CHAPTER 44
Diversity Analysis using R

A diversity index is a numerical measure that quantifies the number of distinct types (such as species) in a dataset (a community) while also accounting for evolutionary relationships among the individuals distributed throughout those types, such as richness, divergence, and evenness. These indicators are numerical representations of biodiversity in a variety of ways (richness, evenness, and dominance). The amount of distinct species present in a community is referred to as species diversity (a dataset).

The effective number of species is the number of equally abundant species required to achieve the same mean proportional species abundance as seen in the dataset under consideration (where all species may not be equally abundant). Using diversity analysis, questions like "how many species are in a sample?" and "how similar are these two samples?" are investigated. The number of species recorded within a region is referred to as alpha diversity, while beta diversity is defined as the number of species not common to the two regions being compared is referred to as beta diversity and gamma diversity is defined as the total number of species within all regions. Species richness, taxonomic or phylogenetic diversity, and/or species evenness are all examples of species diversity. The term "species richness" refers to the number of species present. The genetic link between distinct groupings of animals is taxonomic or phylogenetic diversity. Species evenness measures how evenly the species' abundances are distributed.

Several packages are available in R for calculating the diversity indices, and the vegan package is more popular.

The "vegan" Package in R

To install Vegan package

```
install.packages("vegan")
```

The majority of diversity approaches presume that data is in the form of individual counts. Other data types are employed in the procedures, and some claim that biomass or cover are better than counts of individuals of varying sizes.

This package uses the data set with stem counts of trees on 1 ha plots in the Barro Colorado Island.

To view the data used:
library (vegan)
data("BCI")
fix(BCI)

1. Diversity Indices

The Shannon index is calculated with:
H <- diversity(BCI)

The evenness (equitability) an be obtained using Pielou's evenness index and can be obtained using:
J <- H/log(specnumber(BCI))

The R´enyi diversities can be calculated using:
# to select six locations randomly from the data set
k <- sample(nrow(BCI), 3)

# R`enyi diversities
R <- renyi(BCI[k,])
plot(R)

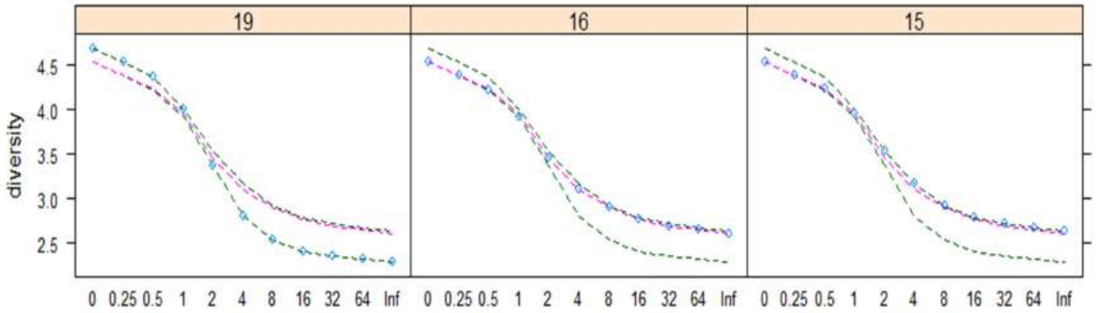


Figure 1: R´enyi diversities in 3 randomly selected plots. The dots represents the values for sites, and the lines the extremes and median in the data set.

A site is more diverse if all of its R´enyi diversities are higher than another site.

Fisher's alpha diversity index:
alpha <- fisher.alpha(BCI)

Species richness rises with sample size, and discrepancies in richness may result from sample size differences. One option is to strive to rarefy species richness while maintaining the same number of individuals to address this issue.

To express richness for the same number of individuals:

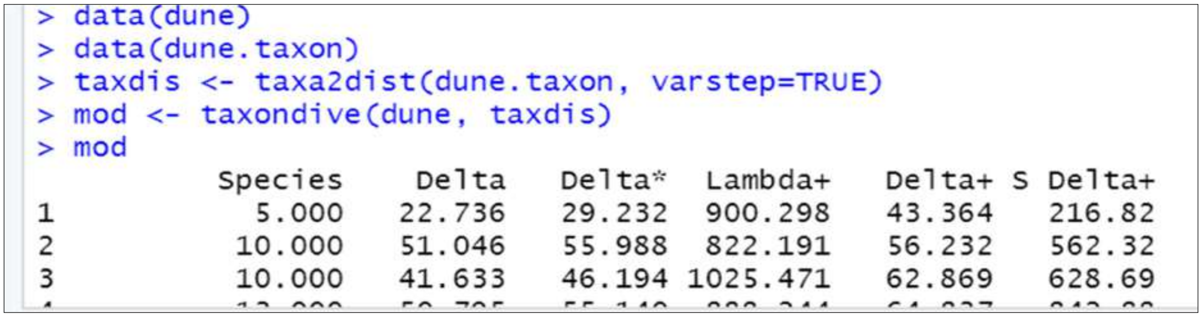
Srar <- rarefy(BCI, min(rowSums(BCI)))

Simple diversity indices consider species identity: all species are equally unique. Taxonomic and functional diversity indexes, on the other hand, assess the distinctions between species. Although taxonomic and functional diversities are utilised in distinct disciplines of science, they both follow the same logic and can be used to taxonomic or functional properties of species.

2. Taxonomic Diversity

In taxonomic diversity the primary data were taxonomic trees which were transformed to pairwise distances among species.

```
data(dune)
data(dune.taxon) # Taxomic trees
taxdis <- taxa2dist(dune.taxon, varstep=TRUE)
mod <- taxondive(dune, taxdis)
mod
```



```
> data(dune)
> data(dune.taxon)
> taxdis <- taxa2dist(dune.taxon, varstep=TRUE)
> mod <- taxondive(dune, taxdis)
> mod
          Species    Delta   Delta*  Lambda+   Delta+ S Delta+
1           5.000   22.736   29.232  900.298   43.364   216.82
2          10.000   51.046   55.988  822.191   56.232   562.32
3          10.000   41.633   46.194 1025.471   62.869   628.69
```

Figure 2: R output

3. Functional Diversity

In functional diversity the data associated with species attributes are translated to pairwise distances among species and futher grouping them.

```
tr <- hclust(taxdis, "aver")
mod <- treedive(dune, tr)
```

4. Species abundance models

Diversity indices can be thought of as variance measures for species abundance distribution. One might want to look at abundance distributions more closely. Vegan includes routines for Fisher's log-series and Preston's log-normal models and various species abundance distribution models.

```
#Species abundance models
k <- sample(nrow(BCI), 1)
fish <- fisherfit(BCI[k,]) # Fisher's log-series
fish
```

```
plot(fish)
```

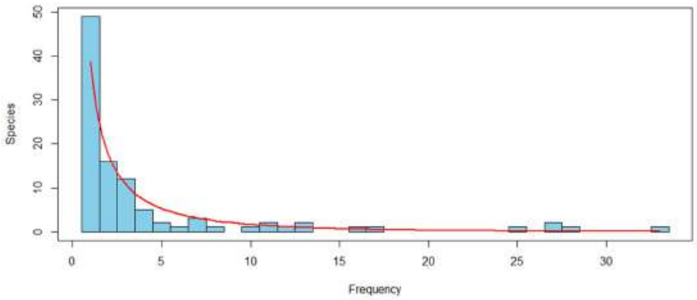


Figure 3: The result of Fisher's log-series fitted to one randomly selected site (Site number=47).

```
Fisher log series model
No. of species: 102
Fisher alpha:   42.56011
```

In Preston's log-normal model, instead of plotting species by frequency, it divides them into increasing frequency groupings. As a result, upper bins with a wide range of frequencies become more prevalent, and the result can resemble a Gaussian distribution truncated on the left in appearance.

```
prest<-prestondistr(BCI[k,])  # Preston's log-normal model
prest
plot (prest)
```

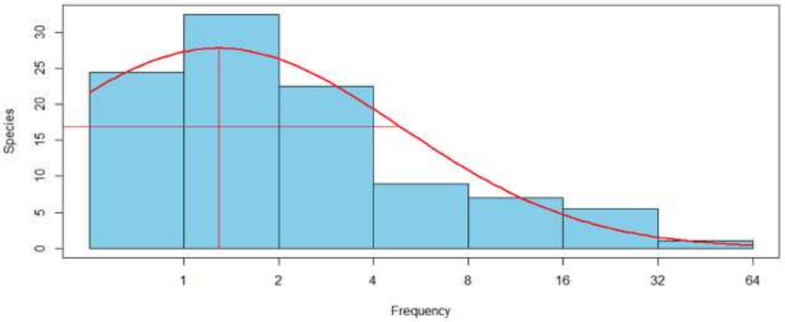


Figure 4: Preston's log-normal model fitted to one randomly selected site (47).

5. Ranked abundance distribution

```
rad <- radfit(BCI[k,]) # ranked abundance
rad
#plot(rad)
radlattice(rad)
```

```
RAD models, family poisson
No. of species 102, total abundance 425

           par1      par2     par3    Deviance AIC      BIC
Null                                  105.2750 384.2921 384.2921
Preemption  0.045509                  81.6840 362.7010 365.3260
```

```
Lognormal   0.7421    1.1905          43.1745 326.1916 331.4415
Zipf        0.1409   -0.85907         48.6464 331.6634 336.9134
Mandelbrot  2.017    -1.5363   6.7022   9.4122 294.4292 302.3042
```

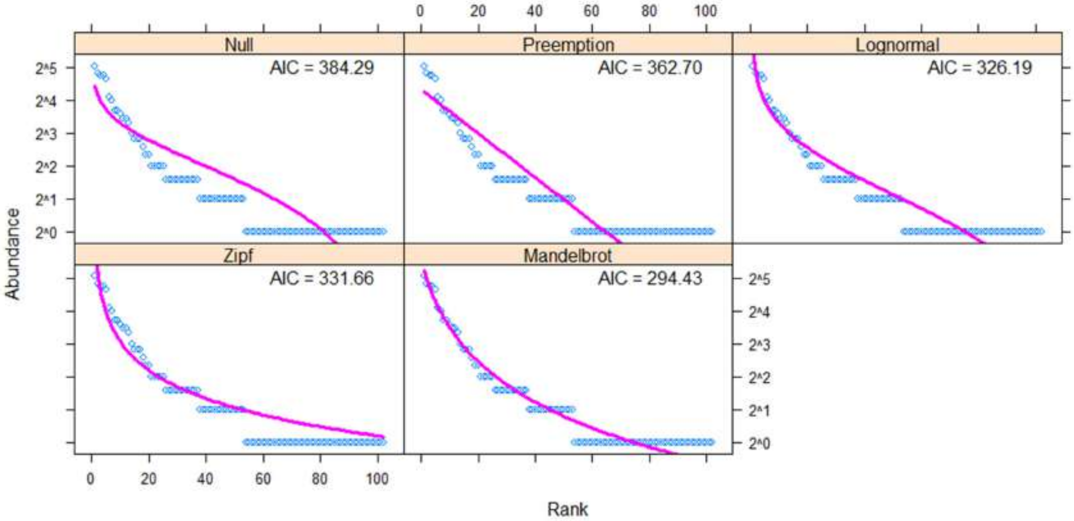


Figure 5: Ranked abundance distribution models for a random plot (no. 47). The best model has the lowest AIC.

6. Species accumulation models

Species accumulation models are similar to rarefaction in that they look at how species accumulate as the number of sites grows. There are a few other options, such as gathering sites in the order they appear and repeating the process randomly.

The recommended is Kindt's exact method

```
sac <- specaccum(BCI) # species accumulation model (Kindt's exact method)
plot(sac, ci.type="polygon", ci.col="green")
```

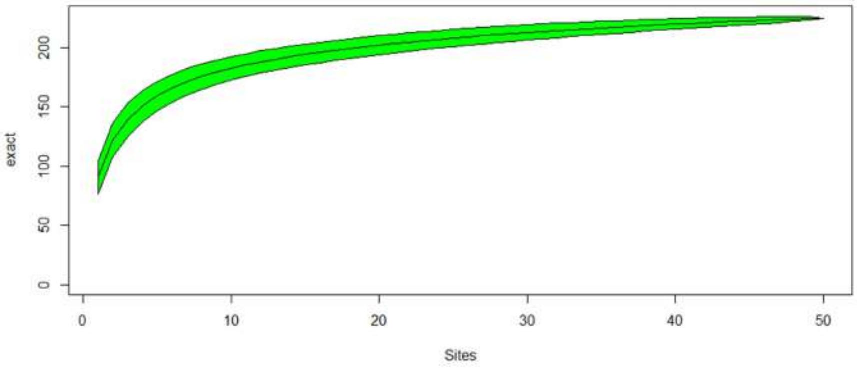


Figure 6: Species accumulation using Kindt's exact method

7. Beta diversity

The most fundamental diversity indices are alpha diversity indices. Whittaker (1960 and 1965) classified diversity into several categories. The most well-known are alpha diversity (diversity in a single location) and beta diversity (diversity over gradients). Although beta diversity should be explored in relation to gradients (Whittaker, 1960 & 1965), practically everyone thinks of it as a measure of general heterogeneity: how many more species are there in a collection of sites than in an average site.
The best-known beta diversity index is based on the ratio of total number of species in a group of sites S to average richness per site α.

```
#Beta diversity
ncol(BCI)/mean(specnumber(BCI)) – 1
```

To know the details of different beta diversities use the function:

```
betadiver(help=TRUE)
```

```
z <- betadiver(BCI, "z")  # To get the diversity measure "z"
```

8. Cluster Analysis (Unsupervised learning)

Unsupervised learning is a machine learning method used to make conclusions from datasets containing unlabeled input data. Cluster analysis is the most frequent unsupervised learning method used for exploratory data analysis to uncover hidden patterns or groupings in data.

Cluster analysis is used to aggregate instances into groups when the group membership is unknown before the study. Cluster analysis is a method for classifying individuals or objects into previously unidentified groups.

8.1 Clustering Methods (Johnson and Wichern, 2006)

The clustering methods commonly used are fall into two general categories.

(i) Hierarchical and
(ii) Non hierarchical.

8.1.1 Hierarchical cluster Analysis

Either a sequence of mergers or a series of sequential divisions is used in hierarchical clustering algorithms. The agglomerative hierarchical technique begins with individual objects, there are as many clusters as there are items. The most similar objects are grouped first, and these groupings are then combined based on their commonalities. As the resemblance between subgroups declines, they eventually merge into a single cluster.

Divisive hierarchical approaches work the other way around. A single group of items is split into two subgroups, with the objects in one subgroup being separated from the ones in the other. These subgroups are then separated into distinct subgroups. The process continues until the number of subgroups equals the number of items or each object forms a group. The findings of both the agglomerative and divisive methods can be shown as a Dendrogram, a two-dimensional figure. The Dendrogram can be seen to depict the mergers or divisions that have occurred at successive levels.

Linkage methods can be used to cluster both items and variables. This isn't always the case with hierarchical agglomerative procedures. The following linking types are now discussed:

(i) Single linkage (minimum distance or nearest neighbour),
(ii) Complete linkage (maximum distance or farthest neighbour) and
(iii) Average linkage (average distances).

Other hierarchical clustering techniques, such as Ward's and Centroid methods, are also documented in the literature.

Hierarchical Cluster analysis: Agglomerative Clustering steps

The steps involved in the agglomerative hierarchical clustering algorithm for groups of N objects (items or variables) are as follows:

(i) Begin with N clusters, each of which contains a single entity and a N×N symmetric distance (or similarity) matrix **D** = $\{d_{ik}\}$.
(ii) Look up the closest (most similar) pair of clusters in the distance matrix. Let d_{uv} be the distance between the two most comparable clusters U and V.
(iii) Combine the U and V clusters. The newly formed cluster should be labelled (UV). Remove the rows and columns pertaining to clusters U and V from the distance matrix and replace them with a row and column indicating the distances between cluster (UV) and the other clusters.
(iv) Repeat steps (ii) and (iii) N-1 times more (All objects will be in a single cluster after the algorithm terminates). Keep track of the merged clusters' identities as well as the levels (distances or similarities) at which they merged.

8.1.2 Non-Hierarchical Clustering Method

Non-hierarchical clustering approaches group things into a collection of K clusters rather than variables. The number of clusters, K, can be set ahead of time or decided during the clustering process. Because the basic data does not need to be saved and a distance matrix does not need to be calculated during the computer run. Non-hierarchical approaches can handle far larger data sets than hierarchical methods can. Non-hierarchical techniques begin with either (1) an initial grouping of items or (2) an initial set of seed points that will form the cluster's nucleus.

8.1.2.1 K means Clustering (Afifi, Clark and Marg, 2004)

The K means clustering is a popular non-hierarchical clustering method. The algorithm proceeds in the following steps for a specified number of clusters K:

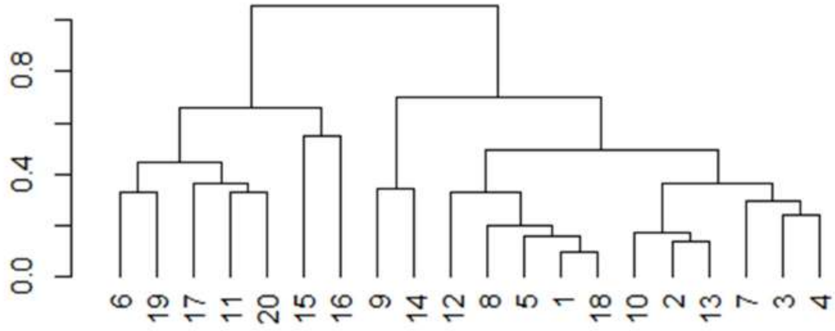
(i) First, divide the data into K clusters. The number of clusters can be set by the user or chosen by the computer according to a random approach.
(ii) Determine the K clusters' means or centroid.
(iii) Calculate the distance between each case's centroid. Leave the case in its own cluster if it is closest to the centroid; otherwise, reassign it to the cluster whose centroid is closest to it.
(iv) For each scenario, repeat step (iii).
(v) Repeat steps (ii), (iii), and (iv) until there are no more cases to assign.

8.2. Dendrogram

The relative size of the proximity coefficients at which cases are joined is shown in a dendrogram, also known as a hierarchical tree diagram or plot. The greater the distance coefficient or, the smaller the similarity coefficient, the more clustering is required, which may be undesirable. Low-distance cases are close together, with a line connecting them a short distance from the left of the Dendrogram, indicating that they have been grouped into a cluster with a low distance coefficient, indicating similarity. When the linking line is to the right of the Dendrogram, on the other hand, the linkage occurs at a high distance coefficient, showing that the cases/clusters were agglomerated despite their differences.

R code for getting a simple dendrogram:

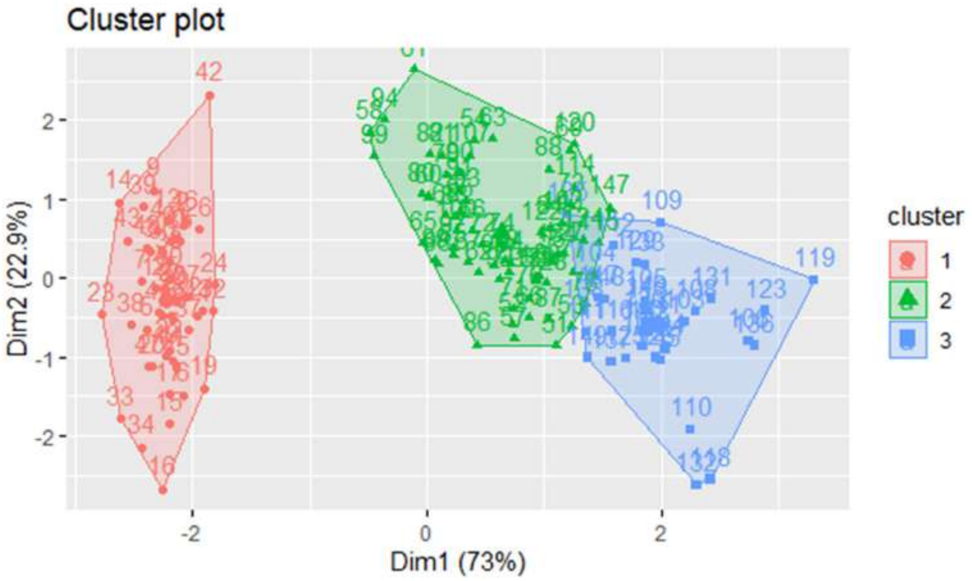
```
attach(iris)
iris1<-iris[1:20 ,-5]  # For selecting a subset data
dist <- dist(iris1, method = "euclidean")
hclust_avg <- hclust(dist, method = 'average')
hcd <- as.dendrogram(hclust_avg)
plot(hcd, main="Main")
```



8.3. PCA based clustering

The R code for the PCA based clustering:

```
library(factoextra)
attach(iris)
iris2<-iris[ ,-5]
dist <- dist(iris2, method = "euclidean")
hclust_avg <- hclust(dist, method = 'average')
sub_grp <- cutree(hclust_avg, k = 3)
fviz_cluster(list(data = iris2, cluster = sub_grp))
```



8.4. Distance Measures

Some distance measures commonly used for assessing spectral similarity/dissimilarity are as follows:

1) Euclidian Distance
2) Mahalanobis D^2
3) City-Block Distance

Some of the R functions used for computing distances between pairs of observations:

- **dist()** R base function [stats package]
- **get_dist()** function [factoextra package]

Compared to the standard dist() function, it supports correlation-based distance measures including "pearson", "kendall" and "spearman" methods.

- **daisy()** function [cluster package]: It can handle different variable types (e.g. nominal, ordinal, (a)symmetric binary). In that case, the Gower's coefficient will be automatically used as the metric. It's one of the most popular proximity measures for mixed data types. Details on the function can be obtained from the R documentation of the daisy() function (?daisy).

For example for Euclidean distance

```
dist.eucl <- dist(data, method = "euclidean")
```

Some of the methods are **"euclidean", "maximum", "manhattan", "canberra", "binary", "minkowski"**

For visualization of distances, following package can be used:

```
library(factoextra)
attach(iris)
iris1<-iris[1:20 ,-5]  # For selecting a subset data
dist <- dist(iris1, method = "euclidean")
fviz_dist(dist, gradient= list(low="green",mid= "white",high= "red"))
```

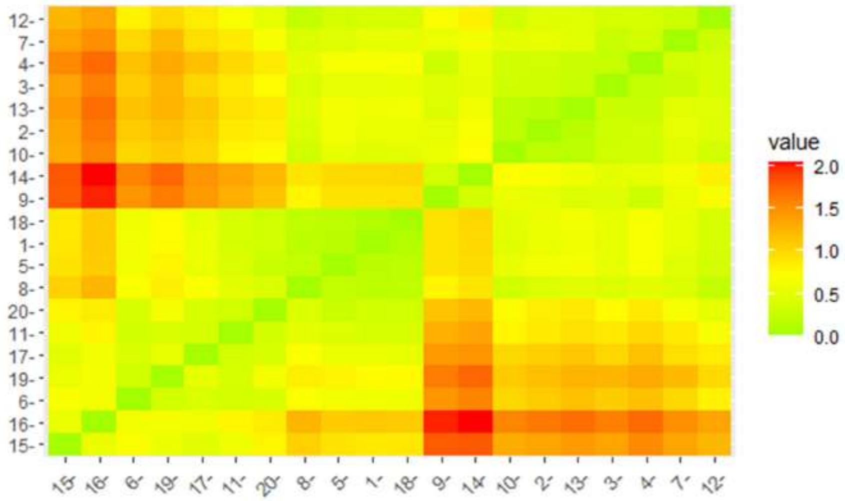


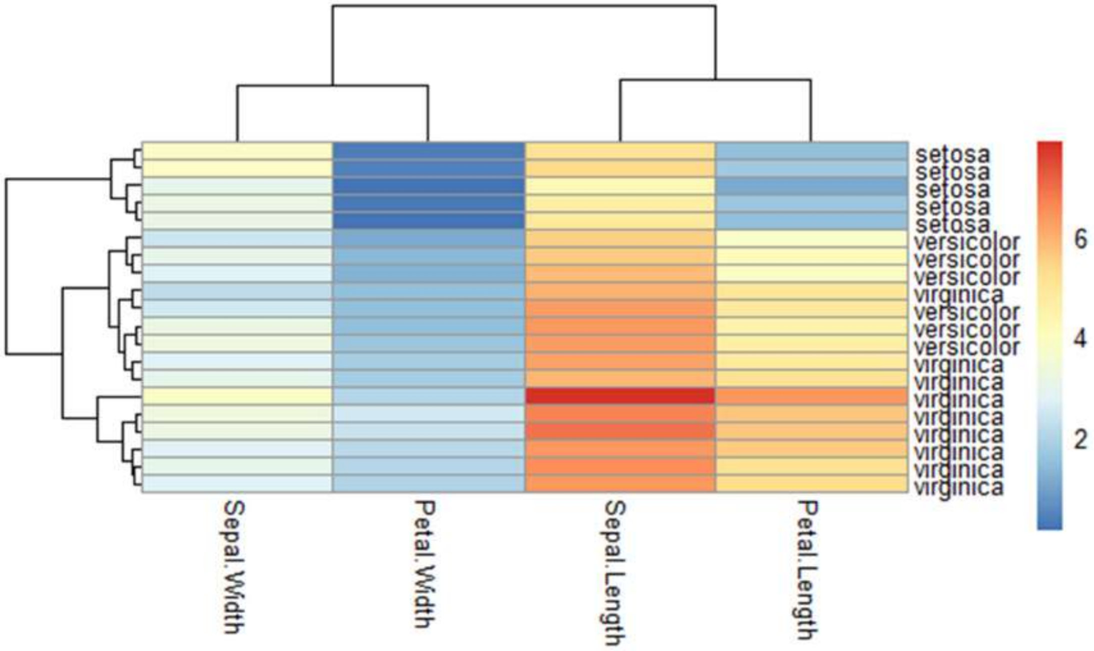
Fig. Distance plot

8.6 Heat Map

A heat map is a data visualization technique that shows the magnitude of a phenomenon as color in two dimensions. "pheatmap" function can draw clustered heatmaps.

The R code for heat map

```
library("pheatmap")
ss <- sample(1:150, 20)   # 30 rows randomly
df <- iris[ss, ]
df1<-as.matrix(df[ ,-5])
rownames(df1)<-as.matrix(df[ ,5])
pheatmap(df1)
```



9. Discriminant Function Analysis (Supervised learning)

Discriminant function analysis is a statistical technique that uses one or more continuous or binary independent variables to predict a categorical dependent variable (also known as a grouping variable) (called predictor variables). Sir Ronald Fisher created the first dichotomous discriminant analysis in 1936. Discriminant function analysis can be used to see if a group of variables is good at predicting membership in a category. Discriminant analysis is utilized when groups are known a priori (unlike in cluster analysis). A score on one or more quantitative predictor measures and a score on a group measure are required for each instance. In simple terms, discriminant function analysis is the act of grouping, classifying, or categorising things into similar groups, classes, or categories.

The assumptions of discriminant analysis are the same as those for MANOVA. The analysis Discriminant analysis is based on the same assumptions as MANOVA. Outliers can be a serious impact on the results and size of the smallest group must be larger than the number of predictor variables. The following are the main assumptions:

- Multivariate normality: For each level of the grouping variable, independent variables are normal.
- Homogeneity of variance/covariance (homoscedasticity): The Box's M statistic can be used to see if the variances of group variables are the same across levels of predictors.
- However, it has been proposed that when covariances are equal, linear discriminant analysis be used, and when covariances are not equal, quadratic discriminant analysis be employed.
- Multicollinearity: As the correlation between predictor variables increases, predictive power decreases.
- Independence: Participants are randomly selected, and a participant's score on one measure is believed to be independent of all other participants' scores on that variable.

It has been proposed that discriminant analysis is reasonably resilient to minor violations of these assumptions, and that discriminant analysis can still be reliable when utilising dichotomous variables (where multivariate normality is often violated).

The discriminant analysis creates a new variable for each function by combining one or more linear combinations of predictors. Discriminant functions are the name given to these functions. The number of functions that can be used is either Ng-1 (number of groups) or p (number of predictors), whichever is less. On that function, the first function maximises the differences across groups. The second function maximises differences on that function, but it can't be associated with the first. This process is repeated for subsequent functions, with the exception that the new function must not be connected with any of the preceding functions.

The following packages and codes are useful for running linear discriminant function analysis in R:

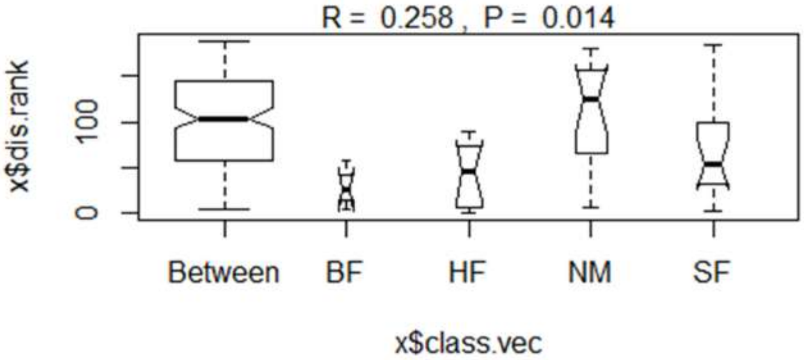
```
library(MASS)
library(tidyverse)
library(caret)
model <- z <- lda(Sp ~ ., Iris, prior = c(1,1,1)/3, subset = train)
```

Further, prediction of group membership and plotting of the membership can also be done.

10. Analysis of Similarities (ANOSIM) using anosim()

Analysis of similarities (ANOSIM) is used to test the significant difference between two or more groups of sampling units.

```
data(dune)
data(dune.env)
dune.dist <- vegdist(x, method="bray", binary=FALSE, diag=FALSE, upper=FALSE,
                na.rm = FALSE)
# method: Dissimilarity index, partial match to "manhattan", "euclidean", "canberra",
"clark", "bray", "kulczynski", "jaccard", "gower", "altGower", "morisita", "horn",
"mountford", "raup", "binomial", "chao", "cao", "mahalanobis", "chisq" or "chord".
attach(dune.env)
dune.ano <- anosim(dune.dist, Management)
summary(dune.ano)
plot(dune.ano)
```



11. Non-metric Multidimensional scaling using metaMDS()

Function metaMDS performs Nonmetric Multidimensional Scaling (NMDS). it standardizes the scaling in the result, so that the configurations are easier to interpret, and adds species scores to the site ordination. The metaMDS function does not provide actual NMDS, but it calls another function for the purpose.

```
mds <- metaMDS(dune, distance = "bray", k = 2)
plot(mds, display = c("sites", "species"))
```

References

- Afifi, A., Clark, V. A. and Marg, S. (2004). *Computer Aided Multivariate Analysis.* USA, Chapman & Hall.
- Anderson, T. W., 1984, An Introduction to applied Multivariate Statistical Analysis, *John Wiley & Sons,* New York.
- Chatfield, C. and Collins, A. J. (1990). *Introduction to Multivariate Analysis.* Chapman and Hall Publications.

- Varghese, E. and George, G. (2017) Classification Techniques for Remotely Sensed Data. In the *Course Manual Winter School on Structure and Functions of Marine Ecosystem: Fisheries* (Eds. Mini, K G, Kuriakose, Somy and Sathianandan, T V, Shafeeque, Muhammed, Monolisha, S , Minu, P and George, Grinson). CMFRI Lecture Note Series No. 12/2017. ISBN-978-93-82263-18-0
- Hair, J. F., Anderson R. E., Tatham, R. L. and Black, W. C. (2006). *Multivariate Data Analysis.* 5th Edn., Pearson Education Inc.
- https://cran.r-project.org/web/packages/vegan/vignettes/diversity-vegan.pdf
- Johnson, R. A. and Wichern, D. W. (2006). *Applied Multivariate Statistical Analysis.* 5th Edn., London, Inc. Pearson Prentice Hall.
- Sathianandan, T. V., Mohamed, K. S. and Vivekanandan, E. (2012). Species diversity in fished taxa along the southeast coast of India and the effect of the Asian Tsunami of 2004 , Mar Biodiv (2012) 42:179–187
- Timm, N. H. (2002). *Applied Multivariate Analysis*. 2nd Edn. , New York, Springer-Verlag
- Whittaker RH (1960). "Vegetation of Siskiyou mountains, Oregon and California." Ecological Monographs, 30, 279–338.
- Whittaker RH (1965). "Dominance and diversity in plant communities." Science, 147, 250–260.