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

Herbert Sizek; Z620: Quantitative Biodiversity, Indiana University

21 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,

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

```
## [1] "D:/GitHub/QB2021_Sizek/2.Worksheets/8.BetaDiversity"
```

```
setwd("D:/GitHub/QB2021_Sizek/2.Worksheets/8.BetaDiversity")
```

3. set your working directory to your "*/8.BetaDiversity*" folder, and
4. load the `vegan` R package (be sure to install if needed).

```
for (i  in c("vegan","ade4", "viridis", "gplots", "BiodiversityR", "indicspecies")){
  require(i,character.only=TRUE)
}
```

```
## Loading required package: vegan
```

```
## Warning: package 'vegan' was built under R version 3.6.3
```

```
## Loading required package: permute
```

```
## Warning: package 'permute' was built under R version 3.6.3
```

```
## Loading required package: lattice
```

```
## This is vegan 2.5-7
```

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

```
## Warning: package 'ade4' was built under R version 3.6.3
```

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

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

```
## Warning: package 'viridisLite' was built under R version 3.6.3
```

```
## Loading required package: gplots
```

```
## Warning: package 'gplots' was built under R version 3.6.3
```

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

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

```
## Loading required package: BiodiversityR
```

```
## Warning: package 'BiodiversityR' was built under R version 3.6.3
```

```
## Loading required package: tcltk
```

```
## 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)
```

```
## Loading required package: indicspecies
```

```
## Warning: package 'indicspecies' was built under R version 3.6.3
```

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)
summary(doubs)
```

```
##         Length Class      Mode
## env     11     data.frame list
## fish    27     data.frame list
## xy       2     data.frame list
## species  4     data.frame list
```

Question 1: Describe some of the attributes of the `doubs` dataset.

a. How many objects are in `doubs` ? > ***Answer 1a***: 4

b. How many fish species are there in the `doubs` dataset? > ***Answer 1b***: 27

c. How many sites are in the `doubs` dataset? > ***Answer 1c***: 30

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? > ***Answer 2a***:

```
S.obs <- function(x=''){
  rowSums(x>0)*1
}
cFisher <-function(x=''){
  fisher.alpha(as.integer(as.vector(x[x>0])))
}
S.obs(doubs$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
```

```
apply(doubs$fish,1,cFisher)
```

```
##          1          2          3          4          5          6          7
##  0.5252543  1.2838726  1.7118462  4.7170847  5.6420718  7.4791167  2.4967465
##          8          9         10         11         12         13         14
##  1.0000000  2.7823859  3.9775884  5.4028053  3.1515726  3.0201644  5.5647670
##         15         16         17         18         19         20         21
##  5.7778860 11.1701802 17.5099120 20.8366287 18.3058149 13.3546197 13.2359384
##         22         23         24         25         26         27         28
## 10.8014143  5.4525555  6.9657122 13.1934641 16.2109829 12.0091137 11.0334863
##         29         30
## 12.5581380  8.6723803
```

b. How does Brown Trout (*Salmo trutta*) abundance vary along the sampled reach of the Doubs River? > ***Answer 2b***: There are regions where there is higher abundance and regions with lower abundance, low - medium - low - medium - high - low - high

c. What do these patterns say about the limitations of using richness when examining patterns of biodiversity?

> ***Answer 2c***: There is spatial variation in richness across space, so if your habitat of interest is heterogeneous the samples might reflect the habitat health or some other variable rather than a large variation in species richness.

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

```
beta.w <- function(site.by.species = "",sitenum1="",sitenum2=""){
  if (sitenum1=="" | sitenum2==""){
    if (xor(sitenum1=="",sitenum2="")){
      warning("beta.w: One site was specified, returning beta for entire SbyS matrix")
    }
    SbyS.pa <- decostand(site.by.species,method = "pa") # convert to presence absence
    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)
  }
  else{
    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)
  }

}
```

2. use this function to analyze various aspects of β-diversity in the Doubs River.

```
beta.w(doubs$fish)
```

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

```
beta.w(doubs$fish,5,26)
```

```
## [1] 0.438
```

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. > *__Answer 3a__: This question is a bit odd, because these three are in a mathamatical relation, being: $= $, for at least whitaker's β diversity. The two components that are directly measured are α and γ, with the prior being the mean local richness (average number of species per site) and the latter being the total number of species across all sites. This means that as $beta$ approaches 1, the species are distributed randomly across sites, that is most sites have the same compositions of species. If there is wide variation between species at sites, γ will increase while α could remain the same, increasing β. If the distribution of species in sites increases then α will fall, increasing β.

b. Is the fish assemblage at site 1 more similar to the one at site 2 or site 10? > ***Answer 3b***:

```
S.obs(doubs$fish[c(1,2,10),])
```

```
##  1  2 10
##  1  3  6
```

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

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

```
beta.w(doubs$fish,1,10)
```

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

> The species at site 1 is more similar to the species at site 2, but this is primarily driven by that site 2 has fewer species than 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 3c***: This would make the measure a bit harder for me to understand value in as the number of sites would effect the outcome more. It would be more like a measure of the first moment of species. I don't know really how I would apply it thoughtfully.

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 matrices are binary matricies while abundance matrices are whole numbers. This means that rare species are of equal value in binary matrices while they are not in abundance matrices.

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),

```
fish <- doubs$fish
fish<- fish[-which(rowSums(fish)==0),] #remove rows with zero values.
```

3. construct a resemblance matrix based on Sørensen's Similarity ("fish.ds"), and

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

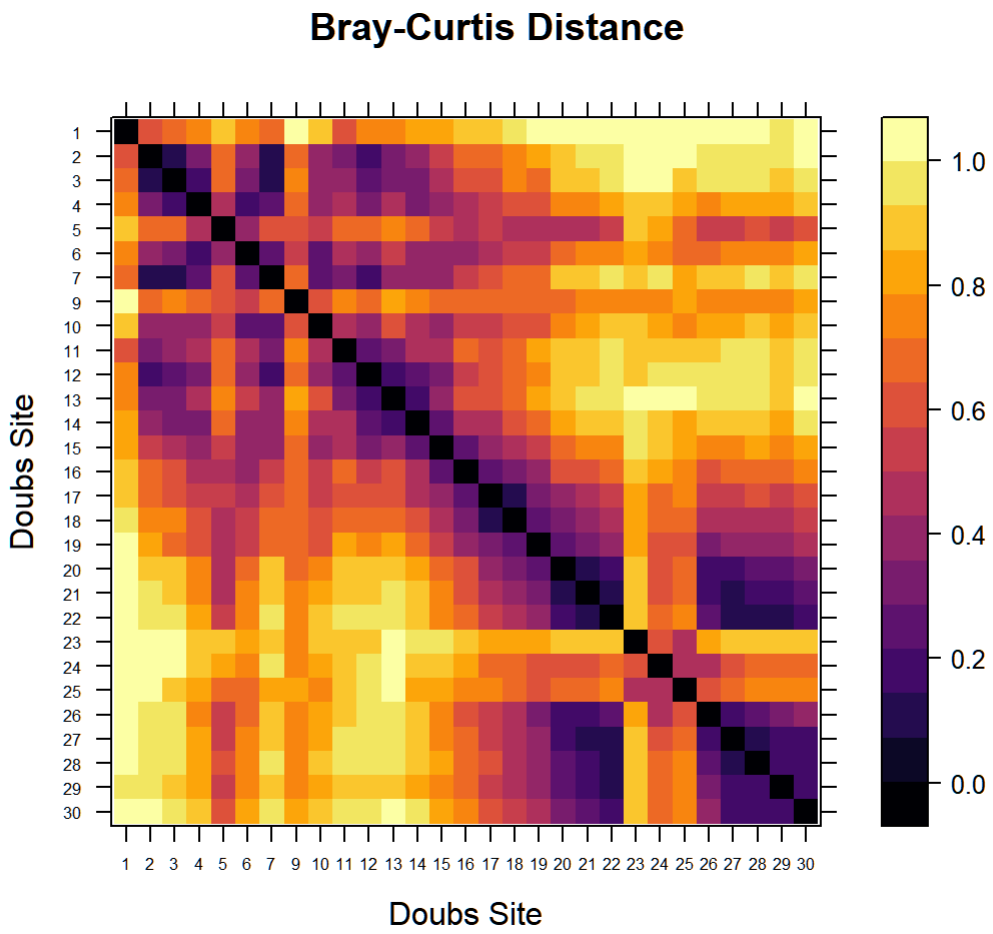
4. construct a resemblance matrix based on Bray-Curtis Distance ("fish.db").

```
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? > ***Answer 5a***: Dissimilarity is closer to 1, because the values close to the diagnal are near zero, while farther away (farther away sites) are near 1.

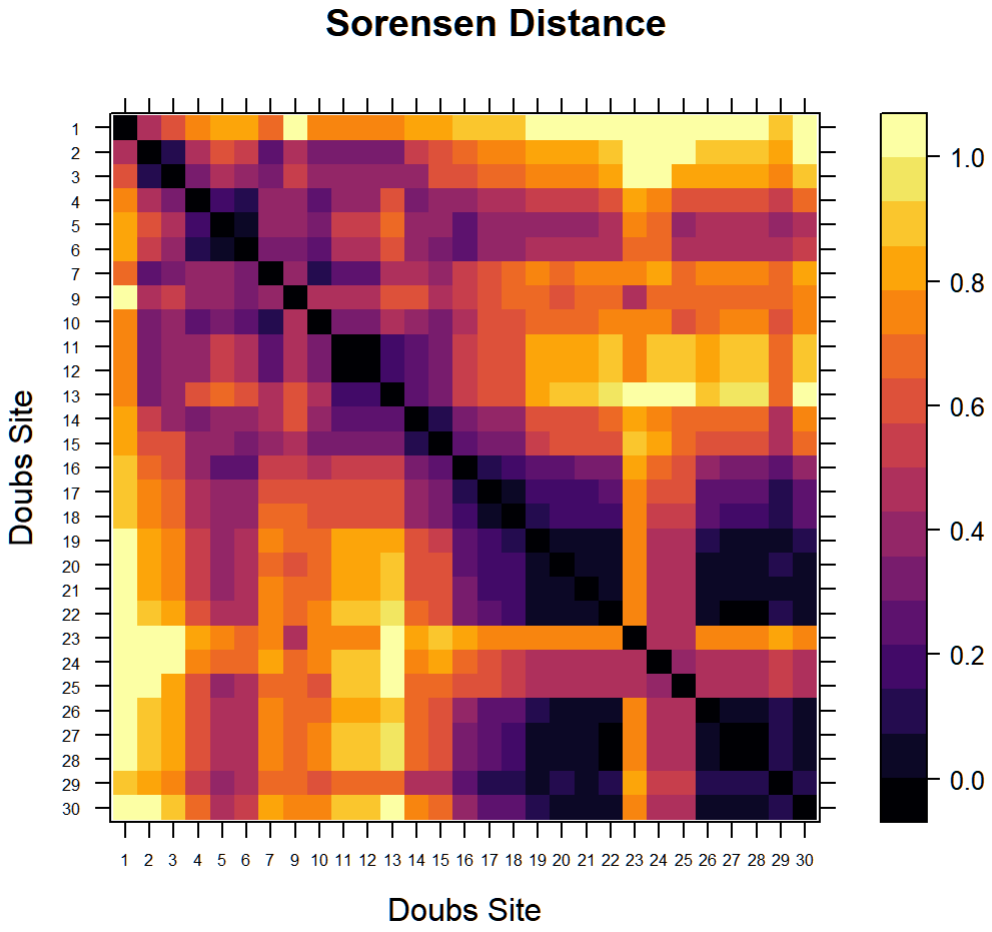
```
order <- rev(attr(fish.db,"Labels"))
levelplot(as.matrix(fish.db)[,order],aspect="iso",col.regions=inferno,xlab="Doubs Site",ylab="Do
ubs Site", main="Bray-Curtis Distance",scales=list(cex=0.5))
```



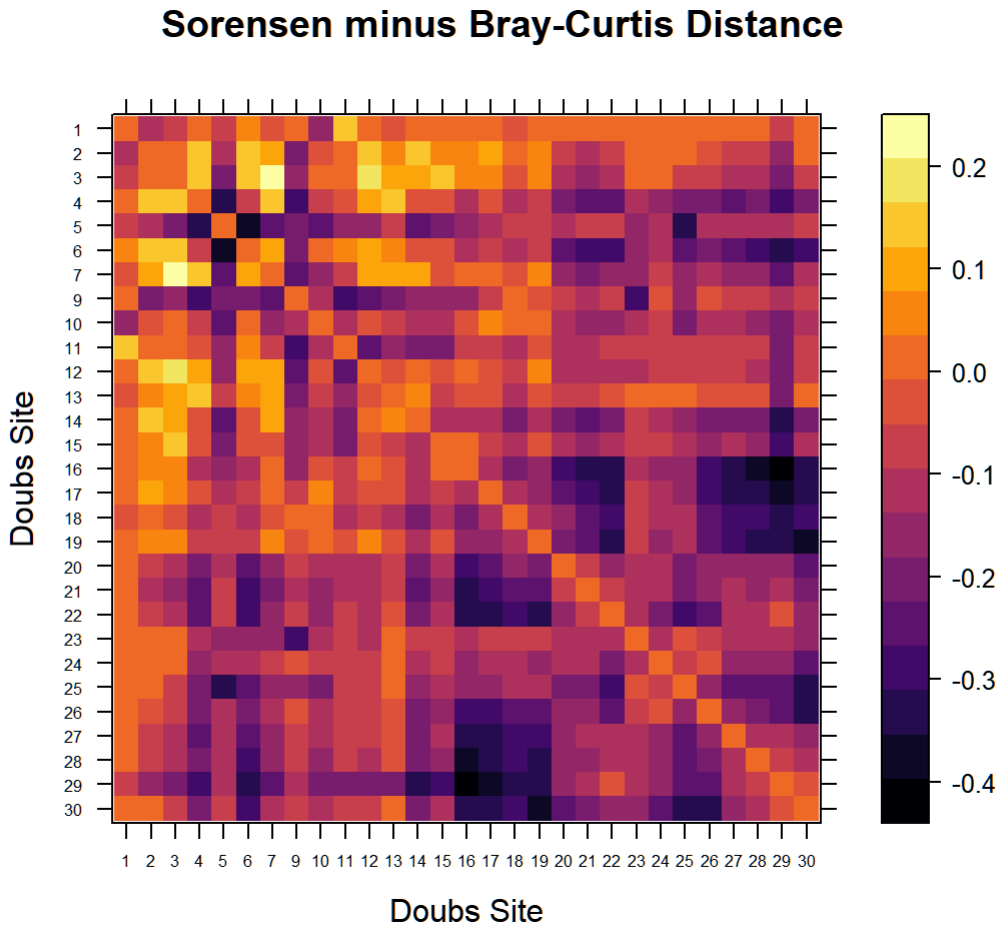
b. Compare the resemblance matrices (`fish.db` or `fish.ds`) you just created. How does the choice of the Sørensen or Bray-Curtis distance influence your interpretation of site (dis)similarity?

> ***Answer 5b***: They are fairly similar, Sorensen makes it seem that the sites are more homogenous then Bray-Curtis. Sorensen generally provides lower values (sites are more similar) in comparison to Bray-Curtis.

```
order <- rev(attr(fish.ds,"Labels"))
levelplot(as.matrix(fish.ds)[,order],aspect="iso",col.regions=inferno,xlab="Doubs Site",ylab="Do
ubs Site", main="Sorensen Distance",scales=list(cex=0.5))
```




```
order <- rev(attr(fish.ds-fish.db,"Labels"))
levelplot(as.matrix(fish.ds-fish.db)[,order],aspect="iso",col.regions=inferno,xlab="Doubs Site",
ylab="Doubs Site", main="Sorensen minus Bray-Curtis Distance",scales=list(cex=0.5))
```



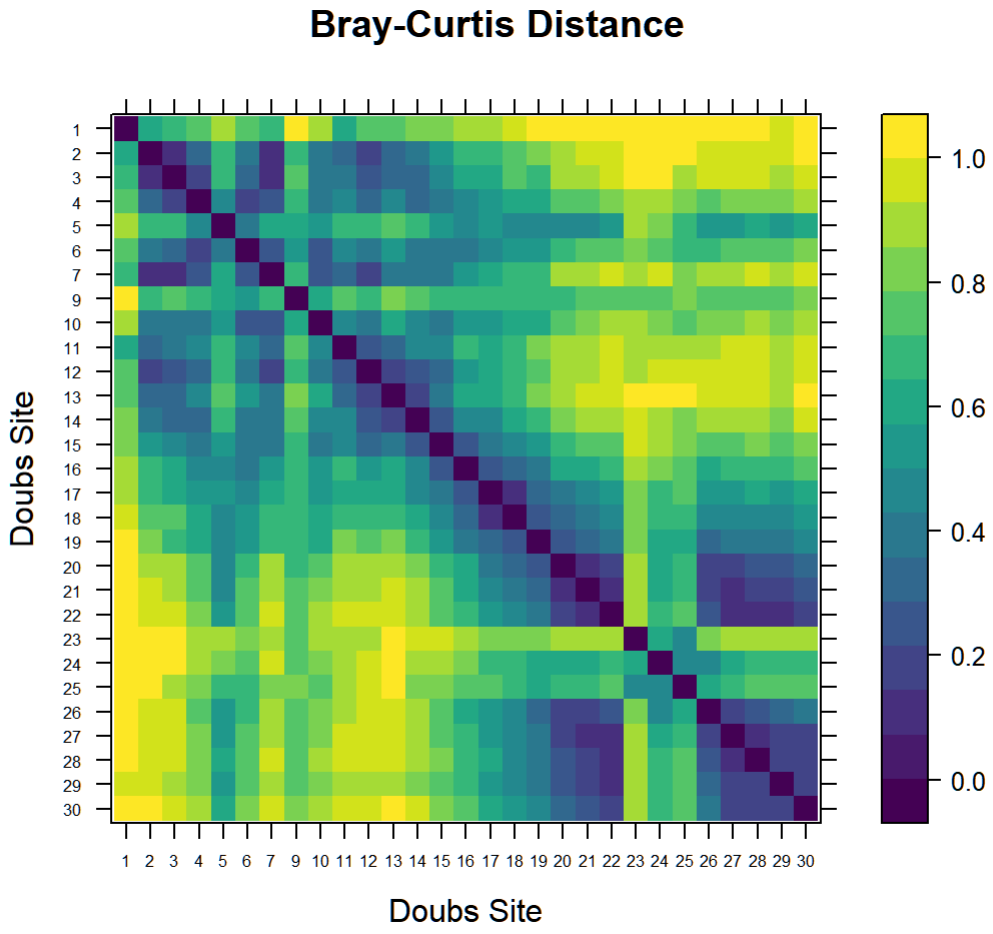
4) VISUALIZING BETA-DIVERSITY

A. Heatmaps

In the R code chunk below, do the following:

1. define a color palette, > Viridis was imported above
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=viridis,xlab="Doubs Site",ylab="Do
ubs Site", main="Bray-Curtis Distance",scales=list(cex=0.5))
```

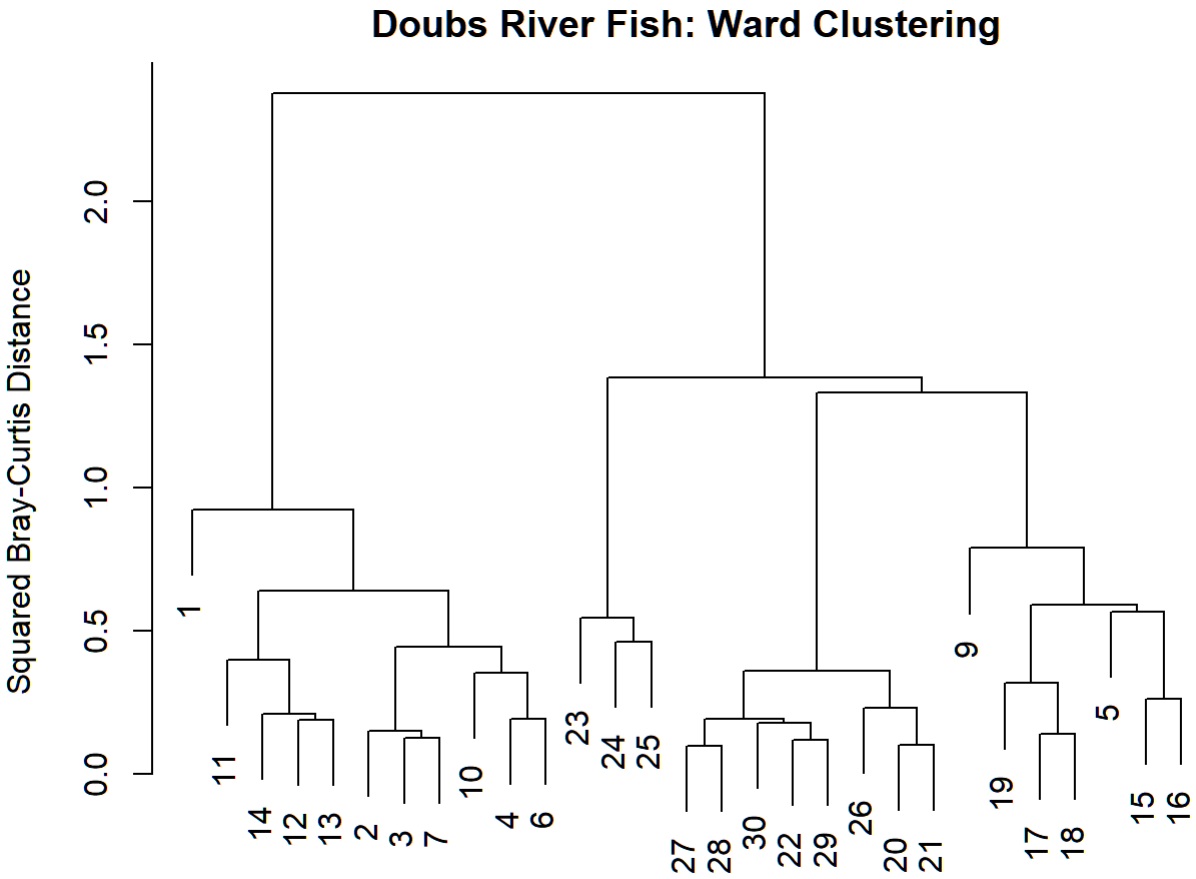


B. Cluster Analysis

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 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 two general habitats, one that is upstream (1-14) and downstream (15 -30), the lower habitat is could be broken down into three pieces, but I would want to test against other clustering algorithms.

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

```
fish.pcoa <- cmdscale(fish.db, eig =TRUE, k=3)
var1 <- round(fish.pcoa$eig[1]/sum(fish.pcoa$eig),3)
var2 <- round(fish.pcoa$eig[2]/sum(fish.pcoa$eig),3)
var3 <- round(fish.pcoa$eig[3]/sum(fish.pcoa$eig),3)
sum.eig <- sum(var1,var2,var3)
```

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.

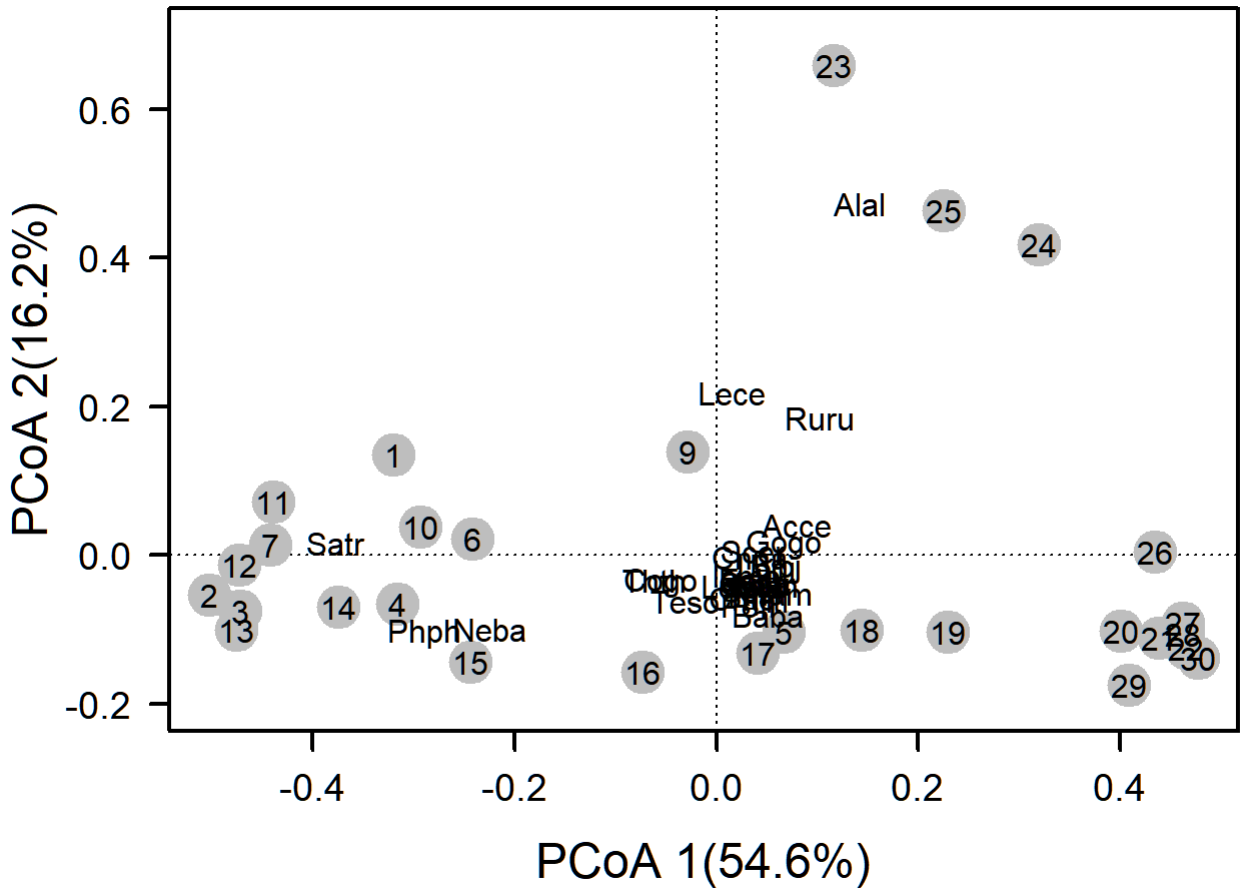
```
fishREL <- fish
for (i in 1:nrow(fish)){
  fishREL[i,] = fish[i,]/sum(fish[i,])
}

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

```
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(",var1*100,"%)",sep=''),
     ylab = paste("PCoA 2(",var2*100,"%)",sep=''),
     pch =16 ,cex =2.0,type ='n',cex.lab=1.5,cex.axis=1.2,axes=FALSE
     )

axis(side =1, labels=TRUE,lwd.ticks=2,cex.axis =1.2,las=1)
axis(side =2, labels=TRUE,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")
```



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
corr.cut <- 0.7
imp.spp <-spe.corr[abs(spe.corr[,1])>=corr.cut | abs(spe.corr[,2])>= corr.cut,]
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. > ***Answer 7a***: Because the PCoA is separating out based on the same principles as Ward Clustering (correlation levels between groups), just with slightly different methods, we see the same results, where there is the general two groups of 1-14 and 15-30 with the exception that 15 is similar to 1-14 on the first two PCs. The separation along the first PC appears mainly to be driven by the presence or absence of three species.

b. Generate a hypothesis about which fish species are potential indicators of river quality. > ***Answer 7b***: This question is too subjective. How do you define river quality? For example, if we are selecting for local diveristy, we might say that Alal is deterimental to the river quality, but it could also be that Alal is more constrained to a specific habitat or are expert niche constructors and exclude other species, but might influence terrestial species differently. We would first have to define a metric for river quality.

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.

```
require(mobsim)
```

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

```
## Warning: package 'mobsim' was built under R version 3.6.3
```

```
comA <- sim_poisson_community(s_pool = 25, n_sim = 1000, sad_type = "lnorm",
        sad_coef = list("meanlog" = 2, "sdlog" = 1))

comB <- sim_thomas_community(s_pool = 25, n_sim = 1000, sad_type = "lnorm",
        sad_coef = list("meanlog" = 2, "sdlog" = 1))
```

Take ten (10) subsamples from each site using the quadrat function and answer the following questions:

```
comm_matA <- sample_quadrats(comA, n_quadrats = 10, quadrat_area = 0.03,
             avoid_overlap = T, plot=F)
```

```
## Warning in sample_quadrats(comA, n_quadrats = 10, quadrat_area = 0.03, avoid_overlap = T, : C
annot find a sampling layout with no overlap.
##                                           Install the package spatstat for an improved meth
od for non-overlapping squares,
##                                           Use less quadrats or smaller quadrat area, or set
avoid_overlap to FALSE.
```

```
comm_matB <- sample_quadrats(comB, n_quadrats = 10, quadrat_area = 0.03,
             avoid_overlap = T, plot=F)
```

```
## Warning in sample_quadrats(comB, n_quadrats = 10, quadrat_area = 0.03, avoid_overlap = T, : C
annot find a sampling layout with no overlap.
##                                           Install the package spatstat for an improved meth
od for non-overlapping squares,
##                                           Use less quadrats or smaller quadrat area, or set
avoid_overlap to FALSE.
```

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).

```
comA.ds <- vegdist(comm_matA$spec_dat,method="bray",binary=TRUE)
comB.ds <- vegdist(comm_matB$spec_dat,method="bray",binary=TRUE)
comA.db <- vegdist(comm_matA$spec_dat,method="bray")
comB.db <- vegdist(comm_matB$spec_dat,method="bray")
print(paste("Random Community's Average Similarity (Sorensen index):",mean(comA.ds)))
```

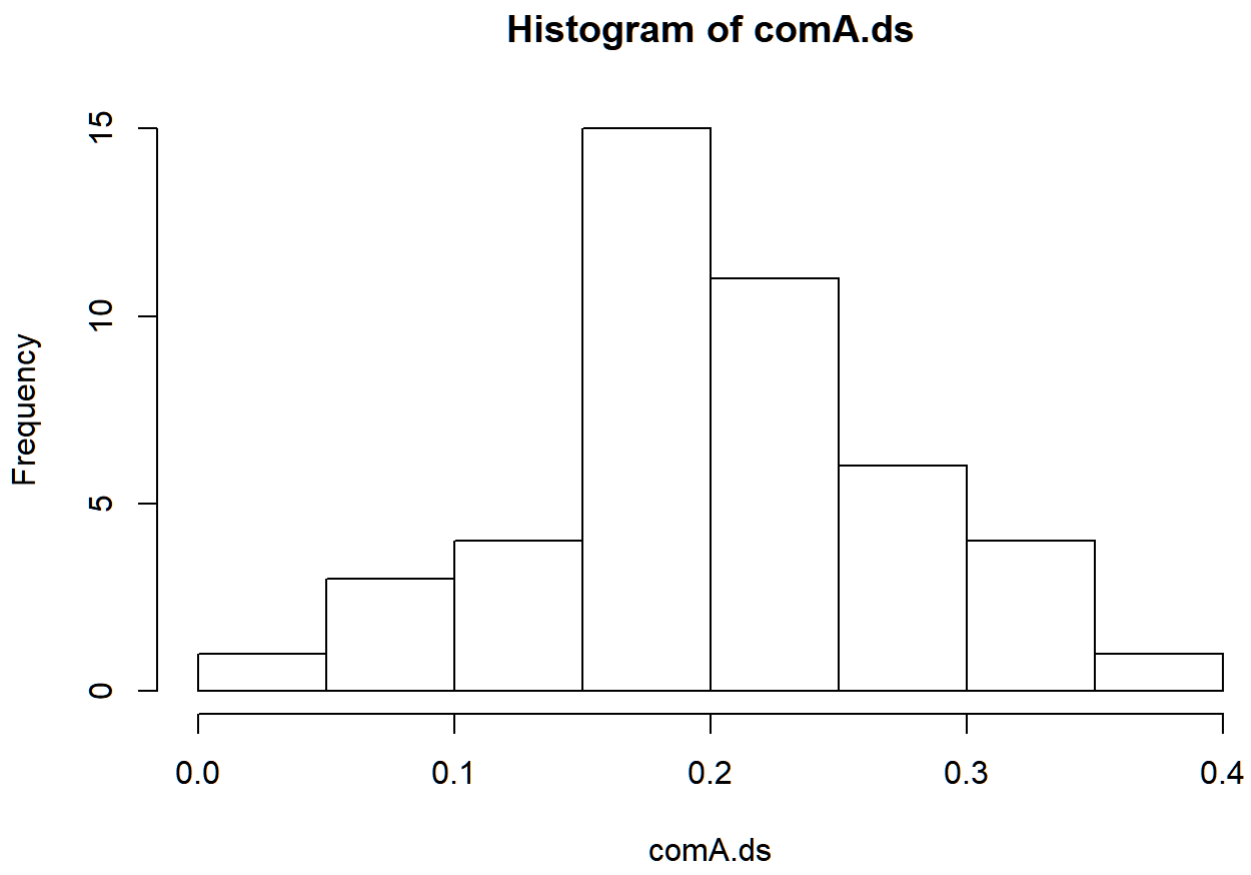
```
## [1] "Random Community's Average Similarity (Sorensen index): 0.202444742993374"
```

```
print(paste("Patchy Community's Average Similarity (Sorensen index):",mean(comB.ds)))
```

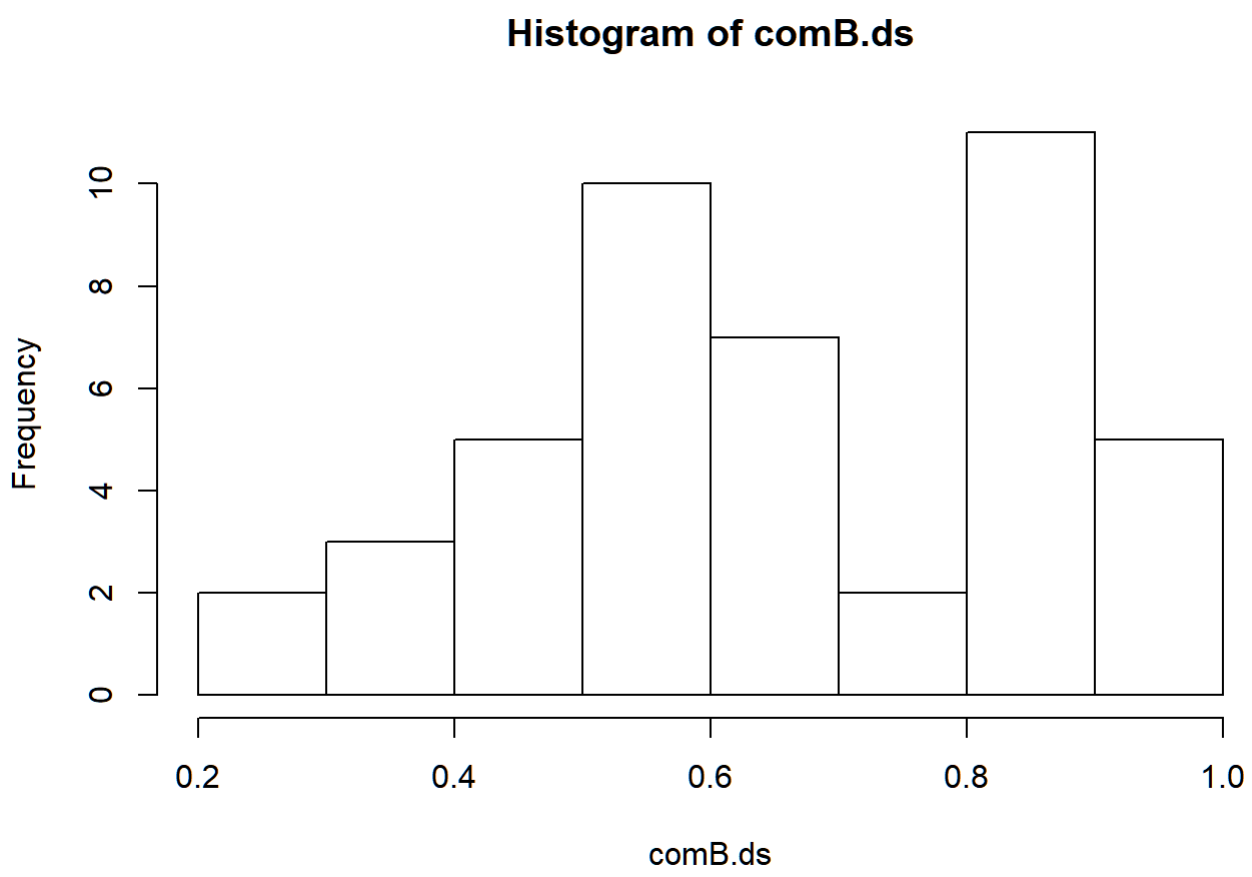
```
## [1] "Patchy Community's Average Similarity (Sorensen index): 0.667763039429706"
```

Use a t-test to determine whether compositional similarity was affected by the spatial distribution. >First we need to look at if the t-test is appropriate in this case.

```
hist(comA.ds,breaks=10)
```




```
hist(comB.ds,breaks=10)
```



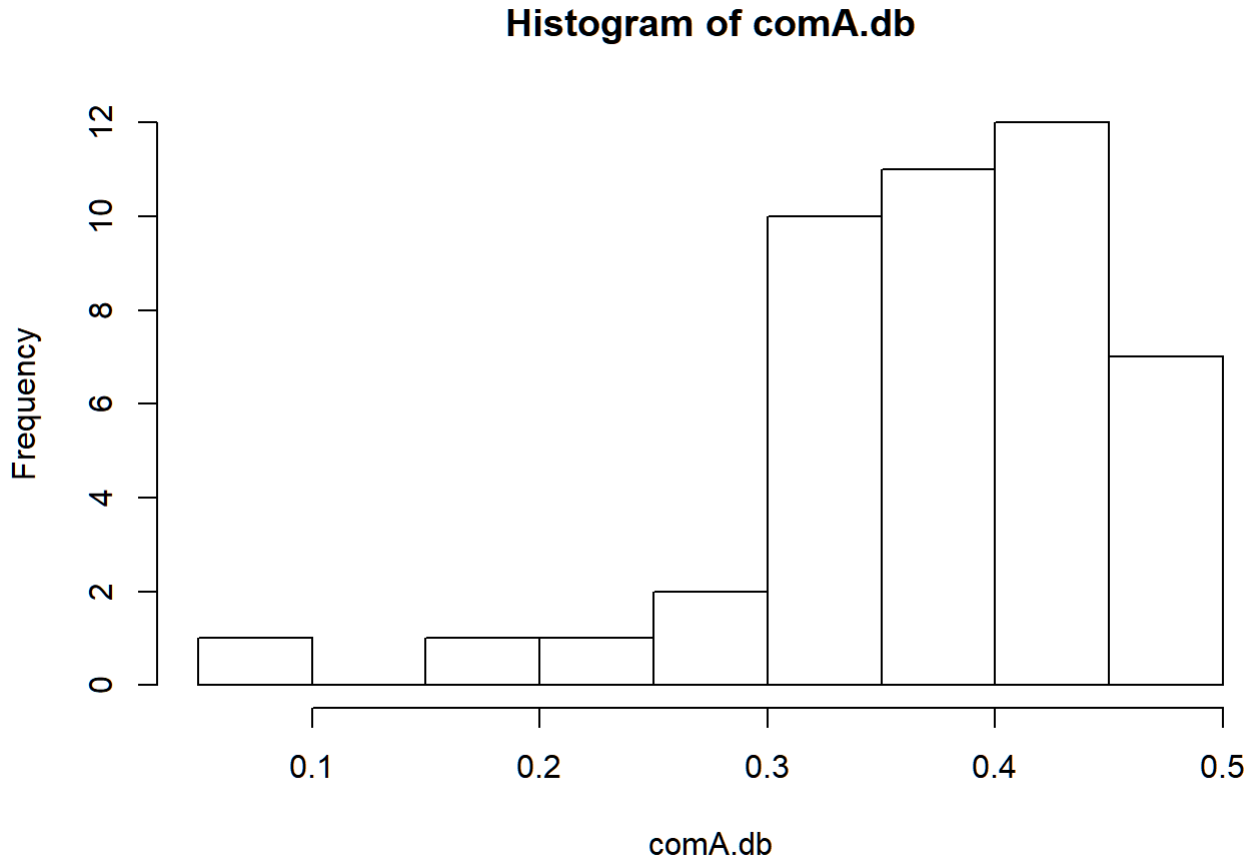
Doesn't look like it from the historgrams because the limit of Sorensen index caps the values. To confirm, an F-test will do.

```
var.test(comA.ds,comB.ds)
```

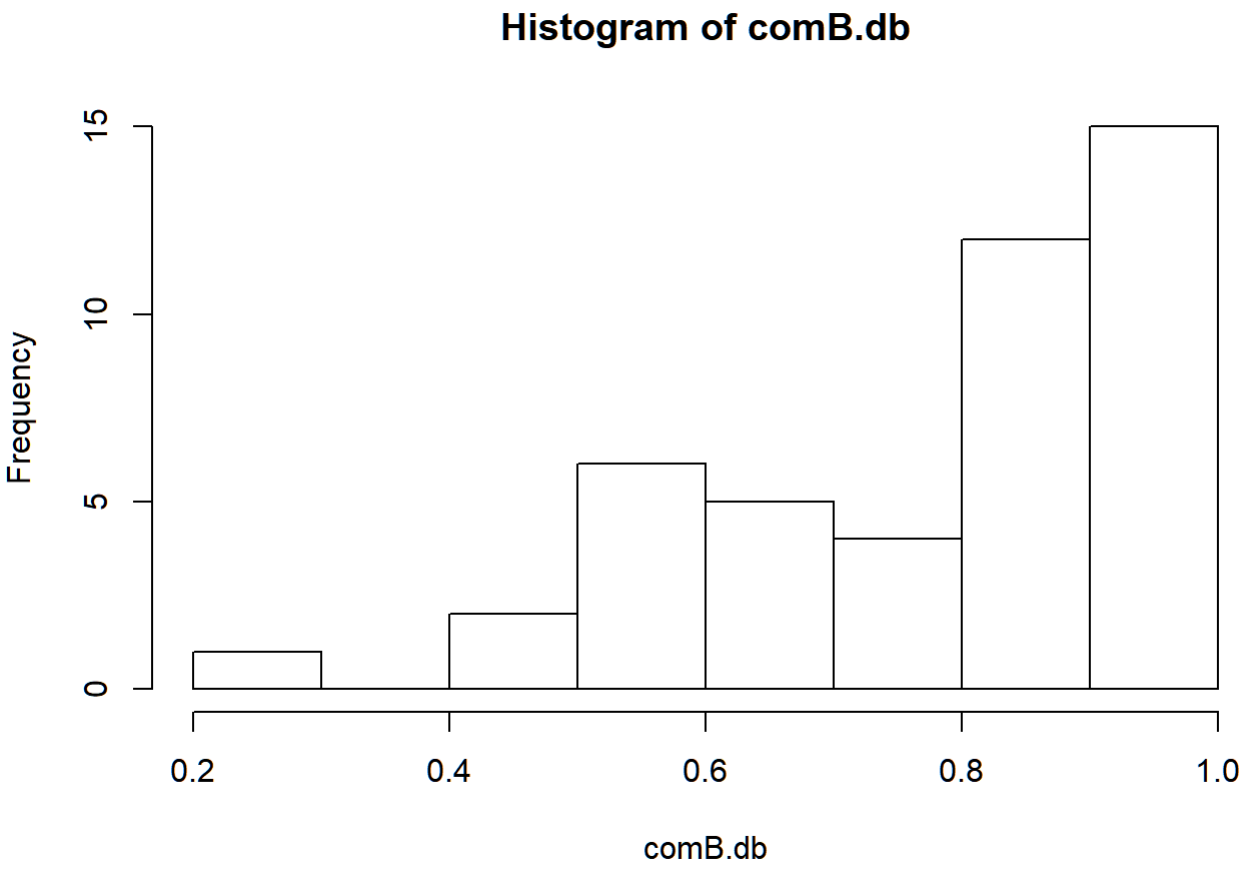
```
##
##  F test to compare two variances
##
## data:  comA.ds and comB.ds
## F = 0.15114, num df = 44, denom df = 44, p-value = 5.061e-09
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.0830551 0.2750231
## sample estimates:
## ratio of variances
##          0.1511359
```

The F-test says that the variances between the datasets are different meaning that the t-test is not appropriate. Let's repeat for BC index.

```
hist(comA.db,breaks=10)
```




```
hist(comB.db,breaks=10)
```




```
var.test(comA.db,comB.db)
```

```
##
##  F test to compare two variances
##
## data:  comA.db and comB.db
## F = 0.19777, num df = 44, denom df = 44, p-value = 3.576e-07
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.1086831 0.3598861
## sample estimates:
## ratio of variances
##          0.1977714
```

Again we see that the t-test is inappropriate. Probably some other test could be used but I don't know frequentist statistics that well. From looking at the histograms, the pairwise comparisons look very different, with the clustered communities having more communities that are completely different from one another.

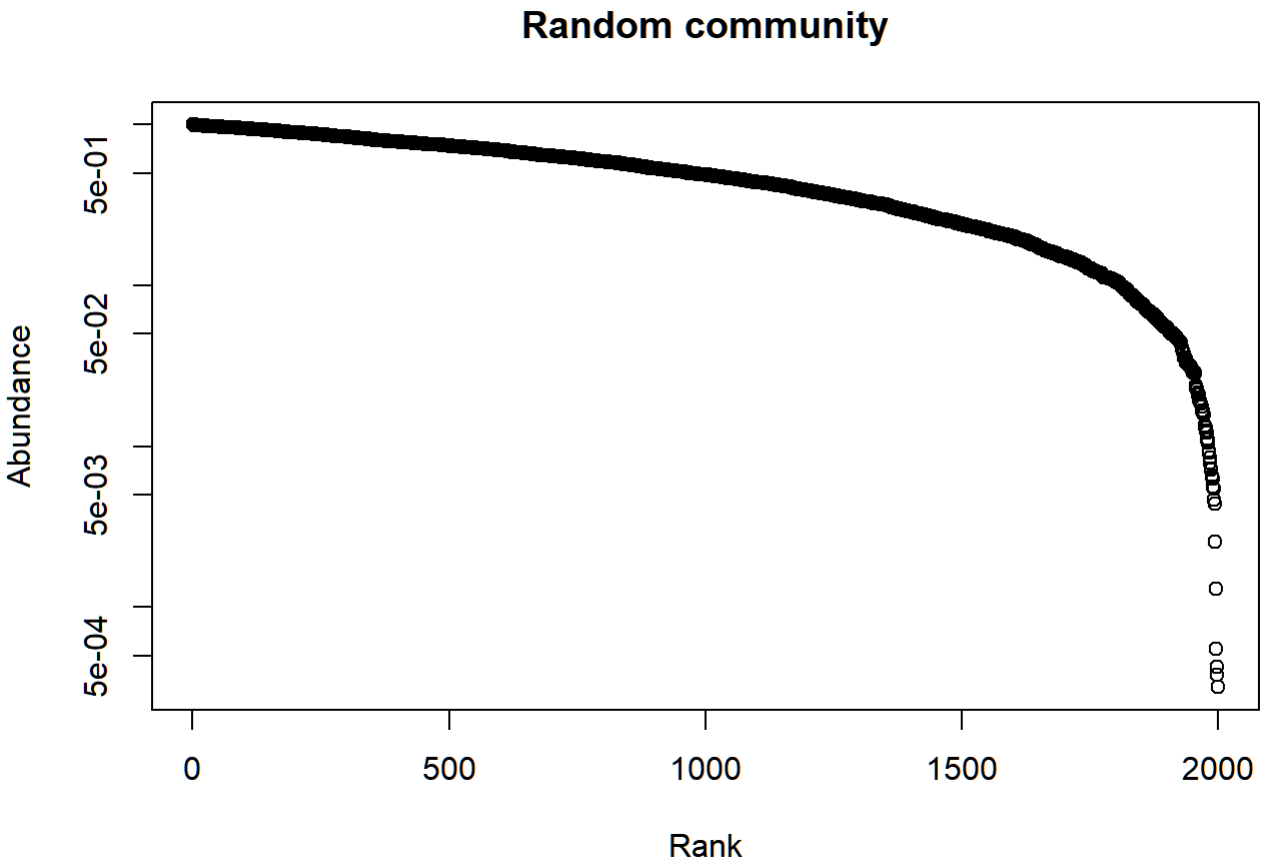
Finally, compare the compositional similarity of site 1 and site 2 to the source community?

```
RACcomA <- rad.lognormal(comA$census)
```

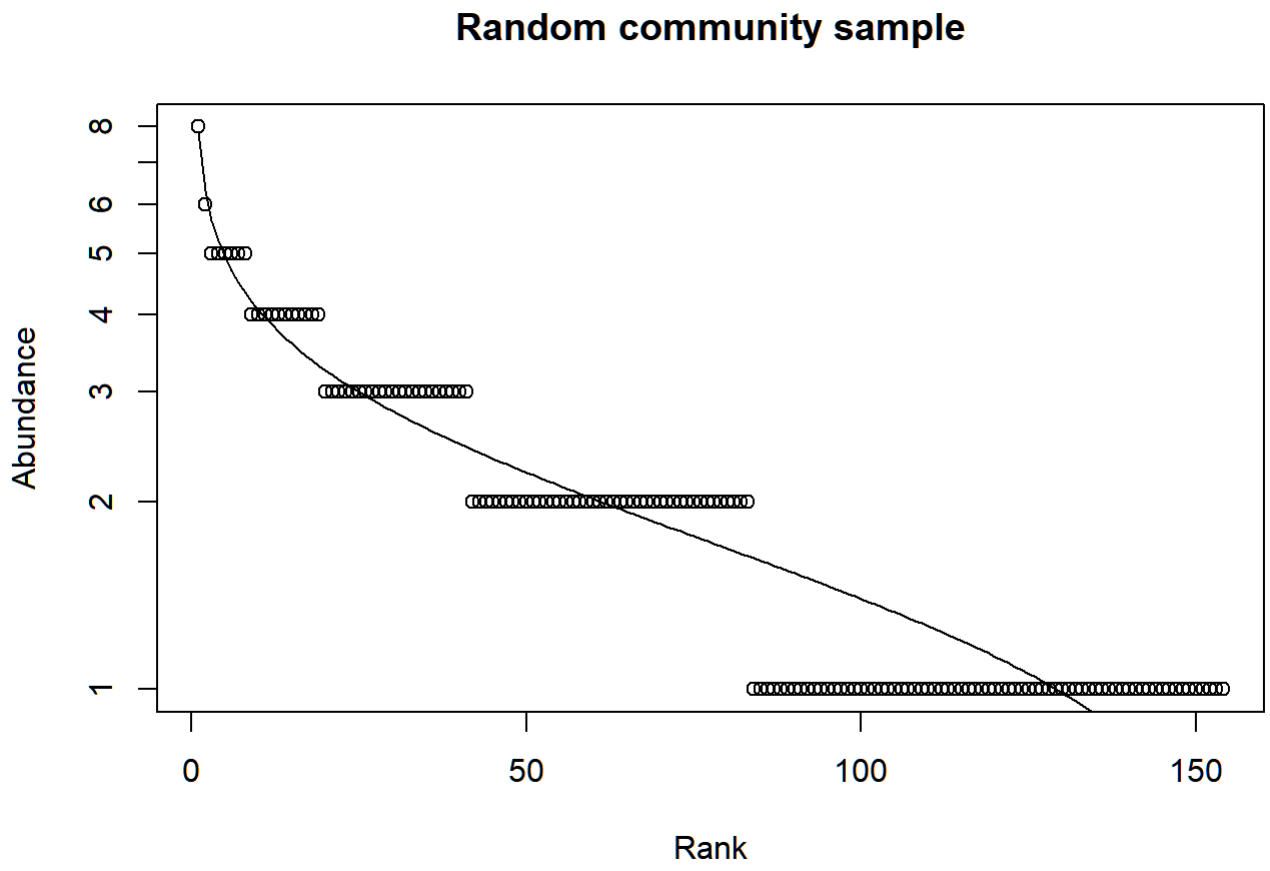
```
## Warning in Ops.factor(left, right): '>' not meaningful for factors
```

```
## Error in y + 0.1 : non-numeric argument to binary operator
```

```
plot(RACcomA, main = "Random community")
```




```
RACcomAsample <- rad.lognormal(comm_matA$spec_dat)
plot(RACcomAsample, main = "Random community sample")
```

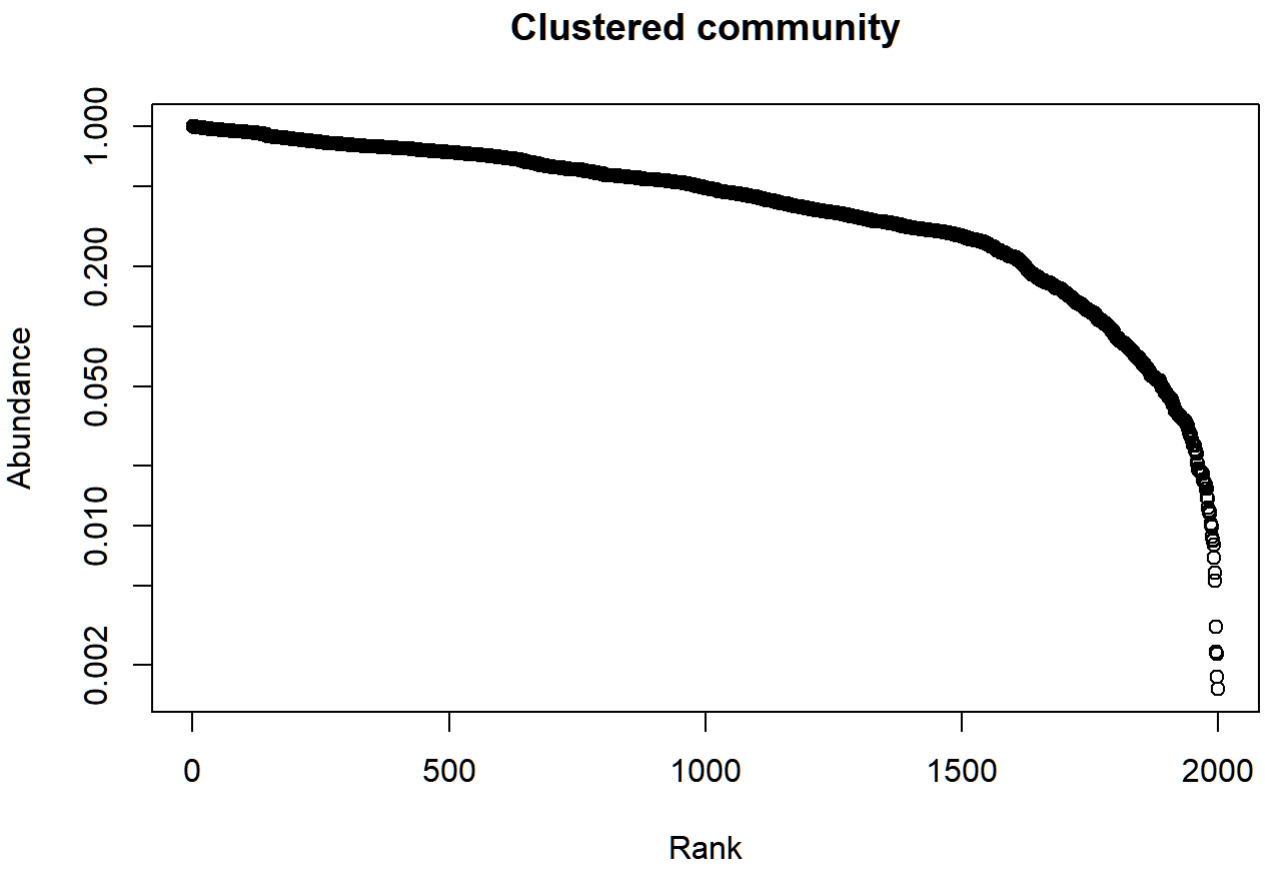



```
RACcomB <- rad.lognormal(comB$census)
```

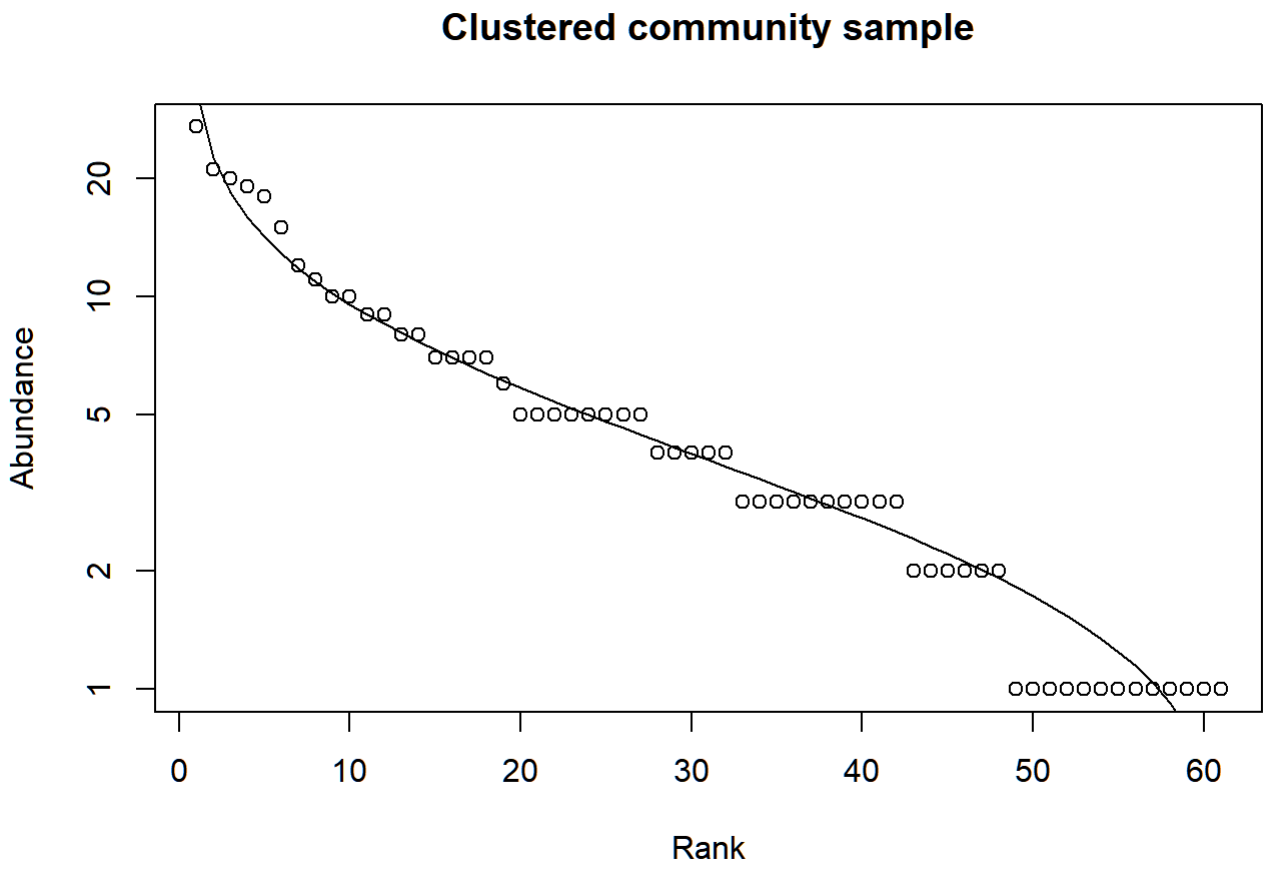
```
## Warning in Ops.factor(left, right): '>' not meaningful for factors
```

```
## Error in y + 0.1 : non-numeric argument to binary operator
```

```
plot(RACcomB, main = "Clustered community")
```



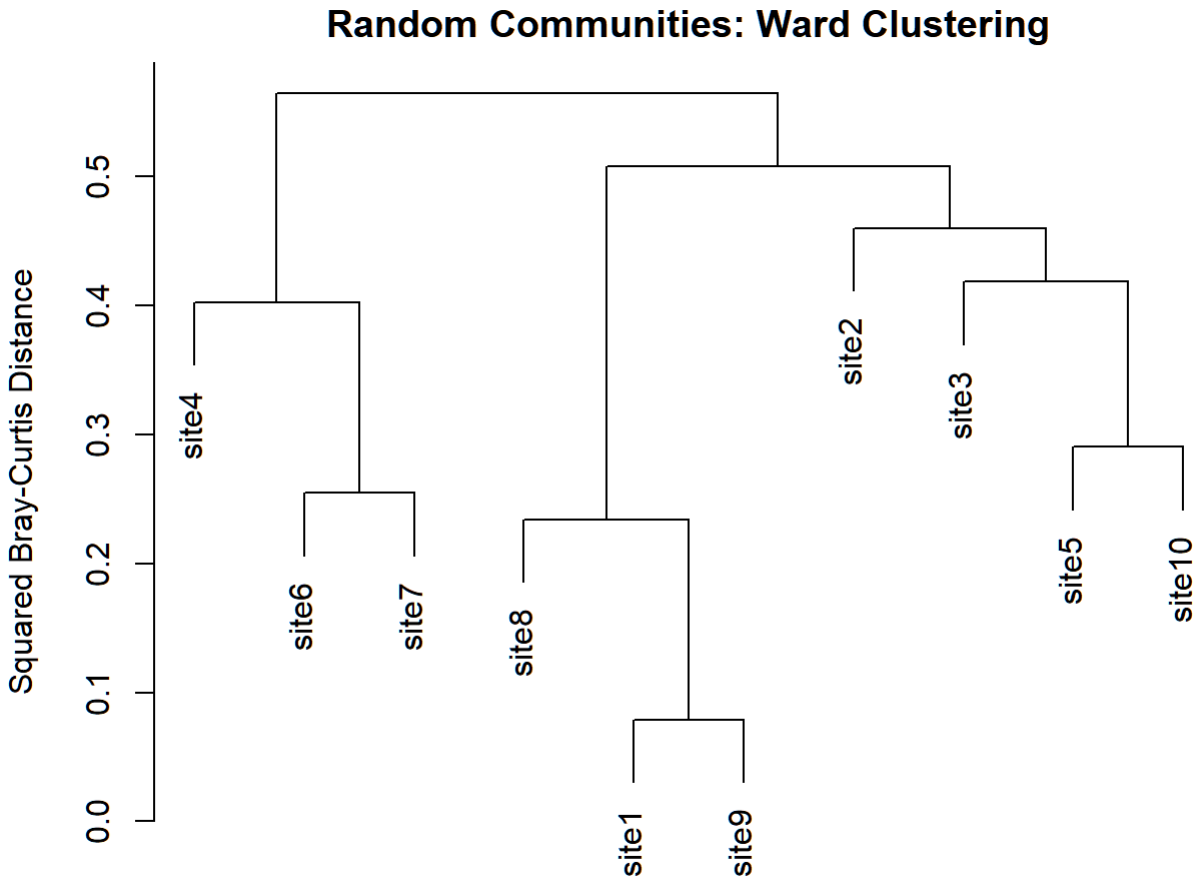

```
RACcomBsample <- rad.lognormal(comm_matB$spec_dat)
plot(RACcomBsample, main = "Clustered community sample")
```



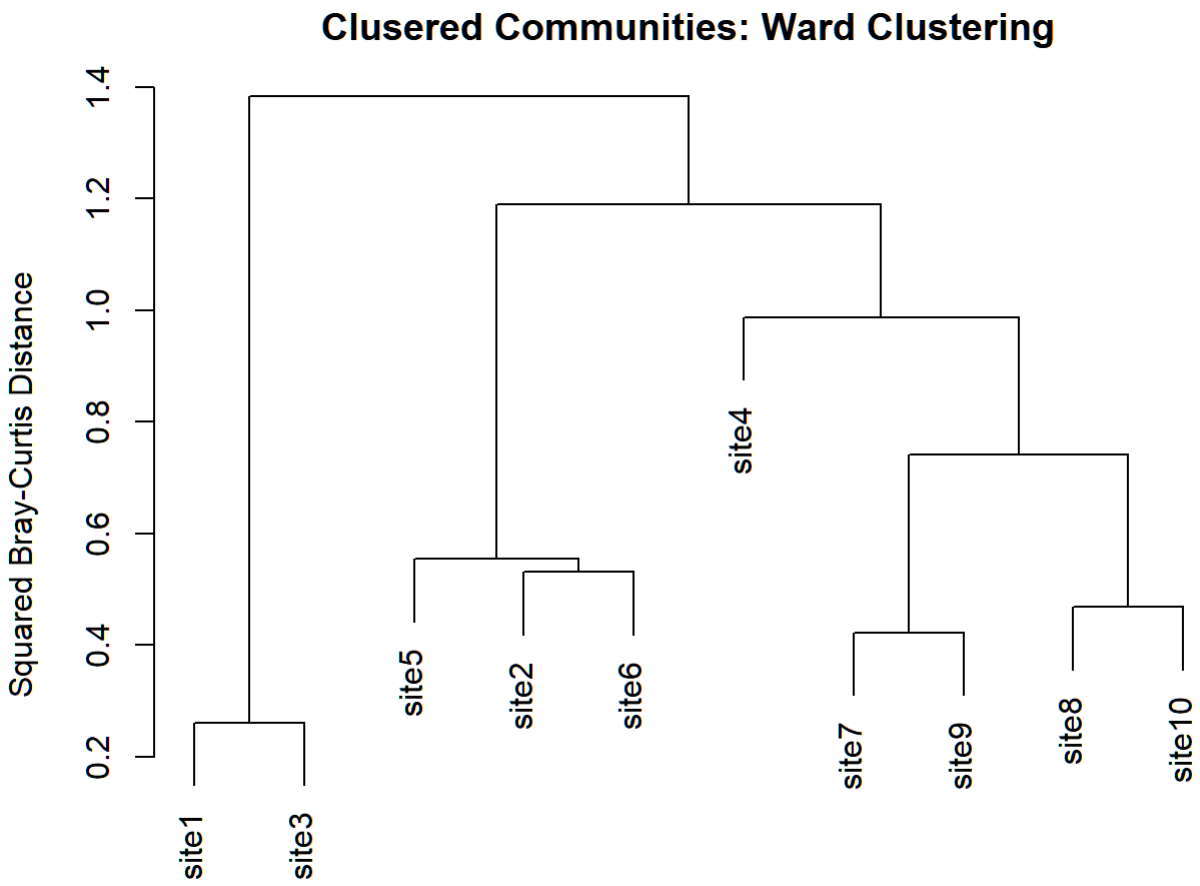
We see that the rank abundance for the actual communities are most likely a geometric/power law distribution. With the sampling we see that the probability of observing a single observation of a species is much lower in the clustering case and that there are fewer species observed in total, this makes sense because of how the communities are generated.

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.

```
comA.ward <- hclust(comA.db, method = "ward.D2")
par(mar = c(1,5,2,2)+0.1)
plot(comA.ward,main = "Random Communities: Ward Clustering",ylab="Squared Bray-Curtis Distance")
```




```
comB.ward <- hclust(comB.db, method = "ward.D2")
par(mar = c(1,5,2,2)+0.1)
plot(comB.ward,main = "Clusered Communities: Ward Clustering",ylab="Squared Bray-Curtis Distanc
e")
```



We see that by looking at the axis values, that the clustered communities are farther apart from one another in composition than the random communities, though it does appear that we have some communities that are close to one another. It would probably be good to do other followup on this, such as does the BC distance correlate to other aspects, such as the Euclidian distance between sites.