HW1

Margarita Onvumere

2024-04-04

Applied Statistics Homework 1

Margarita Onvumere

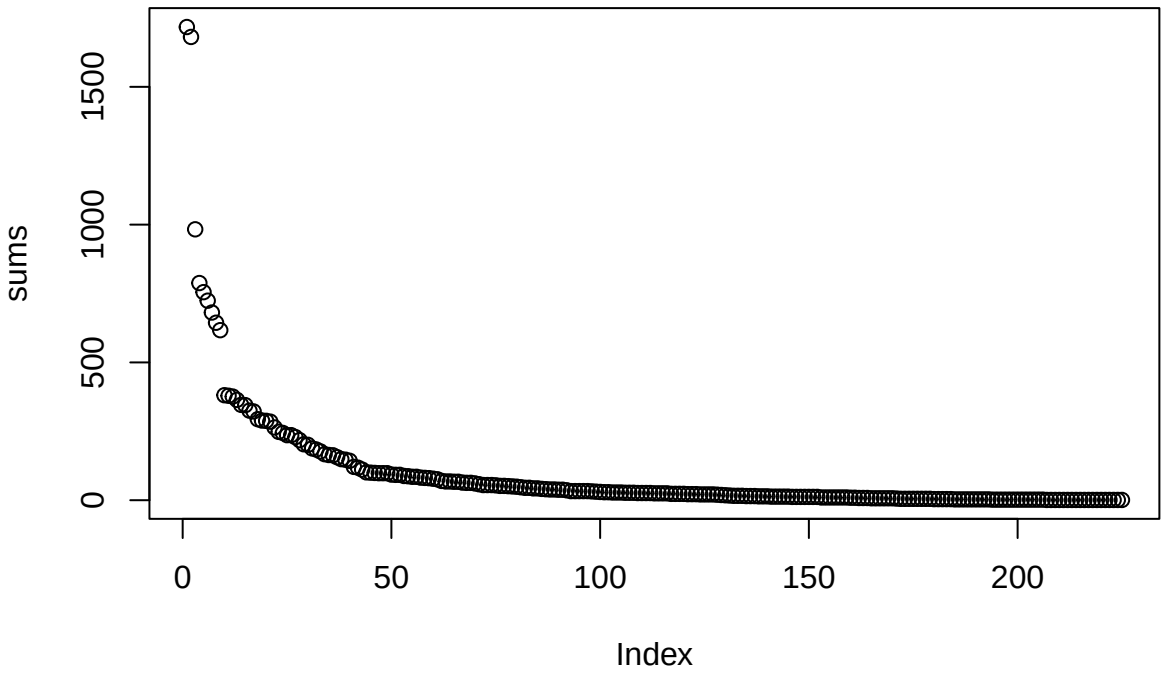
EDA

The pre-installed dataset from the *vegan* package was used:

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

Let's check the distribution of species by total occurrence:

```
sums <- sort(colSums(BCI), decreasing = TRUE)
plot(sums)
```



I wasn't sure what criteria could be used to group some species into one "rare" category, so I left it as is.

Let's look at the environmental data:

```
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
```

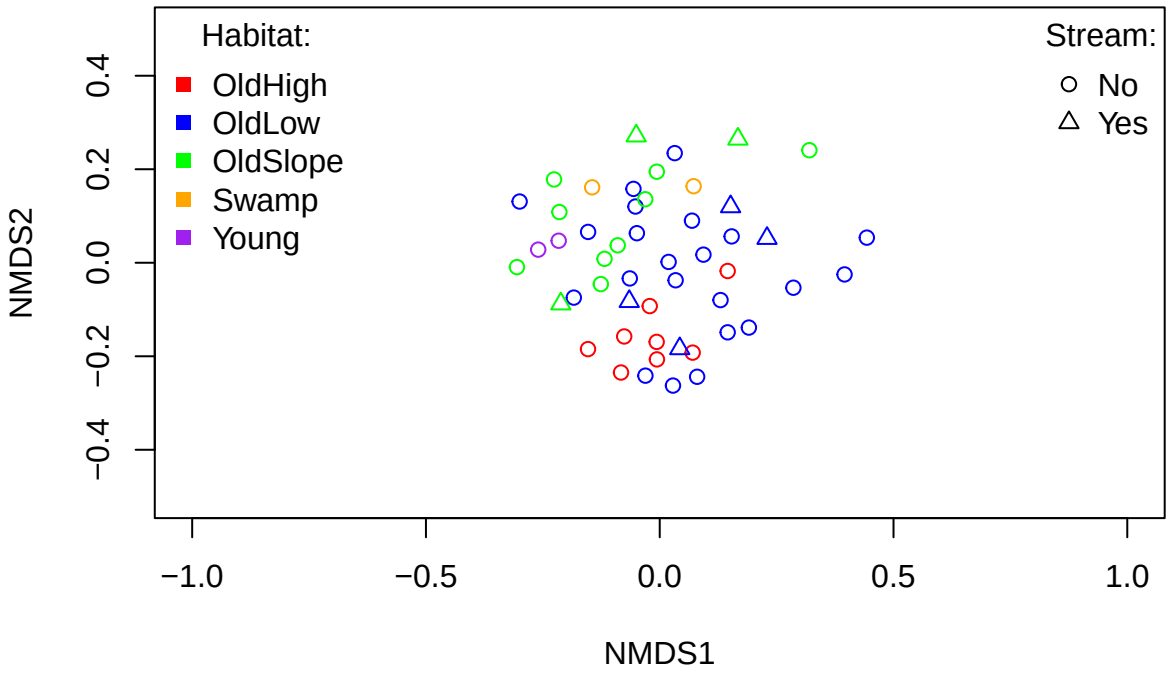
```
##  Median :626204   Median :1011769   Median :2530   Median :120
##  Mean   :626204   Mean   :1011769   Mean   :2530   Mean   :120
##  3rd Qu.:626454   3rd Qu.:1011869   3rd Qu.:2530   3rd Qu.:120
##  Max.   :626654   Max.   :1011969   Max.   :2530   Max.   :120
##  Geology     Habitat   Stream       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
```

There are not very many parameters here that could be used further down in ordisurf(). I added some characteristics from Harms et al. 2001 for visualisation, and converted them into quantitative characteristics for ordisurf().

I also checked to see if there was any additional data in the original article, but there was nothing extra there.

Ordination

```
pal_col <- c("red", "blue", "green", "orange", "purple")
pal_sh <- c(1, 2)

ordiplot(ord, type = "n", xlim = c(-1, 1), ylim = c(-0.4, 0.4))
points(ord, col = pal_col[BCI.env_add_cat$Habitat], pch = pal_sh[BCI.env_add_cat$Stream])
legend("topleft", bty = "n",
       title = "Habitat:",
       legend = levels(BCI.env_add_cat$Habitat), pch = 15, col = pal_col)
legend("topright", bty = "n",
       title = "Stream:",
       legend = levels(BCI.env_add_cat$Stream), pch = pal_sh, col = "black")
```



Interpretation of ordination: envfit()

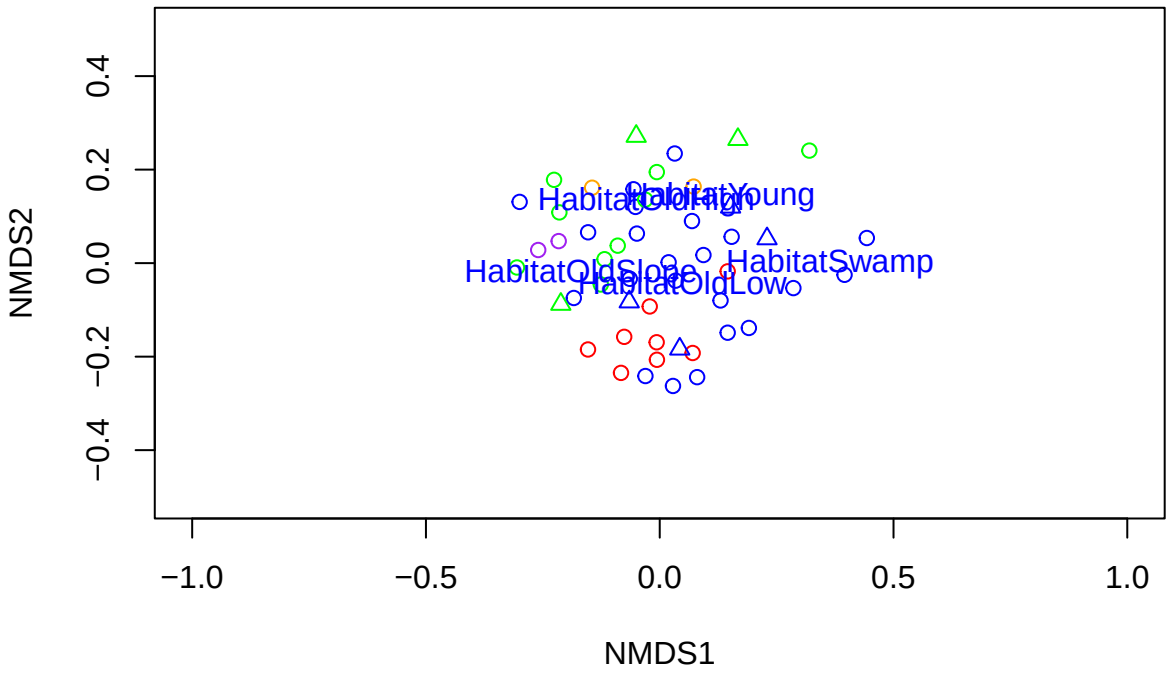
Let's first try to use original dataset:

```
print(envfit_result_0)
```

```
##
## ***VECTORS
##
##                   NMDS1    NMDS2     r2 Pr(>r)
## UTM.EW        -0.38560  0.92267 0.6391  0.001 ***
## UTM.NS         0.81774 -0.57559 0.0084  0.832
## Precipitation  0.00000  0.00000 0.0000  1.000
## Elevation      0.00000  0.00000 0.0000  1.000
## EnvHet        -0.40947  0.91232 0.0141  0.718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
##
## ***FACTORS:
##
## Centroids:
##                     NMDS1   NMDS2
## Age.catc2       -0.0519  0.1203
## Age.catc3        0.0011 -0.0025
## GeologyTb        0.0000  0.0000
## HabitatOldHigh  -0.0284  0.1324
## HabitatOldLow    0.0484 -0.0418
## HabitatOldSlope -0.1684 -0.0212
## HabitatSwamp     0.3644  0.0002
## HabitatYoung     0.1303  0.1410
## StreamNo         0.0010  0.0153
## StreamYes       -0.0061 -0.0939
##
## Goodness of fit:
##              r2 Pr(>r)
## Age.cat 0.0072  0.793
## Geology 0.0000  1.000
## Habitat 0.3836  0.001 ***
## Stream  0.0295  0.234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

All columns except **Habitat** were with $p > 0.05$.

```
envfit_result_0_sel <- envfit(ord, BCI.env[, c("EnvHet", "Habitat", "Stream")])

ordiplot(ord, type = "n", xlim = c(-1, 1), ylim = c(-0.4, 0.4))
points(ord, col = pal_col[BCI.env_add_cat$Habitat], pch = pal_sh[BCI.env_add_cat$Stream])
plot(envfit_result_0_sel, p.max = 0.05)
```



Let's now use dataset with some added parameters:

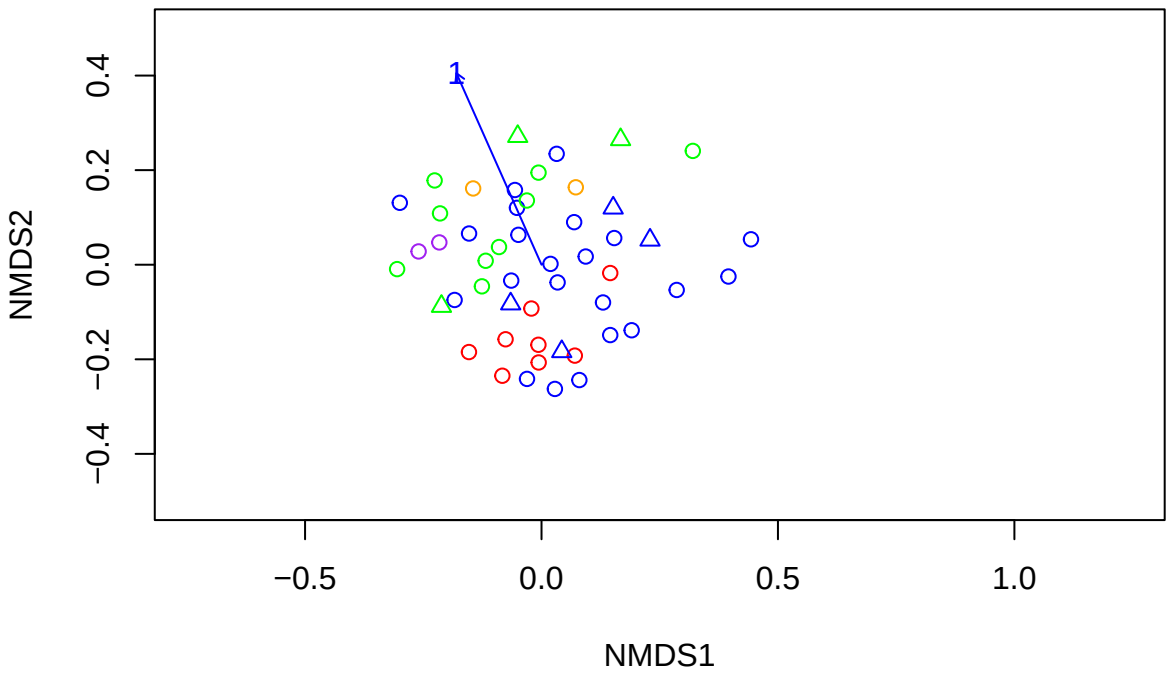
```
print(envfit_result)
```

```
##
## ***VECTORS
##
##                   NMDS1    NMDS2     r2 Pr(>r)
## UTM.EW        -0.94801  0.31823 0.1525  0.022 *
## UTM.NS        -0.88366 -0.46813 0.0048  0.879
## Precipitation  0.00000  0.00000 0.0000  1.000
## Elevation      0.00000  0.00000 0.0000  1.000
## EnvHet        -0.40600  0.91387 0.1491  0.027 *
## Density        0.66171 -0.74976 0.0284  0.506
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
##
## ***FACTORS:
##
## Centroids:
##                     NMDS1   NMDS2
## HabitatOldHigh    -0.0163 -0.1569
## HabitatOldLow      0.0604 -0.0170
## HabitatOldSlope   -0.0743  0.1081
## HabitatSwamp      -0.0360  0.1625
## HabitatYoung      -0.2380  0.0376
## Age.catc2         -0.1534 -0.1847
## Age.catc3          0.0031  0.0038
## GeologyTb          0.0000  0.0000
## StreamNo          -0.0061 -0.0082
## StreamYes          0.0377  0.0506
```

```
## Slope<7             0.0423 -0.0499
## Slope>=7           -0.0743  0.1081
## SlopeAll           -0.1370  0.1001
## Elevation_art<152   0.0604 -0.0170
## Elevation_art>=152 -0.0163 -0.1569
## Elevation_artAll   -0.0900  0.1061
##
## Goodness of fit:
##                    r2 Pr(>r)
## Habitat       0.2778  0.001 ***
## Age.cat       0.0241  0.338
## Geology       0.0000  1.000
## Stream        0.0132  0.519
## Slope         0.1910  0.001 ***
## Elevation_art 0.2498  0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
```

Slope and **Elevation_art** directly represent **Habitat**. Interestingly, **EnvHet** now has p =0.030 *.

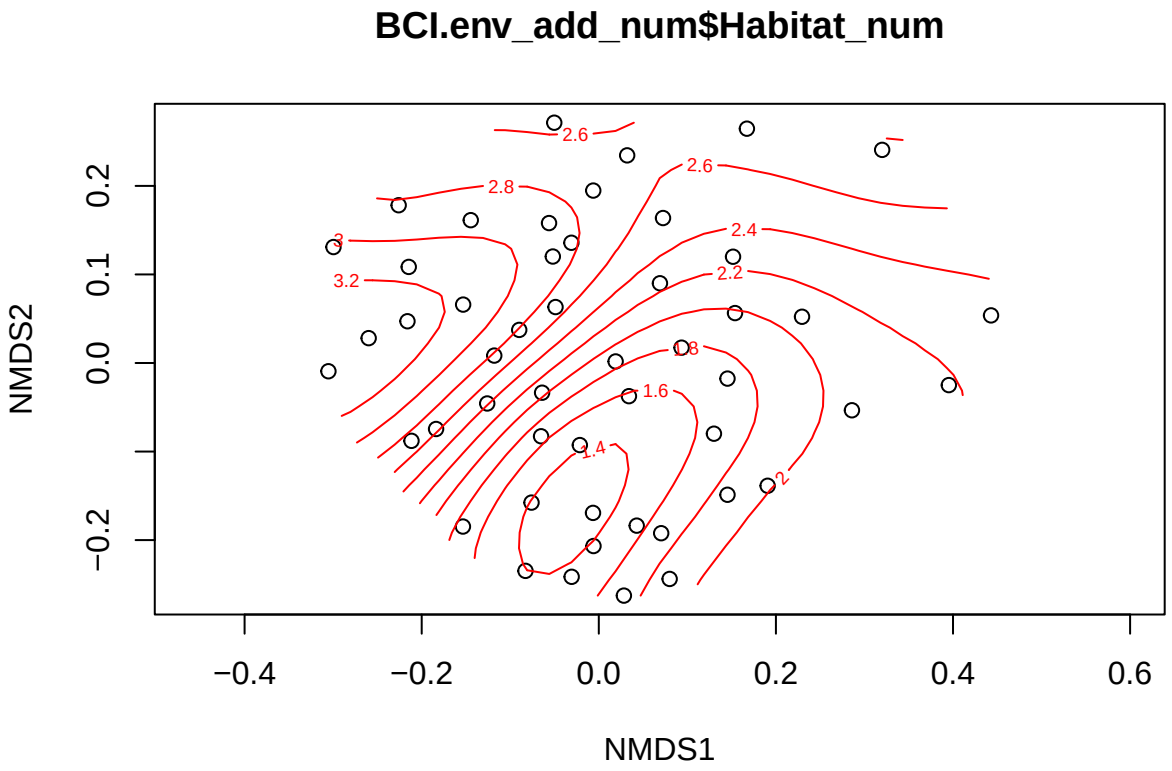
```
envfit_result_sel <- envfit(ord, BCI.env_add_cat[, c("EnvHet")])
# envfit_result_sel <- envfit(ord, BCI.env_add_cat[, c("Elevation_art", "Slope", "EnvHet")]) # it's too
ordiplot(ord, type = "n", xlim = c(-0.5, 1), ylim = c(-0.5, 0.5))
points(ord, col = pal_col[BCI.env_add_cat$Habitat], pch = pal_sh[BCI.env_add_cat$Stream])
plot(envfit_result_sel, p.max = 0.05)
```



Interpretation of ordination: ordisurf()

I tried to interpret **Habitat** using numerical values **Slope_num, Elevation_art_num and Density**, but encoding it numerically in a "naive" way also worked well.

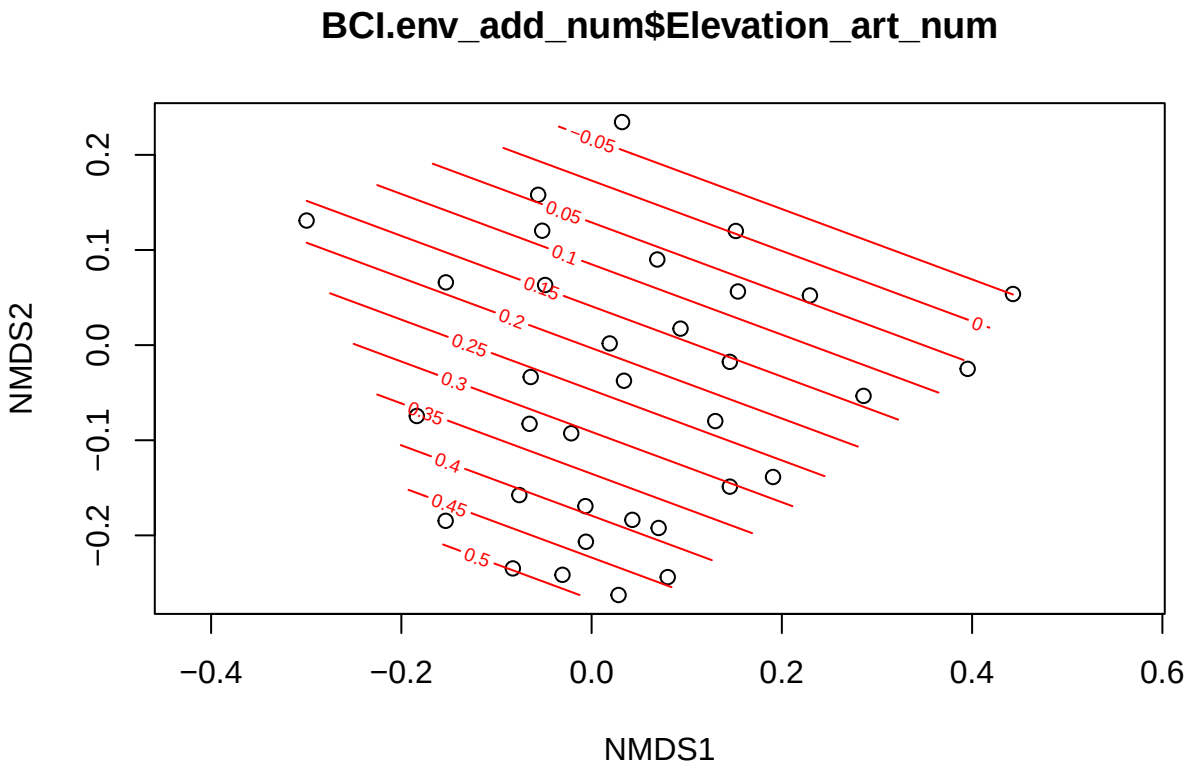
```r
BCI.env_add_num$Habitat_num <- as.integer(BCI.env_add_num$Habitat)
BCI.env_add_num$Stream_num <- as.integer(BCI.env_add_num$Stream)-1

ordisurf_Habitat_num<- ordisurf(ord, BCI.env_add_num$Habitat_num)
```




```r
summary(ordisurf_Habitat_num) # p = 4.35e-06 ***
```

```
##
## 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)  2.28000    0.09388   24.29   <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) 5.847      9 5.166 4.34e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.487   Deviance explained = 54.8%
## -REML = 58.551  Scale est. = 0.44065   n = 50
```

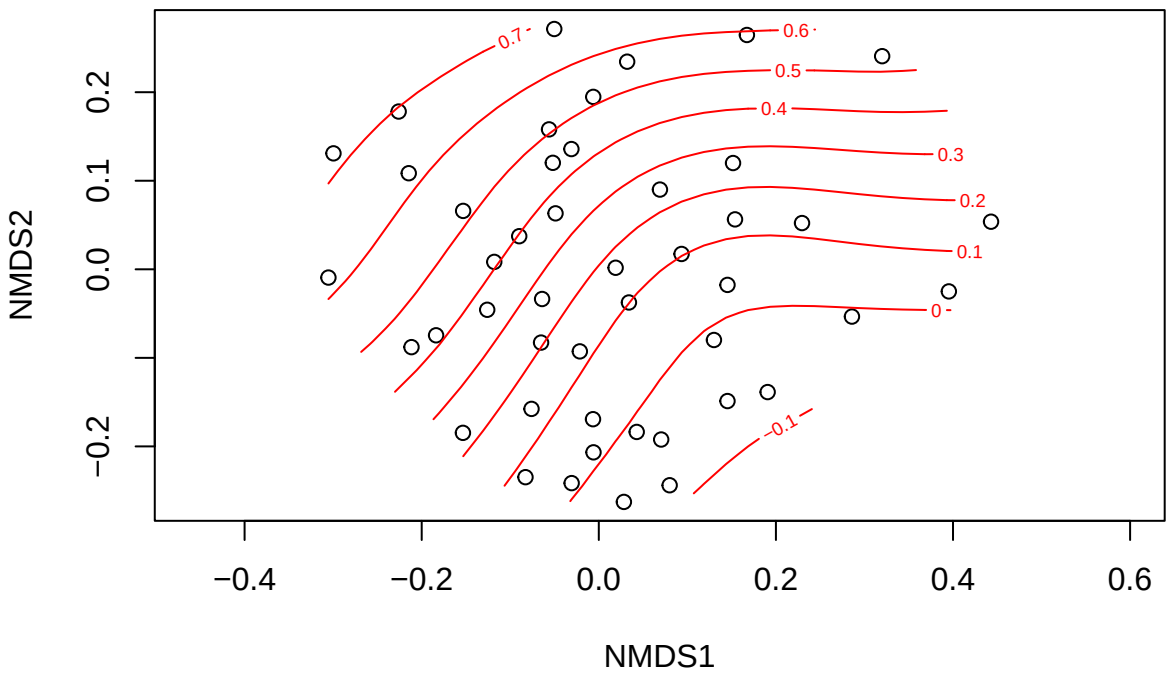
```
ordisurf_Elevation_art_num<- ordisurf(ord, BCI.env_add_num$Elevation_art_num)
```




```
summary(ordisurf_Elevation_art_num) # p = 0.0138 *
```

```
##
## 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)  0.23529    0.06641   3.543  0.00126 **
## ---
## 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) 1.581      9 0.867  0.0138 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.191   Deviance explained =   23%
## -REML = 18.843  Scale est. = 0.14995   n = 34
```

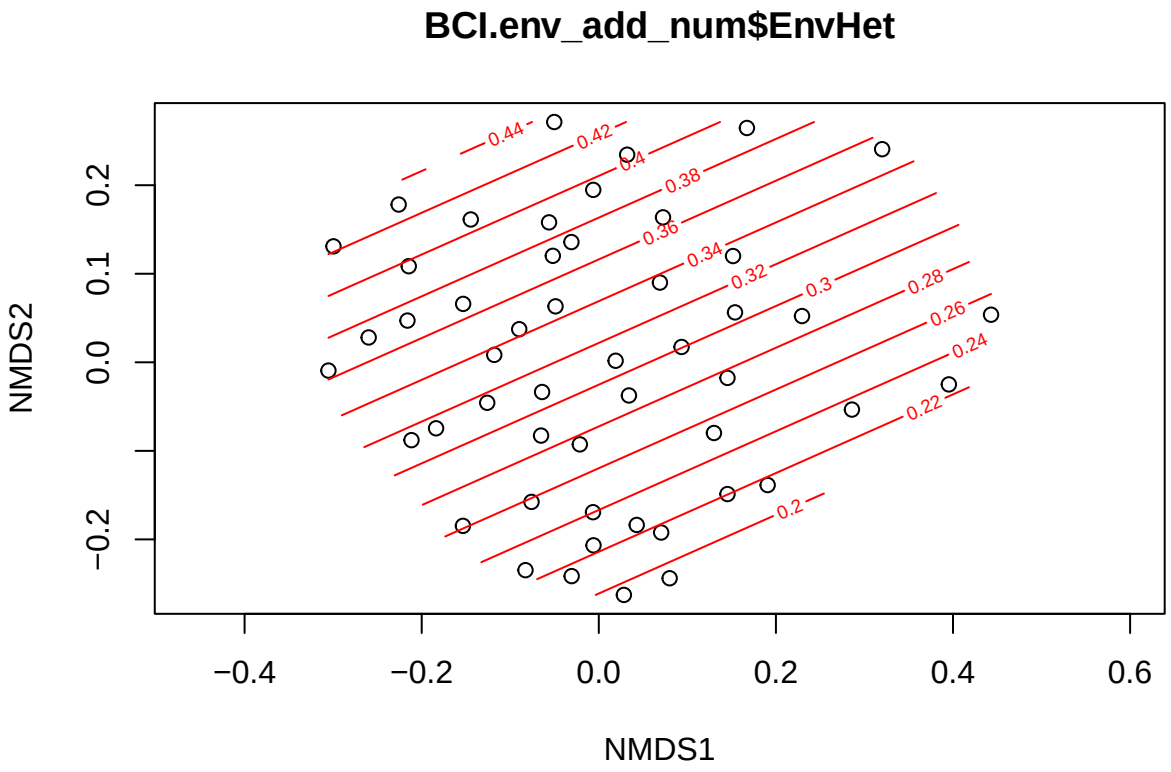
```
ordisurf_Slope_num<- ordisurf(ord, BCI.env_add_num$Slope_num)
```




```
summary(ordisurf_Slope_num) # p = 0.000134 ***
```

```
##
## 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)   0.2609     0.0524   4.978  1.2e-05 ***
## ---
## 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) 4.028      9 2.803 0.000135 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.359   Deviance explained = 41.7%
## -REML = 23.146  Scale est. = 0.1263    n = 46
```

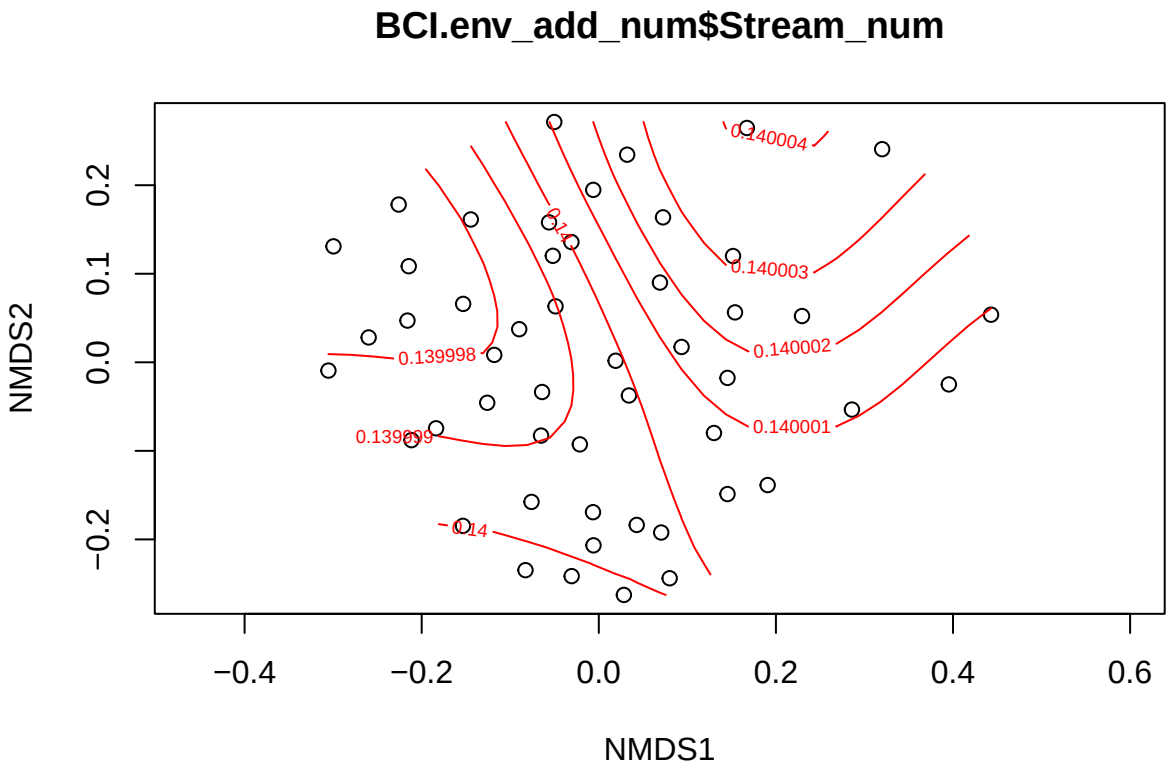
```
ordisurf_EnvHet<- ordisurf(ord, BCI.env_add_num$EnvHet)
```




```
summary(ordisurf_EnvHet) # p = 0.0225 *
```

```
##
## 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    0.03167   9.812 5.27e-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) 1.514      9 0.693  0.0225 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.113   Deviance explained =   14%
## -REML = -0.42832  Scale est. = 0.050139  n = 50
```

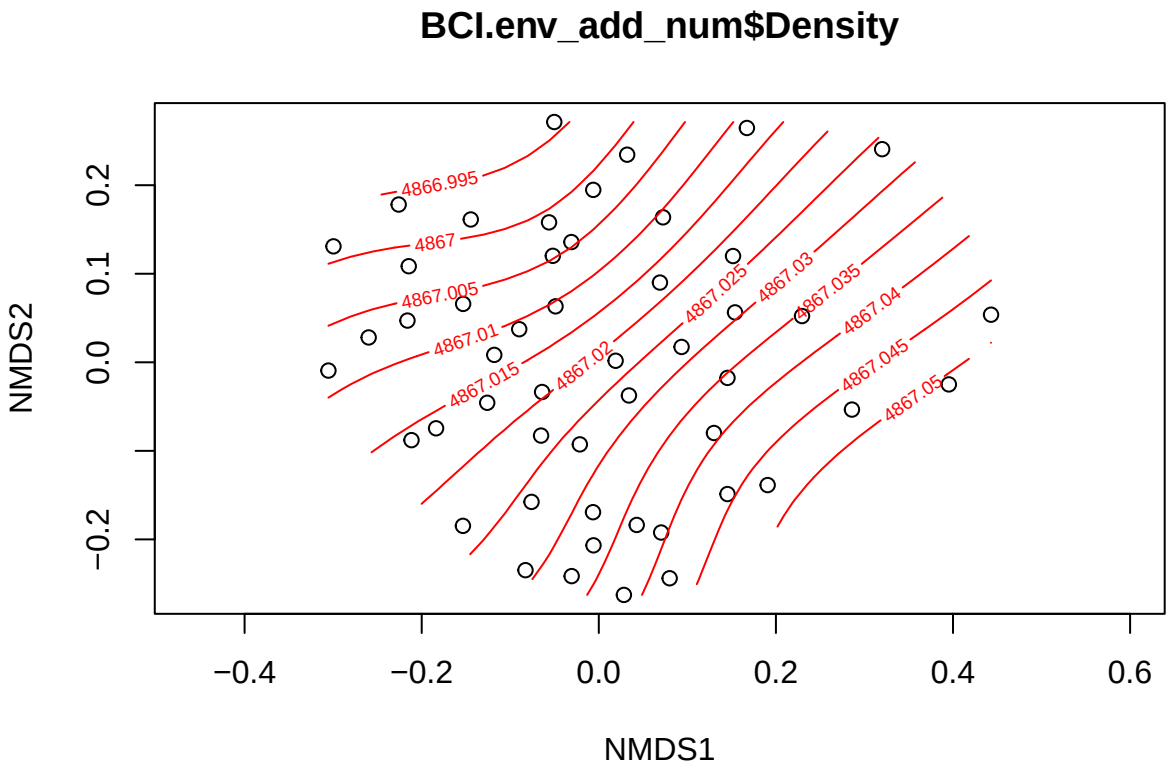
```
ordisurf_Stream_num<- ordisurf(ord, BCI.env_add_num$Stream_num)
```




```
summary(ordisurf_Stream_num) # p = 0.54
```

```
##
## 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)  0.14000    0.04957   2.824  0.00683 **
## ---
## 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) 8.892e-05      9 0    0.54
##
## R-sq.(adj) =  1.11e-06   Deviance explained = 0.000293%
## -REML = 20.114  Scale est. = 0.12286   n = 50
```

```
ordisurf_Density<- ordisurf(ord, BCI.env_add_num$Density)
```




```
summary(ordisurf_Density) # p = 0.601
```

```
##
## 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)  4867.02      65.95    73.8   <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) 0.000529      9 0   0.601
##
## R-sq.(adj) =  2.96e-06   Deviance explained = 0.00138%
## -REML = 372.59  Scale est. = 2.1748e+05  n = 50
```

Conclusion

The most important factors are Habitat, which is determined by the statistically significant parameters Elevation and Slope, and EnvHet.