Lise BELLANGER Richard TOMASSONE

EXPLORATION DE DONNEES
ET
METHODES STATISTIQUES

SCRIPTS R

Le document contient tous les instructions R contenues dans notre livre.

Quelques lignes sont différentes des instructions du livre ; elles sont écrites en rouge.

Il y a 12 scripts :

- script-R-DM02-Outils-V2.r
- script-R-DM03-Avant Traitement-V2.r
- script-R-DM04-Cartes-ACP&MDS-V2.R
- script-R-DM05-Cartes-AFC&AFCM-V2.R
- script-R-DM06-Modèle Factoriel-V2.r
- script-R-DM07-Classes-V2.r
- script-R-DM08-Régression1-V2.r
- script-R-DM09-Régression2-V2.r
- script-R-DM10-Relations-V2.r
- script-R-DM11-Classement1-V2.r
- script-R-DM12-Classement2-V2.r
- script-R-DM13-Arbres-binaires-V2.r

Il contient en outre la description des fichiers et la liste des library utilisées (identique aux parties correspondantes de l'ouvrage)

Certaines fonctions sont utilisées dans plusieurs scripts : c'est le cas de la fonction Dist.forme

```
#   Tracé sur un même graphe de l'histogramme et de la densité de la
# gaussienne correspondante
Dist.forme<-function(x)
 {
 par(mfrow=c(2,2))
 ## pb donnees manquantes
 x.naomit<-na.omit(x)

 hist(x.naomit, col="gray", prob=TRUE,xlim=c(min(x.naomit),
 max(x.naomit)), main="")

 curve(dnorm(x,mean=mean(x.naomit),sd=sd(x.naomit)),add=TRUE,lwd=2,
 col="red")
 boxplot(x.naomit)
 iqd<-summary(x.naomit)[5] - summary(x.naomit)[2]
 points(mean(x.naomit), col = "orange", pch = 18)
 plot(density(x.naomit,width=2*iqd),xlab="x",ylab="",type="l",main="")
 qqnorm(x.naomit)
 qqline(x.naomit)
 }
```

Fichiers utilisés

Fichier	Chapitre d'utilisation	En exercice
Altise (74*3)		11, 13
Amphore-a (15*5)	11	
Asthme (43*2)	8	
BDDrelf (281*21)		13
Bledur (50*12)		5
CalciumRencher86a (10*4)	3	
capitales **(ade4)** (15*15)	4	
Ceram18 (185*132)	2	5, 9, 10
Chazeb-a (23*8)	3, 11, 12	7
chdage (100*3)	12	
Chiens (27*8)		5
Cuisine 29*13)		5
CureThermale (83*12)	7	
cystfibr **(ISwR)** (25*10)		8
Diabete (46*6)		4
dune **(vegan) (**20*30)	10	
dune.env **(vegan) (**20*5)	10	
Eaux1 (20*7)	2, 4, 6, 7	6, 7
Eaux2a (Tab.7B, ch 4) ?		
Eaux2010 (113*9)		11, 12, 13
enseignement.cor (13*13)	6 (lecture pg)	
Gevaudan-a (218*16)		3, 12
Glucose (52*6)		4
Grcs (392*23)		6
Humerus (43*2)		8
ICU (257*13)		12
jumeaux (26*4)	10	
juul (**ISwR**) (1339*6)		8
kanga(**faraway**)	13	
Kangourou (151*21)		11, 13
lnESB162 (162*33)	10	
Loup (43*8)		7, 11
LSA (10*6)		4
malaria **(ISwR)** (100*4)		8
Nematodes (222*9)		11, 13
Os-Griz-a (20*5)	3	3
ozone (**faraway**)	13	9
Patom (16*3)	8	
planete (101*3)	7	
Poumon (72*7)	2	2
PrefConsom (24*4)	5 (lecture*pg p.154)	

procespin (32*11)	9	
Procespinsup (25*11)	9	
Quadra (41*3)		12
RdtFromage (41*17)	8	9
ruspini **(cluster)** (75*2)		7
SocVitesse (24*3)	8	
Spores (15*3)	8	
Tabac (10*2)	3	
TacheMenage (14*5)	5	
usagedrogue.cor (13*13) (*n*=1634)	6	
Vie (31*8)	6	
voix (123*7)		3, 8
voix_acp (116*29)		10
zelazo (**ISwR**)		8

Avec fond jaune: pas sur le site, mais soit dans library soit inclus dans le programme.

CHAPITRE 2
LES OUTILS DE REPRESENTATION D'UN ECHANTILLON

```
#########################################################
#########################################################
#####          METHODES STATISTISTQUES          #####
#####                    ET                     #####
#####          EXPLORATION DE DONNEES           #####
#####                Chapitre 2                 #####
#####  LES OUTILS STATISTIQUE ET MATHEMATIQUE   #####
#####    DE REPRESENTATION D'UN ECHANTILLON     #####
#####          (Version juillet 2013)           #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "Eaux1.txt" 20 * 7 colonnes (6 variables suivies d'un
sigle)
#     "Ceram18.txt"  185 * 132
#     "Eaux2010.txt" 113 * 9
#     "poumon.txt" 72 * 7
# library utiles
#
#  Tableau 4a - Eaux1 : Boîtes à moustaches.
library(MASS); library(car); library(ade4); library(stats);
library(e1071)
Eaux1<-read.table("Eaux1.txt",h=T, row.names=7)
# Tracé des six boîtes à moustaches (Fig.2)
boxplot(Eaux1)
#
# Tableau 4b - Eaux1 : figures 2 et 4.
#  Tracé sur un même graphe de l'histogramme et de la densité
de la gaussienne correspondante
Dist.forme<-function(x)
 {
 par(mfrow=c(2,2))
 ## pb donnees manquantes
 x.naomit<-na.omit(x)

 hist(x.naomit, col="gray", prob=TRUE,xlim=c(min(x.naomit),
 max(x.naomit)),main="")


curve(dnorm(x,mean=mean(x.naomit),sd=sd(x.naomit)),add=TRUE,l
```

```
wd=2,
 col="red")
 boxplot(x.naomit)
 iqd<-summary(x.naomit)[5] - summary(x.naomit)[2]
 points(mean(x.naomit), col = "orange", pch = 18)

plot(density(x.naomit,width=2*iqd),xlab="x",ylab="",type="l",
main="")
 qqnorm(x.naomit)
 qqline(x.naomit)
 }
# Tracé des quatre graphiques réalisés sur les variables HCO3
et SO4
# (Fig.3)
Dist.forme(Eaux1[,1])
Dist.forme(Eaux1[,2])
#
# Tableau 5a - Eaux1 : médiane de HCO3.
Eaux1[,1] # valeurs de HCO3 dans le fichier
rank(Eaux1[,1]) # rang des valeurs de HCO3 dans le fichier
sort(Eaux1[,1]) # valeurs rangées par ordre croissant de HCO3
median(Eaux1[,1]) # médiane de HCO3
#
# Tableau 5b -  Eaux1 : Moyennes et moyennes équeutées
# HCO3 et SO4.
mean(Eaux1[,1])
mean(Eaux1[,1],trim=0.5)
mean(Eaux1[,1],trim=0.05)
mean(Eaux1[,2])
mean(Eaux1[,2],trim=0.5)
mean(Eaux1[,2],trim=0.05)
#
# Tableau 5c - Eaux1 : variances et écart-types
# HCO3 et SO4.
var(Eaux1[,1],na.rm=T)
sd(Eaux1[,1])
var(Eaux1[,2])
sd(Eaux1[,2])
#
# Tableau 6a - Eaux1 : asymétrie et aplatissement.
# Résumé par variable
summary(Eaux1)
# Ecart-type, asymétrie et aplatissement
library(e1071)
# Bibliothèque nécessaire pour les fonctions skewness &
kurtosis
for (i in 1:6  )
```

```
   {
   print(sd(Eaux1[, i],na.rm=T))
   print(skewness(Eaux1[, i],na.rm=T))
   print(kurtosis(Eaux1[,i],na.rm=T))
   }
#
# Tableau 7 - ceram18 : 3 histogrammes et densité.
ceram18<-read.table("ceram18.txt",h=T, row.names=1,sep="\t")
dim(ceram18) ; monnaie=t(ceram18[1,])
summary(na.omit(monnaie))
monnaie<-na.omit(monnaie)/100
par(mfrow=c(2,2))
hist(monnaie,nclass=4,xlab= "4 classes",prob=TRUE,
  col="lightblue", border="pink")
hist(monnaie,nclass=7,xlab= "7 classes",prob=TRUE,
  col="lightblue", border="pink")
hist(monnaie,nclass=10,xlab= "10 classes",prob=TRUE,
  col="lightblue", border="pink")
plot(x=c(3,16),y=c(0,0.7),type="n",bty="l",xlab="Annee/100",
  ylab="Estimation de la densite ")
rug(monnaie)
# lines(density(monnaie,width=
width.SJ(monnaie,method="dpi"),n=200), lty=1)
lines(density(monnaie,bw=0.15),col = "blue")
lines(density(monnaie,bw=0.25),col = "green")
lines(density(monnaie,bw=0.5),col = "red")
lines(density(monnaie,bw=1),col = "pink")
#
# Tableau 8 - Données du fichier Poumon.
Poumon<-read.table("Poumon.txt",h=T,sep="\t")
dim(Poumon)
attach(Poumon)
# Conversion de variables en facteurs
Poumon$Poussiere<-
  factor(Poumon$Poussiere,levels=1:3,labels=c("Haut","Moyen",
  "Bas"))
Poumon$Race<-
factor(Poumon$Race,levels=1:2,labels=c("Blanc","Autre"))
Poumon$Sexe<-
factor(Poumon$Sexe,levels=1:2,labels=c("Homme","Femme"))
Poumon$Tabagie<-factor(Poumon$Tabagie,levels=1:2,
  labels=c("Fumeur","Non-Fumeur"))
Poumon$Tempsemploi<-factor(Poumon$Tempsemploi,levels=1:3,
  labels=c("<10 ans","10-20 ans","> 20 ans"))
xtabs(Oui~Race+Poussiere)
 summary(xtabs(Oui~Race+Poussiere))##        Avec        test
d'independance ...
 xtabs(Non~Race+Poussiere)
```

```
 library(Hmisc)
 describe(Oui~Race+Poussiere)
# Trouve les proportions de malade par Race et Poussiere
 prop<-
tapply(Oui,list(Race,Poussiere),sum)/tapply(Non,list(Race,Pou
ssiere),sum)
 round(prop,3)
# Tracé de la figure 5
 barplot(prop,names=          c("Haut","Moyen","Bas"),legend=
c("Blanc","Autre"))
#
# scale(x, center = TRUE, scale = TRUE)
#
# Tableau 9 - Eaux1 : tracé des caractéristiques
# unidimensionnelles et multidimensionnelles.
#SCATTERPLOT
# Histogrammes sur la diagonale
 panel.hist <- function(x, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE,col="lightblue")
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
}
# Coefficients de corrélation (en valeur absolue) sur la
partie haute,
# Taille de police proportionnelle au coefficient de
correlation
 panel.cor <- function(x, y, digits=2, prefix="", cex.cor,
...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}
# Régression linéaire (lm) sur la partie basse ;
# d'où le signe du coefficient de correlation.
 panel.lm<-function(x,y){
    points(x,y)
    abline(lm(y~x))
    #lines(lowess(x,y),col="red")
}
```

```
# Fig. 6
 pairs(Eaux1, diag.panel=panel.hist, cex.labels = 2,
  font.labels=2, upper.panel=panel.cor,lower.panel=panel.lm )
#
# Tableau  10 - Décomposition singulière de Xa
Xa<-c(1,1,1,1,1,2,3,10)
dim(Xa)<-c(4,2)
A<-svd(Xa)
# L'instruction round(,d) permet de n'imprimer que d
decimals.
round(A$d,5)
round(A$u,5)
round(A$v,5)
round(A$d^2,5)
round(eigen(t(Xa)%*%Xa)$values,5)
round(eigen(t(Xa)%*%Xa)$vectors,5)
t(Xa)%*%Xa
A$v%*%diag(A$d^2)%*%t(A$v)
A$u%*%t(A$u)# Matrice H
Xa%*%solve(t(Xa)%*%Xa)%*%t(Xa)
# Forme directe de la matrice H
solve(t(Xa)%*%Xa)%*%t(Xa)
# Inverse généralisée X+ de X
ya<-c(1,2,3,6)
dim(ya)<-c(4,1)
solve(t(Xa)%*%Xa)%*%t(Xa)%*%ya
# Solution z du système ya=Xz
#
# Tableau 11 - Décomposition singulière de Xb.
Xb<-c(1,1,1,1,1,1,1,1,1,0,0,0,0,0,0,1,1,1)
dim(Xb)<-c(6,3)
B<-svd(Xb)
round(B$d,3)
round(B$u,3)
round(B$v,3)
round(B$v%*%diag(B$d^2)%*%t(B$v),0)# X'X
solve(t(Xb)%*%Xb)# Essai d'inversion de X'X
round(B$u[,1:2]%*%t(B$u[,1:2]),3)# matrice H
round(t(B$u[,1:2]%*%diag(1/B$d[1:2])%*%t(B$v[,1:2])),3)
# Inverse Moore-Penrose
yb<-c(7,36,2,13,44,18)
dim(yb)<-c(6,1)
round(t(yb)%*%B$u[,1:2]%*%diag(1/B$d[1:2])%*%t(B$v[,1:2]),3)
round(ginverse(Xb),3)
#
# EXERCICES : EAUX2010, EAUX1, POUMON
```

CHAPITRE 3
PRATIQUES UTILES AVANT TRAITEMENT

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES           #####
#####                     ET                    #####
#####          EXPLORATION DE DONNEES           #####
#####                Chapitre 3                 #####
#####      PRATIQUES UTILES AVANT TRAITEMENT     #####
#####          (Version juillet 2013)           #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "Eaux1.txt"      20 * 7
#     "Tabac.txt"      10 * 2
#     "Chazeb-a.txt"   23 * 8
#     "Os-Griz-a"      20 * 5
#     "Eaux2010.txt"  113 * 9
#
# library utiles : car, bootstrap, BHH2, DAAG,  DMwR, VIM,
Amelia
#
# Tableau 1 - Eaux1 : Transformations de Box & Cox
 library(car)
 attach(Eaux1)
 summary(L6<-powerTransform(cbind(HCO3,SO4,Cl,Ca,Mg,Na)~1))
 testTransform(L6,c(1,0,0,1,0,0))
# Fig.4
 par(mfrow=c(4,2))
 qqPlot(SO4,main="SO4")
 qqPlot(log(SO4),main="log(SO4)")
# qq.plot(SO4^(-0.3),main="SO4^(-0.3)")
 qqPlot(Cl,main="Cl")
 qqPlot(log(Cl), main="ln(Cl)")
 qqPlot(Mg, main="Mg")
 qqPlot(log(Mg), main="ln(Mg)")
 qqPlot(Na, main="Na")
 qqPlot(log(Na), main="ln(Na)")
#
# Tableau 2b - Tabac : corrélation goudron * nicotine :
Jackknife.
 library(bootstrap)
# Lecture des données Tabac
```

```
# Relation goudron * nicotine sur n = 10 marques de cigarette
 Tabac<-read.table("Tabac.txt",h=T, sep="\t")
 cor(Tabac)
# Jackknife
 xdata<- Tabac
 n<-nrow(Tabac)
 theta <- function(x,xdata)
 {
 cor(xdata[x,1], xdata[x,2])
 }
 JackTabac<-jackknife (1:n,theta,xdata)
# Valeurs obtenues dans la fonction : écart-type, biais, etc.
 JackTabac
# Pseudo-valeurs et estimation
 psTabac<-n*(cor(Tabac[,1],Tabac[,2]))-                    (n-
1)*JackTabac$jack.values
 psTabac
 mean(psTabac)
 cor(Tabac[,1],Tabac[,2])-JackTabac$jack.bias       #       =
mean(psTabac)
sqrt(var(psTabac)/n) # = JackTabac$jack.se
#
# Tableau 3 - Tabac : corrélation goudron * nicotine :
Bootstrap.
# Bootstrap
 xdata<- Tabac
 n<-nrow(Tabac)
 theta <- function(x,xdata)
 {
 cor(xdata[x,1], xdata[x,2])
 }
 B<-1000
 BootTabac<- bootstrap (1:n,nboot=B,theta,xdata)
 mean(BootTabac$thetastar)
 Dist.forme(BootTabac$thetastar)

quantile(BootTabac$thetastar,probs=c(0,0.005,0.025,0.50,0.975
,0.995,1))
#
# Tableau 4 -  ChaZeb-a : test de Student pour les deux
variables carcasse (carc) et os.
 ChaZeb<-read.table("ChaZeb-a.txt",header=T,sep = "\t",
row.names=1)
 dim(ChaZeb)
 names(ChaZeb)
 attach(ChaZeb)
 t.test(Carc~Groupe,data=ChaZeb, var.equal = T)
 t.test(Os~Groupe,data=ChaZeb, var.equal = T)
```

```
#
# Tableau 5 - ChaZeb-a : test de permutation pour carc et os.
# Bibliothèque BHH2
> library(BHH2)
> permtest(Carc[1:12],Carc[13:23])
> permtest(Os[1:12],Os[13:23])
# Bibliothèque DAAG
> library(DAAG)
> twotPermutation(Carc[1:12],Carc[13:23],20000)
> twotPermutation(Os[1:12],Os[13:23],20000)
#
# Tableau 6B - Os-Griz-a : calcul des D2 entre observations.
 Ramus<-read.table("Os-Griz-a.txt",h=T, sep="\t")
 dim(Ramus)
 names(Ramus)
 RamusDon<-Ramus[,2:5]
 colMeans(RamusDon) # c'est ȳ ci-dessous
 SRamus<-cov(RamusDon) # c'est S ci-dessous
 SRamus
 D2Ramus<-      mahalanobis(RamusDon,      colMeans(RamusDon),
cov(RamusDon))
 round(D2Ramus,4)# ce sont les 20 D2 de mahalanobis
 plot(density(D2Ramus, bw=.5), main="Distances de
Mahalanobis, n=20, p=4")
 rug(D2Ramus)
 qqplot(qchisq(ppoints(20), df=4), D2Ramus, main =
  expression("QQ-plot des " * ~D^2 * "de Mahalanobis vs.
  quantiles of" * ~ chi[4]^2))
 abline(0, 1, col = 'red')
#
# Tableau 8a - Eaux2010 : fichier des 113 eaux minérales.
 library(DMwR) ; library(VIM)
 Eaux2010<-read.table("Eaux2010.txt",h=T, row.names=7)
 dim(Eaux2010)
 Eaux2<-Eaux2010[,1:6]
 dim(Eaux2)
 countNA(Eaux2)
# Fig.6, avec un paramétrage des couleurs.
 aggr(Eaux2, col=c("yellow","grey"))
 Eaux2.agr= aggr(Eaux2)
 summary(Eaux2.agr)
 Eaux2[!complete.cases(Eaux2),]
 nrow(Eaux2[!complete.cases(Eaux2),])
 EauxComplete<-na.omit(Eaux2)
 dim(EauxComplete)
#
# TABLEAU 8B -  Eaux2010 : coefficients de corrélation entre
les 6 variables et présentation schématique.
```

```
 round(cor(EauxComplete[,1:6]),3)
 symnum(cor(EauxComplete[,1:6]))
#
# Tableau 8c - Eaux2010 : estimation des données manquantes.
# Estimation des valeurs NA par distance pondérée des 10 obs
les + proches
 Eaux2[,1:6]<-knnImputation(Eaux2[,1:6],k=10)
 EauxReconsMoy<-Eaux2 ; dim(EauxReconsMoy)
# Estimation des valeurs NA par la médiane des k =10 voisins
 Eaux2<-Eaux2010[,1:6]
 Eaux2[,1:6]<-knnImputation(Eaux2[,1:6],k=10,meth="median")
 EauxReconsMed<-Eaux2 ; dim(EauxReconsMed)
#  Collage  des  3  matrices  base(avec  NA)  et  les  deux
reconstituées
 EauxReconsVerif<-cbind(Eaux2010[,1:6],EauxReconsMoy[,1:6],
EauxReconsMed[,1:6])
 nrow(EauxReconsVerif[!complete.cases(EauxReconsVerif),])
#  Impression  des  observations  avec  les  2  estimations  de
valeurs manquantes
 round(EauxReconsVerif[!complete.cases(EauxReconsVerif),],1)
#
#Tableau 8e - Eaux2010 : comparaison aux deux estimations
 for (i in 1:6  )
    {
 a1<-
t.test(EauxReconsVerif[complete.cases(EauxReconsVerif),i]   ,
EauxReconsVerif[!complete.cases(EauxReconsVerif),i+6])
 print(a1$stat)
 print(a1$p.value)
 print(a1$estim)
    }

 for (i in 1:6  )
    {
 a1<-
t.test(EauxReconsVerif[complete.cases(EauxReconsVerif),i]   ,
EauxReconsVerif[!complete.cases(EauxReconsVerif),i+12])
 print(a1$stat)
 print(a1$p.value)
 print(a1$estim)
    }
#
 library(Amelia)
 Eaux2<-Eaux2010[,1:6]
 Eaux2.Amelia.out<-amelia(x=Eaux2[,1:6])
 plot(Eaux2.Amelia.out)

overimpute(Eaux2.Amelia.out,var=1,main="Observation/Estimatio
```

```
n pour la variable HCO3",xlab="Valeurs observées",ylab="Valeurs estimées")

overimpute(Eaux2.Amelia.out,var=4,main="Observation/Estimation pour la variable Ca",xlab="Valeurs observées",ylab="Valeurs estimées")
#
# EXERCICES, fichiers à utiliser : Gevaudan-a, voix
```

CHAPITRE 4
REPRESENTATION D'UN ECHANTILLON
PAR DES CARTES : *ACP*

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES           #####
#####                     ET                    #####
#####          EXPLORATION DE DONNEES           #####
#####                Chapitre 4                 #####
#####       REPRESENTATION D'UN ECHANTILLON      #####
#####           PAR DES CARTES : ACP             #####
#####          (Version juillet 2013)           #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "Eaux1.txt"      20 * 7
#     "capitales(ade4)" 15*15, matrice de corrélation

# library utiles : MASS, car, ade4, stats
#
# Tableau 3 - Eaux1 : Matrice R des coefficients de
corrélation.
 library(MASS)    ;    library(car)    ;    library(ade4)    ;
library(stats), library(bootstrap)
# Le nom des observations est dans la colonne 7
 Eaux1<-read.table("Eaux1.txt",h=T, row.names=7)
 round(cor(Eaux1),3)
#
Eaux1.acp <- dudi.pca(Eaux1,scale=TRUE)
# on choisit de garder 6 axes principaux :
# TAPER 6 avant de poursuivre
# autre possibilité : Eaux1.acp <- dudi.pca(Eaux1,scale=TRUE
, nf = 6 , scannf = FALSE)

# Tableau 4 - Eaux1 : ACPN nombre d'axes principaux.
round(Eaux1.acp$eig,3)# valeurs propres
 round(Eaux1.acp$eig/sum(Eaux1.acp$eig)*100,3) # % valeurs
propres
 round(cumsum(100* Eaux1.acp$eig /sum(Eaux1.acp$eig)),3)#
cumul %
#
# Tableau 5 - Eaux1 : ACP analyse des variables.
 inertie <-inertia.dudi(Eaux1.acp, col.inertia=TRUE)
```

```
# Analyse des variables
  round(Eaux1.acp$co,3)
# Contribution des variables en %
  round(inertie$col.abs/100,3)
# Qualité des variables en %
 round( inertie$col.re/100,3)
#
# Tableau 6 - Eaux1 : ACPN analyse des 20 observations.
# Analyse des observations.
  round(Eaux1.acp$l1,3)   # CP
  inertie <-inertia.dudi(Eaux1.acp, row.inertia=TRUE)
# Contribution des observations en %
  inertie$row.abs/100
# Qualité des observations en %
  inertie$row.re/100
#
# Tableau 7a - Eaux1 : tracé des graphiques du premier plan
principal.
# Cercle des corrélations (Fig2A)
  s.corcircle(Eaux1.acp$co)
# Plan 1 - 2 des observations (Fig2B)
  s.label(Eaux1.acp$li)
#
# Tableau 7b - Eaux1 : observations supplémentaires sur le
premier plan principal.
# Lecture du fichier de 113 observations (Eaux2010) avec 9
colonnes
 Eaux2<-read.table("Eaux2010.txt",h=T, row.names=7)
# Exemple de 4 observations supplémentaires
 Eaux2a<-Eaux2[c(92,46,20,5),1:7]
 Eaux2a
 Eaux2a[4,6]<-12.5
 Eaux2a
 labsup<-c("Tun","Zil","Che","Cub")
 s.arrow(Eaux1.acp$l1, sub = "Graphe ACP Eaux1 + Eaux2",
  possub = "bottomright", csub = 1.5,ylim=c(-7,2))
 s.label(suprow(Eaux1.acp,Eaux2a[,1:6])$lisup,label=labsup,
  add.plot = TRUE, clab = 1.5)
#
# Tableau 8a - Eaux1 : validation par jackknife de la
première valeur propre.
 xdata<-Eaux1
 n<-nrow(Eaux1)
 theta <- function(x,xdata)
 {
 dudi.pca(xdata[x,],scale=TRUE,   scannf   =   FALSE,   nf   =
6)$eig[1]
 }
 Result.VP1<-jackknife(1:n,theta,xdata)
```

```
 Result.VP1
#
# Tableau 8b - Eaux1 : validation par jackknife des six
valeurs propres.
 for (i in 1:6)
 {
  theta <- function(x,xdata)
 {
 dudi.pca(xdata[x,],scale=TRUE,   scannf   =   FALSE,   nf   =
6)$eig[i]
 }
 resultats<-jackknife(1:n,theta,xdata)
 print(i)
 print(resultats$jack.values)
 }
#
# Tableau 8c - Eaux1 : validation par bootstrap de la
première valeur propre.
 theta <- function(x,xdata)
 {
 dudi.pca(xdata[x,],scale=TRUE,   scannf   =   FALSE,   nf   =
6)$eig[1]
 }
 ResultatsbootVP1<- bootstrap(1:n,1000,theta,xdata,func=mean)
 ResultatsbootVP1$func.thetastar
 summary(ResultatsbootVP1$thetastar)
 Dist.forme(ResultatsbootVP1$thetastar)
# Résultats du jackknife after bootstrap et graphique
 summary(ResultatsbootVP1$ jack.boot.val)
 plot(ResultatsbootVP1$ jack.boot.val)
#
# Tableau 9a - Obtention d'un biplot.
 Gauche<-matrix(c(2,2,1,2,-1,1,1,-1,2,-
2),nrow=5,ncol=2,byrow=TRUE, dimnames = list(c("x1","x2"
,"x3","x4","x5"),c("A","B")))
 Droite<-matrix(c(3,2,-1,-2,1,-1,2,-
1),nrow=2,ncol=4,byrow=TRUE,
dimnames = list(c("a","b"),c("y1","y2","y3","y4")))
 biplot (Gauche, t(Droite))
#
# Tableau 9b - Eaux1 : tracé d'un biplot.
 biplot (Eaux1.acp$li[,1:2], Eaux1.acp$co[,1:2],xlab="Axe1",
ylab="Axe2")
#
# Tableau 10 - Eaux1 : MDS de la matrice des distances entre
observations.
 D.Eaux1<- dist(scale(Eaux1))# matrice de distances données
centrées réduites !!!
```

```
 dim(as.matrix(D.Eaux1))#  Dimension  de  la  matrice  de
distances
 Eaux1.mds<-cmdscale(D.Eaux1,k=6,eig=T)
#Analyse factorielle de la matrice de distance
 Eaux1.mds$eig # valeurs propres = (n-1)*valeurs propres de
l'ACP
 colMeans(Eaux1.mds$points) # la matrice est centrée
  round(diag(var(Eaux1.mds$points)),3)
# Variances des colonnes = valeurs propres de l'ACPN
# Ou dans la library ade4
#  Analyse  en  coordonnées  principales  de  la  matrice  de
distance
 Eaux1.pco<-dudi.pco(D.Eaux1,scannf = FALSE, nf=6)
 scatter(Eaux1.pco)  #  coordonnées  dans  le  plan  1-2  des
individus
#
# TABLEAU 11 - capitales : positionnement multidimensionnel.
# data(capitales)
# dim(capitales$df)
# names(capitales$df)
# d0 <- as.dist(capitales$df)
# is.euclid(d0)
# d1 <- cailliez(d0, TRUE)
# is.euclid(d1)
# capitales.mds<-cmdscale(d1,k=2,add=T,eig=T)
#    plot(capitales.mds$points[,1:2],xlab="Axe    1",ylab="Axe
#2",type="n")
#                      text(capitales.mds$points[,1:2],label=
#names(capitales$df),cex=0.7)
#    mtext(outer=T,"Reconstitution    carte    des    capitales
#européennes  à  partir  de  leur  distance",side="3",line=-
#2,cex=1.0)
#nouvelle version de capitale(ade4) d'où pour les version de
# R récentes 3 ….
data(capitales)
dim(as.matrix(capitales$dist))
attr(capitales$dist, "Labels")
d0 <- as.dist(capitales$dist)
is.euclid(d0)
d1 <- cailliez(d0, TRUE)
is.euclid(d1)
capitales.mds<-cmdscale(d1,k=2,add=T,eig=T)
plot(capitales.mds$points[,1:2],xlab="Axe        1",ylab="Axe
2",type="n")
text(capitales.mds$points[,1:2],label=   attr(capitales$dist,
"Labels"),cex=0.7)
mtext(outer=T,"Reconstitution carte des capitales européennes
à partir de leur distance",side="3",line=-2,cex=1.0)
```

```
#
# EXERCICES, , fichiers à utiliser : Eaux2010, Diabete,
Glucose, LSA.
```

CHAPITRE 5

REPRESENTATION D'UN ECHANTILLON PAR DES CARTES : *AFC* ET *AFCM*

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES             #####
#####                     ET                      #####
#####          EXPLORATION DE DONNEES             #####
#####                Chapitre 5                   #####
#####       REPRESENTATION D'UN ECHANTILLON        #####
#####        PAR DES CARTES : AFC ET AFCM          #####
#####          (Version juillet 2013)             #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "TacheMenage.txt"
#     "PrefConsom"    24 * 4
#
# library utile : ade4,
#
# Tableau 5a - TacheMenage : présentation des données.
 library(ade4)
 TacheMenage <-read.table("TacheMenage.txt",h=T, row.names=1)
 TacheMenage
#
# Tableau 5b - TacheMenage : test du  χ².
# Résultats du test du Chi-deux d'indépendance
 chisq.test(TacheMenage)
#
# Tableau 5c - TacheMenage : AFC.
 TacheMenage.afc<-dudi.coa(df = TacheMenage, scannf = F, nf =
3)
# Eboulis des valeurs propres (Fig.2)
 inertie<- TacheMenage.afc$eig/sum(TacheMenage.afc$eig)*100
 barplot(inertie,ylab="%
d'inertie",names.arg=round(inertie,2))
 title("Eboulis des valeurs propres en %")
 round(TacheMenage.afc$eig,2) # Valeurs propres
# Valeurs propres en %
 round(TacheMenage.afc$eig/sum(TacheMenage.afc$eig)*100,2)
```

```
# Lien statistique du test du Chi-deux d'indépendance et
inertie
 sum(TacheMenage.afc$eig)*sum(TacheMenage)
scatter.coa(TacheMenage.afc, method=1,
sub="Tâches ménagères",posieig="none") # Fig.3
```

Tableau 5d - **TacheMenage** : aide graphique à l'interprétation des axes de l'*AFC* et contributions.

```
# Aide graphique à l'interprétation des axes
 par(mfcol=c(1,2))
 score.coa (TacheMenage.afc,xax = 1,dotchart = TRUE)   ##
Fig.4A
 title("Répartition des modalités sur l'axe 1")
 abline(v=0)
 score.coa (TacheMenage.afc,xax = 2,dotchart = TRUE)  ##
Fig.4B
 title("Répartition des modalités sur l'axe 2")
 abline(v=0)
```

#

Tableau 5e - **TacheMenage** : aide à l'interprétation des axes de l'*AFC* : contributions et qualité.

```
# Qualité de représentation qlt et Contribution ctr
# Pour les lignes
 inertieL<-inertia.dudi(TacheMenage.afc, row.inertia=TRUE)
 inertieL$row.abs/100    # ctr des lignes en %
 inertieL$row.rel/100    # qlt des lignes en %
# Pour les colonnes
 inertieC<-inertia.dudi(TacheMenage.afc , col.inertia=TRUE)
 inertieC$col.abs/100    # ctr des colonnes en %
 inertieC$col.rel/100    # qlt des colonnes en %
```

#

Tableau 6 - **TacheMenage** : *AFC* par décomposition en valeurs singulières.

```
 n<-sum(TacheMenage)
 n
 SomL<-as.matrix(rowSums(TacheMenage))
 SomC<-as.matrix(colSums(TacheMenage))
 LI<-SomL/n
 CJ<-SomC/n
 NCr<-
  sqrt(diag(n/rowSums(TacheMenage)))%*%as.matrix(TacheMenage/
  n-(LI)%*%t(CJ))%*%sqrt(diag(n/colSums(TacheMenage)))
 NCr.svd<-svd(NCr)
 round(NCr.svd$d[1:2],5)
 round(NCr.svd$v[,1:2],4)
 round(NCr.svd$u[,1:2],4)
```

#

Tableau 8 - *AFCM* sous R, avec la fonction **dudi.ccm(ade4)**.

```
###### Réaliser l'AFC avec la fonction dudi.coa de la library
ade4
# Création du tableau disjonctif
#  disj<-acm.disjonctif(tableau)
# Résultats de l'AFCM stockés dans afcm
# afcm<-dudi.coa(df = disj, scannf = FALSE, nf = 3)
# L'éboulis des valeurs propres
# inertie<-afcm$eig/sum(afcm$eig)*100
#                                    barplot(inertie,ylab="%
d'inertie",names.arg=round(inertie,2))
# title("Eboulis des valeurs propres en %")
# Extraction des valeurs propres
# afcm$eig
# Etude de l'inertie et calcul des % :
# afcm$eig/sum(afcm$eig)*100
# Représentation graphique du plan factoriel
# scatter.coa(afcm, method = 1, sub = "Blé dur", posieig =
"none")
#
# Tableau 10 - PrefConsom : résultats de l'AFCM.
 Pref<-read.table("PrefConsom.txt",h=T)
 Pref
 summary(Pref)
# Tableau disjonctif complet d'un data frame ne contenant que
les facteurs
# (acm.disjonctif)
 disj<-acm.disjonctif   (Pref[,-1])   #création   du   tableau
disjonctif
 disj
#  AFCM  du  tableau  des  facteurs ;  « row.w »  fournit  les
pondérations,
# ici les effectifs Nb
 Pref.acm<-dudi.acm(df = as.data.frame(Pref[,-1]),
row.w=as.vector(Pref$Nb), scannf = FALSE, nf =6) #l'analyse
factorielle
#  Eboulis des valeurs propres (Fig.5A)
 inertie<-Pref.acm$eig/sum(Pref.acm$eig)*100
 barplot(inertie,ylab="%
d'inertie",names.arg=round(inertie,2))
 title("Eboulis des valeurs propres en %")
#  Valeurs propres
 round(Pref.acm$eig,4)
 round(Pref.acm$eig/sum(Pref.acm$eig)*100,2)   #les   valeurs
propres en %
#  Plans  factoriels  (Nuages  par  modalité  des  facteurs ;
Fig.5B)
 scatter(Pref.acm)
 par(mfrow=c(1,2))  # plan 1-2 et plan 1-3
 s.value(Pref.acm$li, Pref.acm$li[,2])
```

```
 s.value(Pref.acm$li, Pref.acm$li[,3])
# Aide à l'interprétation : axe 1 (Fig.5C : Ch4B-Pref-
AideAxe1-a.jpg)
 modal<-as.data.frame(Pref.acm$co)
 modal<-modal[sort.list(modal$Comp1),]
 dotchart(modal[,1],labels = row.names(modal),cex=0.8)
 title(sub="Répartition des modalités sur l'axe 1") ;
abline(v=0)
# Aide à l'interprétation : axe 2 (Fig.5C : Ch4B-Pref-
AideAxe2-a.jpg)
 modal<-as.data.frame(Pref.acm$co)
 modal<-modal[sort.list(modal$Comp2),]
 dotchart(modal[,2],labels = row.names(modal),cex=0.8)
 title(sub="Répartition des modalités sur l'axe 2") ;
abline(v=0)
# Autre représentation : les variables (Fig.5D : Pref-
Variables-1&2)
 plot(Pref.acm$co[,1],Pref.acm$co[,2],type="n",xlab="Axe
1",ylab="Axe 2", xlim=c(-1.4,1.4))
 text(Pref.acm$co[,1], Pref.acm$co[,2], label=
colnames(disj))
 title("Préférences consommateurs - plan des variables")
 abline(h=0,v=0)
# Autre représentation : les individus (Fig.5D : Pref-Obs-
1&2)
 plot(Pref.acm$li[,1],Pref.acm$li[,2],type="n",xlab="Axe
1",ylab="Axe 2", xlim=c(-1.4,1.4))
 text(Pref.acm$li[,1], Pref.acm$li[,2],
label=row.names(disj))
 title("Préférences consommateurs - Plan des individus") ;
abline(h=0,v=0)
# Contributions (absolues) des modalités des variables
# à la construction de chaque axe
 inertia.dudi(Pref.acm,col.inertia = T)$col.abs
# Contributions (absolues) des individus à la construction de
chaque axe
 inertia.dudi(Pref.acm,row.inertia = T)$row.abs
#
# EXERCICES, , fichiers à utiliser : Cuisine, ceram18,
Chiens, Bledur
```

CHAPITRE 6

ANALYSE FACTORIELLE : LE MODELE FACTORIEL

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES           #####
#####                     ET                    #####
#####          EXPLORATION DE DONNEES           #####
#####                Chapitre 6                 #####
#####  ANALYSE FACTORIELLE : LE MODELE FACTORIEL  #####
#####          (Version juillet 2013)           #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "enseignement.cor"         6 * 6
#     "Vie.txt"                 31 * 9
#     "usagedrogue.cor"        13 * 13
#
# library utiles :

# Tableau 1 - Notation d'enseignement, les données.
 enseignement.cor<- c(1., 0.44, 0.41, 0.29, 0.33, 0.25,
0.44, 1.00, 0.35, 0.35, 0.32, 0.33,
0.41, 0.35, 1.00, 0.16, 0.19, 0.18,
0.29, 0.35, 0.16, 1.00, 0.59, 0.47,
0.33, 0.32, 0.19, 0.59, 1.00, 0.46,
0.25, 0.33, 0.18, 0.47, 0.46, 1.00)
 enseignement.cor<-as.data.frame(enseignement.cor)
 enseignement.cor<-as.matrix(enseignement.cor)
 dim(enseignement.cor)<-c(6,6)
 dimnames(enseignement.cor)<-list(c("Français", "Anglais",
"Histoire", "Arithmétique", "Algèbre", "Géométrie "),
c("Français", "Anglais", "Histoire", "Arithmétique",
"Algèbre", "Géométrie "))
#
# Tableau 2 - Notation d'enseignement, résultats de l'AFCS.
 enseignement.fa12<-
factanal(covmat=enseignement.cor,factors=2,
method="mle",n.obs=1000,rotation="none")
 enseignement.fa12
```

```
 enseignement.fa22<-
factanal(covmat=enseignement.cor,factors=2,
method="mle",n.obs=1000,rotation="varimax")
 enseignement.fa22
#
# Tableau 4 - Notation d'enseignement, graphique des
pondérations sans ou avec rotation varimax.
 par(mfcol=c(2,1))
 plot(enseignement.fa12$loadings[,1],
enseignement.fa12$loadings[,2], type="n",xlab="Facteur 1",
ylab= "Facteur 2")
 text(enseignement.fa12$loadings[,1],
enseignement.fa12$loadings[,2],
label=dimnames(enseignement.cor)[[2]], cex=0.6)
 plot(enseignement.fa22$loadings[,1],
enseignement.fa22$loadings[,2], type="n",xlab="Facteur 1",
ylab= "Facteur 2")
 text(enseignement.fa22$loadings[,1],
enseignement.fa22$loadings[,2],
label=dimnames(enseignement.cor)[[2]], cex=0.6)
 mtext(outer=T, "Haut : sans rotation - Bas : bas rotation",
side=3, line=-2, cex=1.2)
#
# Tableau 6 - Vie : résultats de l'AFCS.
 Vie<-read.table("Vie.txt",h=T, row.names=1)
 Vie.fa1<-factanal(Vie,factors=1,method="mle")
 Vie.fa1
 Vie.fa2<-factanal(Vie,factors=2,method="mle")
 Vie.fa2
 Vie.fa3r1<-factanal(Vie,factors=3,method="mle",   scores   =
c("regression"))
 Vie.fa3r1
#
 round(Vie.fa3r1$scores,2)
#
# Tableau 9 - usagedrogue.cor : matrice des corrélations.
 usagedrogue.cor<-
c(1., 0.447, 0.421, 0.435, 0.114, 0.203, 0.091, 0.082, 0.513,
0.304, 0.245, 0.101, 0.245,
 0.447, 1., 0.619, 0.604, 0.068, 0.146, 0.103, 0.063, 0.445,
0.318, 0.203, 0.088, 0.199,
 0.422, 0.619, 1., 0.583, 0.053, 0.139, 0.110, 0.066, 0.365,
0.240, 0.183, 0.074, 0.184,
 0.435, 0.604, 0.583, 1., 0.115, 0.258, 0.122, 0.097, 0.482,
0.368, 0.255, 0.139, 0.293,
 0.114, 0.068, 0.053, 0.115, 1., 0.349, 0.209, 0.321, 0.186,
0.303, 0.272, 0.279, 0.278,
 0.203, 0.146, 0.139, 0.258, 0.349, 1., 0.221, 0.355, 0.315,
0.377, 0.323, 0.367, 0.545,
```

```
 0.091, 0.103, 0.110, 0.122, 0.209, 0.221, 1., 0.201, 0.150,
0.163, 0.310, 0.232, 0.232,
 0.082, 0.063, 0.066, 0.097, 0.321, 0.355, 0.201, 1., 0.154,
0.219, 0.288, 0.320, 0.314,
 0.513, 0.445, 0.365, 0.482, 0.186, 0.315, 0.150, 0.154, 1.,
0.534, 0.301, 0.204, 0.394,
 0.304, 0.318, 0.240, 0.368, 0.303, 0.377, 0.163, 0.219,
0.534, 1., 0.302, 0.368, 0.467,
 0.245, 0.203, 0.183, 0.255, 0.272, 0.323, 0.310, 0.288,
0.301, 0.302, 1., 0.304, 0.392,
 0.101, 0.088, 0.074, 0.139, 0.279, 0.367, 0.232, 0.320,
0.204, 0.368, 0.304, 1., 0.511,
 0.245, 0.199, 0.184, 0.293, 0.278, 0.545, 0.232, 0.314,
0.394, 0.467, 0.392, 0.511, 1.)
 usagedrogue.cor<-as.matrix(usagedrogue.cor)
 dim(usagedrogue.cor)<-c(13,13)
 dimnames(usagedrogue.cor)<-list(c("cigarettes",      "bière",
"vin",  "alcool",  "cocaïne",  "tranquillisant",  "médicaments",
"héroîne",      "marijuana",      "haschisch",      "inhalants",
"hallucinogène",   "amphétamine"),   c("cigarettes",   "bière",
"vin",  "alcool",  "cocaïne",  "tranquillisant",  "médicaments",
"héroîne",      "marijuana",      "haschisch",      "inhalants",
"hallucinogène", "amphétamine"))
 usagedrogue.fa<-lapply(1:6,function(nf)

factanal(covmat=usagedrogue.cor,factors=nf,method="mle",n.obs
=1634))
#
# Tableau 10 - usagedrogue.cor : résultats partiels de
l'AFCS.
 usagedrogue.fa
#
# Tableau 11a - usagedrogue.cor : reconstitution des
corrélations observées.
 pred<-
usagedrogue.fa[[6]]$loadings%*%t(usagedrogue.fa[[6]]$loadings
)+ diag(usagedrogue.fa[[6]]$uniquenesses)
 round(usagedrogue.cor-pred,digits=3)
#
# Tableau 13 - usagedrogue.cor : tracé du graphe des deux
premiers facteurs.
# Pondération 1 & 2
# Ch5-DrogueLoad-1&2.jpg
 plot(usagedrogue.fa[[6]]$loadings[,1],
usagedrogue.fa[[6]]$loadings[,2], type="n",xlab="Facteur 1",
ylab= "Facteur 2")
```

```
 text(usagedrogue.fa[[6]]$loadings[,1],
usagedrogue.fa[[6]]$loadings[,2],label=dimnames(usagedrogue.c
or)[[2]], cex=0.6)
 mtext(outer=T, "Usage drogue, facteurs 1 & 2",side=3, line=-
2, cex=1.2)
#
# EXERCICES, fichiers à utiliser : Eaux1, grcs
```

CHAPITRE 7

REPRESENTATION D'UN ECHANTILLON PAR DES CLASSES

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES             #####
#####                    ET                       #####
#####          EXPLORATION DE DONNEES             #####
#####                Chapitre 7                   #####
#####       REPRESENTATION D'UN ECHANTILLON       #####
#####                PAR DES CLASSES              #####
#####          (Version juillet 2013)             #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "Eaux1.txt"          20 * 7
#     "CureThermale.txt"   83 * 12
#     "planete"           101 * 4
# library utiles : cluster, ade4, mclust, Hmisc
#
# Les quatre eaux supplémentaires sont dans un fichier Eaux2a
# de même structure que Eaux1 (cf.Tab.7B, chapitre 4)
 Eaux2a<-Eaux2[c(92,46,20,5),1:6] # Eaux2a
 Eaux3<-rbind(Eaux1, Eaux2a) ; dim(Eaux3)
 Eaux3[24,6]<-12.5
#
# Tableau 1a - Eaux3 : résultats de l'algorithme kmeans
(Inertie et centres de gravité).
 Eaux3.cr<-scale(Eaux3)
 Eaux3.kmeans.3<-kmeans(Eaux3.cr,3)
 Eaux3.kmeans.3$withinss
 Eaux3.kmeans.4<-kmeans(Eaux3.cr,4)
 Eaux3.kmeans.4$withinss
 Eaux3.kmeans.5<-kmeans(Eaux3.cr,5)
 Eaux3.kmeans.5$withinss
# Centres de gravité des classes
 round(Eaux3.kmeans.3$centers,1)
 round(Eaux3.kmeans.4$centers,1)
 round(Eaux3.kmeans.5$centers,1)
#
```

```
# Tableau 1c - Eaux3 : critère d'Hartigan pour  le choix du
nombre de classes.
 swkmeans.3<-sum(Eaux3.kmeans.3$withinss)
 swkmeans.4<-sum(Eaux3.kmeans.4$withinss)
 swkmeans.5<-sum(Eaux3.kmeans.5$withinss)
 (swkmeans.3/swkmeans.4-1)*(nrow(Eaux3.cr)-2)
 (swkmeans.4/swkmeans.5-1)*(nrow(Eaux3.cr)-3)
#
# Tableau 2a - Eaux3 : résultats de l'algorithme pam.
 library(cluster)
 Eaux3.pam.3<-pam(Eaux3.cr,3)
 Eaux3.pam.4<-pam(Eaux3.cr,4)
 Eaux3.pam.5<-pam(Eaux3.cr,5)
# Classes fournies dans Tab.2B
 cbind(Eaux3.pam.3$clustering,Eaux3.pam.4$clustering,
Eaux3.pam.5$clustering)
 Eaux3.pam.3$clusinfo
 Eaux3.pam.4$clusinfo
 Eaux3.pam.5$clusinfo
 par(mfrow=c(1,3))
# Graphiques avec une ellipse autour de chaque classe
(Fig.3a)
 plot(Eaux3.pam.3,which=1) ; plot(Eaux3.pam.4,which=1)
 plot(Eaux3.pam.5,which=1)
# Graphiques des silhouettes (Fig.3b)
 par(mfrow=c(1,3))
 plot(Eaux3.pam.3,which=2) ; plot(Eaux3.pam.4,which=2)
 plot(Eaux3.pam.5,which=2)
#
# Tableau 3a - Eaux3 : partitions floues en 3, 4, 5 classes
de fanny.
 Eaux3.fanny.3<-fanny(Eaux3.cr,3)
 Eaux3.fanny.4<-fanny(Eaux3.cr,4)
 Eaux3.fanny.5<-fanny(Eaux3.cr,5)
# Coefficient de Dunn (brut et normalisé)
 Eaux3.fanny.3$coeff
 Eaux3.fanny.4$coeff
 Eaux3.fanny.5$coeff
# Degré d'appartenance des observations à la partition en 5
classes (Tab.3B)
 round(Eaux3.fanny.5$membership,2)
#
# Tableau 6 - Comparaison des histogrammes obtenus par hclust
et agnes. ALM obtenu par mstree(ade4).
 D1<-
matrix(c(0,6,14,6,8,6,0,8,8,2,14,8,0,4,12,6,8,4,0,1,8,2,12,1,
0), nrow=5,ncol=5,byrow=T)
 rownames(D1)<-c("a","b","c","d","e")
 colnames(D1)<- rownames(D1)
```

```
# Comparaison des histogrammes fournis pas hclust et agnes
#  par(mfrow=c(1,2))
#  hc <- hclust(as.dist(D1), "ave")
#  plot(hc)
#  D<-c(6,14,6,8,8,8,2,4,12,1)
#  agnes1<-agnes(D,diss=T, method = "average")
#  plot(agnes1,which=2)
# Impression de l'ALM
 library(ade4)
 neig<-mstree(as.dist(D1))
 s.label(D1,neig=neig)
#
```

Tableau 7 - Programmes permettant d'obtenir les résultats et les dendrogrammes.

```
 D<-c(6,14,6,8,8,8,2,4,12,1)
 agnes1<-agnes(D,diss=T, method = "average")
 agnes1$ac
 agnes1$order.lab<-c("a","b","d","e","c")
 plot(agnes1)
 mtext(outer=T, "Lien moyen (distance moyenne)",side=3,line=-
1,cex=1)
 agnes2<-agnes(D,diss=T, method = "single")
 agnes2$ac
 agnes2$order.lab<-c("a","b","d","e","c")
 agnes3<-agnes(D,diss=T, method = "complete")
 agnes3$ac
 agnes3$order.lab<-c("a","b","d","e","c")
 agnes5<-agnes(D,diss=T, method = "weighted")
 agnes5$ac
 agnes5$order.lab<-c("a","b","d","e","c")
 par(mfrow=c(2,2))
 plot(agnes2,which=2,main="Lien simple")
 plot(agnes3,which=2,main="Lien complet")
 plot(agnes1,which=2,main="Lien moyen")
 plot(agnes5,which=2, main="Lien moyen pondéré")
 mtext(outer=T, "Arbres  par quatre méthodes",side=3,line=-
1,cex=1.2)
#
```

Tableau 9 - **Eaux3** : résultats d'*agnes* pour quatre stratégies.

```
 Eaux3.cr<-scale(Eaux3)
 Eaux3.agnes2<-agnes(Eaux3.cr, method = "single")
 Eaux3.agnes2$ac
 Eaux3.agnes3<-agnes(Eaux3.cr, method = "complete")
 Eaux3.agnes3$ac
 Eaux3.agnes1<-agnes(Eaux3.cr, method = "average")
 Eaux3.agnes1$ac
 Eaux3.agnes4<-agnes(Eaux3.cr, method = "ward")
```

```
 Eaux3.agnes4$ac
 par(mfrow=c(2,2))
 plot(Eaux3.agnes2,which=2,main="Lien simple")
 plot(Eaux3.agnes1,which=2, main="Lien moyen")
 plot(Eaux3.agnes3,which=2,main="Lien complet")
 plot(Eaux3.agnes4,which=2,main="Critere de Ward")
#
# Tableau 10 - Eaux3 : classification par division, résultats
de diana.
 Eaux3.diana1<-diana(Eaux3.cr, metric="euclidean")
 Eaux3.diana1$dc
 par(mfrow=c(1,2))
 plot(Eaux3.diana1)
# mtext(outer=T, "Arbre par hiérarchie divisive et
# distance euclidienne  (Eaux3) ",side=3,line=-1,cex=1.2)
# Tableau 11a - CureThermale : résultats bruts de mona.
 library(cluster)
 Cure<-read.table("CureThermale.txt",header=T,sep="\t")
 dim(Cure)
 Cure1<-Cure[,1:9]
 Cure1.mona<-mona(Cure1)
 names(Cure1.mona)
 Cure1.mona$order
 Cure1.mona$variable
 Cure1.mona$step
#
 plot(Cure1.mona) # Fig.10
#
 library(mclust)
 planete<-read.table("Planete.txt",header=T,sep="\t")
 dim(planete)
 planete<- planete[,2:4]
 planete.clus<-Mclust(planete)
 planete.clus$modelName
 planete.clus$G
 round(planete.clus$parameters$mean,3)
 plot(planete.clus, planete)

# Classification avec recouvrement
 Eaux3.acp <- dudi.pca(Eaux1,scale=TRUE)
 xy<-data.frame(x=Eaux3.acp$l1[,1],y=Eaux3.acp$l1[,2])
 par(mfrow = c(2,2))
 for (k in 1:2) {
   neig=mstree(dist.quant(xy,1),k)
   s.label(Eaux3.acp$l1,neig=neig)
  }
#
# Utilisation des coefficients de corrélation entre variables
# comme des # distances
```

```
 dd <- as.dist((1 - cor(Eaux3))/2)
 round(1000 * dd)
#
# EXERCICES, fichiers à utiliser : Eaux1, ChaZeb-a, Loup
```

CHAPITRE 8

REGRESSION : LES BASES ET LES LIMITES

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES           #####
#####                     ET                    #####
#####          EXPLORATION DE DONNEES           #####
#####               Chapitre 8                  #####
#####    REGRESSION : LES BASES ET LES LIMITES    #####
#####             (Version août 2013)           #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "RdtFromage"      41 * 17
#     "asthme.txt"       43 * 2
#     "SocVitesse.txt"   24 * 3
#     "Spores.txt"       15 * 3
#     "Patom-a"          16 * 3
#
# library utiles :
# stats, Design, leaps, lmtest, car, rms, simpleboot
#
# Tableau 5 - RdtFromage : 7 mesures de routine (type A).
 RdtFromage<-read.table("RdtFromage.txt", h=T)
#  Description du Rdt en fonction des Mesures de routine
(A)
 RdtFromageA<-RdtFromage[,-(8:16)]
 summary(RdtFromageA)
# Tracé sur un même graphe de l'histogramme et
# de la densité de la gaussienne correspondante
 Dist.forme1<-function(x)
  {
 par(mfrow=c(2,2))
 # Problème des donnees manquantes
 x.naomit<-na.omit(x)
 hist(x.naomit, col="gray",
  prob=TRUE,xlim=c(min(x.naomit),max(x.naomit)), main="")
 curve(dnorm(x,mean=mean(x.naomit),sd=sd(x.naomit)),
  add=TRUE,lwd=2,col="red")
 boxplot(x.naomit)
 iqd<-summary(x.naomit)[5] - summary(x.naomit)[2]
 points(mean(x.naomit), col = "orange", pch = 18)
```

```
plot(density(x.naomit,width=2*iqd),xlab="x",ylab="",type="l",
main="")
 qqnorm(x.naomit)
 qqline(x.naomit)
  }
# Tracé des quatre graphiques réalisés sur la variables RFESC
 Dist.forme(RdtFromageA[,8])
#
# Tableau 6 - RdtFromage : modèle avec les 7 mesures de
routine (type A).
 RdtFromageA.lm1<-lm(RFESC~.,data=RdtFromageA[,-1])
 summary(RdtFromageA.lm1)
#
# Tableau 7 - RdtFromage : mesures d'influence du modèle avec
les 7 mesures de routine (type A).
# 1er diagnostic sur le modèle complet
 par(mfrow=c(2,2))
 plot(RdtFromageA.lm1,which=1:4)
# L'observation n° 1 est influente (distance de Cook>1)
#
#Tableau 8 - RdtFromage : modèle avec les 7 mesures de
routine (type A), sans l'observation n°1.
 RdtFromageA.lm2<-lm(RFESC~.,data=RdtFromageA[-1,-1])
 summary(RdtFromageA.lm2)
#
# Tableau 9 - RdtFromage : sélection progressive avec les
mesures de routine, sans l'observation n°1.
# Stepwise
 RdtFromageA.slm2 <- step(RdtFromageA.lm2)
 summary(RdtFromageA.slm2)
# recherche exhaustive
 library(leaps)
 recherche.ex<-regsubsets(RFESC~.,data=RdtFromageA[-1,-1])
 par(mfrow=c(2,2))
  plot(recherche.ex,scale= "bic", main="bic")
  plot(recherche.ex,scale= "Cp", main="Cp")
  plot(recherche.ex,scale= "adjr2", main="R^2_aj")
  plot(recherche.ex,scale= "r2", main="R^2")
#
#Tableau 10 - RdtFromage : vérification des suppositions de
base.
RdtFromageA.lm3<-RdtFromageA.slm2
# Diagnostics
 resRdtFromageA.lm3<-residuals(RdtFromageA.lm3)
 restanRdtFromageA.lm3<-rstudent(RdtFromageA.lm3)
# Graphes
 par(mfrow=c(2,2))
```

```
 plot(RdtFromageA.lm3,which=1:4) # Fig.1e

 influence.measures(RdtFromageA.lm3)    # diagnostic

 plot(rstudent(RdtFromageA.lm3),pch=".",ylab="Résid
studentisés")
 abline(h=c(-2,2))
 lines(lowess(rstudent(RdtFromageA.lm3)))

# Diagnostics : tests
 library(car) ; library(lmtest)
 bptest(RdtFromageA.lm3)   # H0 : Homoscédasticité
 shapiro.test(resRdtFromageA.lm3)  #H0=normalité des résidus
 Box.test(resRdtFromageA.lm3)   #Ho   ="rho=0",   pas   d'auto-
corrélation entre résidus
#
#Tableau 11 – RdtFromage : validation du modèle retenu à
trois régresseurs.
## Validation du modèle à 3 variables
# Par boostrap sur les lignes
 library(simpleboot)
 lboot <- lm.boot(RdtFromageA.lm3, R = 1000) # row=T
 summary(lboot) # resampling rows
 perc.lm(lboot,c(.025, .50, .975))
## Validation externe du modèle à 3 variables
# Par validation croisée
 library(rms)
 RdtFromageA.ols3<- ols(RFESC~CNE + NPN + CAS,
  data=RdtFromageA[-1,-1],x=TRUE, y=TRUE)
# Validation du modèle avec toutes les observations
# Nombre de répétitions boostrap B=40 et calcul de MSE
 validate(RdtFromageA.ols3, method="boot", B=40)
#  Validation  croisée ;  B  =nombre  de  groupes  d'observations
supprimées
 validate(RdtFromageA.ols3, method="crossvalidation" , B=10)
#
#Tableau 14b - asthme : résumé des données.
 asthme<-read.table("asthme.txt",h=T,sep="\t")
 summary(asthme)
#
# Tableau 14c - asthme : ANOVA1.
 summary(aov(mesure~Traitement,data=asthme))
# Tableau 14d - asthme : estimation des paramètres selon les
principales contraintes.
# Table des estimations des paramètres
# Estimations des paramètres sous la contrainte mu=0 : modèle
des moyennes
 asthme.mod1<-lm(mesure~Traitement-1,data=asthme)
```

```
 summary(asthme.mod1)
# Estimation sous la contrainte alpha_A=0 : A est témoin
 asthme.mod2<-lm(mesure~Traitement,data=asthme)
 summary(asthme.mod2)
# Estimation sous la contrainte somme
 summary(lm(mesure~Traitement,data=asthme,
 contrasts=list(Traitement ="contr.sum")))
#
# Tableau 14e - asthme : test de Bonferroni de comparaison
des moyennes.
# Comparaisons multiples : Bonferroni
 attach(asthme)
 pairwise.t.test(mesure,Traitement, p.adj="bonferroni")
 detach(asthme)
#
#  EXERCICES,  fichiers  à  utiliser  :  humerus,  juul(ISwR),
malaria(ISwR), cystfibr(ISwR), voix, zelazo(ISwR), asthme.
```

CHAPITRE 9

LA COLINEARITE : DU DIAGNOSTIC AUX REMEDES

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES           #####
#####                     ET                    #####
#####          EXPLORATION DE DONNEES           #####
#####                Chapitre 9                 #####
##### LA COLINEARITE : DU DIAGNOSTIC AUX REMEDES #####
#####             (Version août 2013)           #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "procespin.txt"      33 * 11
#     "procespinsup.txt"   25 * 12
#
# library utiles :
# MASS; car; ade4; stats; e1071; pls; lars; elasticnet
#
# Tableau 1a - procespin : premières statistiques.
## library à charger:
 library(MASS); library(car); library(ade4); library(stats);
    library(e1071); library(pls)
 options(digits=4)
 procespin0<-read.table("procespin.txt", h=T)
 procespin<- cbind(log(procespin0$y),procespin0)
 dimnames(procespin)[[2]][1]<-c("lny")
 procespin=as.data.frame(procespin)
 dim(procespin)
# Description unidimensionnelles
 summary(procespin)
 sd(procespin) # Ecart-type
# Corrélations
 round(cor(procespin),3)
# Graphe pairs
 pairs(procespin[,-2])
## put histograms on the diagonal
 panel.hist <- function(x, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(usr[1:2], 0, 1.5) )
    h <- hist(x, plot = FALSE)
```

```
    breaks <- h$breaks; nB <- length(breaks)
    y <- h$counts; y <- y/max(y)
    rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
}
## put (absolute) correlations on the upper panels,
## with size proportional to the correlations.
 panel.cor <- function(x, y, digits=2, prefix="", cex.cor,
...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
     txt <- format(c(r, 0.123456789), digits=digits)[1]
    txt <- paste(prefix, txt, sep="")
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}
### Fig.1
 pairs(procespin[,-2], lower.panel=panel.smooth,
  diag.panel =panel.hist,upper.panel=panel.cor)
#
# Tableau 2 - procespin : modèle de régression avec les dix
régresseurs.
# Standardisation des données : travail sur les données
centrées réduites
 procespinstd=as.data.frame(scale(procespin))
 options(digits=3)
 procespin.lm<-lm(lny~.,data=procespinstd[,-2])
 summary(procespin.lm)
# Diagnostics
 par(mfrow=c(2,2))
 plot(procespin.lm, which = c(1:4))
#
# Tableau 3 - procespin : calcul des VIF des dix régresseurs.
 vif(procespin.lm)
 mean(vif(procespin.lm))
#
#Tableau 4a - procespin : ACP des dix régresseurs.
 library(MASS);library(ade4)
 acp.procespin<-dudi.pca(procespin[,-
c(1,2)],scann=FALSE,nf=10) # ACPN
 acp.procespin
 inertia.dudi(acp.procespin, row.inertia=F,col.inertia=T)
# Coordonnées des 33 individus sur les 10 axes
 acp.procespin$li
#
# Tableau 4b - procespin : graphes ACP.
 par(mfrow = c(1,2))
```

```
 s.arrow(acp.procespin$li, xax = 1, yax = 2,
  sub = "Representation des individus" )
 s.corcircle(acp.procespin$co, , xax = 1, yax = 2,
  lab = names(procespin[,-1]),full = TRUE, box = TRUE,
  sub = "Cercle des correlations")
 par(mfrow = c(1,2))
 s.arrow(acp.procespin$li, xax = 5, yax = 7,
  sub = "Representation des individus" )
 s.corcircle(acp.procespin$co, , xax = 5, yax =7,
  lab = names(procespin[,-1]),full = TRUE,
  box = TRUE, sub = "Cercle des correlations")
#
# Tableau 4c - procespin : résultats de la PCR avec les 10
composantes.
 procespin.ACP<-cbind(procespinstd$lny,acp.procespin$l1)
 # l1= the row normed scores
  dimnames(procespin.ACP)[[2]][1]<-c("lny")
  summary(lm(lny~.,data=procespin.ACP))
# l1= the row normed scores
 dimnames(procespin.ACP)[[2]][1]<-c("lny")
 summary(lm(log(y)~.,data=procespin.ACP))
#
# Tableau 4d - procespin : résultats de la PCR avec les 4
meilleures composantes.
# Suppression des composantes non significatives : modèle
descriptif
 summary(lm(lny~RS1+RS5+RS7+RS9,data=procespin.ACP))
#
# Tableau 4e - procespin : modèle à 10 régresseurs avec les 4
meilleures composantes.
 library(pls)
# PCR  sur les 10 composantes sans validation croisée
 pcr0    <-    pcr(lny~.,    ncomp=10,data=procespinstd[,-2],
validation =c("none"))
 coef(pcr0,ncomp=10, intercept=FALSE ) # coef de régression
x1-x10
 rowSums(coef(pcr0,ncomp=10,                comps=c(1,5,7,9),
intercept=FALSE )[,,1:4])
#
# Tableau 4f - procespin : validation croisée du modèle à 10
composantes.
 set.seed(100)
 cvreg<-cvsegments(nrow(procespinstd),k=3,type="random")
# Les 33 observations sont divisées en 3 parties de 11
observations
 pcr1 <- pcr(lny~., ncomp=10,data=procespinstd[,-2],
  validation =c("CV"),segments=cvreg)
 summary(pcr1)
```

```
# Erreur quadratique moyenne de prévision
 msepcv.pcr1<-MSEP(pcr1,intercept=FALSE,
estimate=c("train","CV") )
 par(mfrow=c(1,2)) #fig 4
 plot(msepcv.pcr1,legendpos="topright")
 plot(explvar(pcr1),type="l",main="")
#
# Tableau 4g - procespin : choix du nombre de composantes à
conserver.
 ncomp.pcr<-which.min(msepcv.pcr1$val["CV",,]); ncomp.pcr
 plot(pcr1,ncomp=7,   asp=1,line=TRUE)   #   prévision   par
validation croisée
 plot(pcr1,plottype="scores",comps=1:7) # score plot
 explvar(pcr1)  #  variance  expliquée  par  chaque  composante
principale
 plot(pcr1,"loadings",comps=1:7, legendpos="topleft",
  labels= "numbers", xlab = "nm")
 abline(h=0) # loading plot
#
# Tableau 4h - procespin : prévision pour un nouvel
échantillon (n = 25).
# Modèle à 7 composantes
  pcr2<-pcr(lny~.,ncomp=ncomp.pcr,data=procespinstd[,-2])
 coef(pcr2)
# Nouvelles données centrées et réduites
 procespinsup<-read.table("procespinsup.txt", h=T)
 dim(procespinsup)
 procespinstdsup<-scale(procespinsup[,-c(1,12)],
  center=mean(procespin[,-c(1,2)]), scale=sd(procespin[,-
  c(1,2)]))
 lnyprev.pcr2<-
predict(pcr2,newdata=procespinstdsup)[,1,ncomp.pcr]
 lnyprev.pcr2
#
# Tableau 4i - procespin : validation par jackknife (LOO :
leave one out).
 pcr3 <- pcr(lny~., ncomp=10,data=procespinstd[,-2],
validation =c("LOO"))
 summary(pcr3)
 plot(RMSEP(pcr3),legendpos="topright")
#
# Tableau 5a - procespin : PLSR, recherche du nombre de
composantes.
 set.seed(100)
 cvreg<-cvsegments(nrow(procespinstd),k=3,type="random")
 plsr<-plsr(lny~., ncomp=10,data=procespinstd[,-2],
  validation =c("CV"),segments=cvreg)
# Choix du nombre de composantes
```

```
 msepcv.plsr<-MSEP(plsr,intercept=FALSE,
estimate=c("train","CV") )
 par(mfrow=c(1,2) )
 plot(msepcv.plsr,legendpos="topright") # Fig.7
 plot(explvar(plsr),type="l",main="")
 ncomp.plsr<-which.min(msepcv.plsr$val["CV",,])
 ncomp.plsr
#
# Tableau 5b - procespin : PLSR avec 4 composantes.
# Modèle à 4 composantes PLS
 plsr1<-plsr(lny~.,ncomp=ncomp.plsr,data=procespinstd[,-2])
#  Poids de chaque variables dans la constructions des
composantes pls
 loadings(plsr1)
 plsr1$scores      #les coordonnées des individus sur chaque
composante pls
#  Coefficients de chaque variable initiale de PLSR à 4
composantes
 coef(plsr1)
#
# Tableau 5d - procespin : PLSR avec 25 observations
supplémentaires.
  lnyprev.plsr1<-predict(plsr1,newdata=procespinstdsup,
  ,type=c("response"))[,1,ncomp.plsr]
  lnyprev.plsr1
# Coordonnées de chaque composantes PLS pour les 25 nouvelles
observations
 predict(plsr1,newdata=procespinstdsup, ,type=c("score"))
#
# Tableau 5e - procespin : PLSR prévision et intervalle de
confiance des 25 observations supplémentaires.
 lny<-        procespinstd$lny        ;        plsr1scores<-
as.matrix(plsr1$scores)
 newdataplsr1<-predict(plsr1,newdata=procespinstdsup,
,type=c("score"))
 plsr1scores<-
rbind(as.matrix(plsr1scores),as.matrix(newdataplsr1))
 lny<-as.matrix(c(lny,rep(NA,25)))
# Tableau composantes PLS + lny fichier entier
 plsr1scores=data.frame(lny,plsr1scores)
 newdataplsr1=plsr1scores[34:58,]      #   tableaux   des   25
observations à prévoir
 lm.plsr1=lm(lny~.,data=plsr1scores)

  ICprev=predict(lm.plsr1,newdata=as.data.frame(plsr1scores[3
  4:58,]),
  interval="pred",level=0.95)
 ICprev
```

```
#
# Tableau 5f - procespin : PLSR, validation par jackknife
# Validation externe par Leave-one-out
 plsr2<-plsr(lny~.,          ncomp=10,data=procespinstd[,-2],
validation =c("LOO"))
 summary(plsr2)
 plot(RMSEP(plsr2),legendpos="topright")
 plot(plsr2, ncomp=7,asp=1, line=TRUE)
 plot(plsr2, plottype="scores", comps=1:7) # Fig.9
 explvar(plsr2)
#
# Tableau 6a - procespin : régression Ridge.
 library(MASS) ; library(lars)
# Travail sur les donnees centrees-reduites
 lm.ridge(lny~.,lambda=seq(0,10,1),data=procespinstd[,-2])
 plot(lm.ridge(lny~.,data=procespinstd[,-2],lambda =
seq(0,10,1)))
 select(lm.ridge(lny~.,data=procespinstd[,-2],lambda =
seq(0,10,1)))
 select(lm.ridge(lny    ~.,data=procespinstd[,-2],lambda    =
seq(0,3,0.1)))
 plot(lm.ridge(lny ~.,data=procespinstd[,-2],lambda =
seq(0,3,0.01)))# Fig.10
#
# Tableau 6b. procespin : régression Ridge, choix de k.
 gridge<-lm.ridge(lny~.,data=procespinstd[,-
2],lambda=seq(0,1.7,0.01))
 select(gridge)
 abline(v=1.59)
#
# Tableau 6c. procespin : régression Ridge, calcul de TMSE.
 which.min(gridge$GCV)
 ypredg<-as.matrix(procespinstd[,-c(1,2)])                %*%
gridge$coef[,160]
 tmse<-function(x,y) sqrt(mean((x-y)^ 2))
 tmse(ypredg,procespinstd[,2])
 gridge$coef[,160]
#
# Tableau 6 - procespin : régression Lasso.
 library(MASS) ; library(lars)
# Travail sur les donnees centrees-reduites
 procespinstd.lasso <-lars(as.matrix(procespinstd[,3:12]),
  procespinstd$lny,, type="lasso")
 summary(procespinstd.lasso)
 coef(procespinstd.lasso)
 plot(procespinstd.lasso, breaks=F) # Fig.11
#
# EXERCICES, fichiers à utiliser : RdtFromage, ozone, ceram18
```

CHAPITRE 10

RELATIONS ENTRE DEUX GROUPES DE VARIABLES

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES           #####
#####                     ET                    #####
#####          EXPLORATION DE DONNEES           #####
#####               Chapitre 10                 #####
#####  RELATIONS ENTRE DEUX GROUPES DE VARIABLES  #####
#####             (Version août 2013)           #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "jumeaux.txt"      26 * 4
#     "dune"             20 * 30
#     "dune.env"         20 * 5
#     "lnESB162"        162 * 33
# library utiles : yacca, vegan
#
# Tableau 3 - jumeaux : Graphique des coefficients de
# corrélations.
# Library utiles
 library(yacca)
 jumeaux<-read.table("jumeaux.txt",h=T)
round(cor(jumeaux),3)
# Présentation graphique de la matrice de corrélation
 panel.hist <- function(x, ...)
 {
     usr <- par("usr"); on.exit(par(usr))
     par(usr = c(usr[1:2], 0, 1.5) )
     h <- hist(x, plot = FALSE,col="lightblue")
     breaks <- h$breaks; nB <- length(breaks)
     y <- h$counts; y <- y/max(y)
     rect(breaks[-nB], 0, breaks[-1], y, col="cyan", ...)
 }
# Les coefficients de corrélation sont dans la partie haute
droite,
# avec une taille de police proportionnelle à cette valeur.
 panel.cor <- function(x, y, digits=2, prefix="", cex.cor,
...)
 {
     usr <- par("usr"); on.exit(par(usr))
```

```
     par(usr = c(0, 1, 0, 1))
     r <- abs(cor(x, y))
     txt <- format(c(r, 0.123456789), digits=digits)[1]
     txt <- paste(prefix, txt, sep="")
     if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
     text(0.5, 0.5, txt, cex = cex.cor * r)
 }
# Droites de régression dans la partie basse gauche
 panel.lm=function(x,y){
 points(x,y)
 abline(lm(y~x))
 #lines(lowess(x,y),col="red")
 }
 pairs(jumeaux, diag.panel=panel.hist, cex.labels = 2,
font.labels=2, upper.panel=panel.cor,lower.panel=panel.lm )
#
# Tableau 4b - jumeaux : ACC , nombre d'axes principaux.
 Jumeaux.acc <- cca(jumeaux[,1:2], jumeaux[,3:4]) # avec la
library yacca
 summary(Jumeaux.acc)
# Graphe du premier plan canonique
 plot(Jumeaux.acc$canvarx[,1], Jumeaux.acc$canvary[,1],
type="n",xlab="Composante jumeau 1",ylab=" Composante jumeau
2", main="Premier ensemble canonique : R2= 0.789")
 text(Jumeaux.acc$canvarx[,1], Jumeaux.acc$canvary[,1],)
#
# Tableau 4c - jumeaux : ACC , les coefficients intra-groupe
et leurs graphes.
 Jumeaux.acc$xcoef # facteurs canoniques a¹ et a²
 Jumeaux.acc$canvarx # variables canoniques U¹ et U²
 round(cor(jumeaux[,1:2],Jumeaux.acc$canvarx),3)
 round(Jumeaux.acc$xstructcorr,3)     #     coefficient     de
corrélation intra-groupe X(u)
 round(Jumeaux.acc$xstructcorrsq,3)
 round(Jumeaux.acc$xcanvad ,3) # variance extraite par U¹ et
U²
 Jumeaux.acc$ycoef # facteurs canoniques b¹ et b²
 Jumeaux.acc$canvary # variables canoniques V¹ et V²
 round(cor(jumeaux[,3:4],Jumeaux.acc$canvary),3)
 round(Jumeaux.acc$ystructcorr,3)     #     coefficient     de
corrélation intra-groupe Y(v)
 round(Jumeaux.acc$ycanvad ,3) # variance extraite par V¹ et
V²
# Graphes intra-groupe
 par(mfrow=c(1,2))
 par(pin=c(2.5,2.5))

plot(Jumeaux.acc$xstructcorr[,1],Jumeaux.acc$xstructcorr[,2],
type="n",
```

```
xlab="u1",ylab="u2",xlim=c(-1,1), ylim=c(-1,1),
main="Premier jumeau",cex=0.7)

text(Jumeaux.acc$xstructcorr[,1],Jumeaux.acc$xstructcorr[,2],
label=dimnames(Jumeaux.acc$xstructcorr)[[1]],cex=0.7)
 symbols(0,0,circles=1,inches=F,add=T)
 par(pin=c(2.5,2.5))

plot(Jumeaux.acc$ystructcorr[,1],Jumeaux.acc$ystructcorr[,2],
type="n",
xlab="v1",ylab="v2",xlim=c(-1,1), ylim=c(-1,1),
main="Second jumeau ",cex=0.7)

text(Jumeaux.acc$ystructcorr[,1],Jumeaux.acc$ystructcorr[,2],
label=dimnames(Jumeaux.acc$ystructcorr)[[1]],cex=0.7)
 symbols(0,0,circles=1,inches=F,add=T)
#  mtext(  "  Interprétation  intra  ensembles  ",side=3,line=-
8,outer=T, cex=1.5)
#
# Tableau 4d - jumeaux : ACC , les coefficients intra-groupe
et leurs graphes.
#Corrélation inter-groupe X(v)
 round(cor(jumeaux[,1:2],Jumeaux.acc$canvary),3)
# Corrélation inter-groupe Y(u)
 round(cor(jumeaux[,3:4],Jumeaux.acc$canvarx),3)
## Redondances :
# Groupe X
 round(Jumeaux.acc$xvrd,3) # redondances pour le 1er groupe
 round(Jumeaux.acc$xrd,3)   # redondance totale pour le 1er
groupe
# Groupe Y
 round(Jumeaux.acc$yvrd,3) # redondances pour le 2er groupe
 round(Jumeaux.acc$yrd,3)   # redondance totale pour le 2er
groupe
## Graphes inter-groupes
 par(mfrow=c(1,2))
 par(pin=c(2.5,2.5))
 plot(sqrt(Jumeaux.acc$xcrosscorrsq[,1]),
  sqrt(Jumeaux.acc$xcrosscorrsq[,2]), type="n",  xlab="u1",
  ylab="u2", xlim=c(-1,1), ylim=c(-1,1), main="Jumeau
  2/Jumeau 1",cex=0.7)
 text(sqrt(Jumeaux.acc$xcrosscorrsq[,1]),
  sqrt(Jumeaux.acc$xcrosscorrsq[,2]),
  label= dimnames(Jumeaux.acc$ystructcorr)[[1]],cex=0.7)
 symbols(0,0,circles=1,inches=F,add=T)
 par(pin=c(2.5,2.5))
 plot(sqrt(Jumeaux.acc$ycrosscorrsq[,1]),
  sqrt(Jumeaux.acc$ycrosscorrsq[,2]), type="n", xlab="v1",
```

```
 ylab="v2",xlim=c(-1,1), ylim=c(-1,1), main="Jumeau 1/Jumeau
 2",cex=0.7)
text(sqrt(Jumeaux.acc$ycrosscorrsq[,1]),
 sqrt(Jumeaux.acc$ycrosscorrsq[,2]),
 label= dimnames(Jumeaux.acc$xstructcorr)[[1]],cex=0.7)
symbols(0,0,circles=1,inches=F,add=T)
mtext(outer=T, " Interprétation entre ensembles
 ",side=3,line=-8,cex=1.5)
# Tableau 5 - jumeaux : analyse procustéenne (procuste(ade4)
 et.procustes(vegan)).
# La library ade4 sur l'exemple des 25 jumeaux
library(ade4)
jumeau1<-jumeaux[,1:2]; jumeau2<-jumeaux[,3:4]
jumeaux.procuste<-procuste(jumeau1,jumeau2,scale=TRUE)
# Le graphique (Fig.3A) fournit les n couples de points
# reliés par un trait
plot(jumeaux.procuste)
# La library  vegan sur le même exemple
library(vegan)
jumeaux.proc<-procrustes(jumeau1,jumeau2,scale = FALSE,
symmetric = FALSE)
# scale = F : pas de facteur échelle s pour Y.
# Sortie partielle.
summary(jumeaux.proc)
# Graphique (Fig.3B) des résidus des 25 jumeaux après analyse
procustéenne
plot(jumeaux.proc, kind=2)
#
# Tableau 6 - ACR sous R, avec la fonction rda(vegan)
## Réalisation d'une ACR avec la fonction rda(vegan).
library(vegan)
data(dune)
data(dune.env)
# Utilisation de la variable Manure (engrais) dans le modèle
dune.engrais <- rda(dune ~ Manure, dune.env)
summary(dune.engrais)
plot(dune.engrais,display=c("wa","bp"))
#
# Tableau 7b - lnESB162 : Instructions R pour faire l'ACC et
les différents graphiques (figures 4, 5, 6 et 7).
## ACC avec la fonction cca de la library yacca
library(yacca)
ESB162<-read.table("lnESB162.txt",h=T)
XSol<- ESB162[,4:27]
YBle<- ESB162[,28:36]
dim(XSol)
names(XSol)
dim(YBle)
```

```
 names(YBle)
# ACC
 ESB162.acc<-cca(as.matrix(XSol), as.matrix(YBle))
# Graphiques standard
 plot(ESB162.acc) # Fig.4
# Graphiques spécifiques # Fig.5, plan (U1,V1)
 plot(ESB162.acc$canvarx[,1], ESB162.acc$canvary[,1],
type="n",xlab="Composante Sol 1",ylab=" Composante Blé 1",
main="Premier ensemble canonique : R2= 0.79")
 text(ESB162.acc$canvarx[,1], ESB162.acc$canvary[,1],
label=as.character(ESB162$Pop), cex=0.7)
# Fig.6 Variables Blé sur (V1,V2)
 par(pin=c(5,5))
 plot(sqrt(ESB162.acc$ycrosscorrsq[,1]),
sqrt(ESB162.acc$ycrosscorrsq[,2]),
type="n",  xlab="U1", ylab="U2", xlim=c(-1,1), ylim=c(-1,1),
main="Blé/Sol 1",cex=0.7)
 text(sqrt(ESB162.acc$ycrosscorrsq[,1]),
sqrt(ESB162.acc$ycrosscorrsq[,2]),                    label=
dimnames(ESB162.acc$ystructcorr)[[1]],cex=0.7)
 symbols(0,0,circles=1,inches=F,add=T)
# Fig.7, variables Sol sur (V1,V2)
 par(pin=c(5,5))
 plot(sqrt(ESB162.acc$xcrosscorrsq[,1]),
sqrt(ESB162.acc$xcrosscorrsq[,2]), type="n",  xlab="V1",
ylab="V2", xlim=c(-1,1), ylim=c(-1,1),
main="Sol/Blé 1 & 2",cex=0.7)
 text(sqrt(ESB162.acc$xcrosscorrsq[,1]),
sqrt(ESB162.acc$xcrosscorrsq[,2]), label=
dimnames(ESB162.acc$xstructcorr)[[1]],cex=0.7)
 symbols(0,0,circles=1,inches=F,add=T)
#
#  EXERCICES,  fichiers  à  utiliser  :  voix_acp,  jumeaux,
lnESB162, ceram18
```

CHAPITRE 11

DISCRIMINATION ET CLASSEMENT : I - COMMENT DECRIRE LA SEPARATION DE CLASSES

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES             #####
#####                     ET                      #####
#####          EXPLORATION DE DONNEES             #####
#####               Chapitre 11                   #####
#####       DISCRIMINATION ET CLASSEMENT :        #####
##### I-COMMENT DECRIRE LA SEPARATION DE CLASSES  #####
#####             (Version août 2013)             #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
# courant) :
#     "Amphore-a.txt"      15 * 5
#     "ChaZeb-a.txt"       23 * 8
# library utiles : ade4, MASS, klaR
#
# Tableau 1 - Génération de deux populations Normales de
# 50 observations chacune et tracé graphique.
 library(ade4)
 library(MASS)
 s <- matrix(c(1, 0.8, 0.8, 1), 2)
 set.seed(24122006)
# Génération des 2 échantillons, moyennes différentes,
# même covariances (s)
 x1 <- mvrnorm(50, c(0.3, -0.3), s)
 x2 <- mvrnorm(50, c(-0.3, 0.3), s)
 x <- rbind.data.frame(x1, x2)
 x <- scalewt(x, scale = F)
 fac <- factor(rep(1:2, rep(50, 2)))
 s.class(x, fac, col = c("red", "blue"))
 arrows(0, 0, 2, 0, lwd = 2)
 text(2, 0, "X1", pos = 1, cex = 2)
 arrows(0, 0, sqrt(2), sqrt(2), lwd = 2)
 text(sqrt(2), sqrt(2), "U2", pos = 4, cex = 2)
 arrows(0, 0, 0, 2, lwd = 2)
 text(0, 2, "X2", pos = 2, cex = 2)
 arrows(0, 0, -sqrt(2), sqrt(2), lwd = 2)
 text(-sqrt(2), sqrt(2), "U1", pos = 2, cex = 2)
```

```
#
#Tableau 2 - Amphore-a : AFD réalisée par les deux logiciels
lda(MASS) et discrimin(ade4).
 Amphore<-read.table("Amphore-a.txt",header=T,sep = "\t")
 Amphore$Grp<-as.factor(Amphore$Grp)
 attach(Amphore)
# Vérification de la prise en compte de la variable Grp comme
facteur
 summary(Amphore)
 library(MASS) # Avec la library MASS :
# Fonction lda - Pratique anglaise maximisation du rapport
inter/intra
 Amphore.lda <-lda(Grp~.,data=Amphore)
 names(Amphore.lda)
 Amphore.lda
# Coordonnées des 15 amphores sur les axes discriminants
 predict(Amphore.lda)$x
#### Avec la library ade4 : fonction discrimin :
# "Linear Discriminant Analysis (descriptive statistic)"
# Pratique française maximisation du rapport inter/totale
 library(ade4)
 discrimin(dudi.pca(Amphore[,-5], scan = F), Amphore$Grp,
scan = F)
 plot(discrimin(dudi.pca(Amphore[,-5], scan = F),
Amphore$Grp, scan = F))
#
#Tableau 3a - Amphore-a : décomposition des phases
essentielles des calculs d'une AFD.
 N<-nrow(Amphore)
 T<-(N-1)*cov(Amphore[,1:4])# Matrice Totale T à N-1 degrés
de liberté
 T
 M<-Amphore.lda$means #coordonnées des barycentres g1, g2, g3
des 3 groupes
 M
# Construction de la matrice de "design"
 X_k <- NULL; for(i in 1:3) X_k <- cbind(X_k, as.numeric(Grp
== levels(Grp)[i]))
 X_k
 G<-X_k %*% M
 # Matrice intraclasse ou residuelle W
 W <- t(as.matrix(Amphore[,1:4])-G) %*%
(as.matrix(Amphore[,1:4])-G)
 W
 B<- T-W # Matrice inter classes Between B
 B
 Wcov<-W/(N-ncol(X_k))
 Bcov<-B/(ncol(X_k)-1)
```

```
# Pratique anglaise : maximisation du rapport inter/intra
 eigen(solve(Wcov)%*%Bcov)
# Valeurs de μ1 et μ2 ;
# celles de μ3 et μ4 sont nulles et non négatives
# comme indiquées dans la sortie !
 mu=eigen(solve(Wcov)%*%Bcov)$values
# Relation entre mu et lambda
 lambda=mu/(1+mu)
#
# Tableau 3b - Amphore-a : graphiques important de l'AFD.
 V <- eigen(solve(Wcov)%*%Bcov )$vectors[,1:2]
 V
# Matrice de normalisation v^T W v = 1
 echelle<- sqrt(diag(diag((t(V) %*% Wcov %*% V))))
 echelle
 LD <- V %*% solve(echelle) # LD1 & LD2 de lda (au signe
près)
 LD
 X<-scale(Amphore[,1:4],scale=FALSE) # Données centrées
 LD12<-X%*%LD # Coordonnées des 15 amphores sur les axes
discriminants
 LD12
 MLD1<-    tapply((X%*%LD)[,1],Grp,mean)    #    Coordonnées
barycentres sur LD1
 MLD1
 MLD2<-    tapply((X%*%LD)[,2],Grp,mean)    #    Coordonnées
barycentres sur LD2
 MLD2
# Graphique du premier plan discriminant
 plot(LD12[,1],LD12[,2],xlab="LD1",ylab="LD2",type="n")
 mtext(outer=T, "Discrimination de trois époques d'amphores
  crétoises",side=3, line=-2, cex=1.2)
 points(MLD1[1],MLD2[1],pch=22)
 points(MLD1[2],MLD2[2],pch=23)
 text(LD12[,1],LD12[,2], label=as.character(Grp),cex=1)
 points(MLD1[3],MLD2[3],pch=25)
#
# Tableau 3c - Amphore-a : tracé d'un graphique d'estimation
de densité.
# Estimation de densité sur le premier plan discriminant
 famphore<-kde2d(LD12[,1], LD12[,2],n=50)
 contour(famphore, xlab="LD1",ylab="LD2")
 mtext(outer=T, "Discrimination de trois époques d'amphores
  crétoises ",side=3, line=-2, cex=1.2)
### Liens AFD avec ACC
# Vérification : la racine carrée des valeurs propres de
W^-1 B sont égales
# aux coefficients de corrélation canonique du couple (X, X_K)
```

```
 cancor(X,X_k)$cor

sqrt(eigen(solve(W)%*%B)$values[1:2]/(1+eigen(solve(W)%*%B)$v
alues[1:2]))
#Tableau 4a - Amphore-a : distances de Mahalanobis
 entre les 15 amphores et les barycentres des 3 classes.
 MLD1<- tapply((X%*%LD)[,1],Grp,mean) # première coordonnée
des barycentres
 MLD2<- tapply((X%*%LD)[,2],Grp,mean) # deuxième coordonnée
des barycentres
 MLD12<-cbind(MLD1,MLD2)
 round(dist(rbind(X%*%LD,MLD12)),1)
# Matrice des distances observations * barycentres
#
#Tableau 4d - Amphore-a : tableau des « resubstitutions »dans
les 3 classes.
 pred <- predict(Amphore.lda)$class
 table(Grp, pred)
#
# Tableau 4E - Amphore-a : validation croisée du classement
dans les 3 classes.
 Amphore.lda.jack <-lda(Grp~.,data=Amphore, CV=TRUE)
 table(Grp, Amphore.lda.jack$class)
 binom.test(7, 15)
# Identification des affectations des amphores aux 3 classes
 Amphore.lda.jack$class
#
# Tableau 5 - Amphore-a : discrimination sur une seule
variable discriminante.
# Valeurs de LD1 sur l'échantillon de 15 amphores
 ld1 <- predict(Amphore.lda)$x[,1]
 anova(lm(ld1 ~ Grp))
# Valeurs de LD2 sur l'échantillon de 15 amphores
 ld2 <- predict(Amphore.lda)$x[,2]
 anova(lm(ld2 ~ Grp))
 detach(Amphore)
#
# Tableau 7 - ChaZeb-a : discrimination des Charolais et des
Zébus par lda.
 ChaZeb <- read.table("ChaZeb-a.txt",header=TRUE,
row.names=1) ; ChaZeb
 attach(ChaZeb)
 ChaZeb.lda <- lda(Groupe ~ Vif+Carc+Qual1+Tot+Gras+Os,
ChaZeb)
 ChaZeb.lda
 ChaZeb.lda.pred<-predict(ChaZeb.lda)
 plot(ChaZeb.lda)
#
```

```
# TABLEAU 8 - ChaZeb-a : discrimination des Charolais et des
Zébus, resubstitution des observations.
 tapply(ChaZeb.lda.pred$x,Groupe,mean)
 dist(tapply(ChaZeb.lda.pred$x,Groupe,mean))^2
 pred<-predict(ChaZeb.lda)
 table(Groupe, pred$class)
#
# Tableau 9 - ChaZeb-a : discrimination des Charolais  et des
Zébus par le programme de régression lm.
 Pop=c(rep(-11/23,12),rep(12/23,11))
 ChaZeb.lm <- lm(Pop ~ Vif+Carc+Qual1+Tot+Gras+Os, ChaZeb)
 summary(ChaZeb.lm)
#
#Tableau 10 - Discrimination des Charolais  et des Zébus
comparaison des résultats avec ceux d'une régression.
 gCha <- ChaZeb.lda$means[1,]
 gCha
 gZeb <- ChaZeb.lda$means[2,]
 gZeb
 N<-nrow(ChaZeb)
 M<-ChaZeb.lda$means# coordonnées des moyennes des 3 groupes
 G <- NULL; for(i in 1:2) G <- cbind(G, as.numeric(Groupe ==
levels(Groupe)[i]))
 W <- t(as.matrix(ChaZeb[,1:6]) - G %*% M) %*%
(as.matrix(ChaZeb[,1:6]) - G %*% M) # Matrice Résiduelle W
 Wcov<-W/(N-ncol(G))
 aChaZeb<-solve(Wcov)%*%(as.matrix(gCha-gZeb))
 aChaZeb
 aChaZeb/ChaZeb.lm$coef[2:7]
#
# Tableau 11a - ChaZeb-a : discrimination des Charolais  et
des Zébus sur deux variables (Carc et Qual).
 library(klaR)
# Stepclass : Selection de variables forward/stepward
 resdisc=stepclass(x=ChaZeb[,-7],
grouping=Groupe,method="lda",
direction="both",start.vars = "Carc")
 resdisc
 plot(resdisc)
 ChaZeb.lda2 <- lda(Groupe ~ Carc+Qual1, ChaZeb)
 ChaZeb.lda2.pred<-predict(ChaZeb.lda2)
 tapply(ChaZeb.lda2.pred$x,Groupe,mean)
 dist(tapply(ChaZeb.lda2.pred$x,Groupe,mean))^2
 df(1.725442,4,16)
#
# Tableau 11b - ChaZeb-a : resubstitution des Charolais  et
des Zébus sur deux variables (Vif et Qual).
 pred2<-predict(ChaZeb.lda2)
```

```
 table(Groupe, pred2$class)
 ChaZeb.lda2j<-lda(Groupe ~ Carc+Qual1, ChaZeb, CV=TRUE)
 pred2j<-predict(ChaZeb.lda2)
 table(Groupe, pred2j$class)
"
# EXERCICES, fichiers à utiliser : LOUP, ALTISE, KANGOUROU,
NEMATODES, EAUX2010
```

CHAPITRE 12

DISCRIMINATION ET CLASSEMENT : II - COMMENT AFFECTER DES OBSERVATIONS A DES CLASSES

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES             #####
#####                     ET                      #####
#####          EXPLORATION DE DONNEES             #####
#####               Chapitre 12                   #####
#####       DISCRIMINATION ET CLASSEMENT :        #####
#####    II-COMMENT AFFECTER DES OBSERVATIONS     #####
#####               A DES CLASSES                 #####
#####             (Version août 2013)             #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "Quadra.txt"     40 * 3
#     "Chazeb-a"       23 * 8
#     "chdage.txt"    100 * 2
#
# library utiles : MASS, PresenceAbsence, Design
#
# Tableau 1a - Quadra : exemple fictif de discrimination
quadratique pour deux populations.
 Quadra <- read.table("Quadra.txt", head= T,sep = "\t")
 attach(Quadra)
 library(MASS)
 par(mfrow=c(1,2))
 plot(X,Y,xlab="X",ylab="Y",type="n")
 text(X,Y,label=as.character(Pop),cex=0.8)
 mtext(outer=T, "Discrimination de deux populations",side=3,
line=-2, cex=1.2)
 f1<-kde2d(X,Y,n=50)
 contour(f1, xlab="X",ylab="Y")
 mtext(outer=T, "Discrimination de deux populations",side=3,
line=-2, cex=1.2)
#
#Tableau 1b - Quadra : comparaison lda et qda.
# Discrimination linéaire classique, fonction lda
 Quadra1lda <-lda(Pop~.,data=Quadra)
 Quadra1lda$svd
```

```
 predQuadra1lda<-predict(Quadra1lda)
 table(Pop, predQuadra1lda$class)
# Discrimination quadratique, fonction qda
 Quadra1qda <-qda(Pop~.,data=Quadra)
 predQuadra1qda<-predict(Quadra1qda)
 table(Pop, predQuadra1qda$class)
#
#Tableau 2a - ChaZeb-a : discrimination des Charolais et des
# Zébus : anova (aov) et régression logistique (glm) de la
# variable Gras.
 ChaZeb <- read.table("ChaZeb-a.txt",header=TRUE,
row.names=1)
 attach(ChaZeb)
 round(tapply(Gras,Groupe,mean),1)
 summary(aov(Gras~Groupe))
 modChazebGras <- glm(Groupe ~ Gras,
  family=binomial(link=logit), data=ChaZeb)
 summary(modChazebGras)
 round(cbind(Gras,modChazebGras$fitted),2) # sortie arrangée
(Tab III)
#
# Tableau 4a - chdage : regroupement des observations et
graphe.
 library(MASS)
 library(PresenceAbsence)
 chd<-read.table("chdage.txt",header=T,sep = "\t")
 attach(chd)
 dim(chd)
 names(chd)
# Regroupement en classes d'âges
 x <- c(19.999,29,34,39,44,49,54,59,70)
 mid <- c((x[2:9]+x[1:8])/2)
# Donne la valeur moyenne de CHD aux classes, on construit un
vecteur GRAGE
# qui répartit les valeurs dans des classes en utilisant la
fonction cut
 GRAGE <- cut(AGE, breaks=x)
 y <- tapply(CHD, GRAGE, mean)
 round(y,2)
 plot(AGE, CHD)
# On rajoute les valeurs de y sur le graphiques des valeurs
d'origine
 points(mid, y, col="red", pch=3) # % des classes
#
# Tableau 4b - chdage : Régression logistique binaire sur la
variable AGE.
# Régression logistique binaire simple
 mod <- glm(CHD ~ AGE, family=binomial(link=logit))
```

```
 summary(mod)
# log-vraisemblance
 logLik(mod)
## Estimation de prob(CHD=1/age=x)
 mod$fitted.values
 plot(AGE, mod$fitted) # estimation du modèle
#
# Tableau 4c - chdage : tracé du graphe avec intervalle de
confiance 95%.
# Représentation de l'IC à 95% de la régression (Fig.3)
 grid <-(20:69) #  domaine de la régression
 se <- predict(mod, data.frame(AGE=grid), se=TRUE)
 gl <- binomial(link=logit) #fonction logit est la fonction
de lien utilisé
 plot(mid, y, col="red", pch=3, ylim=c(0,1), ylab="CHD",
xlab="AGE")
 lines(grid, gl$linkinv(se$fit))
 lines(grid, gl$linkinv(se$fit-1.96*se$se), col="red", lty=2)
 lines(grid, gl$linkinv(se$fit+1.96*se$se), col="red", lty=2)
# Estimation matrice variance-covariance
V<-vcov(mod) ; V
# Calcul de l'IC 95%(pi) correspondant à la valeur de AGE=50
# par exemple :
 x <- 50
 sqrt(V[1,1] + x^2 * V[2,2] + 2*x*V[1,2])
#[1] 0.2542218
# Plus simple !
 predict(mod, data.frame(AGE=50), se=TRUE)
#
# Tableau 4d - chdage : calcul de la déviance de la variable
AGE.
# Test de la déviance :
#  Comparaison modèle avec variable AGE au modèle avec
constante
 anova(mod, test="Chisq")
#
# Tableau 4e - chdage : calcul de la déviance de la variable
AGE.
 library(Design)
 mod <- lrm(CHD ~ AGE, x = TRUE, y = TRUE)
# Resultats identiques à glm
 mod
 anova(mod) # statistique de Wald
 par(mfrow=c(3,3))
 resid(mod, 'partial', pl=TRUE)          # Graphe des résidus
partiels
 resid(mod, 'gof')           #test GOF séparément pour chaque
niveau
 plot(resid(mod, "lp1"))
```

```
 resid(mod,'gof') # test global de qualité d'ajustement
 which.influence(mod,.2)
 vif(mod) # indice vif pour multicolinearité
# Validation de l'ajustement du modèle complet
 validate(mod,B=100)
# Validation des index d'ajustement des modèles ajustés
# Bootstrap : Covariance et Distribution des coefficients de
régression
# B=500 répétitions bootstrap
 g <- bootcov(mod,B=500, pr=TRUE, coef.reps=TRUE)
# Impression des quantiles bootstrap
# Normalité des estimations de la régression
 par(mfrow=c(2,2))
 bootplot(g,which=1:2,what="qq")
# Graphes des histogrammes et densités estimatées de tous les
coefficients
 w <- bootplot(g, which=1:2)
# Impression des quantiles bootstrap
 round(w$quantiles,3)
#
#Tableau 7 - chdage : étude de la qualité des résultats.
 library(PresenceAbsence)
 mod <- glm(CHD ~ AGE, family=binomial(link=logit))
 DATA.ROCR.chd.lg<-matrix(0,nrow=nrow(chd),ncol=3)
 DATA.ROCR.chd.lg[,1]<-1:nrow(chd)
 DATA.ROCR.chd.lg[,2]<-as.numeric(mod$y)
 DATA.ROCR.chd.lg[,3]<-mod$fit
 DATA.ROCR.chd.lg<-as.data.frame(DATA.ROCR.chd.lg)
 dimnames(DATA.ROCR.chd.lg)[[2]]<-c("ID"
,"Observed","Predicted")
 dimnames(DATA.ROCR.chd.lg)[[2]]
# Matrice de confusion : fonction cmx( )
# La césure (threshold) a pour valeur 0.5
 cmx(DATA.ROCR.chd.lg,threshold=0.5)
# Calcul de la specificite et de la sensibilite
 sensitivity(cmx(DATA.ROCR.chd.lg,threshold=0.5))
 specificity(cmx(DATA.ROCR.chd.lg,threshold=0.5))
# Courbe ROC pour le modele logistique CHD
 roc.plot.calculate(DATA.ROCR.chd.lg)
 auc.roc.plot(DATA.ROCR.chd.lg) # graphe courbe ROC, cf.
Fig.10.
 auc(DATA.ROCR.chd.lg) # calcul AUC
#
# Tableau 8 - chdage : calcul de différents seuils.
 presence.absence.summary(DATA.ROCR.chd.lg)
 optimal.thresholds(DATA.ROCR.chd.lg)
#
```

```
 presence.absence.hist( DATA.ROCR.chd.lg, na.rm=TRUE,
model.names=c("Modèle 1"), N.bars=10,truncate.tallest=TRUE,
opt.thresholds=TRUE,xlab="Probabilité prédite",
ylab="nombre d'observations")
#
# EXERCICES, fichiers à utiliser : Loup, ICU, Eaux2010,
Gevaudan-a.
```

CHAPITRE 13

ARBRES BINAIRES

```
#########################################################
#########################################################
#####           METHODES STATISTIQUES           #####
#####                     ET                    #####
#####          EXPLORATION DE DONNEES           #####
#####               Chapitre 13                 #####
#####             ARBRES BINAIRES               #####
#####           (Version août 2013)             #####
#########################################################
#########################################################
#  Fichiers de données utilisés (stockés dans le répertoire
courant) :
#     "ozone.txt"      330 * 10
#     "kanga.txt"      148 * 20
#
# library utiles : rpart, MASS, faraway
#
# Tableau 2B - ozone : graphiques par paire des 10 variables.
 library(MASS)
 library(faraway)
 library(rpart)
 data(ozone)
 dim(ozone)
 dimnames(ozone)[[2]]
 summary(ozone)
 pairs(ozone,pch=".")
#
# Tableau 3 - ozone : arbre de régression de la concentration
d'O3.
 (roz <- rpart(O3 ~ .,ozone))
#
# Tableau 4 - ozone : graphe de l'arbre de régression avec
# ses deux options et diagnostics.
 plot(roz,margin=.10)
 text(roz)
# Autre option de sortie
 plot(roz,compress=T,uniform=T,branch=0.4,margin=.10)
 text(roz)
# Diagnostics analogues à ceux de la régression linéaire
 plot(predict(roz),residuals(roz),
xlab="Ajusté",ylab="Résidus")
```

```
 qqnorm(residuals(roz))
#
# Tableau 5 - ozone : prévision à l'aide des résultats
# fournis par l'arbre de régression.
 (x0 <- apply(ozone[,-1],2,median))
 predict(roz,data.frame(t(x0)))
 post(roz,filename="")
#
Tableau 6a - ozone : élagage de l'arbre de régression.
 roze <- rpart(O3 ~ .,ozone,cp=0.001)
 printcp(roze)
#
# Tableau 6b - ozone : élagage de l'arbre de régression
 summary(roze, cp = 0.1)
# Graphes des R2 en fonction des divisions
 plotcp(roze)
#
# Tableau 7 - kanga : données 3 espèces de kangourou.
# Le fichier des données (kanga) est supposé chargé.
 library(MASS)
 library(faraway)
 library(rpart)
 data(kanga)
 dim(kanga)
 dimnames(kanga)[[2]]
# Mesures du spécimen à identifier
 x0 <-
c(1115,NA,748,182,NA,NA,178,311,756,226,NA,NA,NA,48,1009,NA,2
04,593)
#
Tableau 8 - kanga : individu à identifier.
 kanga <- kanga[,c(T,F,!is.na(x0))]
# Valeurs des 2 premières observations
 kanga[1:2,]
# Identification des valeurs manquantes dans l'échantillon
 apply(kanga,2,function(x) sum(is.na(x)))
#
# Tableau 9 - kanga : coefficients de corrélation de
# palate.width et mandible.length avec les autres variables.
 round(cor(kanga[,-
1],use="pairwise.complete.obs")[,c(3,9)],2)
#
#TABLEAU 10 - Comparaison des dimensions des fichiers selon
# les variables supprimées fichier initial : kanga ;
# fichier utilisé : newko.
# Avec suppression de palate.width et mandible.length
 newko <- na.omit(kanga[,-c(4,10)])
 dim(newko)
```

```
# Avec suppression de toutes les observations incomplètes
 dim(na.omit(kanga))
#
# TABLEAU 11 - newko : graphe de l'arbre de classement.
 kt <- rpart(species ~ ., data=newko)
 plot(kt,compress=T,uniform=T,branch=0.4,margin=0.1)
 text(kt)
 kt <- rpart(species ~ ., data=newko,cp=0.001)
# Arbre optimal - validation croisée
 printcp(kt)
#
# Tableau 12 - newko : arbre de classement.
 ktp <- prune(kt,cp=0.0105)
 ktp
 plot(ktp,compress=T,uniform=T,branch=0.4,margin=0.1)
 text(ktp)
# Calcul de l'erreur de mauvais classement
 (tt <-
table(actual=newko$species,predicted=predict(ktp,type="class"
)))
#
1-sum(diag(tt))/sum(tt) # taux de mal-classés
#
#Tableau 13 - newko : arbre de classement avec
# les composantes principales.
 pck <- princomp(newko[,-1])
 pcdf <- data.frame(species=newko$species,pck$scores)
 kt <- rpart(species ~ ., pcdf,cp=0.001)
 printcp(kt)
#
#Tableau 14 - newko : classement du spécimen inconnu (x0).
 nx0 <- x0[!is.na(x0)]
 nx0 <- nx0[-c(3,9)]
 nx0 <- (nx0-pck$center)/pck$scale
 nx0 %*% pck$loadings
 ktp <- prune.rpart(kt,0.0421) # arbre optimal choisi
 ktp
#
#Tableau 15 - newko : classement du spécimen inconnu (x0).
 plot(ktp,compress=T,uniform=T,branch=0.4,margin=0.1)
 text(ktp)
 (tt <- table(newko$species,predict(ktp,type="class")))
 1-sum(diag(tt))/sum(tt) # tx de mauvais classement diminue !
# EXERCICES, fichiers à utiliser : Loup, Altise, Kangourou,
# Nematodes, Eaux2010, BDDrelf.
```

ANNEXE
LOGICIEL R ET DONNEES

Fichiers utilisés + Questions

Fichier	Chapitre d'utilisation	En exercice
Altise (74*3)		11, 13
Amphore-a (15*5)	11	
asthme (43*2)	8	
BDDrelf (281*21)		13
Bledur (50*12)		5
CalciumRencher86a (10*4)	3	
Capitales **(ade4)** (15*15)	4	
ceram18 (185*132)	2	5, 9, 10
ChaZeb-a (23*8)	3, 11, 12	7
chdage (100*3)	12	
Chiens (27*8)		5
Cuisine 29*13)		5
CureThermale (83*12)	7	
Cystfibr **(ISwR)** (25*10)		8
Diabete (46*6)		4
dune **(vegan)** (20*30)	10	
dune.env **(vegan)** (20*5)	10	
Eaux1 (20*7)	2, 4, 6, 7	6, 7
Eaux2a (Tab.7B, ch 4)		
Eaux2010 (113*9) ou 113*8, cf p.96		11, 12, 13
enseignement.cor (13*13)	6 (lecture pg)	
Gevaudan-a (218*16)		3, 12
Glucose (52*6)		4
grcs (392*23)		6
humerus (43*2)		8
ICU (257*13)		12
jumeaux (25*4)	10	
juul **(ISwR)** (1339*6)		8
kanga (**faraway**)	13	
Kangourou (151*21)		11, 13
lnESB162 (162*33)	10	
Loup (43*8)		7, 11
LSA (10*6)		4
malaria **(ISwR)** (100*4)		8
Nematodes (222*9)		11, 13
Os-Griz-a (20*5)	3	3

ozone (**faraway**)	13	9
Patom (16*3)	8	
planete (101*3)	7	
Poumon (72*7)	2	2
PrefConsom (24*4)	5 (lecture pg p.154)	
procespin (32*11)	9	
Procespinsup (25*11)	9	
Quadra (41*3)		12
RdtFromage (41*17)	8	9
Ruspini **(cluster)** (75*2)		7
SocVitesse (24*3)	8	
Spores (15*3)	8	
Tabac (10*2)	3	
TacheMenage (14*5)	5	
usagedrogue.cor (13*13) (*n*=1634)	6	
Vie (31*8)	6	
voix (123*7)		3, 8
voix_acp (116*29)		10
zelazo (**ISwR**)		8

DESCRIPTIONS DES FICHIERS DE DONNEES TRAITEES

Tous les fichiers que nous décrivons ont été, soit traités dans le texte, soit proposés en exercice complémentaire. Ils sont classés par ordre alphabétique sur le nom du fichier. Pour les lire dans un programme, il suffit de les stocker dans le répertoire courant de R et d'utiliser l'instruction **read.table** présentée ci-dessus. Certains fichiers de petite dimension sont listés dans les instructions R.

Nom de l'étude	Nom du fichier

Familles de puces	**Altise** (74*3)

Thème et description des données : 3 familles de puces caractérisées par l'angle et la largeur de leur organe de reproduction, appelé *aedeagus*. La variable Espèce indique la famille d'appartenance de chaque puce, il en existe 3 {Con – *Concinna*, Hei – *Heikertingeri*, Hep - *Heptapotamica*}. Les puces sont décrites à l'aide de deux variables continues : la largeur et l'angle de leur *aedeagus*.

Largeur	Angle	Espece

Origine des données : fichier *Flea Beetles* du site DASL référencé en fin d'annexe.

Amphores crétoises	**Amphore-a** (15*5)

Thème et description des données : Données chapitre 11, §2.3, tableau 1.

1. x1 : hauteur de l'amphore jusqu'au sommet de la poignée

2. x2 : hauteur à la base de la poignée
3. x3 : la hauteur totale de l'amphore
4. x4 : largeur de l'ouverture
5. Grp : groupe d'appartenance

x1	x2	x3	x4	Grp

Origine des données :

Dell'Omodarme, M. (2007) *Esercitazioni di statistica biomedica, alcune note su R*, matt dell@email.it

Traitements contre l'asthme	**asthme** (43*2)

Thème et description des données : durée avant l'apparition d'une crise d'asthme en fonction de trois traitements. Données : chapitre 8, §4.3, tableau 15A.

Traitement	mesure

Origine des données :

Prum, B. (1996) *Modèle linéaire : comparaison de groupes et régression*. Les éditions INSERM, Paris.

Impact de la pollution atmosphérique sur la santé des franciliens	**BDDrelf** (282*21)

Thème et description des données : étude de la caractérisation des automobilistes franciliens à la pollution atmosphérique lors de leur trajet « Domicile-Travail »

1. NO2 : Concentration moyenne de NO2 sur le trajet en µg/m3
2. PtC : Nombre de particules inférieures à 1µm en milliers par cm3
3. Trajet : Nom du trajet
4. Vitesse : Vitesse moyenne sur le trajet en km/h
5. Typo : Typologie : type de trajet Domicile-Travail en fonction de zone géographiques
6. Auto : Temps relatif passé sur Autoroute (pour le trajet effectué)
7. Com : Temps relatif passé sur des Communales (pour le trajet effectué)
8. Dep : Temps relatif passé sur Départementales (pour le trajet effectué)
9. GBd : Temps relatif passé sur les Grands Boulevards (pour le trajet effectué)
10. Nat : Temps relatif passé sur les Nationales (pour le trajet effectué)
11. Periph : Temps relatif passé sur autoroute (pour le trajet effectué)
12. Tun : Temps relatif passé dans des tunnels (pour le trajet effectué)
13. Agglo : Temps relatif passé en agglomération (pour le trajet effectué)
14. GCR : Temps relatif passé en Grande-Couronne (pour le trajet effectué)
15. ParisR : Temps relatif passé dans Paris (pour le trajet effectué)
16. PCR : Temps relatif passé en Petite-Couronne (pour le trajet effectué)
17. P10PA1H : Concentration moyenne de PM10 à la station de fond de Paris-Les Halles en µg/m3 (même période que le trajet effectué)
18. P2PA1H : Concentration moyenne de PM2.5 à la station de fond de Paris-Les Halles en µg/m3 (même période que le trajet effectué)

19. N2PA1H : Concentration moyenne de NO2 à la station de fond de Paris-Les Halles en µg/m3 (même période que le trajet effectué)
20. Saison : ete/hiver
21. Période : matin/soir

Origine des données : `http://www.airparif.asso.fr/`

Rendement de blé dur	**Bledur** (50*12)

Thème et description des données : étude de la relation entre le rendement (RDT) d'un blé dur en fonction de 10 caractéristiques de la parcelle sur laquelle il a poussé.

1. RDT : Rendement en grains
2. PLM :Nb de plantes par m²
3. ZON : Zone géographique
4. ARG : Taux d'argile de la parcelle
5. LIM : Taux de limon de la parcelle
6. SAB : Taux de sable de la parcelle
7. VRT : Variété cultivée
8. PGM : Poids de 1000 grains
9. MST : Matière sèche totale à la récolte
10. AZP : Azote dans la plante à la récolte
11. VRTC : Variété cultivée (codée en 3 classes)

Numero	RDT	PLM	ZON	ARG	LIM	SAB	VRT	PGM	MST	AZP	VRTC

Origine des données : INRA (2006)

Calcium dans le sol et dans du navet	**CalciumRencher86a** (10*4)

Thème et description des données : valeur du calcium en milliéquivavent par 100 grammes.

1. Lieu
2. Calcium disponible dans le sol
3. Calcium échangeable dans le sol
4. Calcium dans feuilles de navet

Lieu	y1	y2	y3

Origine des données :

Kramer, C.Y., Jensen, D.R. (1969) Fundamentals of Multivariate Analysis, Part I. Inference about Means. *Journal of Quality Technology*, **1**(2), 120-133.

Distances entre villes	**Capitales** (`ade4`) (15*15)

Thème et description des données : matrice des distances entre 15 capitales européennes.

Origine des données : `capitales(ade4)`.

Datation de contextes archéologiques	**ceram18** (185*132)

Thème et description des données : Etude d'assemblages céramiques correspondant à une collection d'objets (récipients) issus de contextes archéologiques (fosse à déchet ; niveau d'occupation…) provenant de différent sites de Tours et de ses environs. Chaque objet (récipient ou fragment) est analysé en fonction de critères typologiques : forme du récipient et groupe technique aussi nommé production. Ce dernier critère se définit par la nature de l'argile avec laquelle le vase est réalisé et le choix des couvertes (glaçure, vernis, émail, engobe…) qui lui sont surimposées. Le groupe technique est le seul critère quantifié retenu dans la présente étude (choix ici de la technique du Nombre Minimum d'Individus : *NMI*). Le fichier est constitué de la façon suivante :

- En ligne : la ***date de la monnaie***, puis valeur du *NMI* pour chacun des 183 groupes techniques
- En colonne : les 132 contextes (ou ensembles stratigraphiques) étudiés

ens	X34	X35	X36	X37	X38	…	R11
monnaie	353	NA	NA	NA	NA	…	NA
pfe2	3	0	1	1	0	…	0
⋮	⋮	⋮	⋮	⋮	⋮		⋮
21ca	0	0	0	0	0	…	0

Objectif : datation des 132 contextes archéologiques à l'aide des groupes techniques (productions) pour mieux connaître l'évolution générale des sites archéologiques dans la longue durée (parfois 15 siècles) et obtenir une date pour les contextes non datés par les monnaies.

Origine des données : UMR 7324 CITERES, Laboratoire Archéologie et Territoires, CNRS, Université de Tours.

Bellanger, L., Husi, P., Tomassone, R. (2008) A Statistical Approach for Dating Archaeological Contexts**.** *Journal of Data Science* **6**, 135-154.

Mensurations Charolais/Zébu	**ChaZeb-a** (23*8)

Thème et description des données : mensurations de deux groupes de bovins (12 Charolais et 11 Zébus) élevés à Cuba dans les années 1970 :

1. n : numéro de l' observation
2. Vif : poids vif
3. Carc : poids de la carcasse
4. Qual1 : poids de la viande de première qualité
5. Tot : poids total
6. Gras : poids du gras
7. Os : poids des os
8. Groupe : Char (Charolais) ou Zebu (Zébu).

n	Vif	Carc	Qual1	Tot	Gras	Os	Groupe

Origine des données :

Tomassone, R., Danzart, M., Daudin, J-J., Masson, J-P. (1988) *Discrimination et classement*. Masson, Paris.

Le fichier **chazeb**(**ade4**) est très similaire à **ChaZeb-a** utilisé dans cet ouvrage.

Maladie coronarienne	**chdage** (100*3)

Thème et description des données : présence ou absence d'une maladie coronarienne (CHD) en fonction de l'AGE. Données chapitre 12, §2.3, tableau 3.

ID	AGE	CHD

Origine des données :

Hosmer, D.W. and Lemeshow, S. (1989) *Applied Logistic Regression*, Wiley.

Disponible sur : `www.umass.edu/statdata/statdata/statlogistic.html`

Comparaison de races de chien	**Chiens** (27*8)

Thème et description des données : étude de 27 races de chiens en fonction de 7 critère qualitatif

1. Race : 27
2. Taille (Ta) : petite, moyenne, grande
3. Poids (Po) : petit, moyen, lourd
4. Vélocité (Ve :) : lent, assez rapide, très rapide
5. Intelligence (In) : médiocre, moyenne, forte
6. Affection (Af) : peu affectueux, affectueux
7. Agressivité (Ag) : peu agressif, agressif
8. Fonction (Fn) variable supplémentaire : compagnie, chasse, utilité

Origine des données :

Abdallah, H., Saporta, G. (2003) Mesures de distances entre modalités de variables qualitatives : application à la classification. *R.S.A.* **LI** (2), 78-90.

Saporta, G. (1990) *Probabilités, Analyse des données et Statistique*. Technip, Paris, pp.232-239.

Recettes de cuisine	**Cuisine** (29*13)

Thème et description des données : les auteurs de recettes de cuisine choisissent des types de recettes fort différents ; ainsi pour les recettes de 29 catégories de plats, 13 auteurs ont fait le choix décrit dans le tableau de données avec en ligne les recettes et en colonnes les auteurs. Les auteurs ou les régions concernées sont les suivants :

1. 29 types de recette
2. Les 13 auteurs (A1 : Escoffier ;A2 : Pellaprat ;A3 : Bretagne ;A4 : Occitane ;A5 : Périgord ;A6 : Bocuse cuisine ;A7 : Bocuse marché ;A8 : Guérard gourmande ;A9 : Guérard minceur ;A10 : Olympe ;A12 : Troisgros ;A13 : Vergé)

Recette	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	A11	A12	A13

Origine des données : Cahiers de l'Analyse des Données.

Villes thermales françaises	**CureThermale** (83*12)

Thème et description des données : 83 stations thermales françaises sont classées selon 9 types de maladie qu'elles permettent de soigner (variables binaires 0/1) :

1. GYP : maladies gynécologiques
2. NER : maladies du système nerveux

3. DER : maladies dermatologiques
4. REI : maladies des reins
5. VRE : maladies des voies respiratoires
6. RHU : maladies rhumatismales
7. NUT : maladies liées à la nutrition
8. DIG : maladies du système digestif
9. CIR : maladies de la circulation
10. Station : nom de la station thermale
11. Dep : numéro du département
12. Reg : 5 classes 1 : Sud-Ouest, Pyrénées, Languedoc ; 2 : Sud-Est, Rhône, Alpes, Provence ; 3 : Centre, Auvergne ; 4 : Est, Vosges, Jura ; 5 : Normandie, Région parisienne et autres.

GYP	NER	DER	REI	VRE	RHU	NUT	DIG	CIR	Station	Dep	Reg

Origine des données : quotidien *Le Monde*.

Mucoviscidose	**cystfibr (ISwR)** (25*10)

Thème et description des données : données concernant la mucoviscidose de patients entre 7 et 23 ans.

1. age : âge en années
2. sexe : 0 si homme, 1 si femme
3. height : taille (cm)
4. weight : poids (kg)
5. bmp : masses corporelle (% de la normale)
6. fev1 : volume expiratoire maximal
7. rv : volume résiduel
8. frc : capacité fonctionelle résiduelle
9. tlc : capacité pulmonaire totale
10. pemax : pression expiratoire maximale

age	sex	height	weight	bmp	fev1	rv	frc	tlc	pemax

Origine des données :

O'Neill et *al.* (1983), The effects of chronic hyperinflation, nutritional status, and posture on respiratory muscle strength in cystic fibrosis, *Am. Rev. Respir. Dis.,* 128:1051–1054.

Etude du diabète	**Diabete** (46*6)

Thème et description des données : les données sont constituées par 46 observations et 5 variables ; dans une étude sur le diabète, elles concernent des patients normaux.

Deux variables d'intérêt limité :

1. Obs : n° observation
2. Y1 : poids relatif
3. Y2 : glycémie à jeun

Trois des variables sont d'un intérêt primordial :

4. X1 : intolérance au glucose
5. X2 : réponse insulinique au glucose oral
6. X3 : résistance à l'insuline

Obs	Y1	Y2	X1	X2	X3

Origine des données : données originales de 1979, reprises ultérieurement en 1985 et 1995.

Andrews, D.F. & Herzberg, A.M. (1985) *Data*, New York, Springer Verlag.

Reaven, G.M. & Miller, R.G. (1979) An Attempt to Define the Nature of Chemical Diabetes Using a Multidimensional Analysis. *Diabetologia*, **16**, 17-24.

Relation sol*végétation	**dune (vegan) (**20*30) **dune.env (vegan) (**20*5)

Thème et description des données :

dune : 30 espèces sur 20 sites

dune.env : 20 sites avec 5 variables :

1. A1 : épaisseur de l'horizon A1 du sol
2. Moisture : niveau d'humidité du sol à 5 niveaux (1 < 2 < 4 < 5)
3. Management : exploitation à 4 niveaux : BF (culture biologique), HF (culture hobby), NM (exploitation conservatrice), and SF (culture standard)
4. Use : usage du sol à 3 niveaux : Hayfield (foin) < Haypastu < Pasture (pâturage)
5. Manure : engrais à 5 niveaux : 0 < 1 < 2 < 3 < 4

A1	Moisture	Management	Use	Manure

Origine des données : `library(vegan)`

Jongman, R.H.G, ter Braak, C.J.F & van Tongeren, O.F.R. (1987). *Data Analysis in Community and Landscape Ecology*. Pudoc, Wageningen.

Eaux minérales	**Eaux1** (20*7) & **Eaux2010** (113*9)

Thème et description des données : fichier déjà partiellement analysé dans l'ouvrage cité. Il a été complété par de nouvelles données récoltées au cours des voyages des auteurs ; il est constitué de la façon suivante :

1. HCO3 : bicarbonates (en mg/l)
2. SO4 : sulfates (en mg/l)
3. Cl : Chlore (en mg/l)
4. Ca : Calcium (en mg/l)
5. Mg : Magnesium (en mg/l)
6. Na : Sodium (en mg/l)
7. Sigle : trois caractères
8. Pays : deux lettres {dans Eaux2010 seulement}
9. Nature : plat (eau plate), gaz (eau gazeuse) {dans Eaux2010 seulement}

Eaux1:	HCO3	SO4	Cl	Ca	Mg	Na	Sigle		
Eaux2010:	HCO3	SO4	Cl	Ca	Mg	Na	Sigle	Pays	Nature

Origine des données :

Tomassone, R., Dervin, C., Masson, J-P. (1993) *Biométrie, Modélisation de phénomènes biologiques.* Masson, Paris (2ème éd.)

Remarque : apparaît sous forme de deux fichiers Eaux1 et Eaux2010. Le fichier Eaux1 se trouve au chapitre 2 (Tab.3). On peut consulter de rares sites contenant la composition chimique d'eaux minérales, par exemple :
http://deschosesetdautres.free.fr/index.htm?EAUX
www.astrosurf.com/luxorion/Sciences/eaux-minerales.xls

Notation d'enseignement	**enseignement.cor** (13*13)

Thème et description des données : exemple sur 6 notations sur *n*=1000 élèves, cf. chapitre 6, §4.1. Seule la matrice des coefficients de corrélation est donnée.

Bête du Gévaudan	**Gevaudan-a** (218*16)

Thème et description des données : pendant plus de trois ans, entre le 15 juin 1764 et le 18 juin 1767, plus d'une centaine de femmes et d'enfants furent tués dans le Gévaudan aux confins de l'Auvergne et du Languedoc. Louis XV fut ridiculisé pour son impuissance, surtout dans les journaux anglais. Il envoya ses dragons et ses louvetiers ; il y eut des battues de milliers d'hommes, sans succès. Le mystère règne encore…

1. n°: numéro des observations
2. Del : observations à supprimer si on élimine les observations douteuses ou celle pour lesquelles la victime est rescapée. (OK ou Del)
3. Doute (31) : observation douteuse (1) à supprimer, sinon (0)
4. Per : Période, de P1 à P6
   P1 : de la première victime à la fin des attaques aux confins du Vivarais (n°14, 07/1764))
   P2 : jusqu'à la mise à mort du loup des frères Martel (n°90, 01/05/1765)
   P3 : jusqu'à l'incarcération de la famille Chastel (n°123, 16/08/1765)
   P4 : jusqu'à la mise à mort du grand loup par Antoine de Beauterne (n°132, 21/09/1765) et à la libération des Chastel (n°134, 8/11/1765)
   P5 : jusqu'à l'époque où la bête paraît avoir son point d'attaque au Mont Mouchet (n°172, fin 1766)
   P6 : jusqu'à la mise à mort de la bête de Jean Chastel
5. P : Pa (période P1, P4, P5, P6) ou Pb (période P2 et P3)
6. Jour : an-mois-jour (**/**/**)
7. h : heure
8. Sexe : F (femme) ou H (Homme)
9. Age : âge (en années), ou enf : enfant ; ado : adolescent ; adu : adulte
10. A : enf, ado, adu
11. Occu : occupation b si berger (55)
12. Lieu : b (13) : bois ; c (6) : champs ; j (2) : jardin ; m (14) : maison ; p (15) : pâturage ; r (5) : chemin ; v (18) : village

13. Bilan : B (41) : blessé ; BG (25) : blessé grave ; BL (2) : blessé leger ; M (108) : mort ; R (42) : Rescapé
14. Dévoré (77) :
15. Décapité (23) :
16. Cou (45) : blessure au cou.

n°	Del	Doute	Per	P	jour	h	Sexe	Age	A
Occu	Lieu	Bilan	Dev	Dec	Cou				

Origine des données : *La Bête du Gévaudan, autopsie d'un mythe* de David Teyssandier, 2002, France 3. Cf. sur Internet « Bête du Gévaudan ».

Glucose chez la femme enceinte	**Glucose** (52*6)

Thème et description des données : résultats d'analyse du glucose dans le sang à trois occasions pour 52 femmes à jeun et une heure après avoir consommé du sucre.

1. Trois mesures à jeun : Y1, Y2, Y3
2. Trois mesures une heure après consommation de sucre : X1, X2, X3

Y1	Y2	Y3	X1	X2	X3

Origine des données :

O'Sullivan, J.B., Mahan, C.M. (1966) Glucose Tolerance Test: Variability in Pregnant and Non-pregnant Women. *American Journal of Clinical Nutrition*, **19**, 345-351.

Rencher, A.C. (1995) *Methods of Multivariate Analysis.* Wiley, New York.

Etude de questionnaires	**grcs** (393*23)

Thème et description des données : Corpus de 392 personnes ayant répondu aux 23 items du questionnaire GRCS (Gambling Related Cognitions Scale) d'une échelle d'évaluation des cognitions liées (GRCS traduite et adaptée). Ces items sont codés de 1 à 7 en fonction de l'intensité de l'accord avec la proposition indiquée (1 : Désaccord total, 2 : Désaccord fort, 3 : Désaccord moyen, 4 : Ni accord, ni désaccord, 5 : Accord moyen, 6 : Accord fort, 7 : Accord total), succinctement, ils étaient définis par :

1. Jouer me rend plus heureux
2. Je ne peux pas fonctionner sans jouer
3. Prier m'aide à gagner
4. Les pertes au jeu doivent être suivies par une série de gains
5. Relier mes gains à mon adresse et mes capacités me fait continuer à jouer
6. Jouer améliore l'apparence des choses
7. Il m'est difficile d'arrêter de jouer étant donner que je perds le contrôle
8. Des nombres ou des couleurs particulières peuvent aider à augmenter mes chances de gagner
9. Une série de pertes me procurera un apprentissage qui m'aidera à gagner par la suite
10. Relier mes pertes à de la malchance ou de mauvaises circonstances me fait continuer à jouer
11. Jouer rend l'avenir plus prometteur
12. Mon désir de jouer est tellement plus fort que moi

13. Je collectionne des objets particuliers qui aident à augmenter mes chances de gagner
14. Lorsque je gagne une fois, je gagnerai sûrement encore
15. Relier mes pertes aux probabilités me fait continuer à jouer
16. Etre en train de jouer aide à réduire la tension et le stress
17. Je ne suis pas suffisamment fort pour arrêter de jouer
18. J'ai des rituels et des comportements particuliers qui augmentent mes chances de gagner
19. Il y a des moments où je me sens chanceux(se) et je ne joue qu'à ces moments-là
20. Me souvenir de la somme que j'ai gagnée la dernière fois me fait continuer à jouer
21. Je ne serai jamais capable d'arrêter de jouer
22. Je possède une certaine capacité à prédire mes gains au jeu
23. Si je change tout le temps mes numéros, j'ai moins de chance de gagner que si je conserve les mêmes numéros à chaque fois

Le fichier est constitué de la manière suivante :

- GABS1-GABS23 : réponses des patients

NB : par jeu, on entend les jeux de hasard et d'argent, tels que les jeux de cartes, de dés, les machines à sous, ou tous les types de jeux pour lesquels on mise de l'argent ou on fait des paris.

Origine des données : Centre de Recherche sur le Jeu Excessif – CHU Nantes.

Marie Grall-Bronnec, Gaëlle Bouju, Jean-Damien Le Bourvellec, Philip Gorwood, Claude Boutin, Jean-Luc Vénisse, Jean-Benoît Hardouin (2012) A French adaptation of the Gambling Related Cognitions Scale (GRCS). *Journal of Gambling Issues*, 27.

Raylu, N., Oei, T.P.S., 2004. The Gambling Related Cognition Scale (GRCS) development, confirmatory factor validation and psychometric properties. *Addiction* 99,757-769.

Relation densité*âge	**humerus** (43*2)

Thème et description des données : les données ont été faites pour analyser la relation entre la densité de cendres d'humérus (g/cm2) en fonction de l'âge sur un échantillon de 43 femmes.

Densité	Age

Origine des données : Tomassone et *al.* (1992).

Unité de soins intensifs	**ICU** (257*13)

Thème et description des données : fichier relatif à différents facteurs de risque liés au décès en unité de soins intensifs (données semi artificielles tirées en partie d'un fichier provenant du site **statlib**) :

http://www.math.siu.edu/olive/ICU.lsp
http://lib.stat.cmu.edu/DASL/Datafiles/ICU.html

1. DECEDE : 0 :non ; 1 :oui
2. AGE : valeur en années
3. SEXE : 0 : femme, 1 : homme
4. CHIR_MED : service à l'admission, 0 = Medical, 1 = Surgical

5. INF_J0 : Infection probable à l'admission (0 = oui, 1 = non)
6. TA_SYS : pression sanguine systolique à l'admission (en mm Hg)
7. FC : fréquence cardiaque
8. URG_NURG : 0/1
9. PO2 : gaz dans le sang, seuil : 0 =>60, 1 <= 60
10. PH : seuil : 0 >= 7.25, 1 <7.25
11. BICAR : bicarbonatémie, seuil : 0 > 18, 1 = <18
12. CONSC : niveau de conscience à l'admission (0 = no coma or stupor, 1= deep stupor, 2 = coma)
13. GLASGOW : score de Glasgow

Origine des données : `http://hebergement.u-psud.fr/biostatistiques/`

Falissard, B. (2005) *Comprendre et utiliser les statistiques dans les sciences de la vie.* 3° édition, Masson, Paris (380 pages).

Hérédité et caractéristiques du visage	**jumeaux** (25*4)

Thème et description des données : Dans une étude sur la recherche du caractère du visage qui parait le plus héréditaire, on a mesure sur 25 jumeaux (cf Tab.2 chapitre 10) hauteur (H) et largeur (L) du visage du premier né et du second né soit deux couples (H1,L1) et (H2,L2).

H1	L1	H2	L2

Origine des données : Anderson (1958) et Frets(1921).

Etude d'un facteur de croissance	**juul (`ISwR`)** (1339*6)

Thème et description des données : échantillon de référence d'un facteur de croissance analogue à l'insuline (IGF-I), une observation par sujet, à des âges variés :

1. age : en années
2. menarche : Est-ce que la première période de menstruation a eu lieu ?(1 : non, 2 : oui)
3. sex : 1 : garçon, 2 : fille
4. igf1 : facteur de croissance (microgramme par litre)
5. tanner : stages de la puberté, codes 1–5
6. testvol : volume testiculaire (ml)

age	menarche	sex	igf1	tanner	testvol

Origine des données : `library(ISwR)`

Squelettes de Kangourous	**Kangourou** (151*21)

Thème et description des données : Des mesures ont été faites sur 18 caractéristiques du squelette de kangourous appartenant à trois espèces pour des mâles et des femelles : Mg : *Macropus giganteus* (25 mâles, 25 femelles), Mm :*Macropus fuliginosus melanops* (23 mâles, 25 femelles), Mf :*Macropus fuliginosus fuliginosus* (25 mâles, 25 femelles). On dispose en outre de mesures incomplètes sur trois spécimens « historiques » se trouvant dans des musées européens ; ainsi, celui de Paris a été capturé en 1803. Le fichier est constitué de la manière suivante :

1. n : identification numérique de l'échantillon. Pour les trois spécimens, l'origine est : A (British Museum of Natural History, Londres, male, Mg), B (Muséum National d'Histoire Naturelle, Paris, male, Mg), C (Rijksmuseum van Natuurlijkee, Leiden, femelle, Mf)
2. Sexe : M ou F
3. Espece : Mg, Mm et Mf
4. Caractéristiques : X1, X2, X3, X4, X5, X6, X7, X8, X9, X10, X11, X12, X13, X14, X15, X16, X17, X18. Toutes les mesures sont exprimées en millimètre multiplié par 10.

Le fichier **kanga(`faraway`)**(148*20) utilisé au chapitre 13 contient les mêmes 18 variables mais n'a pas de numéro d'identification. Les noms de ces variables sont les suivants : *« species», «sex», «basilar length», «occipitonasal length», «palate length», «palate width», «nasal length», «nasal width», «squamosal depth», «lacrymal width», «zygomatic width», «orbital width», « rostral width», «occipital depth», «crest width», «foramina length», «mandible length», «mandible width», «mandible depth», «ramus height».*

Origine des données :

Poole, W.E. (1976) Breeding biology and current status of the grey kangaroo, Macropus *fuliginosus fuliginosus* , of Kangaroo lsland, South Australia. *Aust.J.Zool*. **24**, 169-187.

Relation Sol/Blé	**lnESB162** (162*33)

Thème et description des données : Le corpus de données Sol/Blé est constitué par un échantillon de *n=162* sites étudiés selon le même protocole dans diverses régions de France. Il s'agit de sols agricoles « ordinaires », c'est-à-dire non pollués et n'ayant pas reçu de boues d'épuration (sauf une douzaine de cas particuliers, Courbe et *al.* 2002). Ils appartiennent à *18* familles pédo-géologiques contrastées. Sur chaque site, des grains de blé ont été récoltés à maturité sur *1m²* (variété « Soissons » ou « Trémie »). Au pied du blé ainsi récolté, l'horizon de surface labouré du sol a été également prélevé. Sur des échantillons séchés et tamisés à *2mm* de ces horizons de surface, ont été déterminées.

1. *9* variables caractéristiques des propriétés agro-pédologiques classiques : granulométrie 5 fractions (argile : A; limon fin et grossier : LF, LG; sable fin et grossier : SF, SG), le carbone organique (CS), le pH mesuré après agitation dans l'eau (pH), le calcaire ($CaCO_3$) et la capacité d'échange cationique (CEC) ; ces variables sont des teneurs, sauf le pH et la CEC.

row.names	Pop	i	A	LF	LG	SF	SG	CEC	CO3Ca	CS	pH

2. *8* variables représentant les concentrations totales des métaux du sol obtenues après mise en solution par les acides fluorhydrique et perchlorique selon la norme NF ISO 14869-1 : FeS, MnS, CdS, CrS, CuS, NiS, PbS et ZnS, et

CdS	CrS	CuS	FeS	MnS	NiS	PbS	ZnS

3. *7* variables qui sont les concentrations en métaux extraits par deux réactifs, DTPA (DiéthylèneTriamine-PentaAcétique) et NH_4NO_3 (nitrate d'ammonium), choisis pour leur capacité à atteindre seulement les formes chimiques les plus réactives et les plus susceptibles d'être absorbées par les racines des plantes. Les quantités extraites au DTPA correspondraient plutôt aux métaux associés aux matières organiques et aux oxydes de fer, tandis que celles extraites par le NH_4NO_3 correspondraient plutôt aux formes métalliques échangeables, les plus phyto-disponibles. Soit : CdD, CuD, PbD et ZnD (pour DTPA), CdN, CuN et ZnN (pour NH_4NO_3).

CdD	CuD	PbD	ZnD	CdN	CuN	ZnN

4. *9* variables représentant les concentrations dans les grains de blé en CdB, CrB, CuB, FeB, MgB, MnB, NiB, PbB et ZnB.

CdB	CrB	CuB	FeB	MgB	MnB	NiB	PbB	ZnB

Origine des données :

Baize, D., Bellanger, L., Tomassone, R. (2008) Relationships between concentrations of trace metals in wheat grains and soil. *Agron. Sustain. Dev.* 1-16.

Chien ou Loup ?	**Loup** (43*8)

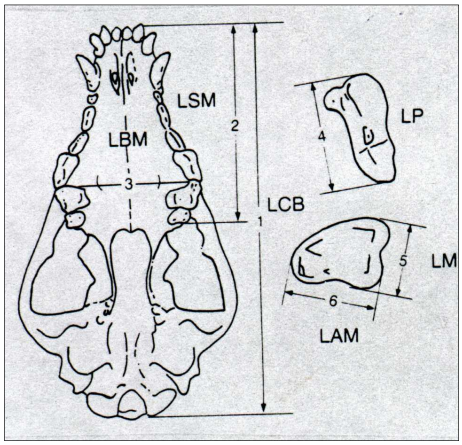
Thème et description des données : les mensurations des crânes de loup et de chien sont relativement voisines. Le fichier est constitué de la manière suivante :

1. n : identification
2. Pop : Chien, Loup, ? (crâne fossile)
3. LCB : longueur condylo-basale
4. LSM : longueur de la mâchoire supérieure
5. LBM : largeur bi-maxillaire
6. LP : longueur de la carnassière supérieure
7. LM : longueur de la première molaire supérieure
8. LAM : largeur de la première molaire supérieure

Une dernière observation est mal identifiée ? Est-ce un chien ou un loup ?

n	Pop	LCB	LSM	LBM	LP	LM	LAM

Origine des données : *Cahier de l'Analyse des Données*



Analyse latente sémantique	**LSA** (10*6)

Thème et description des données : sous le vocable d'analyse latente sémantique (en anglais *latent semantic analysis* ou *LSA*), on a imaginé un traitement numérique de phrases codées ; à partir des cinq phrases suivantes :

1. d1: a bank will protect your money
2. d2: a guard will protect a bank
3. d3: your bank shot is money
4. d4: a bank shot is lucky
5. d5: bank guard

On fait le codage suivant où la présence d'un mot est codée 1, son absence 0.

Mot	d1	d2	d3	d4	d5

Une décomposition en valeurs singulières (*DVS*) de la matrice à 10 lignes et 4 colonnes (colonne 2 à 4) permet d'étudier la proximité des quatre premières phrases. On situe la cinquième phrase comme variable supplémentaire.

Origine des données : Drăghici (2012).

Enfants atteints de malaria	**malaria (`ISwR`)** (100*4)

Thème et description des données : échantillon aléatoire de 100 enfants entre 3 et 15 ans d'un village du Ghana. Les enfants sont suivis pendant une période de 8 mois en mesurant un anticorps particulier ; ils sont classés en deux catégories avec ou sans symptôme de malaria :

1. subject : code du sujet
2. age : en années
3. ab : niveau de l'anticorps
4. mal : Malaria: 0 : non, 1 : oui

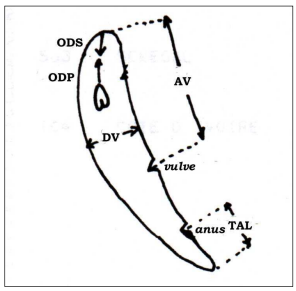
subject	age	ab	mal

Origine des données : `library(ISwR)`

Variabilité de *Xiphinema elongatum*.	**Nematodes** (222*9)

Thème et description des données : Les nématodes sont des vers vivant dans le sol ou en parasite de l'homme et des mammifères ; ils jouent un rôle important en agriculture. Dans une étude portant sur *Xiphinema elongatum*, on a mesuré sept caractéristiques classiques de la morphologie des nématodes correspondant au schéma ci-dessous :

1. BOL : longueur du corps
2. DMA : Coefficient de Man « a » = BOL/DV
3. DMV : Coefficient de Man « c » = 100*(AV/BOL)
4. TAL : Longueur de la queue
5. TLC : Coefficient C = TAL/DV
6. ODS : Position du stylet (1)
7. ODP : Position du stylet (2)
8. Pays : Localisations (19)
9. n : Numéro d'échantillon



Pour chaque localisation il existe plusieurs échantillons. En général une localisation est associée à un pays ; toutefois certains pays sont représentés par plusieurs localisations.

BOL	DMA	DMV	TAL	TLC	ODS	ODP	Pays	n

Origine des données : Michel Luc (Muséum National d'Histoire Naturelle & ORSTOM).

Os bassin de garçons	**Os-Griz-a** (20*5)

Thème et description des données : longueur d'os du bassin de 20 garçons à quatre âges différents.

1. Obs : numéro de l' observation
2. Y1, Y2, Y3, Y4 : mensurations

Obs	Y1	Y2	Y3	Y4

Origine des données :

Elston, R.C. & Grizzle, J.E. (1962) Estimation of Time-response Curves and Their Confidence Bands, *Biometrics*, **18**, 148-159.

Variables climatiques et taux d'ozone au sol	**Ozone (`faraway`)** (330*10)

Thème et description des données :

1. O3 : variable réponse ozone au niveau du sol à Los Angeles
2. vh : altitude à laquelle la pression est 500 millibars
3. wind : vitesse du vent (miles par heure)
4. humid : humidité (%)
5. temp : temperature(°F)
6. ibh : the temperature inversion base height (feet)
7. dpg : gradient de pression (mm Hg)
8. ibt : the inversion base temperature (degrees F)
9. vis : visibilité miles)

10. doy : jour de l'année

O3	vh	wind	humid	temp	ibh	dpg	ibt	vis	doy

Origine des données : `library(faraway)`

Breiman, L., Friedman, J., Olshen, R., Stone, C. (1984) *CART: Classification and Regression Trees*, Wadsworth International.

Poids atomique de l'iode	**Patom** (16*3)

Thème et description des données : Chapitre 8, §4.7, tableau 20A.

Poids	Argent	Iode

Origine des données :

Cox, D.R., Snell, E.J. (1981) *Applied Statistics, Principles and Examples*. Chapman and Hall, Londres.

Etude exoplanètes	**planete** (101*3)

Thème et description des données : donnes décrites au chapitre 7, §5.2. Elles constituées de 101 exoplanètes sur lesquelles trois variables ont étéobservées :

1. x_1 : masse en % de celle de Jupiter
2. x_2 : période en jours terrestres
3. x_3 : excentricité

n°	Masse	Periode	Excentricite

Origine des données :

Mayor, M., Frei, P-Y. (2001) *Les nouveaux mondes du cosmos, A la découverte des exoplanètes*. Editions du Seuil, Paris.

Byssinosis	**Poumon** (72*7)

Thème et description des données : la byssinosis » est une maladie propre aux personnes ayant travaillé sur le coton dans une atmosphère mal ventilée. Le fichier comprend 72 lignes et 7 colonnes (2 variables et 5 facteurs) ; les 72 lignes du fichier représentent chacune une combinaison des niveaux des 5 facteurs. Les sept colonnes représentent :

1. Oui : nombre de personnes malades
2. Non : nombre de personnes non malades
3. Poussiere : Haut, Moyen, Bas
4. Race : Blanc, Autre
5. Sexe : Homme, Femme
6. Tabagie : Fumeur, Non-Fumeur
7. Tempsemploi : <10 ans, 10-20 ans, >20 ans

Oui	Non	Poussiere	Race	Sexe	Tabagie	Tempsemploi

Origine des données :

Everitt, B. & Rabe-Hesketh, S. (2001) *Analyzing Medical Data using S-PLUS*. Springer, New York.

Préférences des consommateurs	**PrefConsom** (24*4)

Thème et description des données : résultat d'une enquête sur les préférence de consommateur pour 4 produits en fonction du sexe et de la classe d'âge.

1. Nb : nombre de personnes
2. Produits à 4 niveaux A, B, C, D
3. Sexe : M ou F
4. Age : 3 classes d'âge A1, A2, A3

Nb	Produit	Sexe	Age

Origine des données : Cahiers de l'Analyse des Données.

Processionnaire du pin	**procespin** (32*11) **procespinsup** (25*11)

Thème et description des données : dans l'étude à l'origine de cet exemple, les expérimentateurs souhaitaient connaître l'influence de certaines caractéristiques de peuplements forestiers sur le développement de la processionnaire. L'unité qui représente ici l'observation, est une parcelle forestière, dont la surface de 10 hectares est d'un seul tenant. Cette parcelle est considérée comme homogène par rapport aux variables étudiées. Les valeurs de ces variables ont été obtenues comme des moyennes de valeurs mesurées sur des placettes échantillon de 5 ares. La variable à expliquer ($\boldsymbol{y}$) est le nombre de nids de processionnaires par arbre d'une placette. Pour des raisons partiellement statistiques, nous chercherons à analyser la relation entre ln $\boldsymbol{y}$ et les dix variables explicatives suivantes mesurées sur chaque placette :

1. $\boldsymbol{x}^1$ **:** altitude (en mètres)
2. $\boldsymbol{x}^2$ **:** pente (en degrés)
3. $\boldsymbol{x}^3$ **:** nombre de pins dans une placette de 5 ares
4. $\boldsymbol{x}^4$ **:** hauteur de l'arbre échantillonné au centre de la placette (en mètres)
5. $\boldsymbol{x}^5$ **:** diamètre de cet arbre (en centimètres)
6. $\boldsymbol{x}^6$ **:** note de densité du peuplement (échelle de 1 à 4)
7. $\boldsymbol{x}^7$ **:** orientation de la placette (1= orientation vers le sud, 2= autre)
8. $\boldsymbol{x}^8$ **:** hauteur des arbres dominants (en mètres)
9. $\boldsymbol{x}^9$ **:** nombre de strates de végétation
10. $\boldsymbol{x}^{10}$ **:** mélange du peuplement (1= pas mélangé; 2= mélangé)

y	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10

Origine des données :

Tomassone, R., Audrain, S., Lesquoy-de Turckheim, E., Millier, C. (1992) *La régression : nouveaux regards sur une ancienne méthode statistique*. Masson, Paris. 2eme ed.

Même structure pour le fichier de données supplémentaires **procespinsup** (25*11).

Exemple fictif pour l'*AFD* quadratique	**Quadra** (41*3)

Thème et description des données : : donnes fictives décrites dans l'exemple présenté au chapitre 12, §1.2 et construites pour montrer la nécessité de choisir une règle quadratique en *AFD* dans certains cas. Le fichier comporte 40 observations faites sur deux

variables quatitatives X et Y, ainsi qu'une variable qualitative à deux modalités (Pop) représentant la classe d'appartenance (A ou B).

X	Y	Pop

Rendement fromager	**RdtFromage** (41*17)

Thème et description des données : le rendement et la qualité d'un processus de fabrication d'un fromage est naturellement lié à la composition des laits qui permettent de le fabriquer. Les technologues ont défini un rendement fromager (RFESC) et ils souhaitent connaître l'influence de différentes caractéristiques du lait dans le cas de la fabrication de fromages à pâte pressée. Ces caractéristiques sont définies par trois types de mesures : a) celles que toute usine de fabrication utilise ; b) celles qui demandent de faire appel à un laboratoire spécialisé ; c) celles qui permettent de faire des mesures en continu dites rhéologiques. Le fichier est constitué de la façon suivante :

- a) mesures de routine (7) :
  - MAT : matière azotée totale (g/l)
  - CNE : concentration en caséines (g/l)
  - NPN : azote non protéique (g/l)
  - CAT : concentration en calcium total (mg/l)
  - CAS : concentration en calcium soluble (mg/l)
  - CAI : concentration en calcium ionique (mg/l)
  - ES : extrait sec du lait (%) standardisé à 25g MG/l
- b) mesures de laboratoire (6) :
  - Ktot : proportion de caséine K totale (%)
  - Astot : proportion de caséine α_s (%)
  - DMM : diamètre moyen des micelles (nm)
  - D10 : 10% des micelles ont un diamètre inférieur (nm)
  - D90 : 10% des micelles ont un diamètre supérieur (nm)
  - CIS : concentration en citrate soluble (mM)
- c) mesures rhéologiques :
  - TPR : temps de prise (mn)
  - VRG : vitesse de raffermissement du gel présure (mV/mn)
  - FMG : fermeté maximale du gel présure (mV)
- Enfin le rendement fromager :
  - RFESC

MAT	CNE	NPN	CAT	CAS	CAI	ES
Ktot	AStot	DMM	D10	D90	CIS	
TPR	VRG	FMG				
RFESC						

Origine des données : Laboratoire de technologie laitière de l'INRA à Grignon dans le cadre d'un contrat ARILAIT (1985).

Meule de Parmesan (Rada in Chianti , 11 septembre 2010). Fichier : **RdtFromage**

Exemple historique	**Ruspini (`cluster`)** (75*2)

Thème et description des données : données dans un plan souvent utilisées comme exemple pour des algorithmes de classification.

Origine des données : `library(cluster)` ;

E. H. Ruspini (1970): Numerical methods for fuzzy clustering. *Inform. Sci.*, **2**, 319–350.

Résistance à l'avancement d'une charrue	**SocVitesse** (24*3)

Thème et description des données : étude de la résistance à l'avancement (Y) de 3 types de soc de charrue (Soc) en fonction de la vitesse (Vitesse). Données : chapitre 8, §4.4.3, tableau 17A.

Y	Soc	Vitesse

Origine des données :

Tomassone, R., Dervin, C., Masson, J-P. (1993) *Biométrie, Modélisation de Phénomènes biologiques*. Masson, Paris.

Abondance de spores	**Spores** (15*3)

Thème et description des données : abondance de spores avant (X) et après (Y) trois traitements fongicides (A, B, C). . Données chapitre 8, §4.6, tableau 19A.

Tr	X	Y

Origine des données :

Tomassone, R., Dervin, C., Masson, J-P. (1993) *Biométrie, Modélisation de Phénomènes biologiques*. Masson, Paris.

Relation Goudron*Nicotine	**Tabac** (10*2)

Thème et description des données : poids dans 10 marques de cigarettes de :

1. Goudron (en mg)
2. Nicotine (en mg)

goudron	nicotine

Origine des données :

Tomassone, R., Audrain, S., Lesquoy-de Turckheim, E., Millier, C. (1992) *La régression : nouveaux regards sur une ancienne méthode statistique*. Masson, Paris, 2ième édition révisée.

Répartition des tâches ménagères	**TacheMenage (**14*5)

Thème et description des données : Le fichier correspond à une table de contingence avec des effectifs où les lignes sont 13 tâches ménagères et les colonnes indiquent si elles sont réalisées par la femme, alternativement, par l'homme ou de concert (cf chapitre 5, Tab.5A).

Tache	Femme	Alternativement	Homme	Ensemble

Origine des données : Kroonenberg et Lombardo, R. (1999)

Les données se trouvent aussi dans la library **ade4** sous le nom **data(housetasks)**.

Consommation de drogue	**usagedrogue.cor** (13*13) (*n*=1634)

Thème et description des données : consommation de drogue en milieu étudiant sur *n*=1634 élèves, cf. chapitre 6, §4.3. Seule la matrice des coefficients de corrélation est donnée.

Origine des données :

Huba, G.J., Wingard, J.A., Bentler, P.M. (1981) A comparison of two latent variable causal models for adolescent drug use, *Journal of Personality and Social Psychology*, 40, 180-193.

Espérance de vie	**Vie** (31*8)

Thème et description des données : les données représentent les espérances de vie pour des hommes et des femmes à la naissance, à 25, 50 et 75 ans dans 31 pays. Cf. chapitre 6, §4.2, tableau 5.

Pays	h0	h25	h50	h75	f0	f25	f50	f75

Origine des données :

Keyfitz, N. & Flieger, W. (1971) *Population : The Facts and Methods of Demography.* W.H.Freeman, San Francisco.

Dépression et monotonie de la voix	**voix** (123*7)

Thème et description des données : Le corpus contient les données mesurées sur 101 patients déprimés hospitalisés en psychiatrie. On désire tester, à l'entrée à l'hôpital, l'existence d'une association entre la monotonie de la voix et l'intensité de la dépression. La voix étant sensiblement modifiée par les médicaments psychotropes, la consommation de différents antidépresseurs a aussi été mesurée. Le fichier est constitué de la manière suivante :

1. af0s : paramètre permettant d'évaluer la variabilité de la hauteur de la voix
2. tricyc : antidépresseur tricyclique (oui=1, non=0)

3. serot : antidépresseur sérotoninergique (oui=1, non=0)
4. neurol : neuroleptique (oui=1, non=0)
5. bz : benzodiazépine (oui=1, non=0)
6. ham : intensité de la dépression, mesurée par l'échelle de dépression de Hamilton (HDRS)
7. tyrer : anxiété mesurée à l'aide l'échelle de Tyrer

Af0s	tricyc	serot	neurol	bz	ham	tyrer

Origine des données : `http://hebergement.u-psud.fr/biostatistiques/`

Mises à disposition par l'équipe de R. Jouvent.

Falissard, B. (2005) *Comprendre et utiliser les statistiques dans les sciences de la vie*. Masson, Paris. 3eme ed.

Modifications de la voix dans la dépression	**voix_acp** (116*29)

Thème et description des données : Comparaison des modifications de la voix dans la dépression en analysant d'un côté des paramètres physiologiques (acoustiques) et de l'autre des particularités sémiologiques de l'état dépressif (échelles d'évaluation). Scores d'anxiété et de dépression et caractéristiques de la voix (données non publiées, aimablement). Le fichier est constitué de la manière suivante :

Profil acoustique : conserver les colonnes 1, 2, 4, 8, 14, 16 reflétant des aspects variés de la voix :

1. MOYF0 : hauteur moyenne de la voix
2. SDF0 : variabilité globale de la hauteur de la voix
3. AF0S : variabilité locale de la hauteur de la voix
4. DURTOT : temps mis pour lire trois phrases standard
5. SDENERG : écart-type de l'énergie du signal vocal normalisé ; mesure la variabilité de l'énergie utilisée pour la diction
6. DUR_COMP : temps mis pour compter de 1 à 10

Profil clinique : trois scores ont été retenus, correspondant aux colonnes 18, 19 et 21 :

1. HAMILTON : échelle d'Hamilton (mesurant une intensité dépressive globale)
2. COVI : échelle de Covi (évaluant plus spécialement les éléments anxieux)
3. RALENT : échelle de ralentissement (évaluant la perte d'énergie)

MOYF0	SDF0	SDF0RC	AF0S	MAXF0	MINF0	RANGEF0	
DURTOT	DURPH1_2	DURPH2_3	EN_PEN	PEN_SEZ	PUISV_NV	SDENERG	
SD_E_PIC	DUR_COMP	ACCELERA	HAMILTON	COVI	TYR	RALENT	
ATAYLOR	ANHED	IRRIT	HUM_EXP	TRISTES	ANXIETE	EMOUSST	CONTROL

Origine des données : `http://hebergement.u-psud.fr/biostatistiques/`

Mises à disposition par l'équipe de R. Jouvent.

Falissard, B. (2005) *Comprendre et utiliser les statistiques dans les sciences de la vie*. Masson, Paris. 3eme ed.

BIBLIOTHEQUES UTILISEES

Dans le tableau suivant l'adresse des sites Internet permet d'obtenir davantage d'informations sur les library, les auteurs et les personnes qui assurent la maintenance.

ade4	Analysis of Ecological Data : Exploratory and Euclidean methods in Environmental sciences http://pbil.univ-lyon1.fr/ADE-4/
amap	Another Multidimensional Analysis Package http://mulcyber.toulouse.inra.fr/projects/amap/
Amelia	*bootstrap* EM algorithme sur données incomplètes crée des corpus avec données estimées ; http://gking.harvard.edu/amelia/
BHH2	Useful Functions for **B**ox, **H**unter and **H**unter II
bootstrap	Logiciel (*bootstrap*, validation croisée, jackknife) et données ; Efron, B. & Tibshirani, R. (1993) *Bootstrap Methods and Their Applications*, Chapman and Hall.
boot	A. C. Davison & D. V. Hinkley (1997) *Bootstrap Methods and Their Applications*, CUP). http://statwww.epfl.ch/davison/BMA/library.html
car	Companion to Applied Regression Fox, J. & Weisberg, S., (2011) *An R Companion to Applied Regression*, Sage.
cluster	Cluster Analysis Extended Rousseeuw et *al.* Martin Maechler, maechler@stat.math.ethz.ch
ClustOfVar	Clustering of variables marie.chavent@math.u-bordeaux1.fr
coin	**Co**nditional **In**ference Procedures in a Permutation Test Framework ; http://www.jstatsoft.org/v28/i08/
DAAG	**D**ata **A**nalysis **A**nd **G**raphics data and functions ; http://www.stats.uwo.ca/DAAG
DMwR	Functions and data for the book "**D**ata **M**ining **w**ith **R**"
e1071	Misc Functions of the Department of Statistics, TU Wien
elasticnet	Estimation et *ACP sparse* http://www.stat.umn.edu/~hzou
faraway	Functions and datasets for books by Julian Faraway http://www.maths.bath.ac.uk/~jjf23/
Hmisc	Mélange de Harrell pour la fonction impute ; http://biostat.mc.vanderbilt.edu/trac/Hmisc
ISwR	Données et scripts avec exemples et exercices du livre Dalgaard, P. (2008) *Introductory Statistics with R*, 2nd ed., Springer Verlag,
klaR	Classification and visualization http://www.statistik.tu-dortmund.de
lars	Least Angle Regression, Lasso and Forward Stagewise http://www-stat.stanford.edu/~hastie/Papers/#LARS
leaps	Regression subset selection including exhaustive search; tlumley@u.washington.edu
lmtest	Test de modèles de régression linéaire; Achim.Zeileis@R-project.org
MASS	Venables and Ripley, *Modern Applied Statistics with S* (2002) ; http://www.stats.ox.ac.uk/pub/MASS4/
mclust	Normal Mixture Modeling for Model-Based Clustering, Classification, and

	Density Estimation http://www.stat.washington.edu/research/reports/2006/tr504.pdf
mice	**M**ultivariate **I**mputation by **C**hained **E**quations ; http://www.stefvanbuuren.nl/publications/MICE in R - Draft.pdf
outliers	Tests for outliers http://www.komsta.net/
pls	Partial Least Squares and Principal Component regression http://mevik.net/work/software/pls.html
PresenceAbsence	Evaluation de modèles de présence/absence ; eafreeman@fs.fed.us
rms	Regression Modeling Strategies; Frank E Harrell Jr <f.harrell@vanderbilt.edu> http://biostat.mc.vanderbilt.edu/rms
rpart	Recursive partitioning and regression trees http://mayoresearch.mayo.edu/mayo/research/biostat/splusfunctions.cfm
simpleboot	programmes de *bootstrap*; rpeng@jhsph.edu
stats	fonctions statistiques de R ; R-core@r-project.org
vegan	Community Ecology Package ; http://vegan.r-forge.r-project.org/
VIM	**V**isualization and **I**mputation of **M**issing Values ; http://cran.r-project.org/package=VIM
yacca	Yet Another Canonical Correlation Analysis Package ; Mardia, K.V. Kent, J.T. & Bibby, J. M. (1979) *Multivariate Analysis*. London, Academic Press ; buttsc@uci.edu