

Applied_statistics_HW1

Ganaeva Daria

01 04 2024

1. Install necessary packages

```
#install.packages("vegan")  
# install.packages("devtools")  
#devtools::install_github("gavinsimpson/ggvegan")
```

2. Upload necessary packages

```
library(vegan)  
library(ggvegan)
```

3. Upload the data

```
data(BCI)  
data(BCI.env)
```

4. Exploratory data analysis (explore data structure; check for missing values, outliers; collect summary data; filter out low abundance and non-significant data)

```
#head(BCI)  
head(BCI.env)
```

```
##   UTM.EW  UTM.NS Precipitation Elevation Age.cat Geology  Habitat Stream EnvHet  
## 1 625754 1011569          2530        120    c3      Tb  OldSlope  Yes 0.6272  
## 2 625754 1011669          2530        120    c3      Tb   OldLow  Yes 0.3936  
## 3 625754 1011769          2530        120    c3      Tb   OldLow  No  0.0000  
## 4 625754 1011869          2530        120    c3      Tb   OldLow  No  0.0000  
## 5 625754 1011969          2530        120    c3      Tb  OldSlope  No  0.4608  
## 6 625854 1011569          2530        120    c3      Tb   OldLow  No  0.0768
```

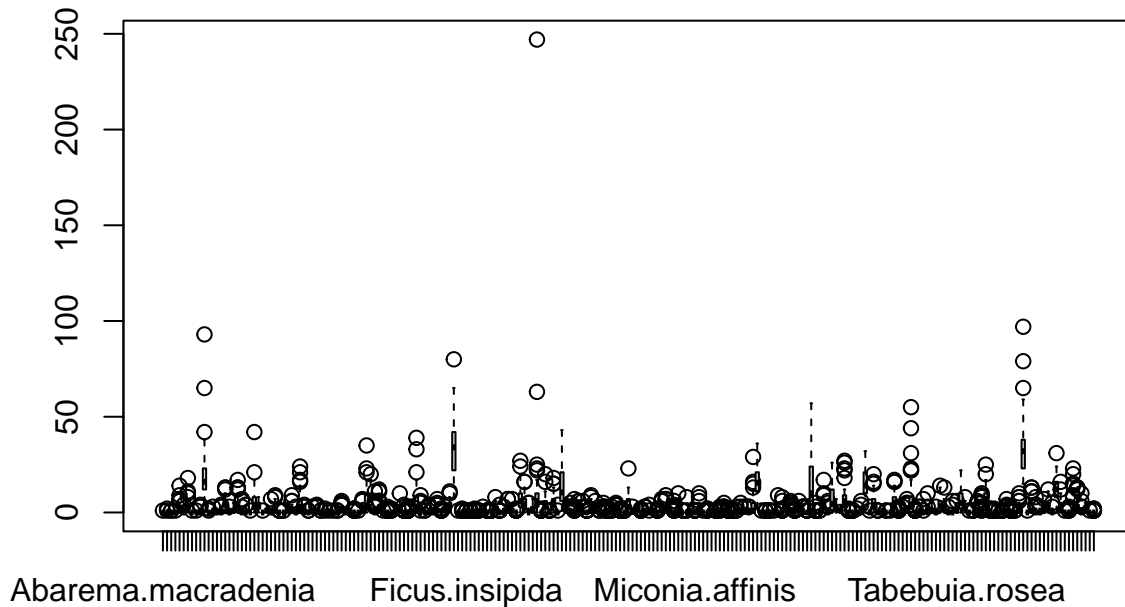
```
length(which(is.na(BCI)))
```

```
## [1] 0
```

```
length(which(is.na(BCI.env)))
```

```
## [1] 0
```

```
boxplot(BCI)
```



```
#summary(BCI)
BCI_f <- BCI[colMeans(BCI)>=1]

summary(BCI.env)
```

```
##      UTM.EW      UTM.NS      Precipitation      Elevation      Age.cat
## Min.   :625754   Min.   :1011569   Min.   :2530   Min.   :120   c2: 1
## 1st Qu.:625954   1st Qu.:1011669   1st Qu.:2530   1st Qu.:120   c3:49
## Median :626204   Median :1011769   Median :2530   Median :120
## Mean   :626204   Mean   :1011769   Mean   :2530   Mean   :120
## 3rd Qu.:626454   3rd Qu.:1011869   3rd Qu.:2530   3rd Qu.:120
## Max.   :626654   Max.   :1011969   Max.   :2530   Max.   :120
## Geology      Habitat      Stream      EnvHet
## Tb:50      OldHigh : 8      No :43      Min.   :0.0000
##              OldLow  :26      Yes: 7      1st Qu.:0.0768
##              OldSlope:12              Median :0.3536
##              Swamp   : 2              Mean   :0.3107
##              Young   : 2              3rd Qu.:0.4848
##              Max.    :0.7264
```

```
BCI.env_f <- BCI.env[, -c(3,4,6)]
```

5. Ordination of objects

```
ordi <- metaMDS(BCI_f, dist = "bray", autotransform = FALSE) #stress < 0.2 - we can continue
```

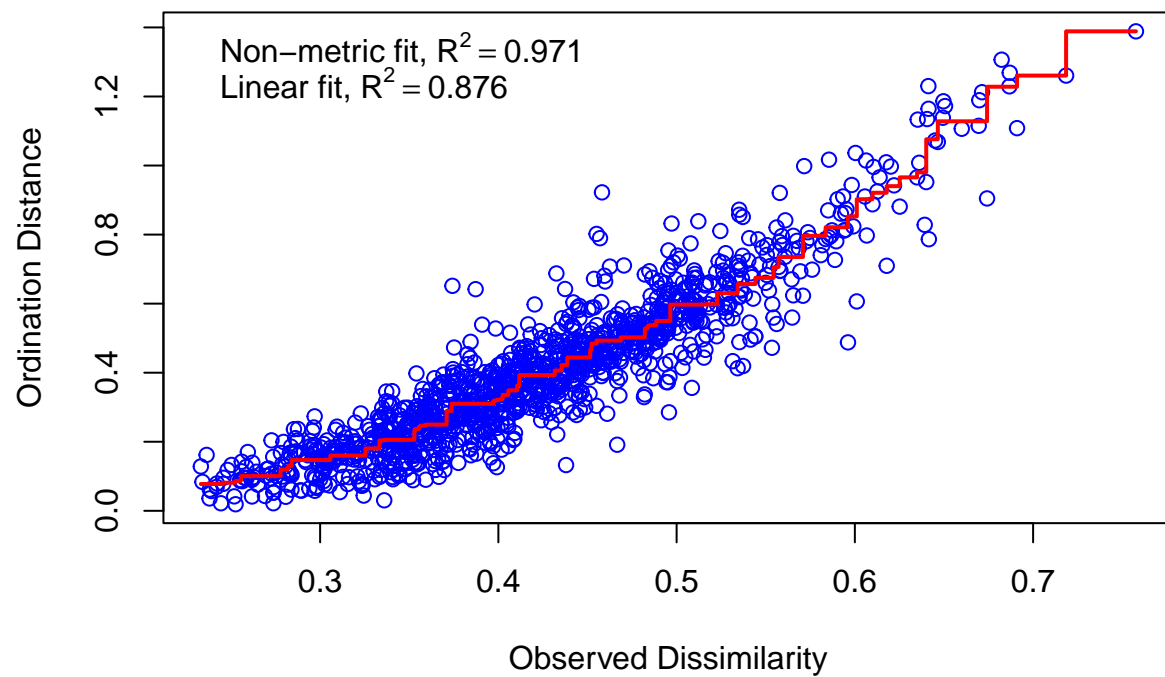
```
## Run 0 stress 0.1711008
## Run 1 stress 0.1732791
## Run 2 stress 0.1729598
## Run 3 stress 0.1909393
## Run 4 stress 0.1950467
## Run 5 stress 0.1732789
## Run 6 stress 0.1711008
## ... Procrustes: rmse 7.060397e-05  max resid 0.0004464199
## ... Similar to previous best
## Run 7 stress 0.185983
## Run 8 stress 0.1861105
## Run 9 stress 0.1711008
## ... Procrustes: rmse 3.537512e-05  max resid 0.0002209148
## ... Similar to previous best
## Run 10 stress 0.2096425
## Run 11 stress 0.1861492
## Run 12 stress 0.1845794
## Run 13 stress 0.1711008
## ... Procrustes: rmse 0.0001264952  max resid 0.0008005168
## ... Similar to previous best
## Run 14 stress 0.198767
## Run 15 stress 0.1729598
## Run 16 stress 0.1768077
## Run 17 stress 0.1711008
## ... New best solution
## ... Procrustes: rmse 2.085936e-05  max resid 0.0001310921
## ... Similar to previous best
## Run 18 stress 0.1711008
## ... Procrustes: rmse 1.893957e-05  max resid 0.0001177631
## ... Similar to previous best
## Run 19 stress 0.1729598
## Run 20 stress 0.1918013
## *** Best solution repeated 2 times
```

```
#point coordinates; merge with environmental data
ordi_pt <- data.frame(BCI.env_f, scores(ordi, display = "sites"))

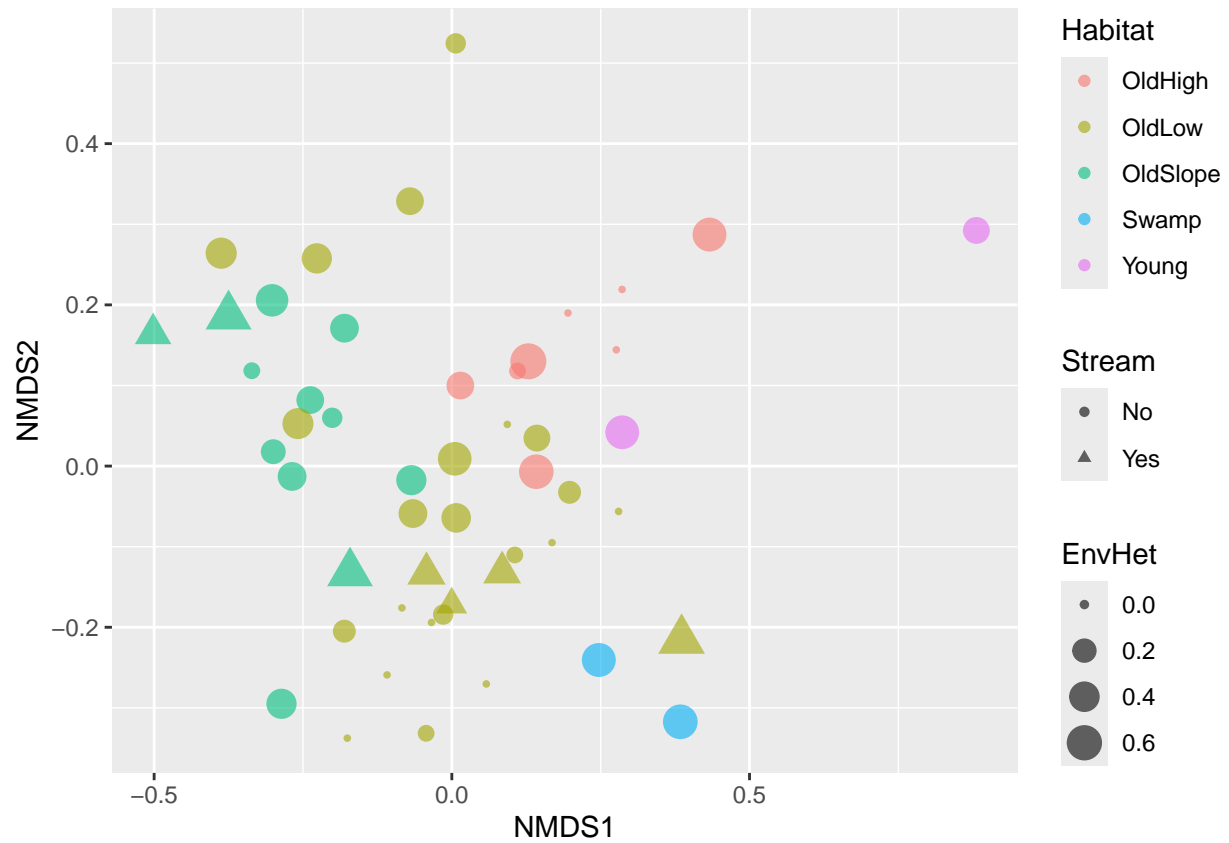
#species
ordi_sp <- data.frame(scores(ordi, display = "species"))
ordi_sp$Species <- rownames(ordi_sp)
```

6. Visualization

```
# Shepherd's diagram
stressplot(ordi)
```

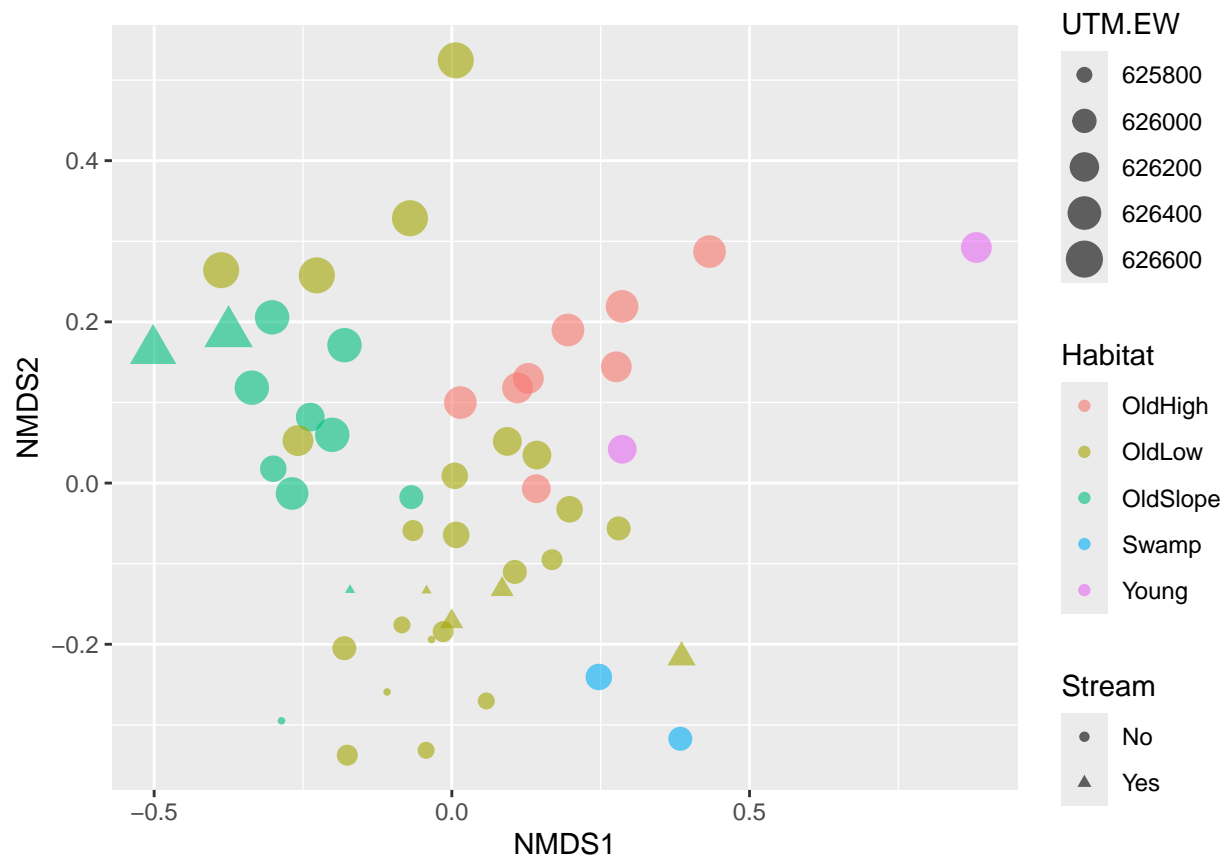


```
#visualize using ggplot()
gg_oridi <- ggplot() +
  geom_point(data = ordi_pt,
             aes(x = NMDS1, y = NMDS2, colour = Habitat,
                 shape = Stream, size = EnvHet), alpha = 0.6)
gg_oridi
```

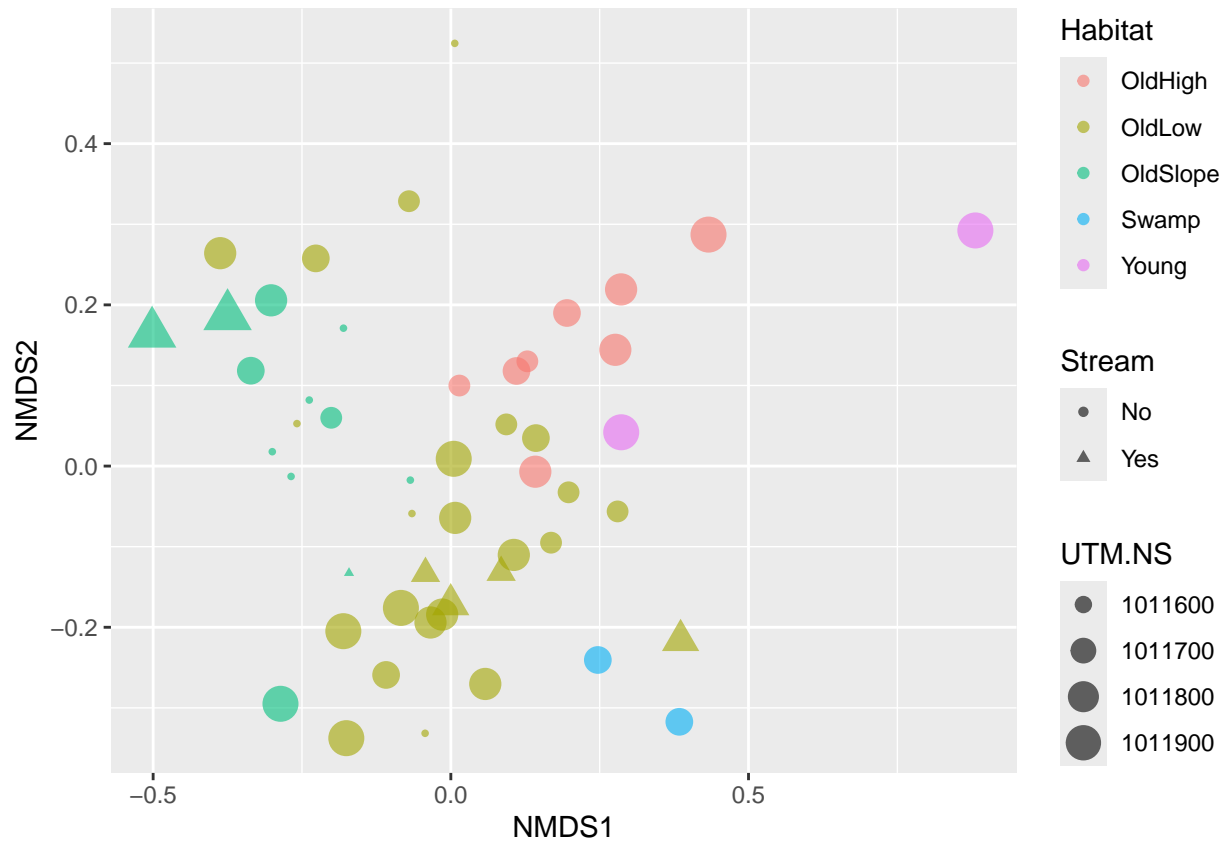


```
#add species labels
gg_ordi_sp <- gg_ordi +
  geom_text(data = ordi_sp,
    aes(x = NMDS1, y = NMDS2,
      label = Species), size=2)

gg_ordi_sp
```

```
gg_ordi <- ggplot() +
  geom_point(data = ordi_pt,
    aes(x = NMDS1, y = NMDS2, colour = Habitat,
      size = UTM.NS, shape=Stream), alpha = 0.6)
gg_ordi
```



7. Interpretation of ordination

```
efit <- envfit(ordi, BCI.env_f[, -c(3)])
```

```
efit$vectors
```

```
##          NMDS1    NMDS2    r2 Pr(>r)
## UTM.EW -0.12697  0.99191 0.7804 0.001 ***
## UTM.NS  0.87624 -0.48188 0.0260 0.544
## EnvHet -0.10502  0.99447 0.0175 0.662
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

```
efit$factors
```

```
## Centroids:
##          NMDS1    NMDS2
## HabitatOldHigh  0.1981  0.1477
## HabitatOldLow   -0.0060 -0.0579
## HabitatOldSlope -0.2689  0.0456
## HabitatSwamp    0.3154 -0.2789
## HabitatYoung    0.5836  0.1671
```



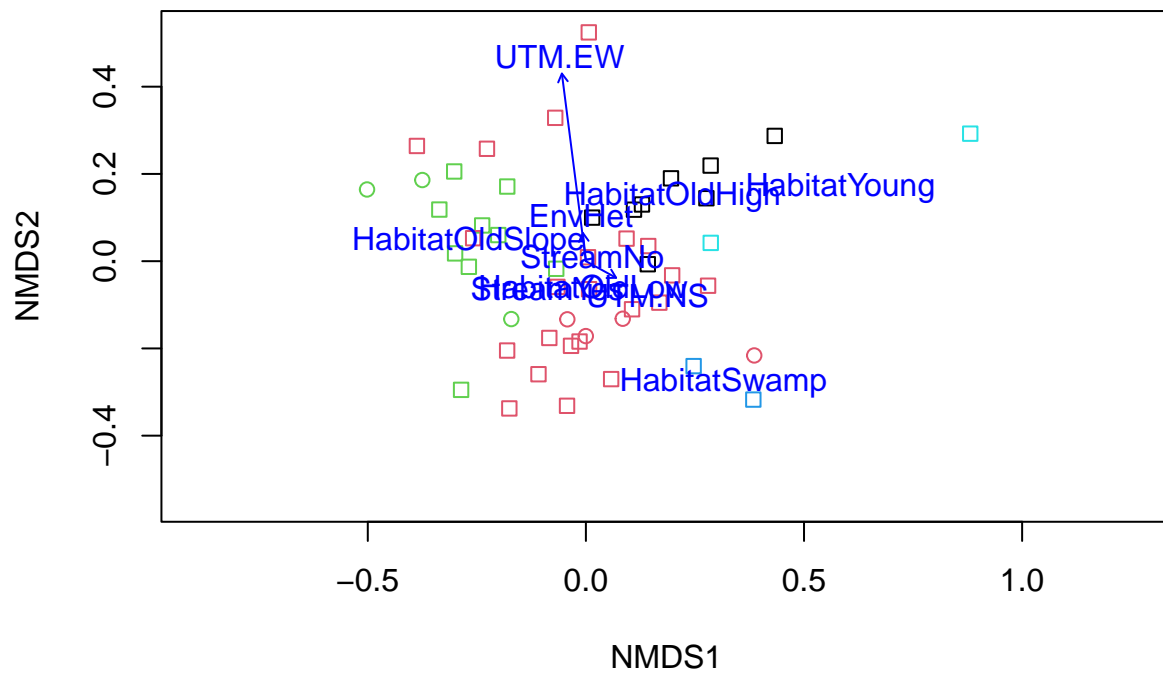
```
## StreamNo          0.0144  0.0101
## StreamYes         -0.0886 -0.0622
##
## Goodness of fit:
##           r2 Pr(>r)
## Habitat 0.4993  0.001 ***
## Stream  0.0186  0.391
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

#statistical significance observed for UTM.EW and Habitat (P-value <0.001)

8. Visualization

```
pal_col <- c("red", "green", "blue", "black", "orange")
pal_sh <- c(0,1)

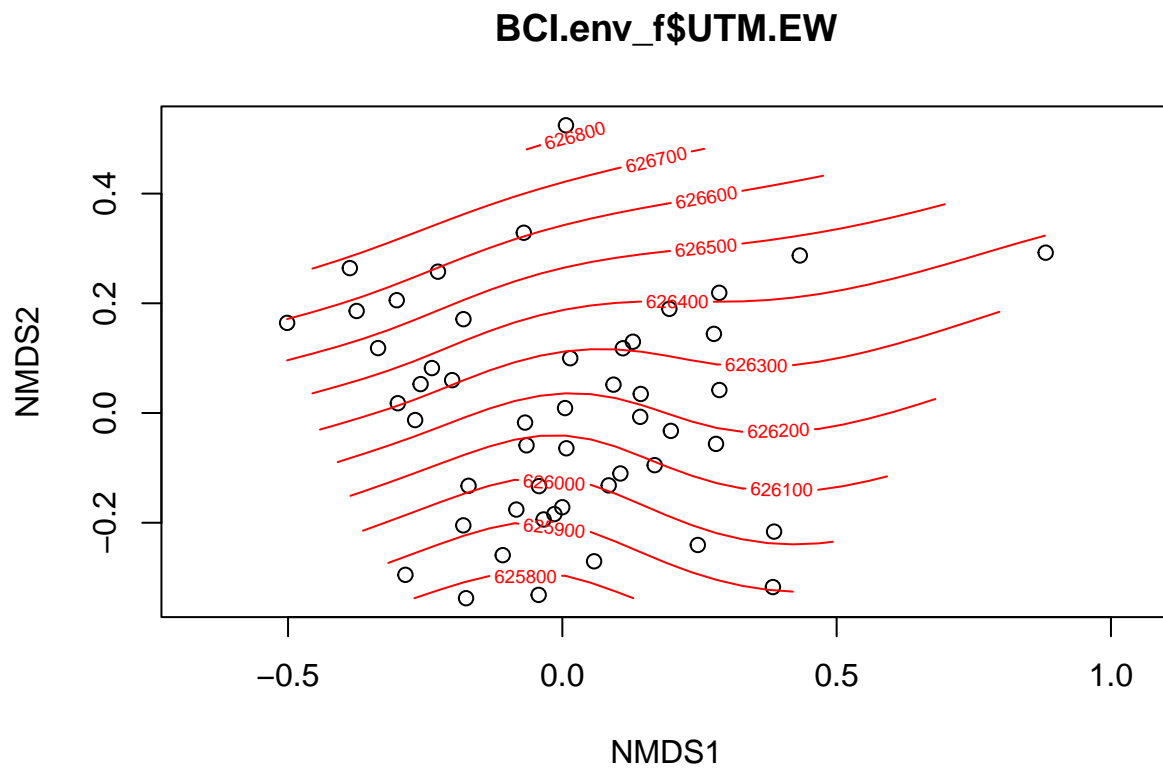
ordiplot(ordi, type = "n")
points(ordi, col = BCI.env_f$Habitat, pch = pal_sh[BCI.env_f$Stream])
plot(efit)
```



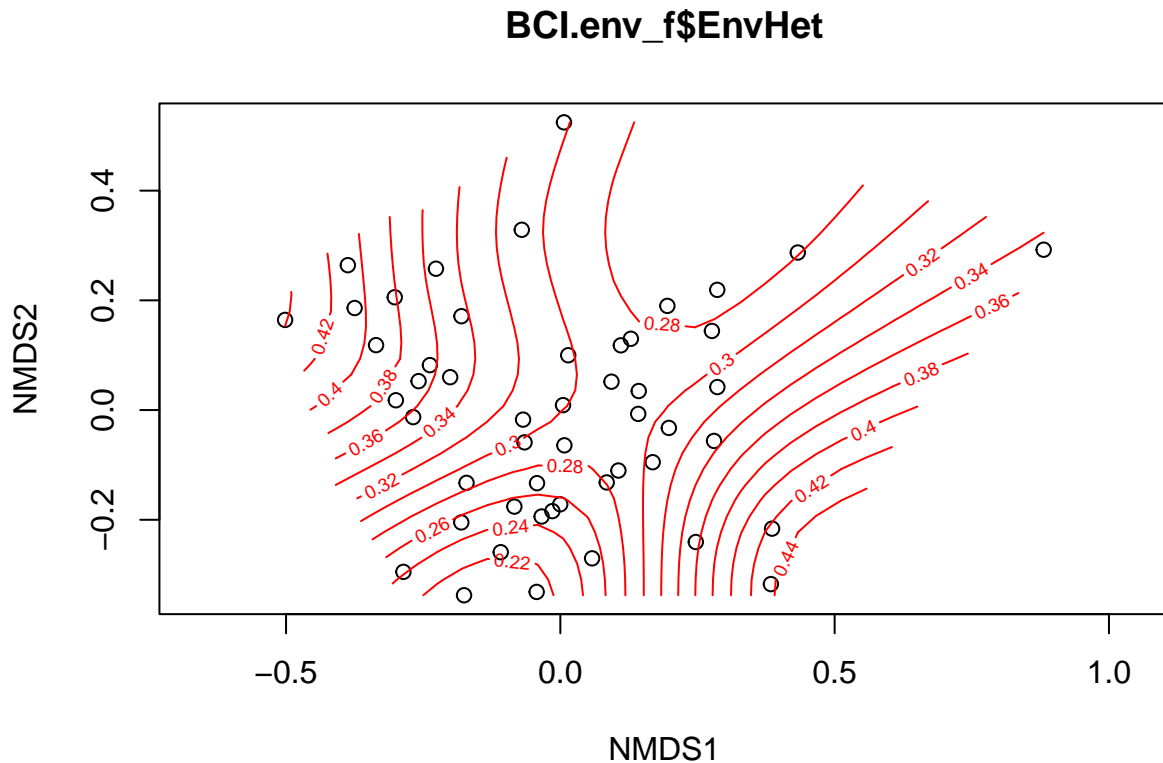
```
#correlation observed for UTM.EW, EnvHet and Habitat OldLow
```

9. ordisurf() and visualization

```
ordisurf_UE <- ordisurf(ordi, BCI.env_f$UTM.EW, method = "REML")
```



```
ordisurf_EH <- ordisurf(ordi, BCI.env_f$EnvHet, method = "REML")
```



```
summary(ordisurf_UE)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ s(x1, x2, k = 10, bs = "tp", fx = FALSE)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 626203.97    17.19   36437   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df    F p-value
## s(x1,x2)  5.945     9 25.59 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.825   Deviance explained = 84.6%
## -REML = 314.61   Scale est. = 14768      n = 50
```

```
summary(ordisurf_EH)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ s(x1, x2, k = 10, bs = "tp", fx = FALSE)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.31072    0.03165   9.818 7.34e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##             edf Ref.df   F p-value
## s(x1,x2)  3.045     9 0.7  0.0794 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.114   Deviance explained = 16.9%
## -REML = 0.49292   Scale est. = 0.050076   n = 50
```

```
#correlation for UTM.EW, with community is statistically significant; for EnvHet - not
```

10. Conclusions

We discovered the relationship between tropical tree species beta-diversity and different environmental factors. Correlation with statistical significance was observed for such factors as Habitat and UTM (East-Western coordinates).