### BING-F4002 Acquisition et analyse de données

## Fiche TP 5 : Expliquer les répartitions

Dufrêne M. - Gilbert M.

(avec la collaboration initiale de - Barbier N. - Deblauwe V.)

Version d'octobre 2015

# Solutions proposées (mais il y en d'autres ...)

#### TP: Q1

Construire d'abord une distribution d'abondance théorique d'abondance d'une espèce en fonction d'un facteur écologique (relation linéaire et gaussienne) puis utilisation des outils statistiques connus pour reconstruire le modèle à priori.

Q1a : Comment interprétez-vous les résultats des estimations des paramètres a et b ? Augmenter la proportion de « hasard » et analyser son impact sur les résultats.

#### # Solution potentielle pour la Q1a

L'impact du caractère aléatoire (hasard=5, 10, 20, ...) se marque d'abord sur le paramètre intercept b puis c'est le paramètre pente a qui devient non significatif quand son rôle dans la prédiction des abondances devient minime. Il est aussi possible d'imposer une variation aléatoire sur la pente.

Q1b : Comment feriez-vous pour faire varier aléatoirement la pente cette fois ?

#### # Solution potentielle pour la Q1b

```
taxon = (a + hasard * alea ) * facteur1 * + b
```

Q1c : Appliquer cette approche à un modèle de régression multiple avec un deuxième facteur écologique.

#### # Solution potentielle pour la Q1c

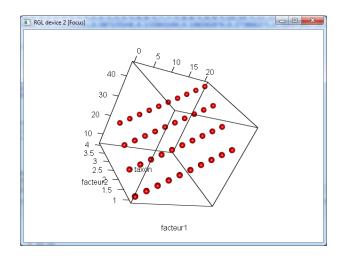
```
rm(list=ls())
```

```
# 1. Creation d'une liste de valeurs pour un facteur écologique fact1 <- seq(0.5, 20, by = 0.5) fact2 <- seq(1, 4, by = 1) taxon = 0 fichier = data.frame(facteur1=fact1, facteur2=fact2, taxon)

# 2. Construction d'un modèle linéaire multiple Y = b0 + b1*X1 + b2*X2 b0 = 3; b1 = 1.5; b2 = 3 fichier$taxon = b0 + b1 * fichier$facteur1 + b2 * fichier$facteur2 plot(fichier$facteur1, fichier$taxon)
```

#### plot(fichier\$facteur2, fichier\$taxon)

```
# plot3D
library(rgl)
with(data = fichier,
    plot3d(x = facteur1,
    y = facteur2,
    z = taxon,
    type = "s",
    col = "red",
    size = 2)
```



#### # 4. Régression linéaire myReg = Im(fichier\$taxon ~ fichier\$facteur1 + fichier\$facteur2) summary(myReg)

```
Call.
```

lm(formula = fichier\$taxon ~ fichier\$facteur1 + fichier\$facteur2)

#### Residuals:

```
Min 1Q Median 3Q Max -3.537e-15 -1.745e-15 -9.457e-16 6.281e-16 2.200e-14
```

#### Coefficients:

```
Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.000e+00 1.923e-15 1.560e+15 <2e-16 ***
fichier$facteur1 1.500e+00 1.151e-16 1.303e+16 <2e-16 ***
fichier$facteur2 3.000e+00 5.944e-16 5.047e+15 <2e-16 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 4.184e-15 on 37 degrees of freedom Multiple R-squared: 1, Adjusted R-squared: 1 F-statistic: 1.049e+32 on 2 and 37 DF, p-value: < 2.2e-16

Les paramètres b0, b1 et b2 sont bien égaux à 3.0, 1.5 et 3.0.

Q1d : Appliquer cette approche à un modèle de régression gaussienne avec un facteur écologique.

# Solution potentielle pour la Q1d

```
Équation gaussienne :
taxon = c * exp^{[-1/2 * (x-u)^2]} / t^2
```

avec u = l'optimum, c = l'abondance maximum à l'optimum et t = tolérance

# Solution potentielle pour la Q1c

rm(list=ls())

# 1. Creation d'une liste d'abondance pour un facteur écologique

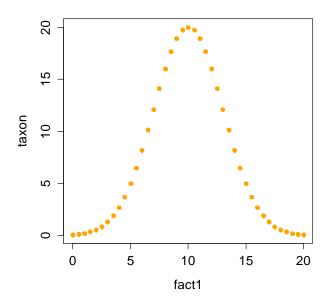
```
fact1 <- seq(0.0, 20, by = 0.5)
```

```
# parametres de l'equation gaussienne

c = 20; t = 3; u = 10;

taxon = c * exp(-0.5 * (fact1 - u)*(fact1 - u) /(t*t))
```

```
plot(fact1, taxon,
pch = 19, # forme du symbole
cex = 1, # taille du symbole
col = "orange") # couleur du symbole
```

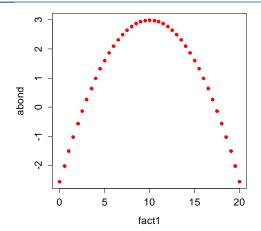


# transformation pour la regression gaussienne

# log neperien des abondances pour obtenir une parabole qui va permettre d'utiliser une regression polynomiale d'ordre 2

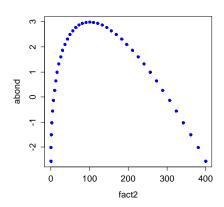
```
abond = log(taxon)
fact2 = fact1*fact1

plot(fact1, abond,
pch = 19,  # forme du symbole
cex = 1,  # taille du symbole
col = "red") # couleur du symbole
```



plot(fact2, abond, pch = 19, cex = 1, col = "blue")

# forme du symbole # taille du symbole



#### # Regression polynomiale fichier = data.frame(facteur1=fact1, facteur2=fact2, abond) myReg = Im(abond ~ fact1 + fact2, fichier) summary(myReg)

```
call:
```

lm(formula = abond ~ fact1 + fact2, data = fichier)

#### Residuals:

Min 1Q Median 3Q Max -6.741e-15 -5.070e-16 -1.360e-17 2.970e-16 1.368e-14

#### Coefficients:

Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.560e+00 1.153e-15 -2.220e+15 <2e-16 \*\*\*
fact1 1.111e+00 2.667e-16 4.165e+15 <2e-16 \*\*\*
fact2 -5.556e-02 1.289e-17 -4.309e+15 <2e-16 \*\*\*
--Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.582e-15 on 38 degrees of freedom Multiple R-squared: 1, Adjusted R-squared: 1 F-statistic: 9.282e+30 on 2 and 38 DF, p-value: < 2.2e-16

# param = data.frame(coefficients(myReg)) param

```
coefficients.myReg.
(Intercept) -2.55982328
fact1 1.11111111
fact2 -0.05555556

b0 = param[1,1];
b1 = param[2,1];
b2 = param[3,1];
```

```
# plot3D
library(rgl)
with(data = fichier,
    plot3d(x = fact1,
    y = fact2,
    z = taxon,
    type = "s",
    col = "red",
    size = 2)
)
```

# Calcul des parametres initiaux de la gaussienne a partir des parametres de la regression polynomiale

```
u = - b1/(2*b2); u
t = 1/sqrt(-2*b2); t
c = exp(b0 + b1*u + b2*u*u); c
> u = - b1/(2*b2); u
[1] 10
> t = 1/sqrt(-2*b2); t
[1] 3
> c = exp(b0 + b1*u + b2*u*u); c
[1] 20
```

1. Expliquer l'ordination AFC du fichier des Carabides par les variables écologiques disponibles

```
TP: Q2
```

On a réalisé une AFC sur le fichier des Carabides. Comment expliquer la position des stations sur les axes de l'AFC par des variables écologiques ?

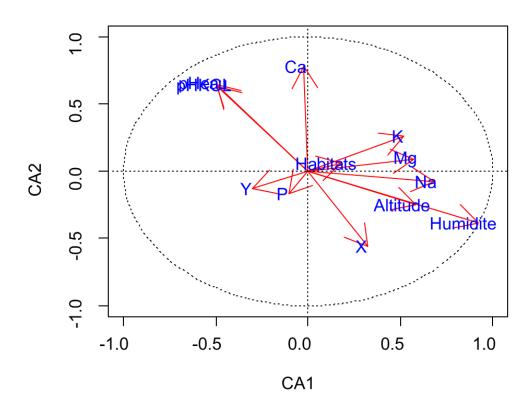
Le principe de ce TP est expliqué ci-dessus.

```
# Lecture des jeux de donnees
rm(list=ls())
# Definition de la directory de travail
# Attention !!! Pas d'accents et inversion des \
setwd("D:/Dufrene/GXABT Enseignements/ULB BING-F4002 Acquisition et analyse de
donnees/TP5 - Expliquer/Donnees")
# Pour verifier:
getwd()
[1] "D:/Dufrene/GXABT Enseignements/ULB BING-F4002 Acquisition et analyse d
e donnees/TP5 - Expliquer/Donnees"
# Lecture de donnees
eco=read.table('carabides_32sta_12eco.txt', h=T, sep="\t", row.names="Stations")
esp=read.table('carabides_32sta_103esp.txt', h=T, sep="\t", row.names="Stations")
# realise une AFC sur le fichier especes
library(vegan)
esp.ca <- cca(esp)
summary(esp.ca)
call:
cca(X = esp)
Partitioning of mean squared contingency coefficient:
               Inertia Proportion
                 7.239
Total
                                  1
                 7.239
                                  1
Unconstrained
Eigenvalues, and their contribution to the mean squared contingency coeffic
ient
Importance of components:
                                   CA2
                                            CA3
                                                     CA4
                                                             CA5
                           CA1
                                                                      CA6
                                                                               CA7
                        0.9138 0.7984 0.69291 0.59967 0.5357 0.51243 0.46014
Eigenvalue
Proportion Explained 0.1262 0.1103 0.09571 0.08283 0.0740 0.07078 0.06356
Cumulative Proportion 0.1262 0.2365 0.33223 0.41506 0.4891 0.55984 0.62340
Scaling 2 for species and site scores
* Species are scaled proportional to eigenvalues
* Sites are unscaled: weighted dispersion equal on all dimensions
# recupere les coordonnees des stations (lignes) et des especes (colonnes) sur les axes CA1 et
coordolig <-as.data.frame(scores(esp.ca, scaling=1, display="wa"))
coordocol <-as.data.frame(scores(esp.ca, scaling=1, display="spe"))
```

```
# Graphique amélioré
plot(esp.ca, type = "n") # plot vide
# coordonnées des colonnes - espèces
points(coordocol$CA1, coordocol$CA2,
   col='darkred',
    pch = 1)
colvec <- c("blue", "darkcyan", "yellow", "orange", "red", "grey", "green", "brown")</pre>
points(coordolig$CA1, coordolig$CA2,
   col = colvec[eco$Habitats],
   pch = 16
text(coordolig$CA1, coordolig$CA2,
  labels = rownames(coordolig), # label de stations sur le graphique
  col = colvec[eco$Habitats],
                                     # couleur du label (peut être un vecteur)
                       # position du label
  pos = 3,
  cex = 0.6)
                        # taille du label
                        က
                        7
                        0
                            -3
                                  -2
                                               0
                                                                 3
                                         -1
                                              CA1
# Cercle des correlations
# ===========
# calcul de la correlation
correlation = data.frame(cor(eco, coordolig))
correlation
# representation du cercle des correlations
a \leftarrow seq(0,2*pi,length=100)
plot(cos(a), sin(a),
    type = 'I', Ity = 3,
    xlab = 'CA1', ylab = 'CA2',
    main = "Correlation circle")
arrows(0,0, correlation$CA1, correlation$CA2, col='red')
text(correlation$CA1, correlation$CA2,rownames(correlation), col='blue')
abline(h=0, lty=3) # axe horizontal (lty = type de la ligne)
```

abline(v=0, lty=3) # axe vertical (lty = type de la ligne)

### **Correlation circle**



**TP: Q3** 

Réalisez une PCoA sur le fichier des Carabides et expliquer la position des stations sur les axes de l'AFC par des variables écologiques.

# 

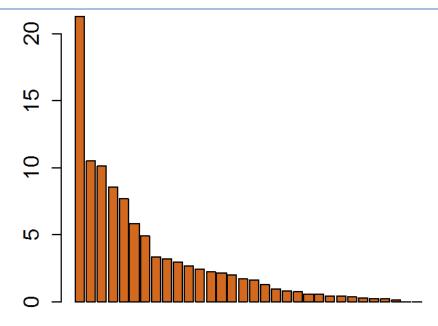
library(vegan)

# Transformation log logesp=log(esp+1)

# Matrice de distance de Bray-Curtis d14 = vegdist (esp, method = "bray")

# PCoA esp.pcoa <- cmdscale(d14, add= TRUE, eig= TRUE)

# Calcul des valeurs propres valeurs\_propres = round(esp.pcoa\$eig/sum(esp.pcoa\$eig)\*100, digits = 2) barplot(valeurs\_propres, col = "chocolate")

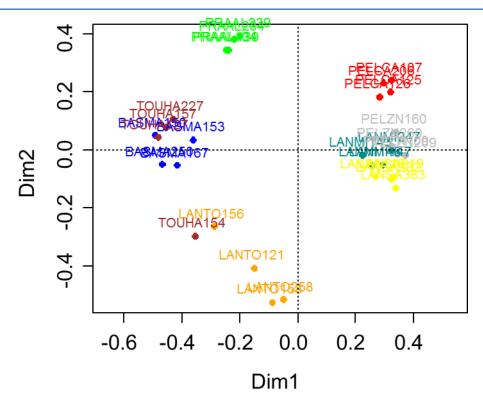


```
# recupere les coordonnees des stations (lignes) sur les axes Dim1 et Dim2
coordolig = data.frame(scores(esp.pcoa))
# Graphique amélioré
# =========
ordiplot(esp.pcoa, type = "n") # plot vide
# Voir les correspondances entre les habitats et les codes stations
table(eco$Habitats, substr(rownames(esp),1,5))
# quand eco$Habitats = 1 => col= "blue", ...
colvec <- c("blue", "darkcyan", "yellow", "orange", "red", "grey", "green", "brown")
points(coordolig$Dim1, coordolig$Dim2,
col = colvec[eco$Habitats],
pch = 16
text(coordolig$Dim1, coordolig$Dim2,
labels = rownames(coordolig), # label de stations sur le graphique
col = colvec[eco$Habitats],
                                   # couleur du label (peut être un vecteur)
                      # position du label
pos = 3,
```

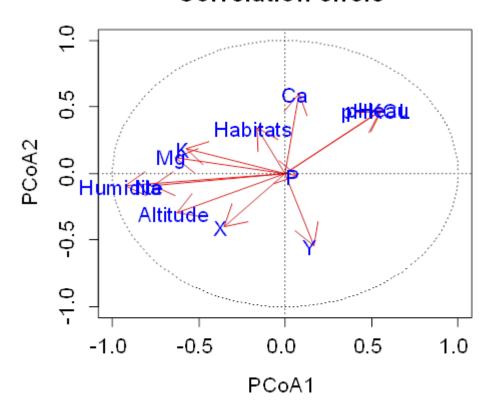
# taille du label

abline(h=0, lty=3) # axe horizontal (lty = type de la ligne) abline(v=0, lty=3) # axe vertical (lty = type de la ligne)

cex = 0.6)



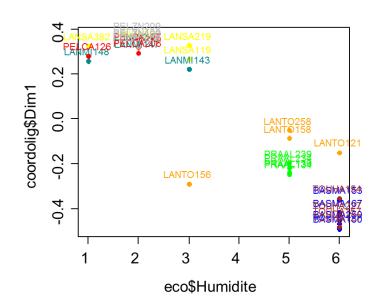
# Correlation circle



#### # Plot d'une variable

# -----

plot(eco\$Humidite, coordolig\$Dim1, col = colvec[eco\$Habitats], pch = 16)
text(eco\$Humidite, coordolig\$Dim1,
 labels = rownames(coordolig), # label de stations sur le graphique
 col = colvec[eco\$Habitats], # couleur du label (peut être un vecteur)
 pos = 3, # position du label
 cex = 0.6) # taille du label

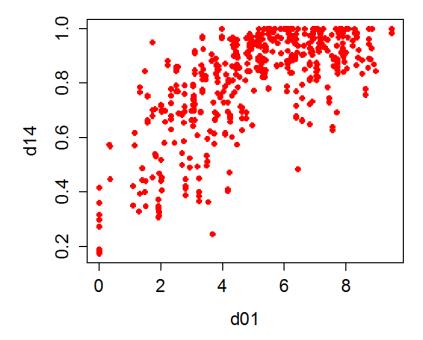


#### **TP:Q4**

Réalisez un test de Mantel (voir cours) entre une matrice de distances biologiques (Carabides) et une matrice de distance écologique (variables écologiques).

Le principe de ce TP est expliqué ci-dessus.

```
# Lecture des jeux de donnees
rm(list=ls())
# Definition de la directory de travail
# Attention !!! Pas d'accents et inversion des \
setwd("D:/Dufrene/GXABT Enseignements/ULB BING-F4002 Acquisition et analyse de
donnees/TP5 - Expliquer/Donnees")
# Pour verifier:
getwd()
[1] "D:/Dufrene/GXABT Enseignements/ULB BING-F4002 Acquisition et analyse d
e donnees/TP5 - Expliquer/Donnees"
# Lecture de donnees
eco=read.table('carabides_32sta_12eco.txt', h=T, sep="\t", row.names="Stations")
esp=read.table('carabides_32sta_103esp.txt', h=T, sep="\t", row.names="Stations")
# Transformation log pour les espèces
logesp=log(esp+1)
attach(eco)
logeco=data.frame(pHeau, pHKCL, logP=log(P+1),logK=log(K+1), logCa=log(Ca+1),
         logMg=log(Mg+1), logNa=log(Na+1),Humidite, row.names=rownames(eco))
# Matrice de distance de Bray-Curtis pour les abondances
d14 = vegdist (logesp, method = "bray")
# Matrice de distance euclidienne pour les variables ecologiques
d01 = vegdist (logeco, method = "euclidian")
# Matrice de distance euclidienne pour uniquement l'humidite
d01Hum = vegdist (logeco$Humidite, method = "euclidian")
# Relation entre D01 et D14
plot(d01, d14, col = "red", pch = 16)
```



#### cor(d14, d01)

mantel(d01, d14, method="pearson", permutations=999)

```
Mantel statistic based on Pearson's product-moment correlation
```

mantel(xdis = d01, ydis = d14, method = "pearson", permutations = 999)

Mantel statistic r: 0.6772 Significance: 0.001

Empirical upper confidence limits of r: 90% 95% 97.5% 99%

0.0617 0.0813 0.1032 0.1201

Based on 999 permutations

mantel(d01Hum, d14, method="kendall", permutations=999)

Mantel statistic based on Kendall's rank correlation tau

mantel(xdis = d01Hum, ydis = d14, method = "kendall", permutations = 999)

Mantel statistic r: 0.5462 Significance: 0.001

Empirical upper confidence limits of r:

95% 97.5% 99% 0.0395 0.0570 0.0782 0.0926

Based on 999 permutations

#### **TP: Q5**

Utilisez vos compétences en analyse de la variance pour voir quelles sont les variables écologiques qui expliquent le plus une partition en 4 groupes des stations en se basant sur la répartition des Carabides.

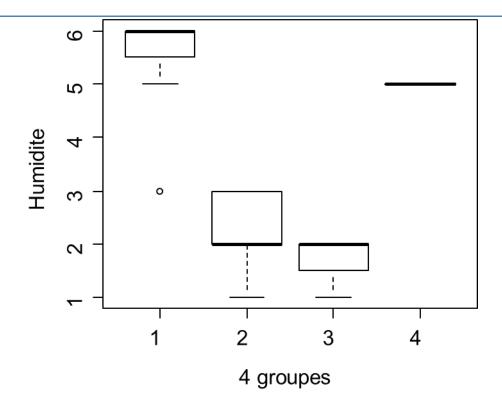
#### Analyse de la variance entre des groupes biologiques

```
# Groupement biologique
```

```
cluster <- hclust(d14, method = "ward")
cluster.4gr <- cutree(cluster, 4)
habitat =substr(rownames(esp),1,5)
table(habitat,cluster.4gr)

cluster.4gr
habitat 1 2 3 4
BASMA 4 0 0 0
LANMI 0 1 3 0
LANSA 0 4 0 0
LANTO 4 0 0 0
PELCA 0 0 4 0
PELZN 0 4 0 0
PRAAL 0 0 0 4
TOUHA 4 0 0 0
```

#### # Anova sur l'Humidite



#### **TP:Q6**

Réalisez une CCA entre le fichier biologique des Carabides et les facteurs écologiques disponibles ?

Le principe de ce TP est expliqué ci-dessus.

```
# Lecture des jeux de donnees
rm(list=ls())
# Definition de la directory de travail
# Attention !!! Pas d'accents et inversion des \
setwd("D:/Dufrene/GXABT Enseignements/ULB BING-F4002 Acquisition et analyse de
donnees/TP5 - Expliquer/Donnees")
# Pour verifier:
getwd()
[1] "D:/Dufrene/GXABT Enseignements/ULB BING-F4002 Acquisition et analyse d
e donnees/TP5 - Expliquer/Donnees"
# Lecture de donnees
eco=read.table('carabides_32sta_12eco.txt', h=T, sep="\t", row.names="Stations")
esp=read.table('carabides_32sta_103esp.txt', h=T, sep="\t", row.names="Stations")
# Transformation log pour les espèces
logesp=log(esp+1)
attach(eco)
logeco=data.frame(pHeau, pHKCL, logP=log(P+1),logK=log(K+1), logCa=log(Ca+1),
         logMg=log(Mg+1), logNa=log(Na+1), Humidite, row.names=rownames(eco))
```

#### # realise une CCA # =======

#### library(vegan)

#### CCAlisting <- cca(esp, logeco) CCAlisting

Call: cca(X = esp, Y = logeco)

Inertia Proportion Rank

Total 7.2395 1.0000 Constrained 3.7189 0.5137 8 Unconstrained 3.5206 0.4863 23

Inertia is mean squared contingency coefficient

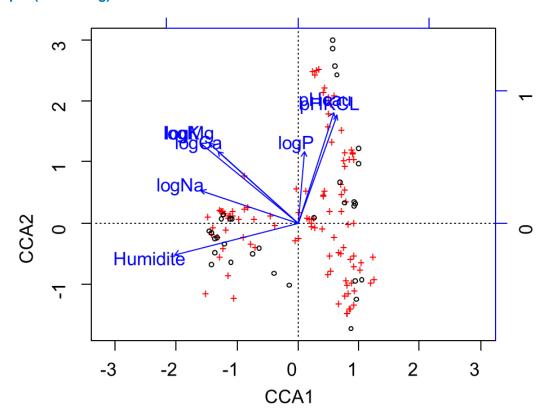
#### Eigenvalues for constrained axes:

CCA1 CCA2 CCA3 CCA4 CCA5 CCA6 CCA7 CCA8 0.8895 0.6636 0.5603 0.4678 0.4264 0.3536 0.2819 0.0759

#### Eigenvalues for unconstrained axes:

CA1 CA2 CA3 CA4 CA5 CA6 CA7 CA8 0.5772 0.5178 0.4439 0.3252 0.2553 0.2485 0.2351 0.2014 (Showed only 8 of all 23 unconstrained eigenvalues)

#### plot(CCAlisting)



```
TP: Q7
```

Réalisez une CCA entre le fichier biologique des Carabides et les facteurs écologiques disponibles ?

Le principe de ce TP est expliqué ci-dessus.

```
# realise une CAP
```

# ========

#### library(yegan)

# CAP sur les log des abondances

cap=capscale(logesp ~ pHeau+pHKCL+logP+logK+logCa+logMg+logNa+Humidite+Altitude, data=logeco,

distance="bray", sqrt.dist=F, add=T, dfun=vegdist)
summary(cap)

#### **Call:**

capscale(formula = logesp ~ pHeau + pHKCL + logP + logK + logCa + logM
g + logNa + Humidite + Altitude, data = logeco, distance = "bray", sqr
t.dist = F, add = T, dfun = vegdist)

Partitioning of squared Bray distance (euclidified):

**Inertia Proportion** 

Total 15.977 1.0000 Constrained 9.520 0.5959 Unconstrained 6.457 0.4041

Eigenvalues, and their contribution to the squared Bray distance (euclidified)

Importance of components:

CAP1 CAP2 CAP3 CAP4 CAP5 CAP6 CAP7

Eigenvalue 4.2271 1.5129 1.12396 0.80227 0.62350 0.51798 0.31940 Proportion Explained 0.2646 0.0947 0.07035 0.05021 0.03903 0.03242 0.01999 Cumulative Proportion 0.2646 0.3593 0.42963 0.47984 0.51887 0.55129 0.57128

CAP8 CAP9 MDS1 MDS2 0.20006 0.19280 1.24758 0.94165 0.01252 0.01207 0.07809 0.05894 0.58380 0.59587 0.67396 0.73289

Accumulated constrained eigenvalues

Importance of components:

CAP1 CAP2 CAP3 CAP4 CAP5 CAP6 ...
Eigenvalue 4.227 1.5129 1.1240 0.80227 0.62350 0.51798
Proportion Explained 0.444 0.1589 0.1181 0.08427 0.06549 0.05441
Cumulative Proportion 0.444 0.6029 0.7210 0.80528 0.87077 0.92518

Scaling 2 for species and site scores

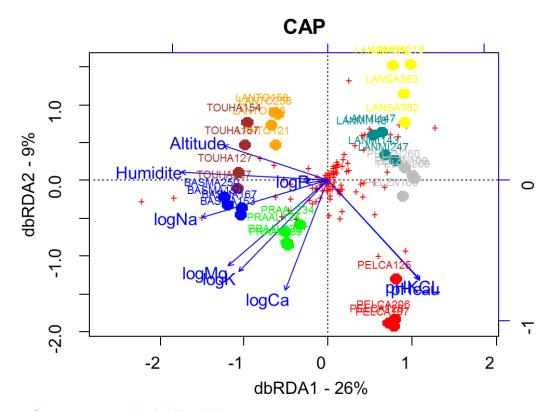
- \* Species are scaled proportional to eigenvalues
- \* Sites are unscaled: weighted dispersion equal on all dimensions
- \* General scaling constant of scores: 4.717504

#### # Graphique

colvec <- c("blue", "darkcyan", "yellow", "orange", "red", "grey", "green", "brown") figure=ordiplot(cap, scaling=2, xlab='dbRDA1 - 26%', ylab='dbRDA2 - 9%', main="CAP") points(figure, "sites", pch=16, col=colvec[eco\$Habitats], cex=1.5) text(figure, "sites",

labels = rownames(esp), # label de stations sur le graphique

```
col = colvec[eco$Habitats], # couleur du label (peut être un vecteur)
pos = 3, # position du label
cex = 0.6) # taille du label
```



# Le modèle complet est-il significatif ? anova(cap, perm=999)

1 0.1928

0.6569

99

CAP9

Permutation test for capscale under reduced model

```
Model: capscale(formula = logesp \sim pHeau + pHKCL + logP + logK + logCa + log
gMg + logNa + Humidite + Altitude, data = logeco, distance = "bray", sqrt.d
ist = F, add = T, dfun = vegdist)
                               Df
                                                   Var
                                                                                  F N.Perm Pr(>F)
                                  9 9.5200 3.6042
                                                                                                   199 0.005 **
Mode1
Residual 22 6.4567
Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
# Les axes contraints sont-il significatifs?
anova(cap, by="axis", perm=999)
Model: capscale(formula = logesp ~ pHeau + pHKCL + logP + logK + logCa +
logMg + logNa + Humidite + Altitude, data = logeco, distance = "bray",
sqrt.dist = F, add = T, dfun = vegdist)
                               Df
                                                   Var
                                                                                      F N.Perm Pr(>F)
                                  1 4.2271 14.4031
CAP1
                                                                                                       199
                                                                                                                       0.005
CAP2
                                   1 1.5129
                                                                  5.1550
                                                                                                       199
                                                                                                                        0.005 **
CAP3
                                   1 1.1240
                                                                    3.8297
                                                                                                       199
                                                                                                                        0.005
CAP4
                                   1 0.8023
                                                                    2.7336
                                                                                                       199
                                                                                                                        0.005 **
                                                                    2.1245
CAP5
                                  1 0.6235
                                                                                                       199
                                                                                                                        0.010 **
CAP6
                                  1 0.5180
                                                                    1.7649
                                                                                                       999
                                                                                                                       0.035 *
CAP7
                                  1 0.3194
                                                                    1.0883
                                                                                                          99
                                                                                                                        0.360
CAP8
                                  1 0.2001
                                                                    0.6817
                                                                                                          99
                                                                                                                        0.770
```

0.790

```
Residual 22 6.4567
Signif. codes: 0 '***' 0.001 '**' 0.05 '.' 0.1 ' ' 1
# Role marginal significatif des variables (compte tenu des autres)
anova(cap, by="margin", perm=999)
Permutation test for capscale under reduced model
Marginal effects of terms
Model: capscale(formula = logesp ~ pHeau + pHKCL + logP + logK + logCa + logMg + logNa + Humidite + Altitude, data = logeco, distance = "bray", sqrt.d
ist = F, add = T, dfun = vegdist)
           Df
                Var
                           F N.Perm Pr(>F)
            1 0.6446 2.1964
pHeau
                                   399 0.01750 *
            1 0.6468 2.2037
1 0.7811 2.6613
pHKCL
                                   399 0.01750 *
logP
                                   199 0.00500 **
            1 0.7588 2.5856
1 0.6879 2.3440
1 0.4815 1.6405
1 0.7112 2.4233
                                   199 0.01000 **
logK
logCa
                                   299 0.01667 *
logMg
                                   999 0.05100
                                   199 0.01000 **
logNa
                                   199 0.00500 **
Humidite 1 0.7131 2.4297
Altitude 1 0.5112 1.7418
                                   999 0.04400 *
Residual 22 6.4567
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

\* \* \*