.6.2
```

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

```
## BiodiversityR 2.12-3: Use command BiodiversityRGUI() to launch the Graphical User Interface;
## to see changes use BiodiversityRGUI(changeLog=TRUE, backward.compatibility.messages=TRUE)
```

```
## Warning: package 'mobsim' was built under R version 3.6.2
```

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, pleae do not print the dataset when submitting
data(doubs)
#print(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: There are four lists in 'doubs'. ***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?

Answer 2a: The fish richness seems overall lower in the upstream samples compared to the downstream samples. ***Answer 2b***: The Brown Trout abundance is low in the downstream samples, and higher in the upstream samples. ***Answer 2c***: This example shows that richness is limited when looking at biodiversity, because while the species richness increased downstream, the abundance of the brown trout decreased, so species richness by itself isn't telling the whole story.

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){
  if (pairwise == TRUE){
    if (sitenum1 == ""|sitenum2 == ""){
      print("Error: please specify sites 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(site.by.species = doubs$fish, sitenum1 = 1 ,sitenum2 = 2, pairwise = TRUE)
```

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

```
beta.w(site.by.species = doubs$fish, sitenum1 = 1 ,sitenum2 = 10, pairwise = TRUE)
```

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

```
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: Local richness and turnover both go into regional fish diversity, with local richness providing information about how many species are present at each site and turnover providing information about how species change from one site to the next. ***Answer 3b***: Fish assemblage at site 1 is more similar to the one at site 2 than site 10. ***Answer 3c***: There is a multiplicative relationship between local (alpha) and regional (gamma) diversity, which quantifies how many more times diverse the regional pool is compared to the average alpha diversity at the sites. If beta diversity was defined additively, then that would mean beta diversity would just quantify the difference between the average alpha diversity and the regional diversity, which I think would undervalue the amount of alpha diversity.

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 are only using presence-absence data, so they don't take into account whether a species is rare at any particular site, but whether that species is shared between sites or not. Abundance-based metrics use abundance data, so information about species that are rare at sites is accounted for, not just rare in terms of which sites it is at.

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.dj<-vegdist(fish, method = "jaccard", binary = TRUE)
fish.db<-vegdist(fish, method = "bray", diag = TRUE)
fish.ds<-vegdist(fish, method = "bray", binary = TRUE, diag = TRUE)

#print(fish.db)
#print(fish.ds)
#fish.compare<-fish.db-fish.ds
#print(fish.compare)
```

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 represents dissimilarity, because a value of 1 means that the two sites do not share any species. If you print the diagonal in the resemblance matrix, it is all zeros, so you know that 0 means completely identical. ***Answer 5b***: It seems like the difference between the two usually changes the beta diversity value between 0 and 0.1. The biggest difference seems to happen when the bray-curtis dissimiliarity value is around 0.5. The Sorensen distances only use presence-absence, so there is more emphasis on the similarity of samples owing to shared sites, whereas the bray-curtis uses abundance data to determine pairwise differences.

4) VISUALIZING BETA-DIVERSITY

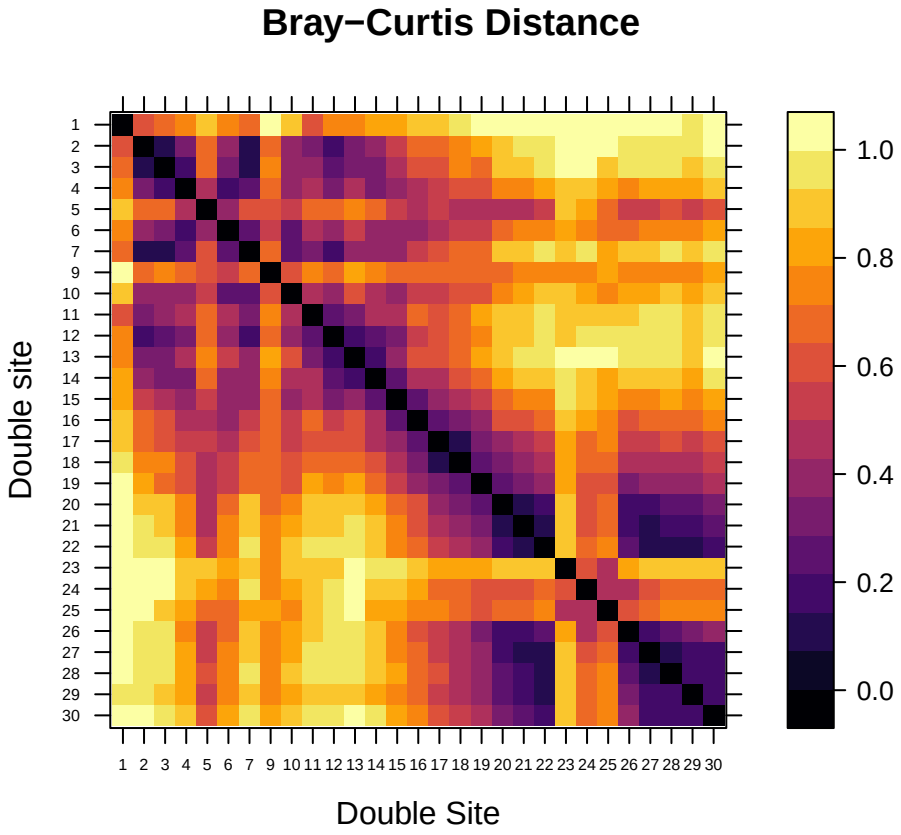
A. Heatmaps

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 = "Double Site",
          ylab = "Double site", scales = list(cex = 0.5), main = "Bray-Curtis Distance")
```

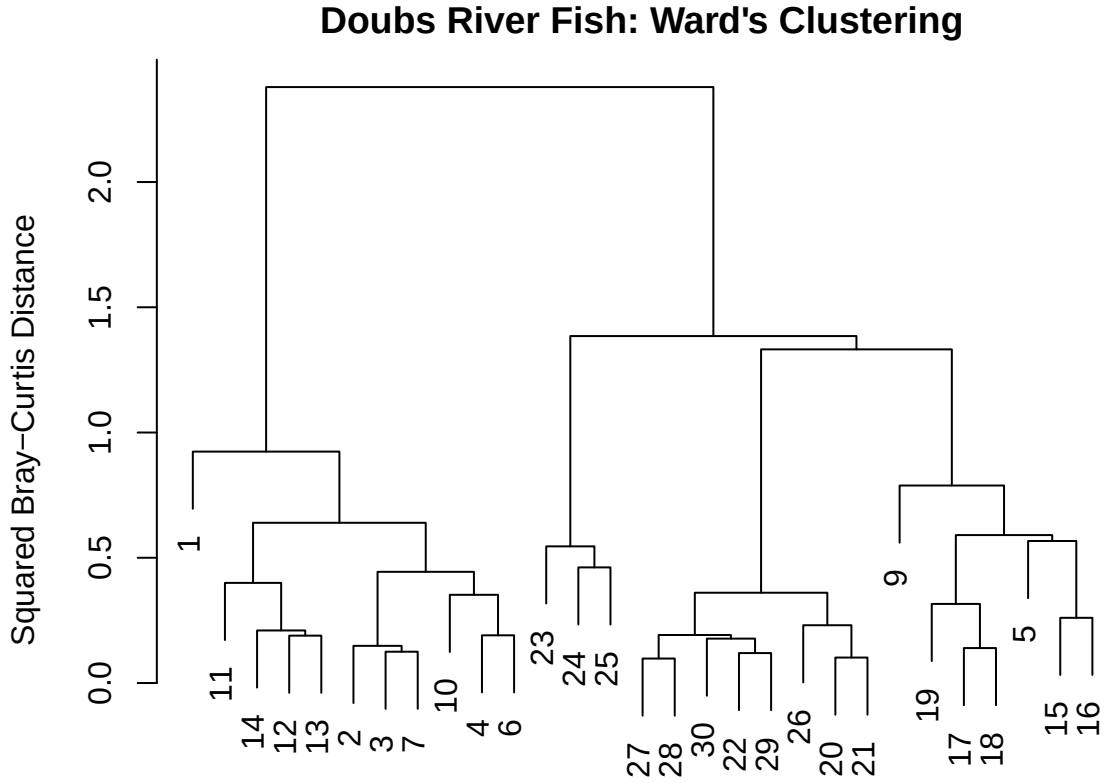


B. Cluster Analysis

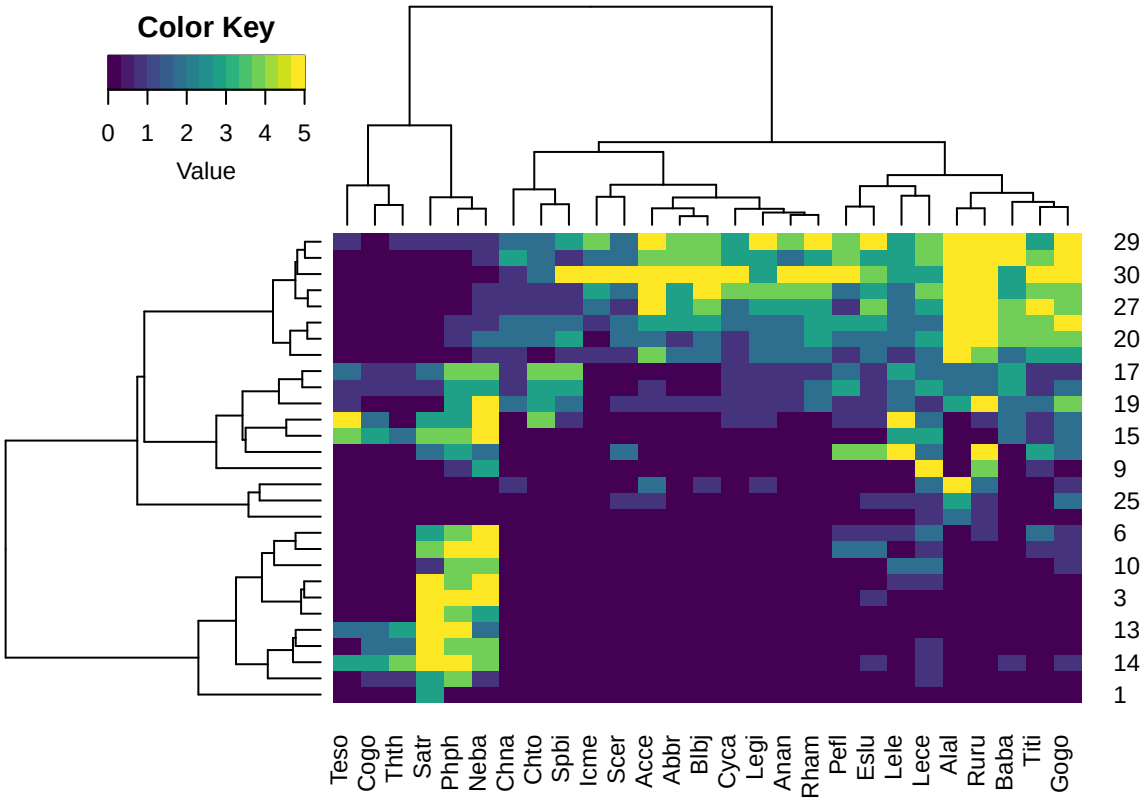
In the R code chunk below, do the following:

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")
```




```
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 fish diversity is related to location of how upstream or downstream the site is on the river, and specifically that some species of fish are more abundant upstream with others abundant downstream. There is some ecological factor that is driving the difference in diversity of fish in upstream and downstream sites which could be based on habitat differences that affect preferred food sources or effects of human presence.

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
```

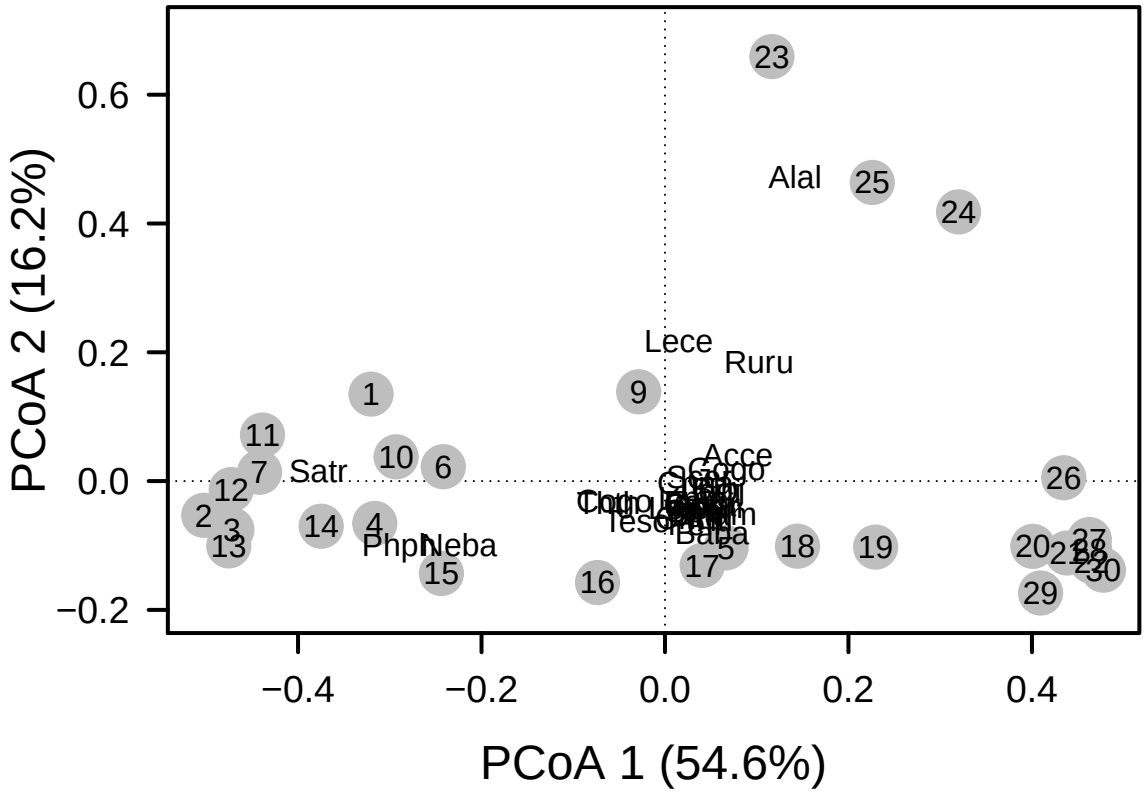
```
sum.eig<-sum(explainvar1, explainvar2, explainvar3)

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

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

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.

```
spe.corr<-add.spec.scores(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)
#print(fit)
```

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

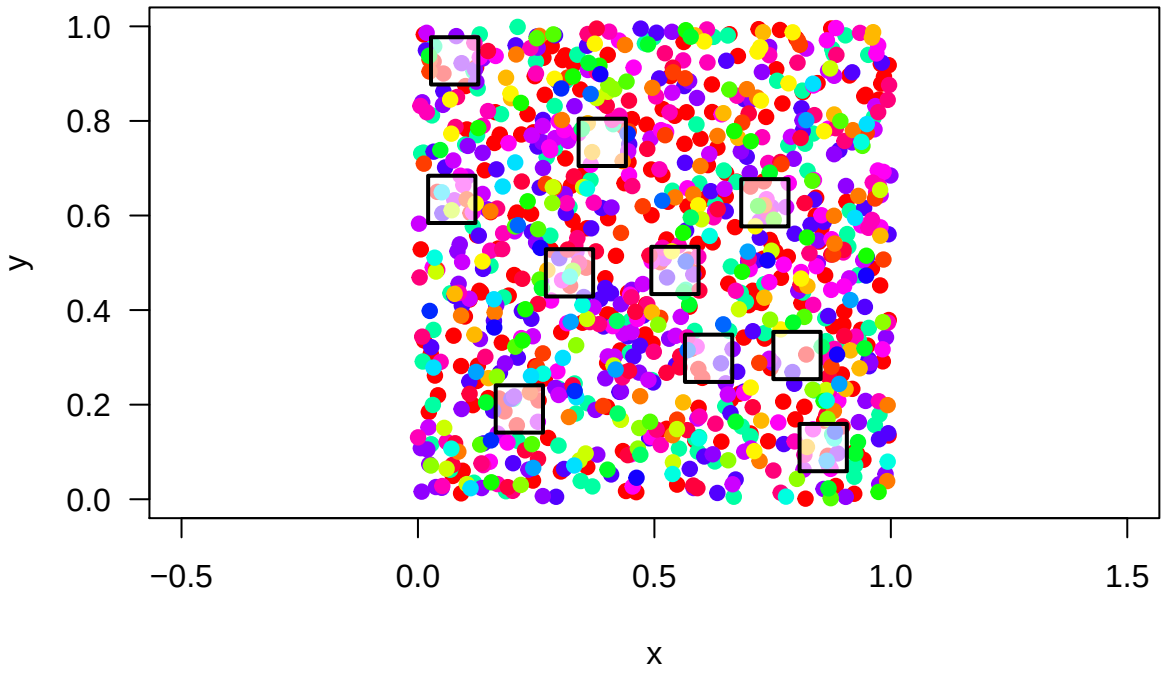
a. Describe the grouping of sites in the Doubs River based on fish community composition.
b. Generate a hypothesis about which fish species are potential indicators of river quality.

> ***Answer 7a***: The Doubs River sites are mostly grouped by site number (which is determined by geographic location along the river). The first axis explains ~55% of the variation in the data, so the primary variable driving the grouping of sites based on fish community composition is how far upstream or downstream the site is. The second axis is maybe separating sites based mostly on the abundance of the species "Alal", and to a lesser extent "Lece" and "Ruru". ***Answer 7b***: The species "Satr" could be an indicator of river quality because it is associated with the lowered numbered sites, so the downstream sites, but from the envfit analysis, the "Alal" and "Rham" species both have R^2>0.8, so their presence explains a lot of the variation.

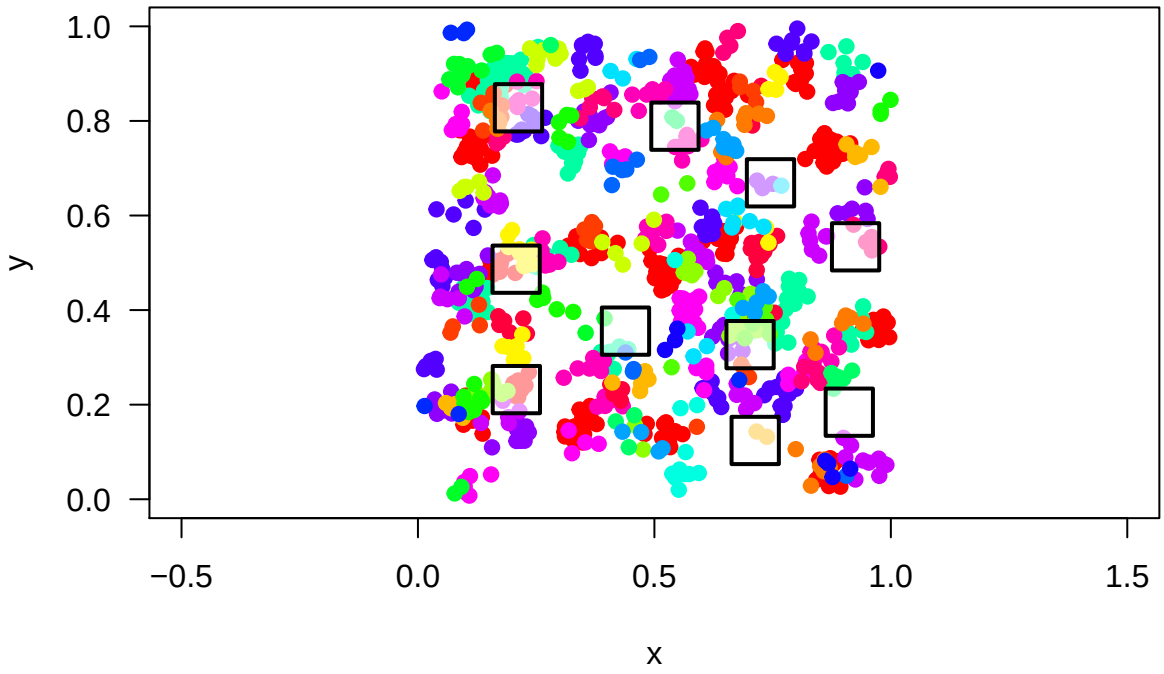
SYNTHESIS

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:

```
com_random <- sim_poisson_community(s_pool = 25, n_sim = 1000, sad_type = "lnorm",
        sad_coef = list("meanlog" = 2, "sdlog" = 1))
#plot(com_random)
com_random_quads <- sample_quadrats(com_random, n_quadrats = 10, quadrat_area = 0.01,
             method = "random", avoid_overlap = T)
```




```
com_patchy <- sim_thomas_community(s_pool = 25, n_sim = 1000, sad_type = "lnorm",
        sad_coef = list("meanlog" = 2, "sdlog" = 1))
#plot(com_patchy)
com_patchy_quads <- sample_quadrats(com_patchy, n_quadrats = 10, quadrat_area = 0.01,
             method = "random", avoid_overlap = T)
```



1) Compare the average pairwise similarity among subsamples in site 1 (random spatial distribution) to the average pairswise 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?

Yes, compositional similarity is affected by the spatial distribution.

```
random_beta<-vegdist(com_random_quads$spec_dat, method = "bray")
patchy_beta<-vegdist(com_patchy_quads$spec_dat, method = "bray")
print(random_beta)
```

```
##           site1     site2     site3     site4     site5     site6     site7
## site2  0.5000000
## site3  0.5833333 0.4615385
## site4  0.6363636 0.5000000 0.5000000
## site5  0.5789474 0.4285714 0.7142857 0.4736842
## site6  0.7000000 0.3636364 0.6363636 0.5000000 0.4117647
## site7  0.5789474 0.5238095 0.5238095 0.6842105 0.6250000 0.7647059
## site8  0.7000000 0.4545455 0.6363636 0.7000000 0.5294118 0.4444444 0.6470588
## site9  0.5454545 0.5000000 0.5833333 0.4545455 0.4736842 0.4000000 0.7894737
## site10 0.8461538 0.7333333 0.8666667 0.6923077 0.6000000 0.6363636 0.8000000
##           site8     site9
## site2
## site3
## site4
```

```
## site5
## site6
## site7
## site8
## site9  0.7000000
## site10 0.8181818 0.6923077
```

```
print(patchy_beta)
```

```
##           site1     site2     site3     site4     site5     site6     site7
## site2  0.7058824
## site3  1.0000000 1.0000000
## site4  0.7500000 1.0000000 1.0000000
## site5  1.0000000 0.9000000 1.0000000 1.0000000
## site6  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## site7  1.0000000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## site8  0.9354839 0.7647059 1.0000000 1.0000000 0.5675676 1.0000000 1.0000000
## site9  1.0000000 0.8125000 1.0000000 1.0000000 1.0000000 1.0000000 1.0000000
## site10 1.0000000 0.8297872 0.7500000 1.0000000 0.8181818 1.0000000 1.0000000
##           site8     site9
## site2
## site3
## site4
## site5
## site6
## site7
## site8
## site9  0.7931034
## site10 0.8636364 0.7600000
```

```
t.test(random_beta, patchy_beta)
```

```
##
##  Welch Two Sample t-test
##
## data:  random_beta and patchy_beta
## t = -13.543, df = 85.398, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.3903770 -0.2904339
## sample estimates:
## mean of x mean of y
## 0.5985023 0.9389077
```

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. > This cluster diagram shows that the compositions between site 1 (random) and site 2 (patchy) have very different compositions and group together mostly (7 out of 10 for each site), even though they were simulated from the same distribution. There are a few sites from each community that group together, but most group based on whether they are from a patchy or randomly distributed community.

```r
random_sites<-(com_random_quads$spec_dat)
rownames(random_sites)<-c("R1","R2","R3","R4","R5","R6","R7","R8","R9","R10")
patchy_sites<-(com_patchy_quads$spec_dat)
rownames(patchy_sites)<-c("P1","P2","P3","P4","P5","P6","P7","P8","P9","P10")
all_sites<-rbind(random_sites,patchy_sites)

all.db<-vegdist(all_sites, method = "bray")
all.ward<-hclust(all.db, method="ward.D2")
par(mar = c(1,5,2,2)+0.1)
plot(all.ward, main = "Randomly distributed community: Ward's clustering", ylab = "Squared Bray-Curtis
```