BING-F4002 Acquisition et analyse de données

Fiche TP 2 : Distances et groupements

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
Lire les fichiers de données brutes du fichier « carabides_32sta_103esp.txt » et « carabides_32sta_12eco.txt » et identifier une manière simple pour vérifier que les données sont *à priori* correctement lues.

```
# Solution potentielle pour la Q1
# Initiation et remise à zéro des données précédentes
rm(list=ls())                                   # on élimine les données stockées
# Lecture de donnees
data=read.table('carabides_32sta_103esp.txt', h=T, sep="\t", row.names="Stations")
sum(data) ; dim(data) ; ls(data)        # verification de la somme totale sous xls = 14512
```

TP : Q2
Analyser les 9 variables du fichier « carabides_32sta_12eco.txt » (hors habitats, X et Y) pour visualiser leur distribution et proposer une transformation adhoc, si nécessaire. Après avoir vu la Q3, comment pourriez-vous montrer l'impact des transformations ?

```
# Lecture de données

eco=read.table('carabides_32sta_12 eco.txt', h=T, sep="\t", row.names="Stations")
sum(eco)                    # verification de la somme totale sous xls = 32765.13
ls(eco)                     # voir la liste des variables
attach(eco)                 # rendre cette liste directement accessible

# representation en histogramme avec environ 20 classes
hist(Altitude, breaks=20, prob=TRUE, xlab="P", col="lightgreen", main = "")
# ajout d'une courbe de distribution normale
curve(dnorm(x, mean=mean(Altitude), sd=sd(Altitude)), col="darkblue", lwd=2, add=TRUE)

# ecriture d'une fonction pour éviter de répéter les ordres
histo<-function(vareco)
```

```
{
hist(vareco, breaks=20, prob=TRUE, xlab= substitute(vareco), col="lightgreen", main = "");
curve(dnorm(x, mean=mean(vareco), sd=sd(vareco)), col="darkblue", lwd=2, add=TRUE);
}

# Mise en page pour voir 2 histogrammes d'un coup
        par(mfrow = c(1, 2))            # plot multi-panneaux

# appel de la fonction
histo(Ca);histo(Humidite);
histo(K) ; histo(Mg);
histo(Na) ; histo(P);
histo(pHeau) ; histo(pHKCL);

# Bilan potentiel (pas de règles absolues)
#    Humidite et Altitude sont OK ;
# pHeau et pHKCL => racine()
sqrt_pHeau=sqrt(pHeau) ; histo(pHeau) ; histo(sqrt_pHeau) ;
sqrt_pHKCL=sqrt(pHKCL) ; histo(pHKCL) ; histo(sqrt_pHKCL) ;
# Ca, K, Mg, Na, P => log base 2 ou népérien
Log_Ca =log(Ca+1, base=2); histo(Ca); histo(Log_Ca) ;
Log_K =log(K+1, base=2); histo(K); histo(Log_K) ;
Log_Mg =log(Mg+1, base=2); histo(Mg); histo(Log_Mg) ;
Log_Na =log(Na +1, base=2); histo(Na); histo(Log_Na) ;
Log_P =log(P+1, base=2); histo(P); histo(Log_P) ;

        # Reconstruction d'un dataframe

        eco_transforme = data.frame(Humidite, Altitude, sqrt_pHeau, sqrt_pHKCL,
            Log_Ca, Log_K, Log_Mg, Log_Na, Log_P, row.names=row.names(eco)) ;

# Solution plus radicale : tout transformer en log base 2 avec la fonction decostand()

library(vegan)

eco_log = decostand (eco,
method  = "log",
logbase = 2,
)

# mais cela implique un écrasement vers la droite de certaines variables qui avaient une distribution normale.

# Comment voir les différences ?

# => Calcul d'une distance euclidienne pour les differents dataframe (eco, eco_transforme et eco_log) et comparaison de matrice de distances (cfr Q3)
```

TP : Q3
Montrer les relations ou les différences entre les indices de distance euclidienne, de Bray-Curtis, de Soerensen (quantitatif et binaire), de Jaccard (quantitatif et binaire) sur des données brutes du fichier « carabides_32sta_103esp.txt ».

```
# Solution potentielle pour la Q3

# Initiation
rm(list=ls())   # on élimine les données stockées

library(vegan)

# Lecture de donnees
data=read.table('carabides_32sta_103esp.txt', h=T, sep="\t", row.names="Stations")
sum(data) ; dim(data) ; ls(data)          # verification de la somme totale sous xls = 14512

# d1 = Euclidienne
  d1 <- vegdist(data,method="euclidian")

# ds7 = 1-s7 = Jaccard
  ds7 <- vegdist(data,method="jaccard")                    # Distance de Jaccard quantitatif
  ds7bin <- vegdist(data,method="jaccard", binary=TRUE)   # Distance de Jaccard abs/pres
  s7 <- designdist(data, method ="(a)/(a+b+c)", abcd= TRUE)

# ds8 = 1-s8 = Soerensen
  ds8 <- vegdist(data,method="bray")
  ds8bin <- vegdist(data,method="bray", binary=TRUE)

# d14 = Bray-Curtis = 1 - S17 = 1 - Steinhaus
  d14 <- vegdist(data,method="bray")

# Comparaison des distances calculées
par(mfrow = c(3, 2)) # plot multi-panneaux
par(mar = c(5, 4, 4, 2)) # réduire les marges
plot( ds7bin, ds7, xlab="D7 – Jaccard - binaire", ylab=" D7 – Jaccard", col="chocolate")
plot( ds7bin, ds8bin, xlab="D7 – Jaccard - binaire", ylab="D8 – Soerensen - binaire", col="red")
plot( d14, d1, xlab="D14 - Brays-Curtis ou Steinhaus", ylab="D1 - Euclidienne", col="blue")
plot( ds7, d14, xlab="D7 – Jaccard", ylab="D14 - Brays-Curtis ou Steinhaus", col="green")
plot( ds7bin, d14, xlab="D7 – Jaccard - binaire", ylab="D14 - Brays-Curtis ou Steinhaus", col="orange")
plot( ds8, d14, xlab=" D8 - Soerensen", ylab="D14 - Brays-Curtis ou Steinhaus", col="brown")

# Les commentaires sont évidents …
```

TP : Q4
Comparer différentes approches de combinaison de calcul de distances (minimum 2 distances) et de méthodes de groupement (minimum deux distances) sur la base du fichier de données brutes du fichier « Demazy_2013_carabides.txt» et montrer les différences.
Analyser les résultats et proposer les arguments pour défendre le résultat qui vous semble le plus convaincant.

On attend vos suggestions par mail.