AFD /ACP AFDM sur données environnementales

Abdelhakim Benechehab - Younes Gueddari

Octobre 2019

Objet d'étude :

Dans le cadre de projet de recherche industrielle, on s'intéresse à la contribution d'un site industriel de traitement de déchets verts par compostage lors de la mise en exploitation, localisé dans la Loire. En effet, un tel processus dans certaines conditions de fonctionnement (entrant important à différentes périodes de l'année, conditions de fermentation anaérobie au lieu de dégradation aérobie, mauvaise gestion du site) peut entrainer l'émission de composés chimiques avec des risques sanitaires potentiels au niveau des populations avoisinantes.

Afin de discriminer la contribution du site par rapport à la présence éventuelle de ces composés avant installation (que l'on appelle bruit de fond) des campagnes de mesure de ces composés ont été effectuées avant (dans le labels les 2 lettres BF) et après la mise en activités du site (lettre CA dans le labels) à différentes périodes de l'année (H pour hiver et E été).

On cherche donc à répondre à certains questionnements comme : - la localisation des m points de mesure autour du site, montre-elle des regroupements de comportement (composés chimiques atmosphériques d'origine industrielle, automobile, milieu urbain, milieu rural...)? - existe-il une différence entre les campagnes hiver/été ? - existe- il une signature entre les individus avant et après la mise en activité du site ?

Description des données fournies

```
library(readxl)
data <- read_excel("TP4_covC1234_DS19_20.xlsx")

data <- data[,2:19]

#visualization d'un échantillon
print(data[1:10,])
```

```
## # A tibble: 10 x 18
##        B     T     E     X `9_ane` `10_ane` `13_ane` `14_ane` `1_M_2_PA`
##    <dbl> <dbl> <dbl> <dbl>   <dbl>    <dbl>    <dbl>    <dbl>      <dbl>
##  1 17.3   83.4  41.6 142.     20.1     30.9     44.4     70.4       8.86
##  2 14.2   81.5  44.5 157.     24.4     35.3     31.3     76.2      23.7
##  3 20.7  196.  121.  403.     30.5     60.3     22.4     30.1      57.5
##  4 20.0  170.   87.7 289.     30.2     42.8     23.5     69.5      28.0
##  5 24.2  127.   75.2 274.     28.9     69.4     19.4     38.4      11.4
##  6  5.70 133.  149.  448.     31.2     65.8     16.9     33.5     532.
##  7  4.33  91.8  24.8  77.5    25.9     42.0      0       13.1       3.62
##  8 12.4   71.4  48.9 166.     23.9     36.0     13.7     56.9       8.80
##  9 13.1   75.9  56.0 198.     17.4     37.2     16.6     24.1      11.1
## 10 13.5   65.7  46.2 157.     27.1     35.7     56.7     89.5       9.65
## # ... with 9 more variables: BTM <dbl>, FormicAcid <dbl>,
## #   aceticacid <dbl>, NonaDecanoicAc <dbl>, Tot_OcNoDecana <dbl>,
## #   TYPE <chr>, SAISON <chr>, Campagne <chr>, Localisation <chr>
```

Etape :1ère

1) Le traitement statistique des données permettra d'évaluer la variabilité :

a. sur l'ensemble de l'échantillon - c. des 6 différentes campagnes.

La première données statistique qu'on peut extraire est le résumé de nos données :

```
summary(data)
```

```
##        B                T                E                X          
##  Min.   :  0.00   Min.   : 14.36   Min.   :   0.0   Min.   :   77.53  
##  1st Qu.: 22.31   1st Qu.:122.59   1st Qu.: 116.2   1st Qu.:  389.02  
##  Median : 51.09   Median :183.44   Median : 151.7   Median :  593.97  
##  Mean   : 55.03   Mean   :207.77   Mean   : 231.3   Mean   :  877.60  
##  3rd Qu.: 75.63   3rd Qu.:253.56   3rd Qu.: 226.5   3rd Qu.:  911.31  
##  Max.   :138.81   Max.   :675.65   Max.   :4845.0   Max.   :12987.88  
##      9_ane            10_ane            13_ane            14_ane       
##  Min.   : 11.66   Min.   :   0.00   Min.   :   0.00   Min.   :    0.00  
##  1st Qu.: 51.78   1st Qu.:  71.38   1st Qu.:  60.08   1st Qu.:   68.85  
##  Median : 74.67   Median : 173.27   Median : 201.96   Median : 1000.35  
##  Mean   :103.65   Mean   : 263.89   Mean   : 958.57   Mean   : 1833.09  
##  3rd Qu.:120.31   3rd Qu.: 319.57   3rd Qu.: 591.36   3rd Qu.: 1899.30  
##  Max.   :702.24   Max.   :5316.46   Max.   :5176.41   Max.   :10996.74  
##     1_M_2_PA            BTM             FormicAcid        aceticacid     
##  Min.   :   0.00   Min.   :   36.62   Min.   :  13.94   Min.   :   0.00  
##  1st Qu.:  66.85   1st Qu.:  210.02   1st Qu.:  56.56   1st Qu.:  50.21  
##  Median : 131.51   Median :  320.62   Median : 305.79   Median : 172.40  
##  Mean   : 312.79   Mean   :  581.40   Mean   : 481.46   Mean   : 360.50  
##  3rd Qu.: 385.74   3rd Qu.:  490.62   3rd Qu.: 754.63   3rd Qu.: 359.43  
##  Max.   :5191.04   Max.   :21891.22   Max.   :3618.02   Max.   :5624.22  
##  NonaDecanoicAc    Tot_OcNoDecana       TYPE              SAISON         
##  Min.   :   0.00   Min.   :   0.0   Length:140         Length:140        
##  1st Qu.:  54.36   1st Qu.: 132.8   Class :character   Class :character  
##  Median : 272.11   Median : 330.1   Mode  :character   Mode  :character  
##  Mean   : 661.42   Mean   : 621.7                                        
##  3rd Qu.:1136.75   3rd Qu.: 704.5                                        
##  Max.   :2981.69   Max.   :4466.8                                        
##    Campagne         Localisation      
##  Length:140         Length:140        
##  Class :character   Class :character  
##  Mode  :character   Mode  :character  
##                                       
##                                       
##
```

Depuis ce résumé, on voit qu'on a 14 variables quantitatives représentant les concentrations de différentes molécules, et 4 variables qualitatives à savoir, la campagne, la saison, le type et la localisation.

Les variables quantitatives dont le min est égale à 0 reprèsentent des données manquantes (une valeure de concentration nulle est considérée comme donnée manquante), pour les variables qualitatives, la seule contenant une donnée manquante est la variable TYPE (comporte un '?').

Afin de contourner le problème des données manquantes, on a décidé de remplacer là où il y a un zéro (donnée manquante) par la moyenne de cette variable pour biaiser le moins possible notre démarche.

```r
moy <- colMeans(data[,1:14])
for (i in 1:14) {
  data[which(data[,i] == 0),i] <- moy[i]
}

summary(data)
```

```
##        B                T                E                X          
##  Min.   :  4.329   Min.   : 14.36   Min.   :  13.57   Min.   :   77.53  
##  1st Qu.: 23.621   1st Qu.:122.59   1st Qu.: 116.55   1st Qu.:  389.02  
##  Median : 51.823   Median :183.44   Median : 152.69   Median :  593.97  
##  Mean   : 55.427   Mean   :207.77   Mean   : 232.94   Mean   :  877.60  
##  3rd Qu.: 75.631   3rd Qu.:253.56   3rd Qu.: 229.09   3rd Qu.:  911.31  
##  Max.   :138.811   Max.   :675.65   Max.   :4844.96   Max.   :12987.88  
##      9_ane            10_ane            13_ane            14_ane       
##  Min.   : 11.66   Min.   :  16.28   Min.   :   6.004   Min.   :    7.396  
##  1st Qu.: 51.78   1st Qu.:  77.45   1st Qu.:  66.953   1st Qu.:   71.995  
##  Median : 74.67   Median : 177.61   Median : 237.402   Median : 1047.972  
##  Mean   :103.65   Mean   : 265.78   Mean   : 992.805   Mean   : 1859.274  
##  3rd Qu.:120.31   3rd Qu.: 319.57   3rd Qu.: 958.570   3rd Qu.: 1899.300  
##  Max.   :702.24   Max.   :5316.46   Max.   :5176.410   Max.   :10996.742  
##     1_M_2_PA            BTM             FormicAcid      
##  Min.   :   3.621   Min.   :   36.62   Min.   :  13.94  
##  1st Qu.:  70.802   1st Qu.:  210.02   1st Qu.:  56.56  
##  Median : 133.221   Median :  320.62   Median : 305.79  
##  Mean   : 315.021   Mean   :  581.40   Mean   : 481.46  
##  3rd Qu.: 385.739   3rd Qu.:  490.62   3rd Qu.: 754.63  
##  Max.   :5191.044   Max.   :21891.22   Max.   :3618.02  
##    aceticacid       NonaDecanoicAc     Tot_OcNoDecana   
##  Min.   :   3.354   Min.   :   8.476   Min.   :   9.236  
##  1st Qu.:  60.878   1st Qu.: 104.777   1st Qu.: 145.380  
##  Median : 184.730   Median : 550.487   Median : 352.448  
##  Mean   : 365.650   Mean   : 737.015   Mean   : 643.942  
##  3rd Qu.: 360.649   3rd Qu.:1136.749   3rd Qu.: 704.506  
##  Max.   :5624.216   Max.   :2981.690   Max.   :4466.796  
##      TYPE              SAISON           Campagne        
##  Length:140         Length:140         Length:140        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
##                                                          
##                                                          
##                                                          
##  Localisation      
##  Length:140        
##  Class :character  
##  Mode  :character  
##                    
##                    
##                    
```

Concernant l'étendu des variables quantitatives, il nous faut un expert du domaine pour pouvoir qualifier les résultats qu'on a, mais intuitivement nous assumons qu'ils sont assez grands pour connoter une bonne variabilité des données.

On va ensuite calculer les moyennes et les écarts types des 14 variables quantitatives pour toutes les observations :

```
quantitative <- data[,1:14]
#extraction des données quantitatives

quantitative <- apply(quantitative,2,as.numeric)

moy <- colMeans(quantitative)
#calcul de la moyenne de chaque variable

ones = rep(1, nrow(quantitative))
Mean = ones %*% t(moy)

XC <- (quantitative-Mean)
#centrer les données

n <- length(quantitative[,1])
V <- (1/n)*t(XC)%*%XC
#Calcul de la matrice de variance/covariance

sd <- sqrt(diag(V))
#écart-type de chacune des variables
```

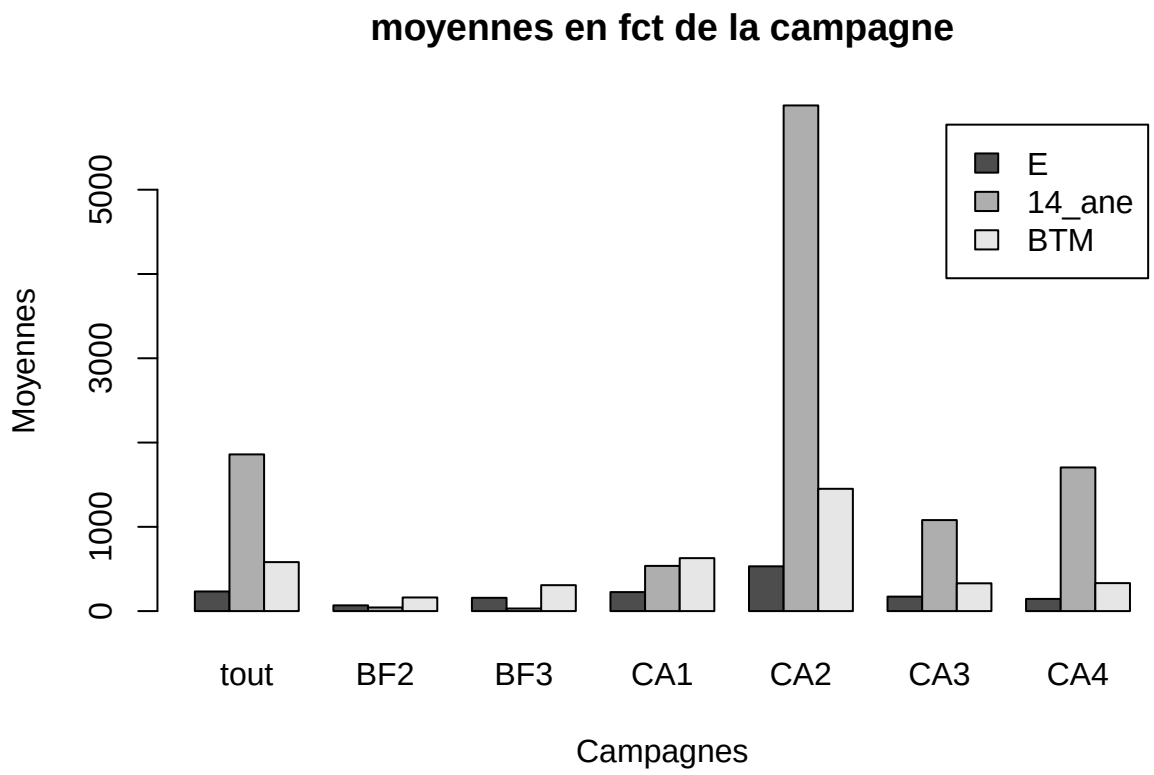
Ensuite pour pouvoir comparer avec les données correspondantes à chaque campagne, on va refaire ce calcule pour chaque campagne, en voici un example pour la première :

```
#Première campagne
quantitativeC1 <- data[data$Campagne=="BF2",1:14]

quantitativeC1 <- apply(quantitativeC1,2,as.numeric)

moyC1 <- colMeans(quantitativeC1)

ones = rep(1, nrow(quantitativeC1))
MeanC1 = ones %*% t(moyC1)

XCC1 <- (quantitativeC1-MeanC1)
nC1 <- length(quantitativeC1[,1])

VC1 <- (1/nC1)*t(XCC1)%*%XCC1

sdC1 <- sqrt(diag(VC1))
```

Après calcul des moyennes et écarts types de hcacune des variables pour chacune des campagnes, on va tracer l'histogramme correspondant aux moyennes de 3 variables choisies au hasard parmi les 14 (les variables choisies sont E, 14_ane et BTM), dans les différentes campagnes pour pouvoir comparer entre elles, en voici le code correspondant :

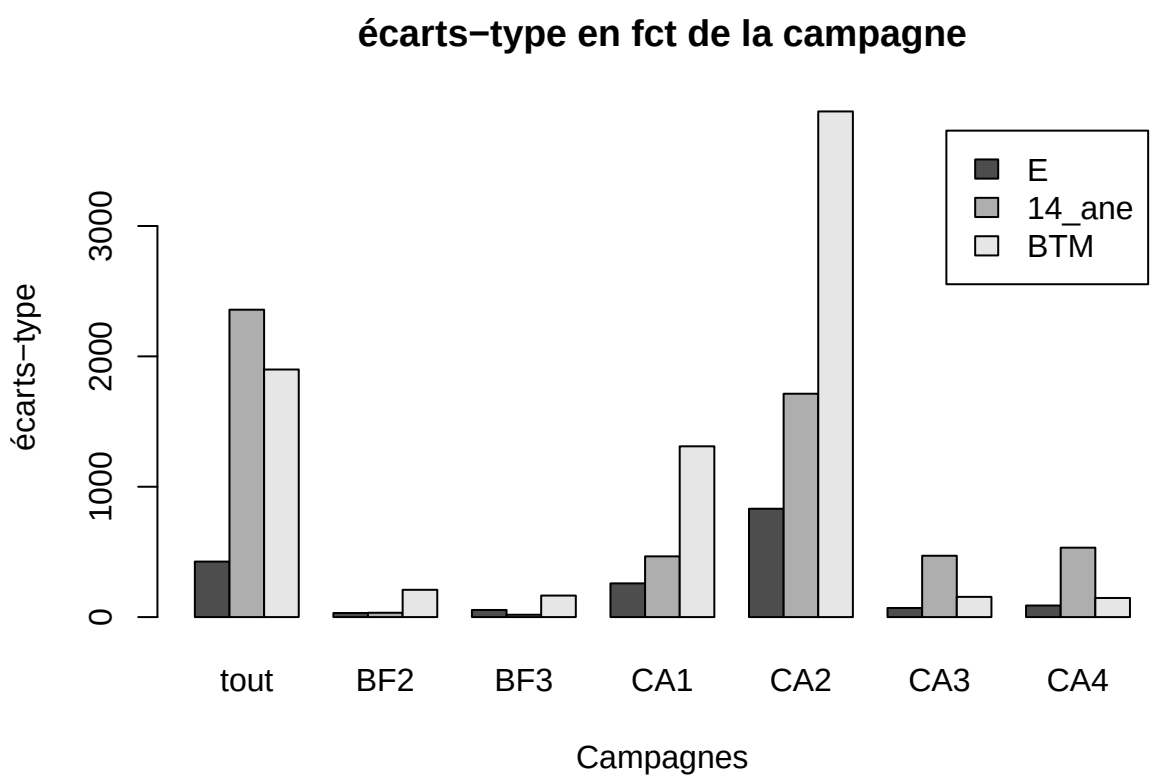
```
E <- c(moy[3],moyC1[3],moyC2[3],moyC3[3],moyC4[3],moyC5[3],moyC6[3])
ane <- c(moy[8],moyC1[8],moyC2[8],moyC3[8],moyC4[8],moyC5[8],moyC6[8])
BTM <- c(moy[10],moyC1[10],moyC2[10],moyC3[10],moyC4[10],moyC5[10],moyC6[10])
#vecteurs représentant les moyennes pour chaque variable
```

```r
M <- as.matrix(cbind(E,ane,BTM))
colnames(M) <- c("E","14_ane","BTM")
rownames(M) <- c("tout","BF2","BF3","CA1","CA2","CA3","CA4")

#fonction barplot pour l'histogramme
barplot(t(M),
        beside = TRUE,
        legend.text = TRUE,
        ylab = "Moyennes",
        xlab = "Campagnes",
        main = "moyennes en fct de la campagne")
```



En analysant cet histogramme, on peut dire que les hausses dans les concentrations des espèces étudiées ont été enregistrés pendant la période après l'ouverture du site, et notamment pendant la deuxième campagne "CA2". ces valeurs étaient relativement faibles pendant les campagnes avant l'ouverture du site.

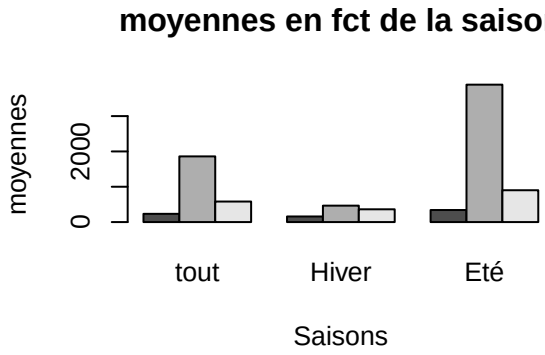
Avant de passer à la question suivante, nous avons décidé de tracer le même histogramme mais pour les valeurs d'écart type cette fois, voila ce que ca donne :

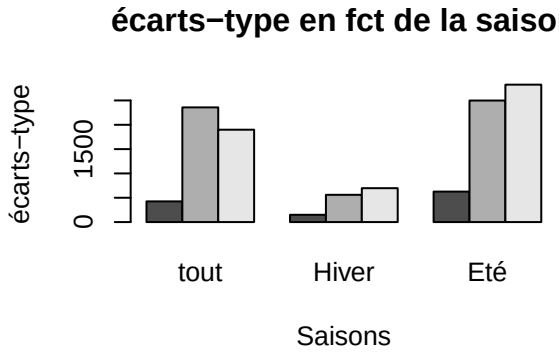


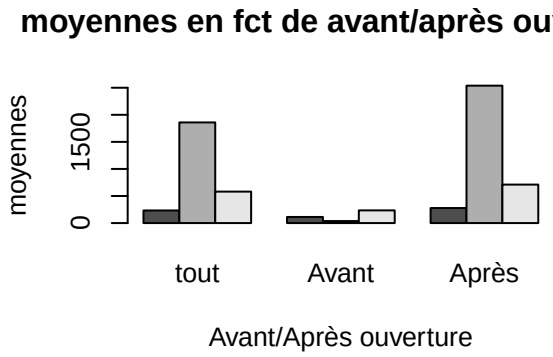
La même remarque encore une fois, pendant les campagnes après ouverture du site et surtout la deuxième on a une variabilité importante par rapport au reste, et c'est cette valeur "outlier" qui biaise l'écart-type moyen de toutes les données et le rend relativement important alors que pendant la plupart des campagnes il est relativement faible.

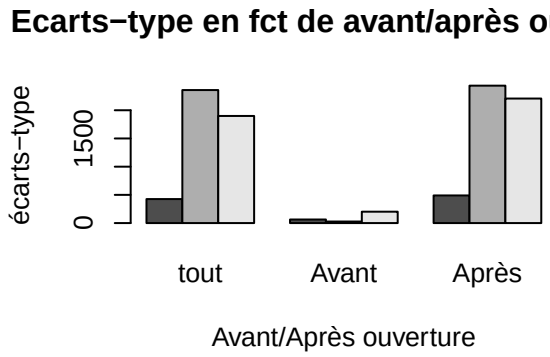
b. pour les campagnes de mesure en hiver et les campagnes en été d. sur les deux campagnes avant ouverture du site (BF) et après ouverture du site (CA)

Même chose dans cette question, les quatres graphes qui suivent representent les moyennes et écarts types des memes variables étudiées dansla question précédente, en hiver/été et avant/après ouverture :









Conclusion : significativement, les concentrations des espèces mesurées ainsi que sa variabilité ont connu une forte hausse pendant l'été et après l'ouverture du site industriel, ce qui confirme les remarques faites dans la question précédente.

Cette conclusion est très utile pour la suite du travail, comme elle va nous orienter et nous donner un recul par rapport à notre stratégie d'analyse de ces données.

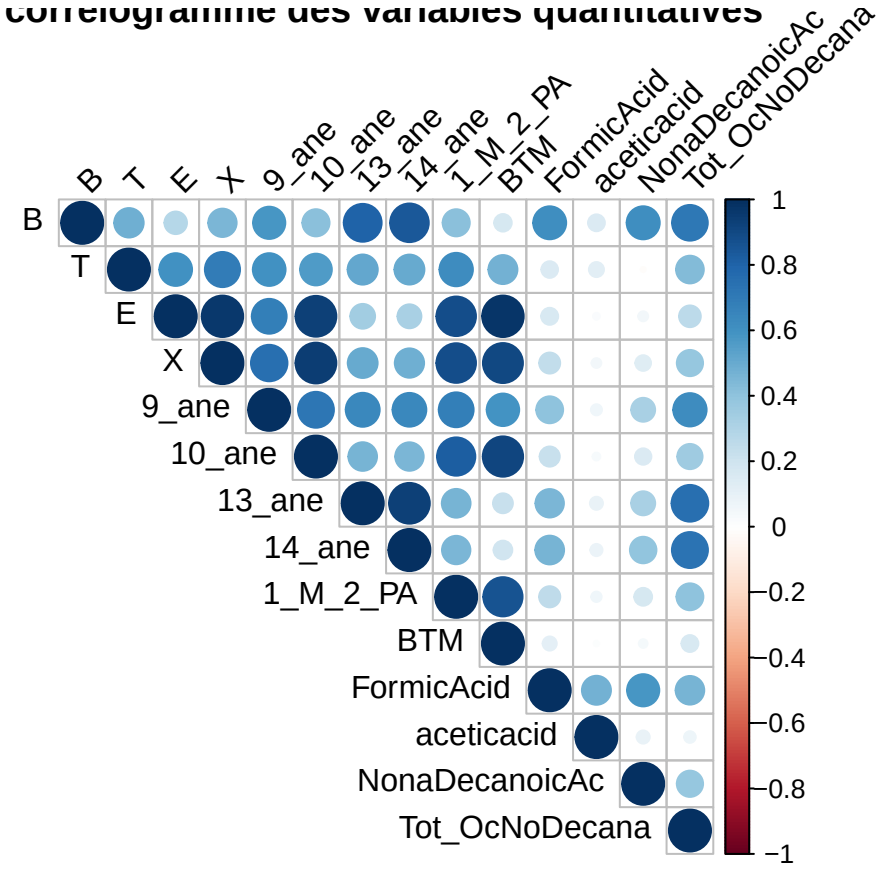
2) L'affichage des corrélations possibles entre les différentes variables : pour chacune des 4 périodes (hiver, été, avant activité, après activité)

Maintenant on va calculer la matrice de corrélation dans le cas de toutes les observations, et la visualiser :

```
Gamma <- cor(quantitative)
library(corrplot)
```

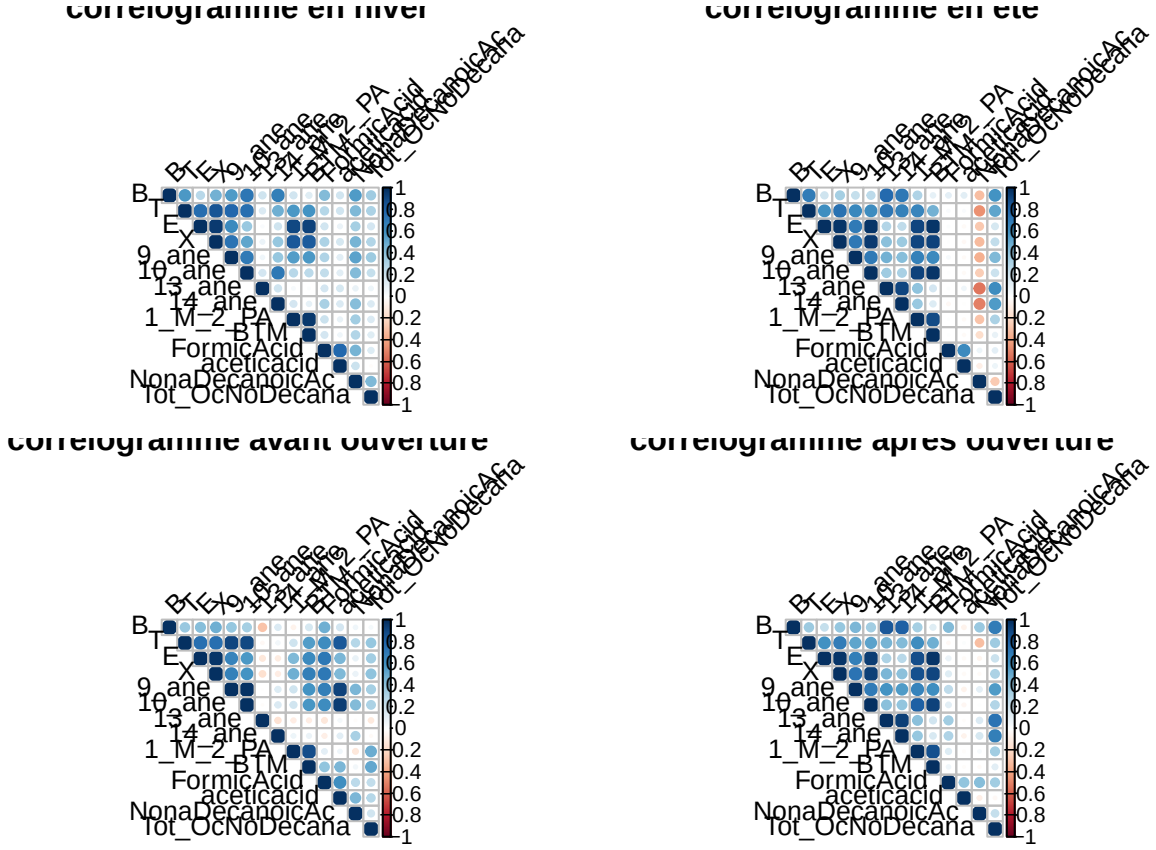
```
## corrplot 0.84 loaded
```

```
corrplot(Gamma, type = "upper", order = "original",
         tl.col = "black", tl.srt = 45, title = "correlogramme des variables quantitatives")
```



Ce corrélogramme montre qu'il y a deja une forte corrélation entre certaines variables, entre la variable 'E' et la variable 'X' par exemple, et que d'autres variables (aceticacid par exemple) ne sont pas corrélées à aucune des autres.

Traçons maintenant le corrélogramme pour les 4 périodes demandées :



L'interpretation de ces résultats est un peu difficile, vu que les 4 périodes montrent des corrélations fortes entre différentes variables à chaque fois.

La seule remarque qu'on a pu en tirer c'est que pour une période donnée, il y a des variables qui ne sont corrélées à aucune des autres, ces variables qu'on va appeler par la suite des variables marqueurs sont très importants pour définir la signature d'une période plutard dans l'ACP.

3) Des traitements statistiques que pouvez-vous en déduire sur les différents périodes hiver/été et avant et après ouvertures du site ?

Récapitulons maintenant tout ce qu'on a pu obtenir des traitements statistiques qu'on vient d'effectuer sur nos données :

- Les concentrations des molécules chimiques contrôlées (variables quantitatives) connaissent une forte hausse et une forte variabilité également pendant la période après l'ouverture du site et surtout la deuxième campagne pendant cette période (CA2)
- Ces concentrations sont plus importantes aussi en été qu'en hiver avec des écarts assez considérables.
- Le corrélogramme sur les données entières montre quelques corrélations non-négligeables entres certaines variables et d'autres non-corrélées à aucune des autres (aceticacid).
- Pendant différentes périodes de l'année (hiver,été,avant ouverture du site,après ouverture du site), le corrélograme change complétement et donne lieu à de nouvelles variables non-corrélées à aucune des autres et qu'on va utiliser pour définir la signature par la suite.

Etape :2éme

4) On cherche à savoir si l'on peut identifier une réduction du nombre de variables par ACP: recherche de composantes principales et si les individus se regroupent ou pas selon ce nouvel espace R (q<p avec p=14). Vous devez mettre en oeuvre l'ACP en justifiant les résultats (Inertie expliquée/inertie totale, qualité de la réduction de dimension et de la qualité des projections de individus sur les 14 variables quantitatives. Quelles variables sont les mieux expliquées ?

Avant de commencer l'analyse, on a mis en place une fonction réalisant l'ACP, en voila l'implementation :

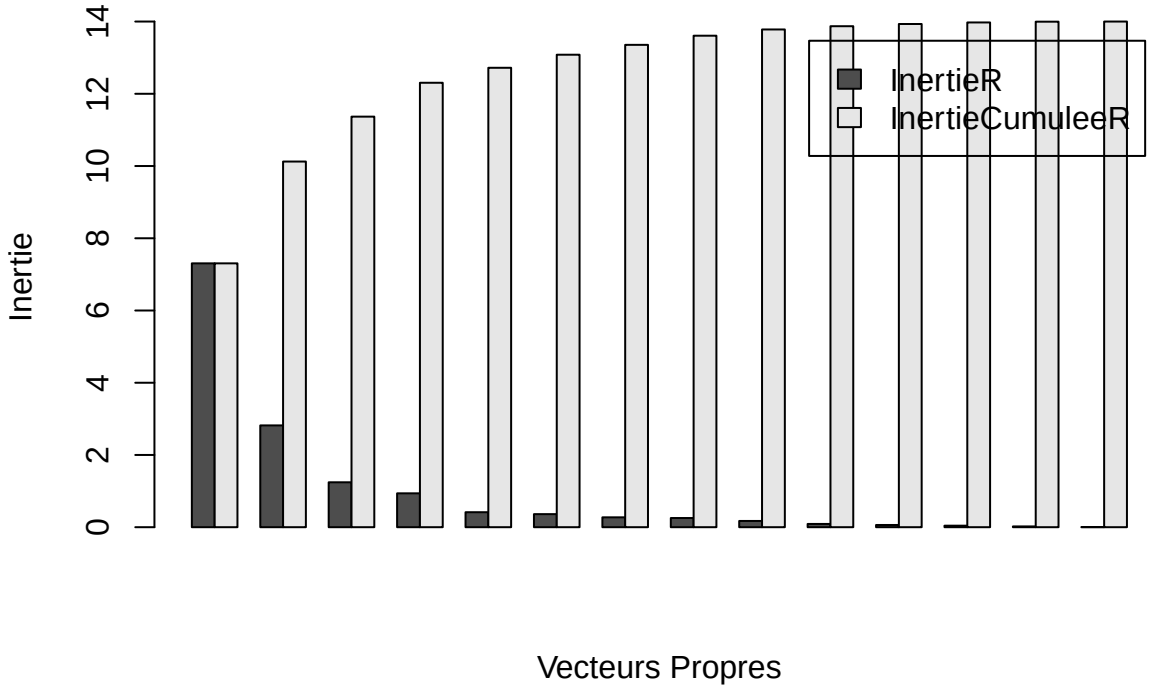
```
AnalyseACP<-function(da){
  #1ére étape : centrer et réduire les données
  moy <- colMeans(da)
  ones = rep(1, nrow(da))
  Mean = ones %*% t(moy)
  XC <- (da-Mean)
  n <- length(da[,1])
  V <- (1/n)*t(XC)%*%XC
  SD <- sqrt(diag(V))
  SD <- ones %*% t(SD)
  daCR <- XC/SD

  #Décomposition spectrale
  VCR <- (1/n)*t(daCR)%*%daCR

  ACPR <- eigen(VCR)

  VecteursPropresR <- ACPR$vectors
  ValeursPropresR <- ACPR$values
  print(c("Les valeurs propres obtenues :",ValeursPropresR))

  #Histogramme de l'inertie
  InertieR <- ValeursPropresR
  InertieCumuleeR <- rep(0,14)
  for (i in 1:14) {
    InertieCumuleeR[i] <- sum(ACPR$values[1:i])
  }
  barplot(t(as.matrix(cbind(InertieR,InertieCumuleeR))),
          beside = TRUE,
          legend.text = TRUE,
          angle=TRUE,
          ylab = "Inertie",
          xlab = "Vecteurs Propres")

  # Projection sur les nouvelles coordonnées
  NewdaCR <- daCR%*%VecteursPropresR

  # RIn = Inertie cumulée jusqu'à l'axe i / Inertie totale
  RIn <- function(i){
    A <- sum((ValeursPropresR))
    B <- sum((ValeursPropresR[1:i]))
    return(B/A)
```

```
  }
  print(c("Inertie cumulée jusqu'au deuxième axe / Inertie totale",RIn(2)))

  #Definition de la fonction Q(nn,i,k) avc nn la matrice des données, qui calcule
  #la qualité de projection de l'individu i en prenant en compte ses k premières
  #coordonnées
  Qual <- function(nn,i,k){
    A <- sum((nn[i,]^2))
    B <- sum((nn[i,1:k]^2))
    return(B/A)
  }

  #Qualité de projection
  KR <- 0
  for (j in 1:length(da[,1])) {
    KR <- KR + Qual(NewdaCR,j,3)
  }
  KR <- KR/length(da[,1])
  print(c("qualité des projections",KR))

  #Contribution de la variable j à l'axe i
  coeffR <- function(i,j){
    return(cor(NewdaCR[,i],daCR[,j]))
  }
  R <- matrix(0,14,14)

  #Matrice des contributions
  for (i in 1:14) {
    for (j in 1:14) {
      R[i,j] <- coeffR(i,j)
    }
  }
  colnames(R) <- c("B","T","E","X","9_ane","10_ane","13_ane","14_ane","1_M_2_PA","BTM",
                   "FormicAcid","aceticacid","NonaDecanoicAc","Tot_OcNoDecana")

  #Histogramme des contributions de chaque variable dans les 2 premièrs axes factoriels

  barplot(sqrt(R[1,]^2+R[2, ]^2),
          axisnames = FALSE,
          col=rainbow(14),
          legend=colnames(R),
          xlim=c(0,27),
          ylim = c(0,1),
          main = "Histogramme des contributions de chaque \n variable dans les 2 premièrs
          axes factoriels")
  return(NewdaCR)

}
```

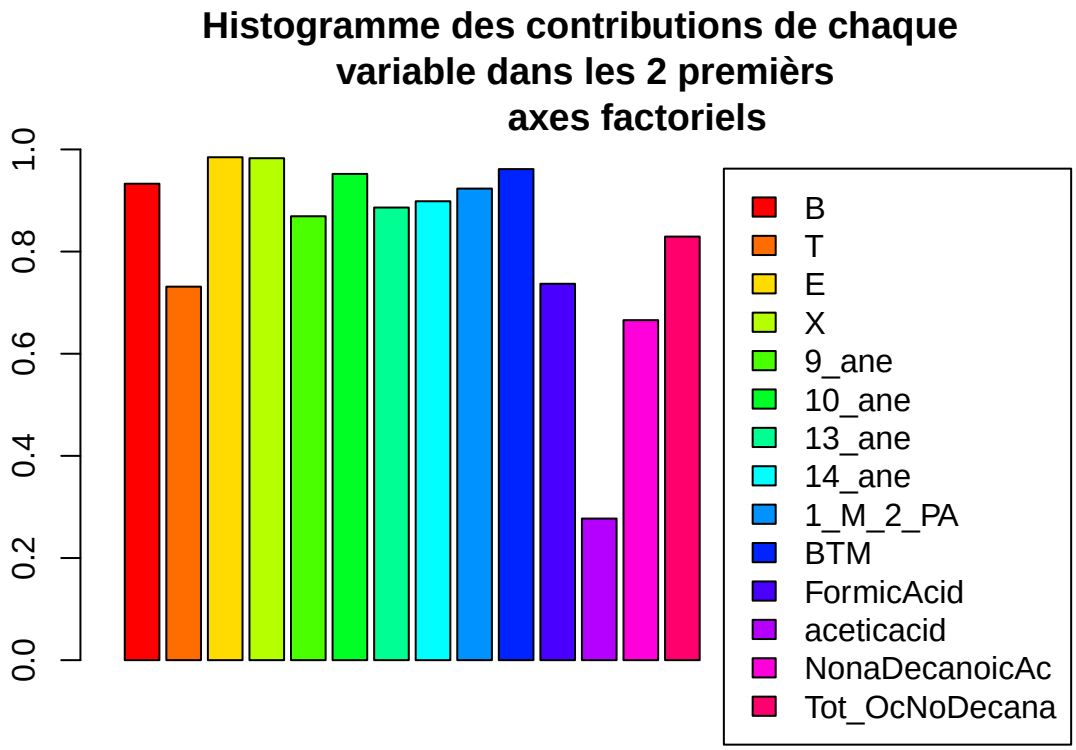
On exécute notre fonction sur la totalité de nos données quantitatives :

```
#On récupère les nouvelles coordonnées sur les nouveaux axes factoriels
NewdaCR <- AnalyseACP(quantitative)
```

```
##  [1] "Les valeurs propres obtenues :" "7.30491703797004"
##  [3] "2.81834323442796"               "1.24358582640429"
##  [5] "0.937566258254316"              "0.414899698091799"
##  [7] "0.36183476685432"               "0.273092308485101"
##  [9] "0.25371065315244"               "0.172422306997817"
## [11] "0.0903761549633187"             "0.061388151039208"
## [13] "0.0424222542916395"             "0.021666527030401"
## [15] "0.00377482203735609"
```

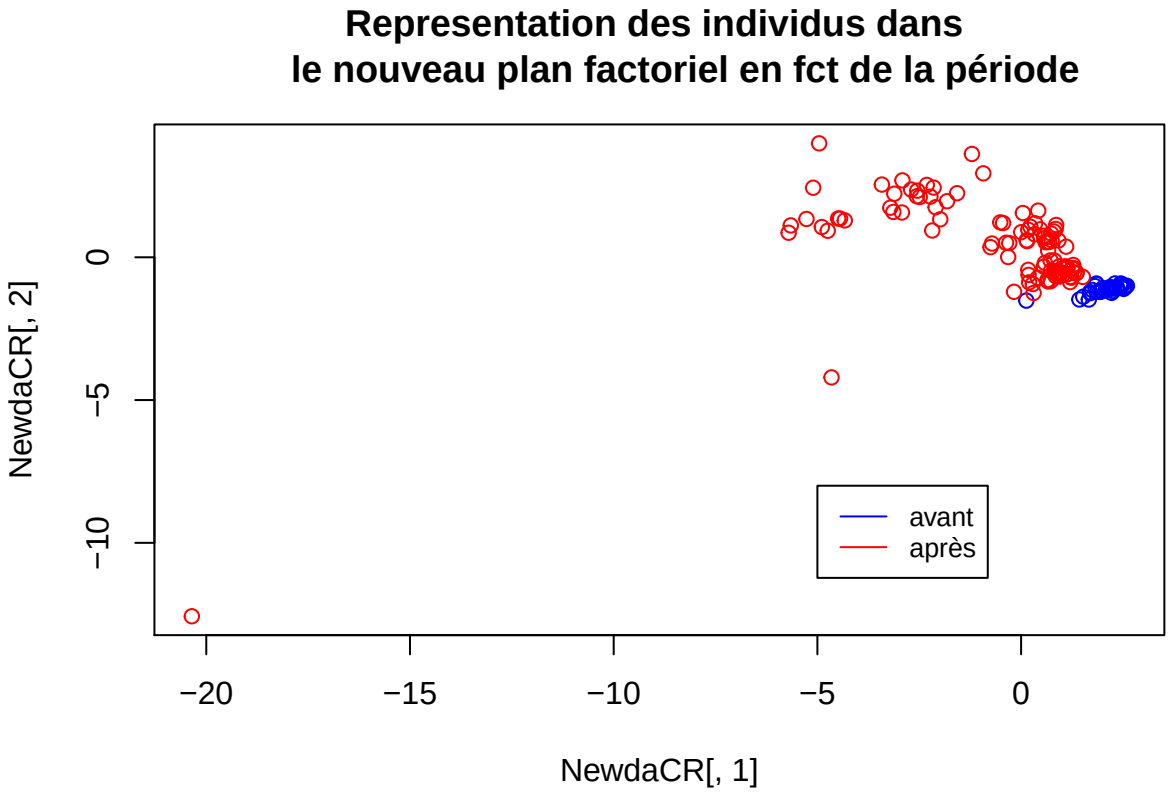



```
## [1] "Inertie cumulée jusqu'au deuxième axe / Inertie totale"
## [2] "0.723090019456999"
## [1] "qualité des projections" "0.693586708572126"
```

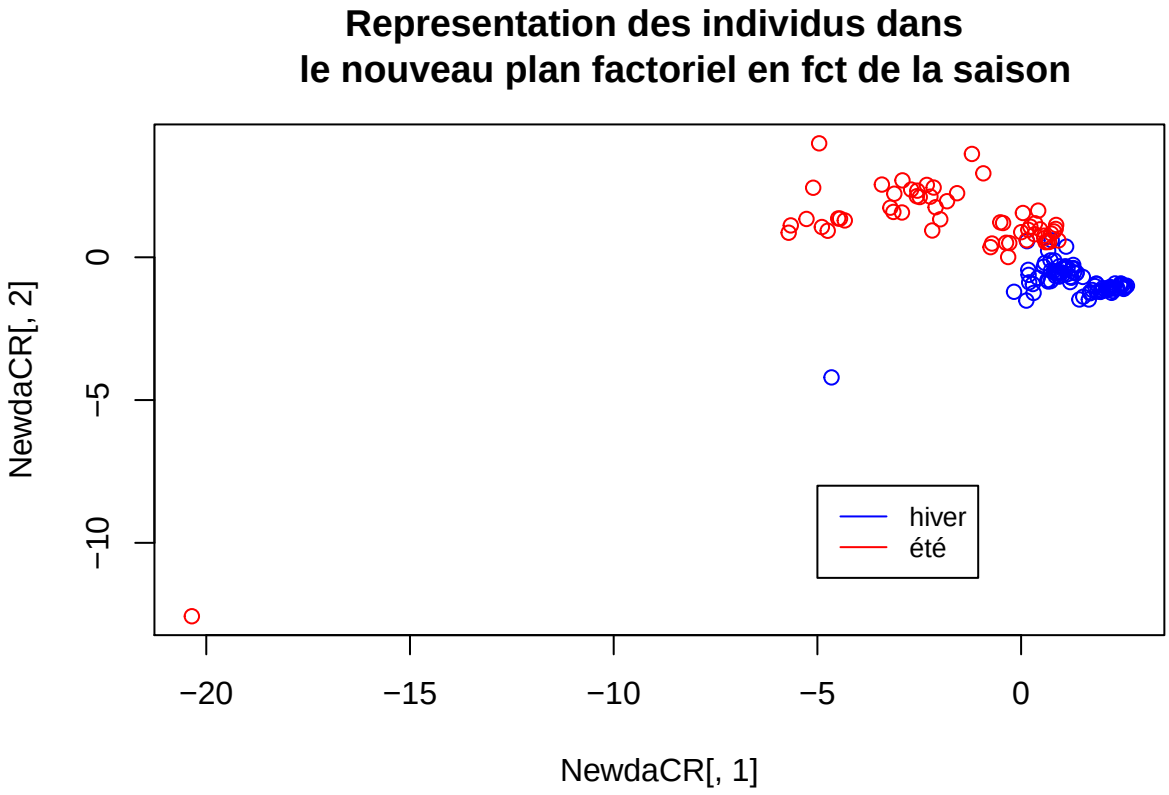


- Si on choisit deux axes principaux on obtient un coefficient supérieur à 66% On peut voir aussi que c'est cohérent avec la méthode du coude puisqu'on remarque un point d'inflexion sur la deuxieme barre.
- Pour le graphe des contributions, on peut voir que l'aceticacid n'a pas une grande contribution et au même temps c'est la variable qui n'est corrélée à aucune des autres d'où la cohérence des résultats.
- Afin de voir est ce que la représentation des individus dans le nouveau plan factoriel permet un bon regroupement des individus, on vas les representer coloriés en fonction de la saison hiver/été, et la période avant/après ouverture du site

```
NewdaCR <- cbind(NewdaCR[,1],NewdaCR[,2],data$Campagne,rep(0,length(data$Campagne)))
NewdaCR[which(substr(NewdaCR[,3], 1, 2) == "BF"),4] <- 4
NewdaCR[which(substr(NewdaCR[,3], 1, 2) == "CA"),4] <- 2
plot(NewdaCR[,1], NewdaCR[,2],col=NewdaCR[,4], main = "Representation des individus dans
     le nouveau plan factoriel en fct de la période")
legend(-5, y=-8, legend=c("avant", "après"),
       col=c(4, 2), lty=1, cex=0.8)
```




```
NewdaCR <- cbind(NewdaCR[,1],NewdaCR[,2],data$SAISON,rep(0,length(data$SAISON)))
NewdaCR[which(NewdaCR[,3] == "hiver"),4] <- 4
NewdaCR[which(NewdaCR[,3] == "été"),4] <- 2
plot(NewdaCR[,1], NewdaCR[,2],col=NewdaCR[,4], main = "Representation des individus dans
     le nouveau plan factoriel en fct de la saison")
legend(-5, y=-8, legend=c("hiver", "été"),
       col=c(4, 2), lty=1, cex=0.8)
```



D'où la bonne representativité des individus dans notre plan factoriel selon les critères qu'on avait jugé qualitativement dans la 1ére étape.

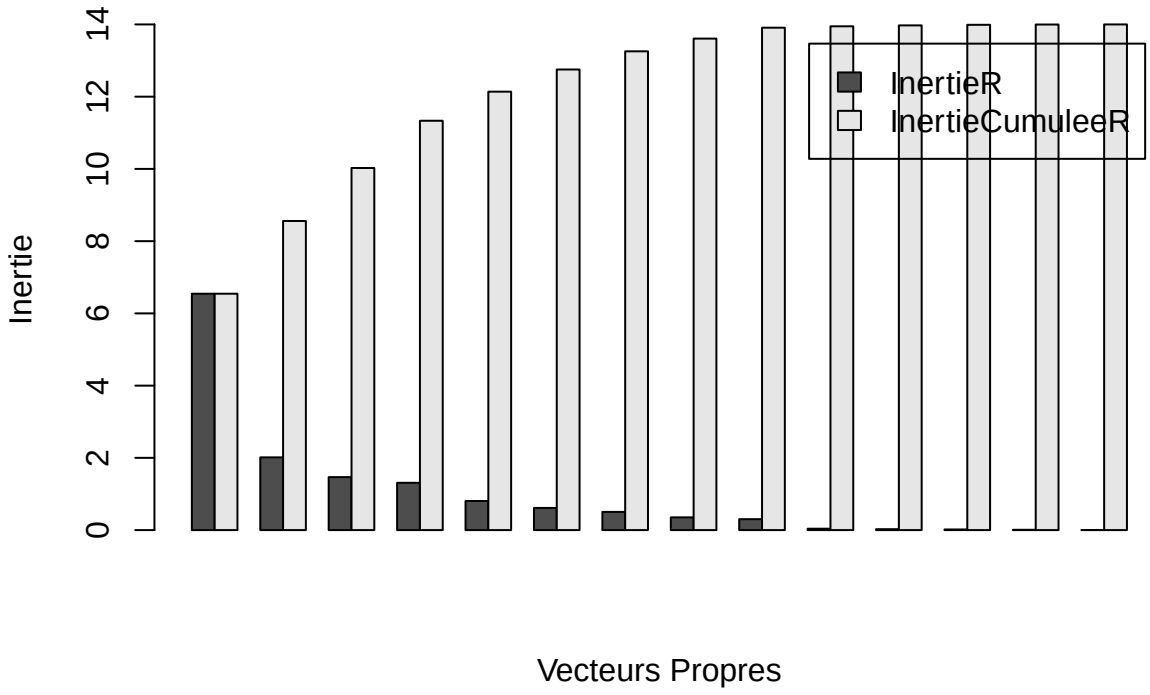
- Pour répondre à la dernière partie de la question, il faut revenir à l'histogramme des contributions, les variables les mieux expliquées sont ceux qui contribuent le maximum aux axes factoriels, et donc c'est pratiquement toutes les variables sauf l'aceticacid comme on a vu avant.

5) On recherche une signature des composants pour chaque période (été, hiver, 1 avant_activité, 1 apres_activité) ; une réduction du nombre de variables par ACP serait-elle une méthode adaptée pour tenter de répondre et comment la mettre en oeuvre ?

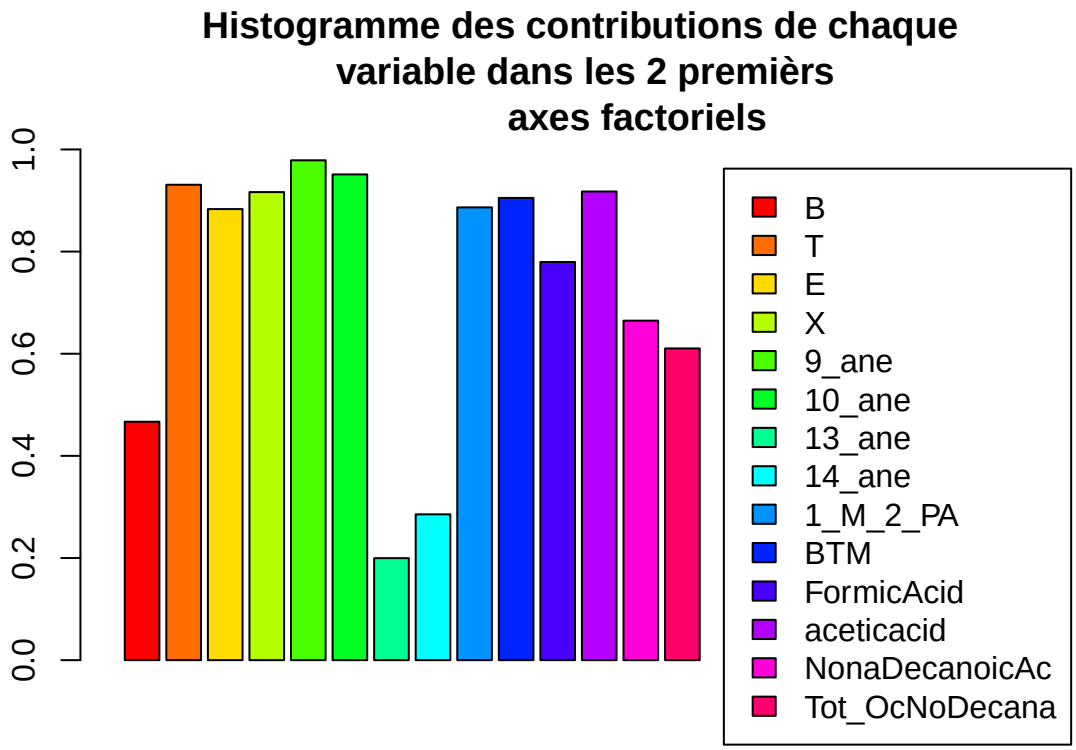
On exécute une analyse ACP sur les données avant, après puis les données hiver/été.

```
# Avant l'ouverture du cite
NewdaCRav <- AnalyseACP(quantitativeav)
```

```
##  [1] "Les valeurs propres obtenues :" "6.54450199091632"
##  [3] "2.01328867709391"               "1.46655551810667"
##  [5] "1.30864503774387"               "0.805196089220692"
##  [7] "0.613853044848849"              "0.503713001812633"
##  [9] "0.352124130138066"              "0.301661276849868"
## [11] "0.0383342888761796"             "0.0249713704706743"
## [13] "0.0167068181606601"             "0.00808090407045819"
## [15] "0.00236785169115027"
```

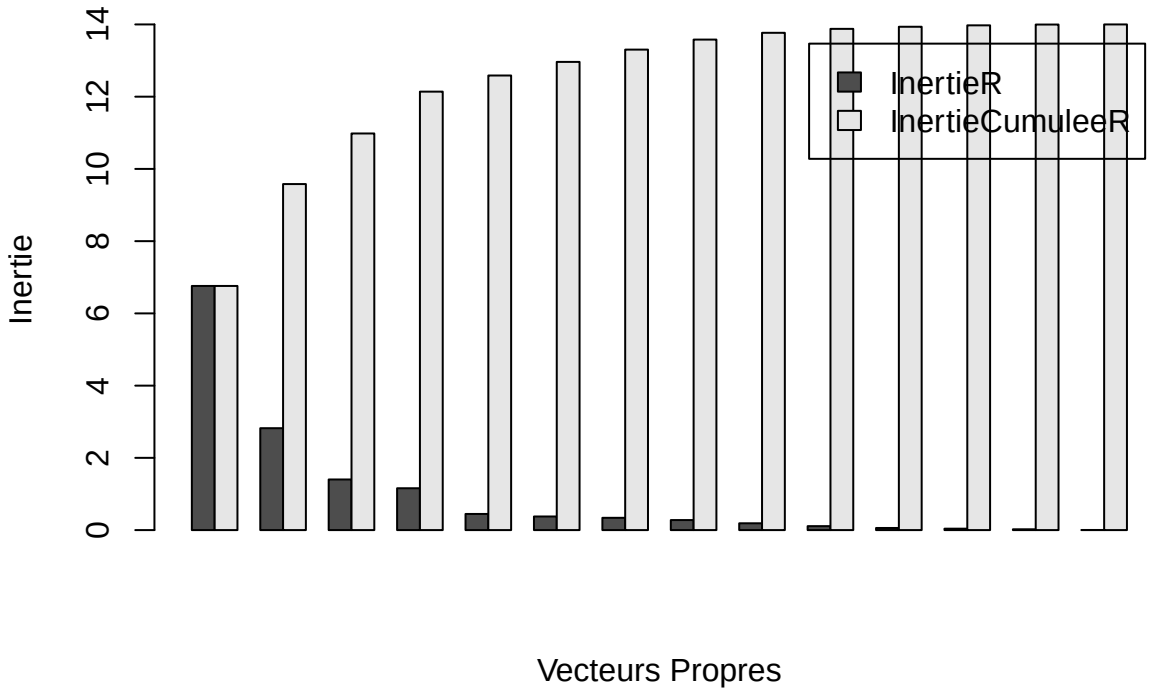



```
## [1] "Inertie cumulée jusqu'au deuxième axe / Inertie totale"
## [2] "0.611270762000731"
## [1] "qualité des projections" "0.578255826842336"
```

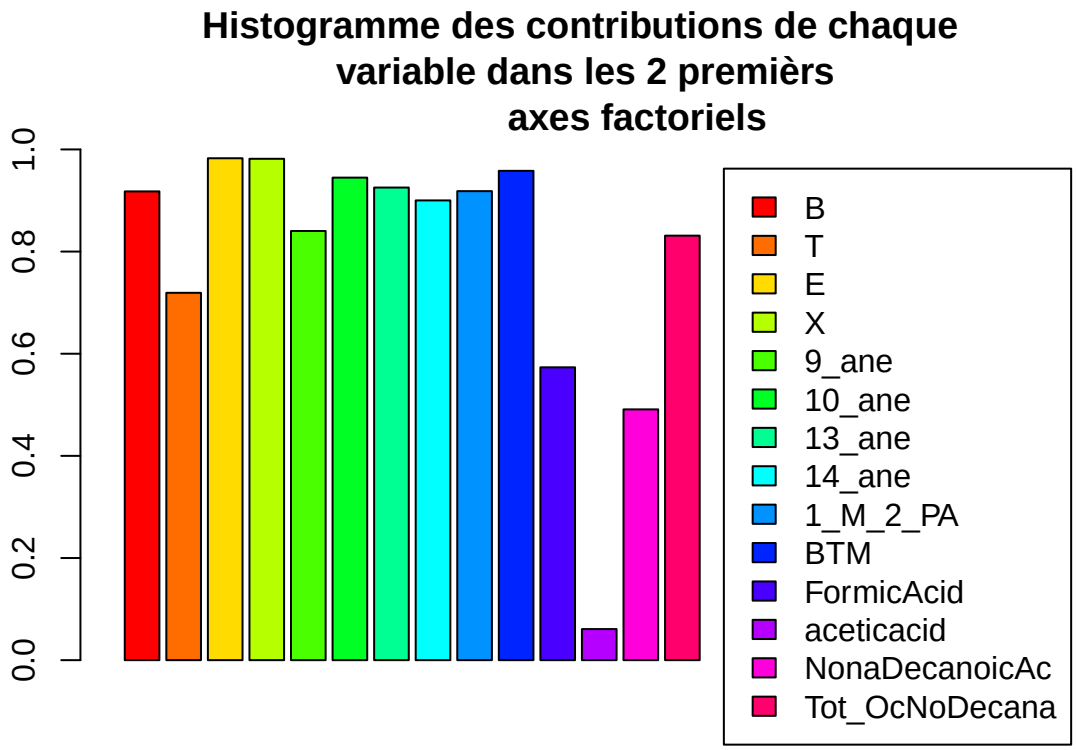



```
# Après l'ouverture du cite
NewdaCRav <- AnalyseACP(quantitativeap)
```

```
##  [1] "Les valeurs propres obtenues :" "6.75919469165772"
##  [3] "2.82096473009676"               "1.40098892088359"
##  [5] "1.15826115223064"               "0.446443046641428"
##  [7] "0.376651468239916"              "0.339914003508642"
##  [9] "0.277692873464561"              "0.187920445551924"
## [11] "0.109329785451229"              "0.0582682087728573"
## [13] "0.0392223363329987"             "0.0218491121388052"
## [15] "0.00329922502892541"
```




```
## [1] "Inertie cumulée jusqu'au deuxième axe / Inertie totale"
## [2] "0.684297101553892"
## [1] "qualité des projections" "0.706571438605668"
```



On peut voir à partir du dernier graphe une difference entre les données avant et les données après:

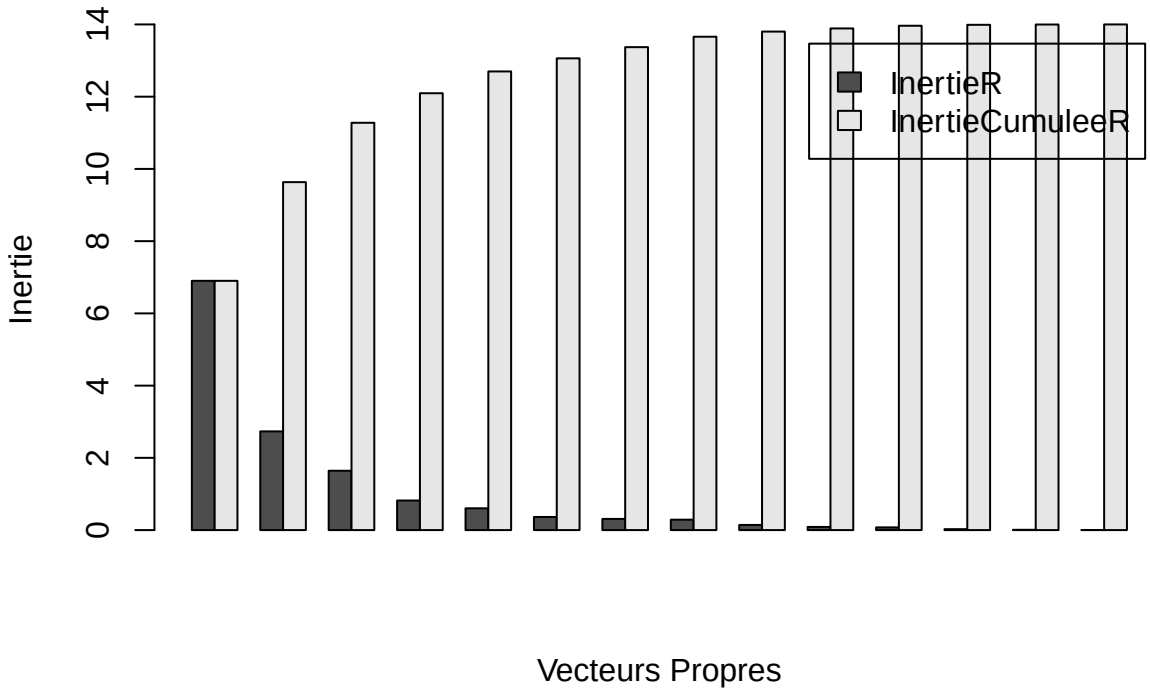
- Avant: B et 14_ane ont un petit pourcentage alors qu'après elles contribuent complètement aux nouveaux axes principaux.
- Après: la contribution de l'aceticacid est très petite et celle du NonaDecanoicAc et du FormicAcid depasse à peine la moitié alors qu'avant elle avait un pourcentage assez grand.

On peut dire que les pourcentages de aceticacid ,NonaDecanoicAc ,FormicAcid ,B et 14_ane forment une signature des données après et avant l'ouverture du site industriel.

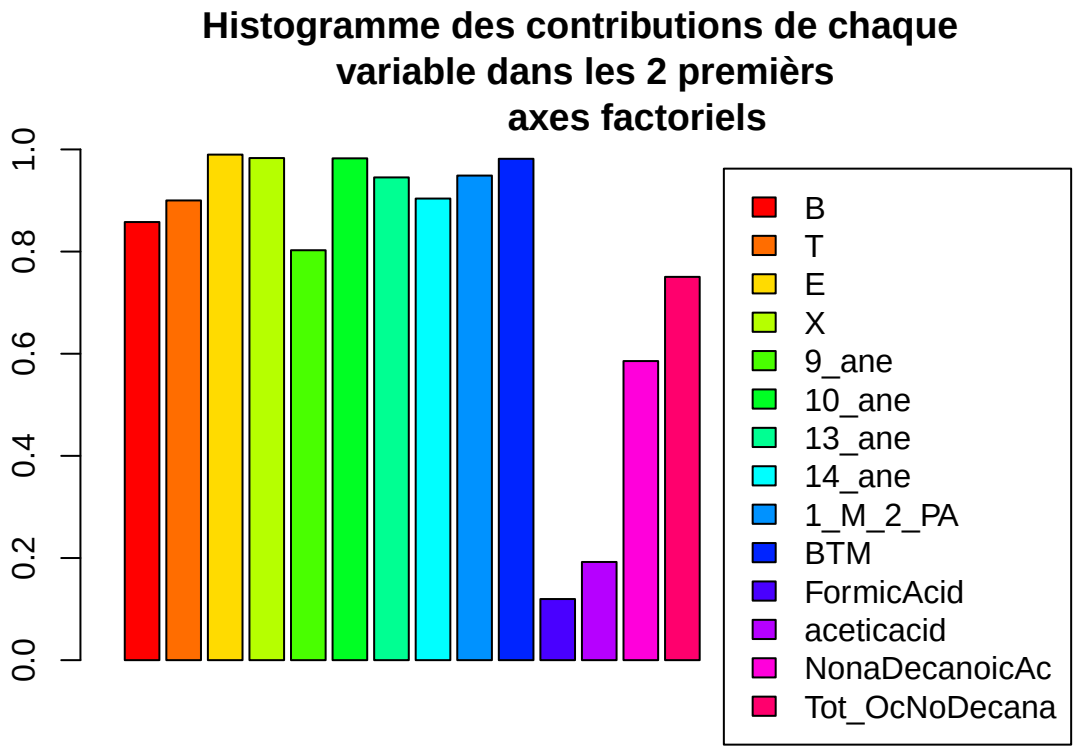
De même pour la saison :

```
# Pendant l'été
NewdaCRav <- AnalyseACP(quantitativeete)
```

```
##  [1] "Les valeurs propres obtenues :" "6.90117226777116"
##  [3] "2.73252971164879"               "1.6430072479776"
##  [5] "0.818502226850043"              "0.602854560624302"
##  [7] "0.363269550291321"              "0.309638952684723"
##  [9] "0.288263876940202"              "0.142566207676341"
## [11] "0.0871693329007088"             "0.0745394115182248"
## [13] "0.0260107457332655"             "0.00830204323621621"
## [15] "0.00217386414709347"
```

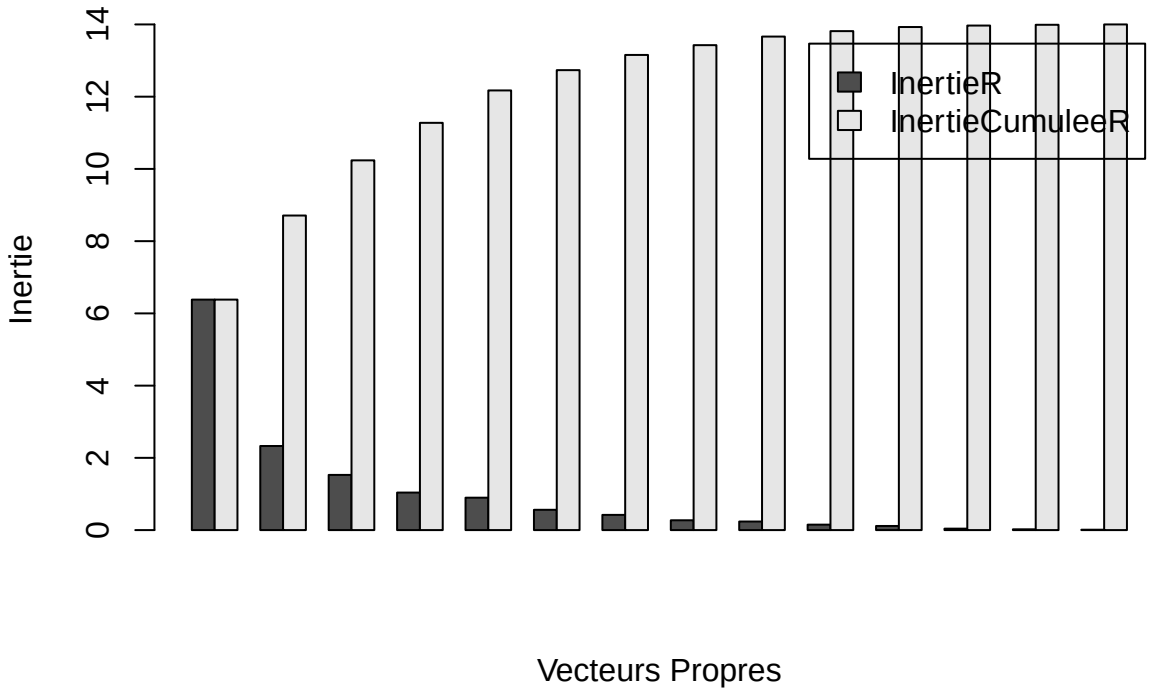



```
## [1] "Inertie cumulée jusqu'au deuxième axe / Inertie totale"
## [2] "0.688121569958569"
## [1] "qualité des projections" "0.706117754292849"
```

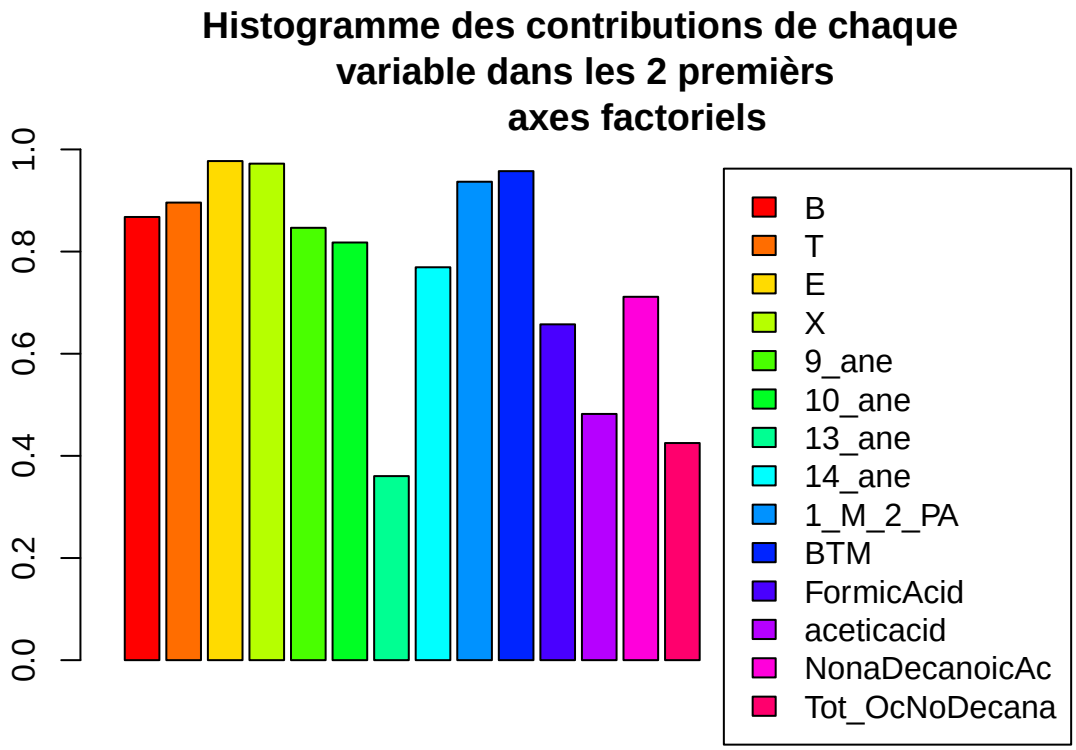



```
# Pendant l'hiver
NewdaCRav <- AnalyseACP(quantitativehiver)
```

```
##  [1] "Les valeurs propres obtenues :" "6.38126859848064"
##  [3] "2.32757413429077"               "1.52763587063936"
##  [5] "1.03865626961708"               "0.896926290277267"
##  [7] "0.562074006911687"              "0.420674419287781"
##  [9] "0.271597366634239"              "0.237054208792386"
## [11] "0.150898946149913"              "0.114017400102446"
## [13] "0.0400584448617704"             "0.0214354754840374"
## [15] "0.0101285684706171"
```




```
## [1] "Inertie cumulée jusqu'au deuxième axe / Inertie totale"
## [2] "0.622060195197958"
## [1] "qualité des projections" "0.638388768210163"
```



- En été: FormicAcid ,aceticacid ont un très petit pourcentage alors qu'en hiver elles representent presque la moitié données.
- En hiver: La contribution de NonaDecanoicAc et Tot_OcNoDecana devient plus petite par rapport à l'été.

On peut dire que les pourcentages de FormicAcid ,aceticacid, NonaDecanoicAc et Tot_OcNoDecana forment une signature des données en été et en hiver.

Pour savoir si la réduction du nombre de variables par cet ACP est judicieuse on va revenir aux representations des individus selon les deux critères qu'on a; On a vu que selon les modalités de la variable SAISON, le nuage obtenu est séparé significativement confirmant nos remarques issus du pré-traitement de la 1ére étape. La même chose pour la période avant/après ouverture du site, donc l'ACP est bien adéquate.

Compte tenu de vos résultats de cette étape : quels sont vos principaux constats, quelles propositions de traitements faites-vous pour chaque période, quelles sont les variables marqueurs ?

Les principaux constats de cette partie sont les suivants :

- L'aceticacid n'est pas bien representé dans le plan factoriel obtenu, alors qu'elle figure parmi les éléments constituants les signatures des deux périodes avant/après ouverture du site. Ainsi on pense qu'il peut être sage de ne pas la considèrer dans nos traitements.
- Pour chaque période étudiée, on propose d'enlever les variables mal expliquées vu qu'elles n'influencent pas significativement sur les axes factoriels.

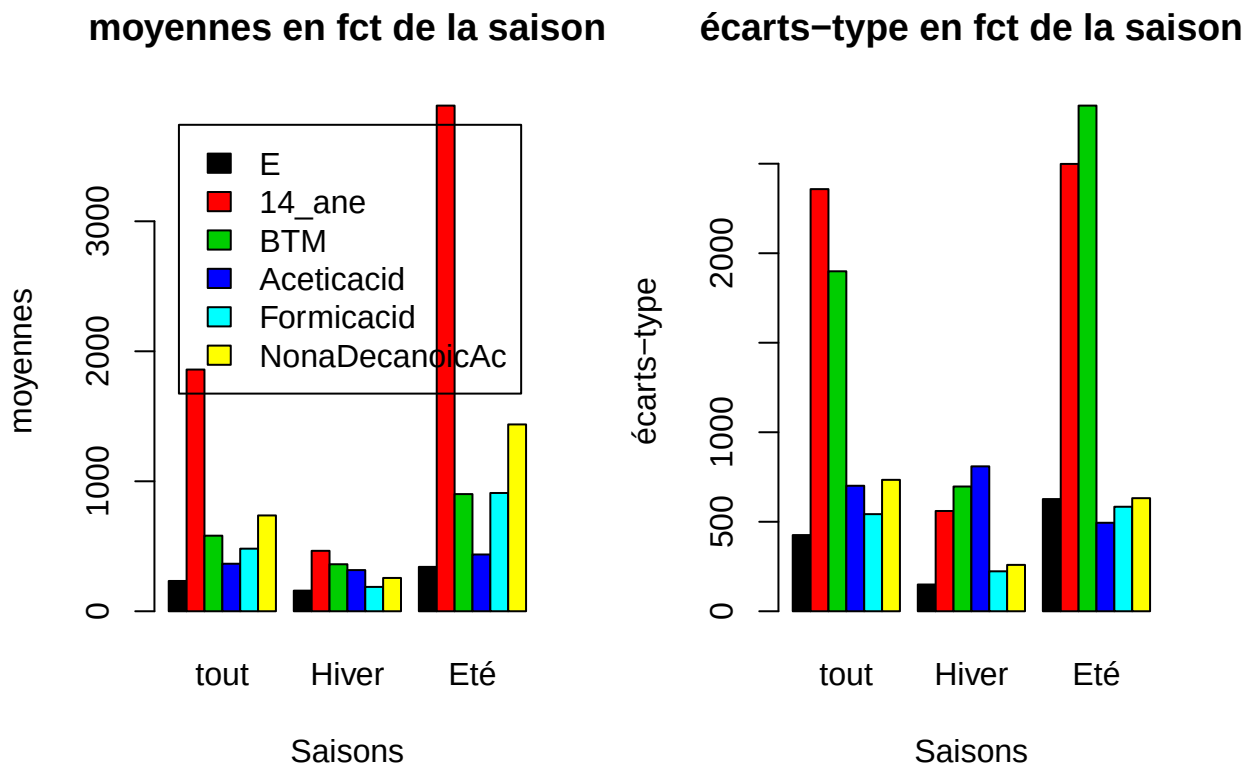
- Ces mêmes variables qui constituent les signatures des périodes étudiées sont des variables marqueurs également (Sauf l'aceticacid vu qu'il est initialement mal-representé)

Etape :3éme

Vous disposez maintenant d'information sur 4 variables de type catégorie : AFD s'intéresse à une (des) variable(s) de type qualitative (période (avant ou après installation du site), saison (été/hiver), localisation, ou le type d'environnement de proximité).

6) Avant de faire une AFD , les statistiques par variable sur les données selon les deux types de modalités hiver/été de la variable saison : y-a-t-il des différences entre les groupes (H/E) ?

Cette question est un retour vers la 1ère étape et notamment la deuxième partie de la première question où nous avons comparé les valeurs moyennes et les écarts types do nos variables quantitatives pendant les deux saisons hiver/été, rappelons le résultat obtenu :



Cette fois-ci on a utilisé 6 variables pour une vision plus générale, les 6 variables qu'on a choisi sont assez significatives vu que c'est elles qui ont representé des corrélations assez différentes sur le corrélogramme, et le résultat est le suivant :

- Quasiment toutes les variables connaissent une hausse très importante de leur valeur moyenne et de leur variabilité en été plutôt qu'en hiver, c'est ces mêmes variables qui sont corrélées entre elles et avec plein d'autres (logique !!)

- Pour la variable Aceticacid qui n'était corrélée à aucune des autres, c'est la seule qui connaît plutôt l'invrese, une lègere hausse de la moyenne et de la variabilité en hiver cette fois.

7) Afin de discriminer au mieux les groupes a) saison (hiver/été) on vous propose de mettre en oeuvre une AFD sur la variable qualitative concernée: qu'observez-vous ? que vaut le critère donné

$$\eta$$

(variance interclasse/ variance totale) pour les axes discriminants Y de valeurs propres

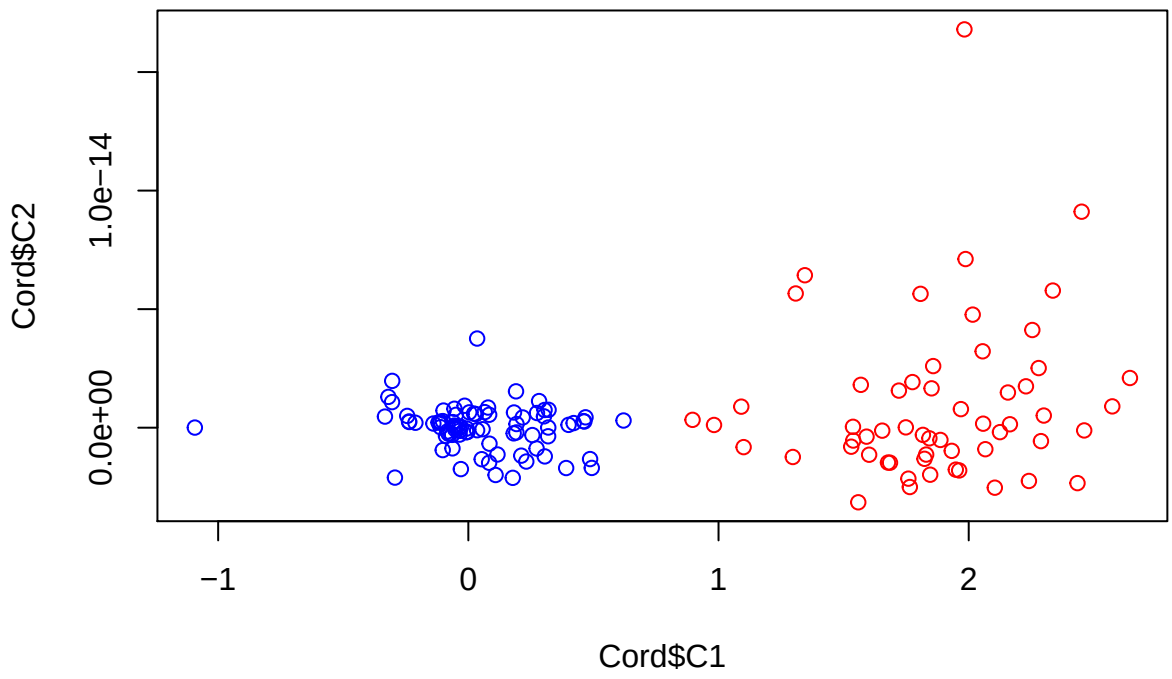
$$\lambda$$

principaux : que retenez-vous pour l'interprétation ?

Afin d'effectuer l'AFD on a repris le programme réalisé en TP3, en voila l'issue :

```
#nombre d'observations
n <- 140

# Calcul de B
B <- (1/n)*(83*(moyhiver-moy)%*%t(moyhiver-moy)+57*(moyete-moy)%*%t(moyete-moy))

# Calcul de W
n1 <- 83
Whiver <- (quantitativehiver-Meanhiver)
W1 <- (1/n1)*t(Whiver)%*%Whiver
#variance intraclasse pour la modalité hiver


n2 <- 57
Wete <- (quantitativeete-Meanete)
W2 <- (1/n2)*t(Wete)%*%Wete
#variance intraclasse pour la modalité été


W <- (1/n)*(n1*W1+n2*W2)
# variance intra classe totale

#V-B-W
#vérification de la formule : c bien vérifié

#Décomposition pour assurer la diogonalisibité
C <- matrix(0,nrow=14,ncol=2)

C[,1] <- sqrt(n1/n)*(moyhiver-moy)
C[,2] <- sqrt(n2/n)*(moyete-moy)

#Matrice à diagonaliser
A <- t(C) %*% solve(V) %*% C

decomp1 <- eigen(A)
values1 <- decomp1$values
vectors1 <- decomp1$vectors
```

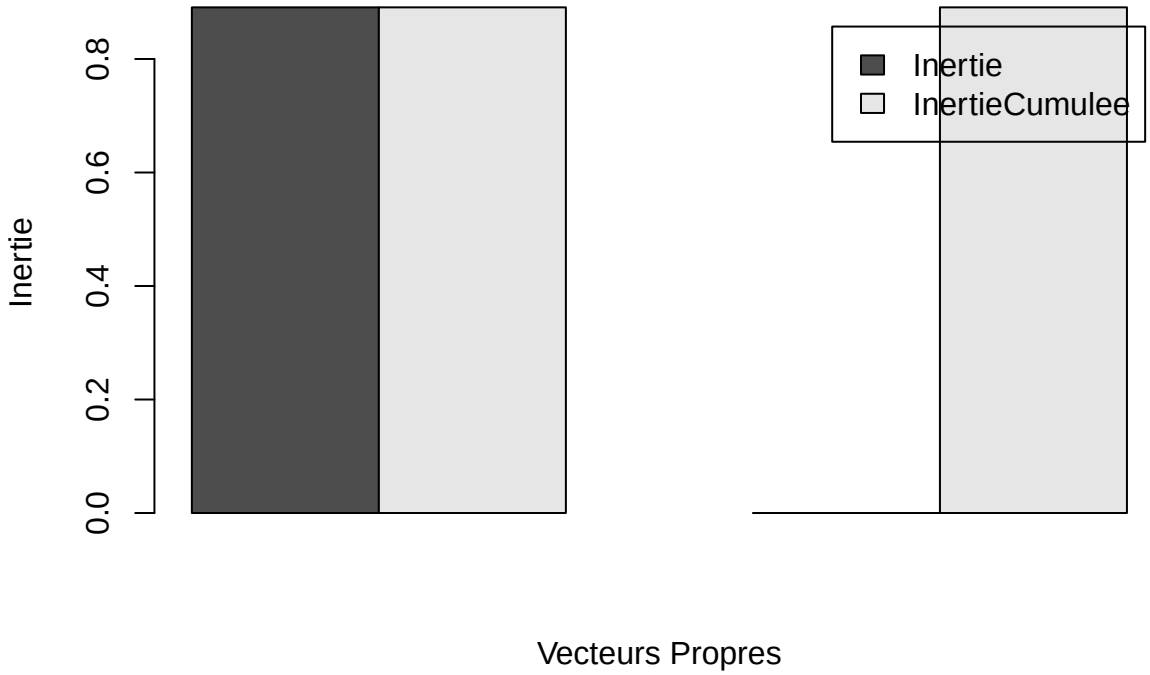
```r
vectors1 <- solve(V) %*% C %*% vectors1

#Les nouvelles coordonnées des individus sur le nouveau plan factoriel
Cord <- matrix(0,nrow=140,ncol=4)
colnames(Cord) <- c("C1","C2","SAISON","binary")

Cord <- as.data.frame(Cord)

Cord[,1:2] <- quantitative %*% vectors1

Cord[,3] <- data$SAISON

#La colonne binaire pour différencier la couleur
Cord[which(Cord$SAISON == "hiver"),4] <- 4
Cord[which(Cord$SAISON == "été"),4] <- 2

plot(Cord$C1,Cord$C2,col=Cord$binary)
legend(1, y=-2e-14, legend=c("hiver", "été"),
       col=c(4, 2), lty=1, cex=0.8)
```



Ce graphe est assez parlant pour la suite, on peut d'emblée dire que l'AFD est bien réussi vu que la séparation est assez parlante et donne lieu à une droite qui distingue les deux modalités de la variable SAISON sans interférences.

Confirmant cela par l'histogramme des inerties et inerties cummulées pour chaque axe :

```r
#Inertie
Inertie<-values1
InertieCumulee<-rep(0,length(values1))
for (i in 1:length(values1)) {
  InertieCumulee[i]<-sum(values1[1:i])
}

barplot(t(as.matrix(cbind(Inertie,InertieCumulee))),
        beside = TRUE,
        legend.text = TRUE,
        ylab = "Inertie",
        xlab = "Vecteurs Propres")
```



On a que preque toute l'inertie est contenue dans le premier axe d'où la précision du graphe obtenu.

Ensuite il nous a été demandé de calculer le critère de projection SCE/SCT, en voici le calcul :

```r
rapp = rep(1, 2)

for (i in 1:2) {
  v <- vectors1[,i]

  SCT <- t(v) %*% V %*% v
  SCR <- t(v) %*% W %*% v
  SCE <- t(v) %*% B %*% v
```

```
  rapp[i] <- SCE/SCT
}

recap <- matrix(0,nrow=2,ncol=2)
colnames(recap) <- c("Axe1","Axe2")
rownames(recap) <- c("Valeur propre","critère éta")

recap[1,] <- values1
recap[2,] <- rapp

print(recap)
```

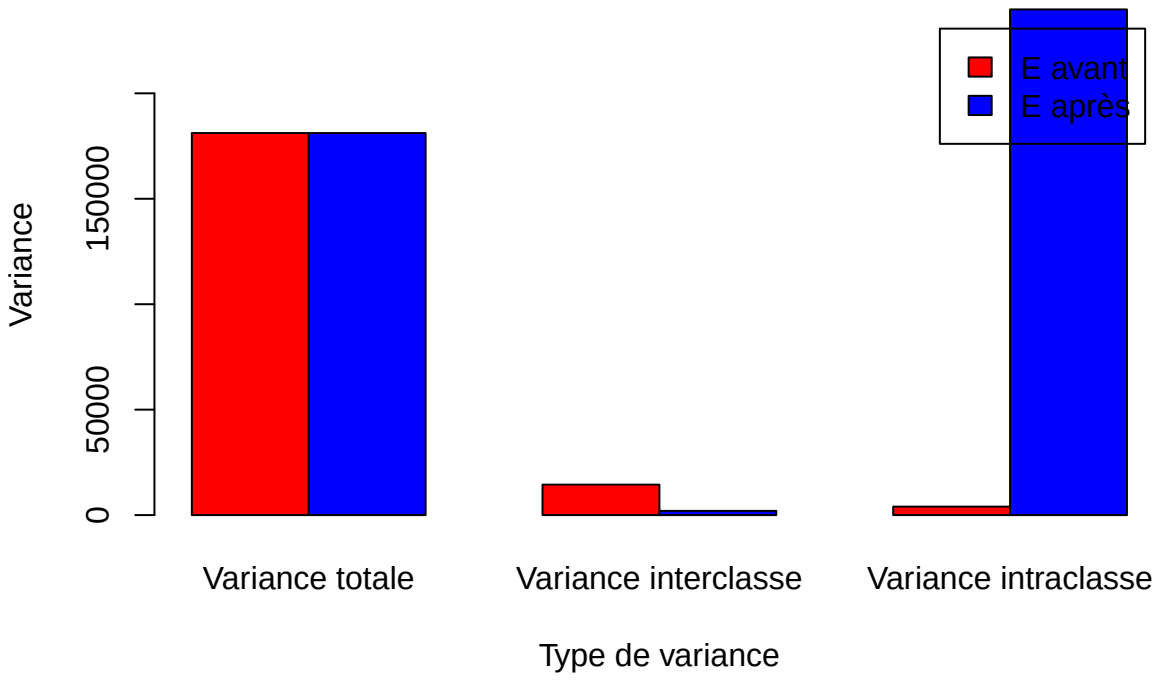
```
##                     Axe1        Axe2
## Valeur propre 0.8909709 1.94289e-16
## critère éta   0.8909709 4.28469e-02
```

Le critère donné est assez proche de 1 ainsi le plan factoriel obtenu est adéquat et contient presque toute l'information (l'inertie) du nuage initial.

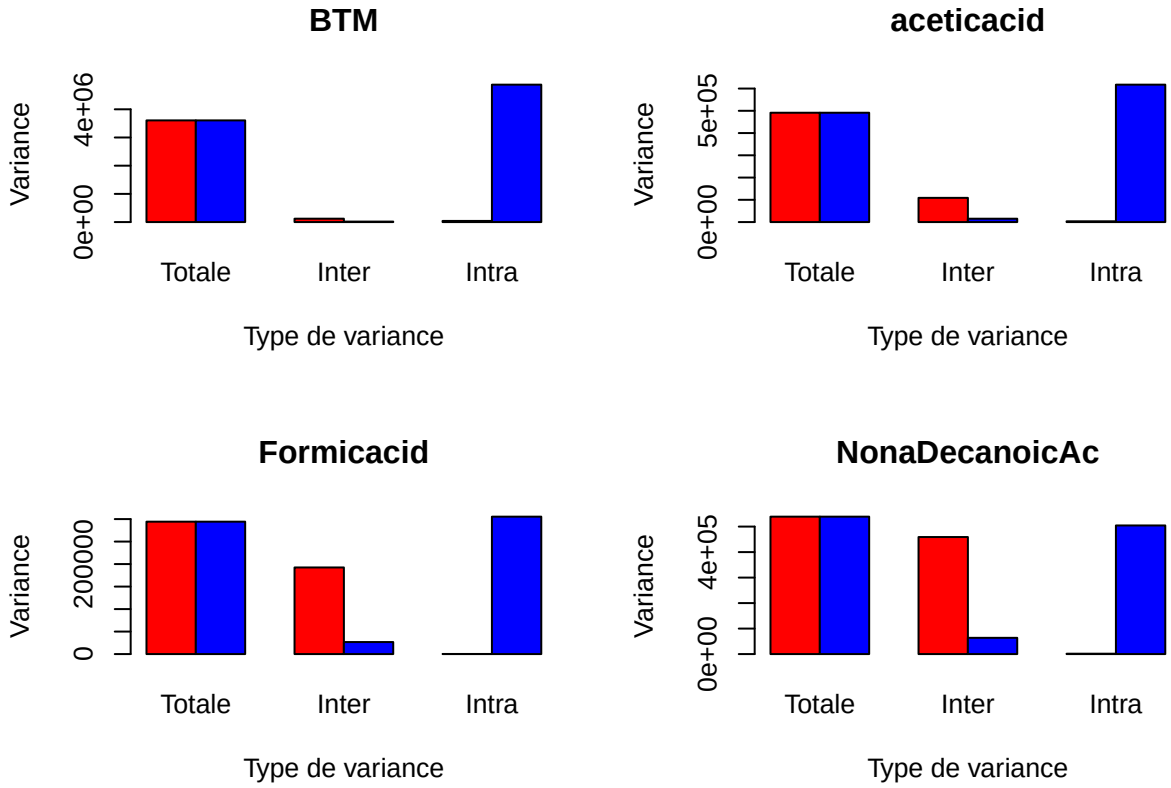
8) Les statistiques sur les données selon les deux types de modalités (avant installation (BF) ou après installation du site de compostage (CA)) montraient-elles des différences entre les groupes en termes de variance totale et variances inter et intraclasses?

Comparons les valeurs des variances totale/inter/intra-classes entre les périodes avant et après ouverture du site à travers les 6 molécules étudiées avant :

```
# Calcul de la variance interclasse
Bav <- (moyav-moy)%*%t(moyav-moy)
Bap <- (moyap-moy)%*%t(moyap-moy)

# Variance intraclasse deja calculée : Vav et Vap

#Variance totale/inter/intra avant et après ouverture du site
Eav <- c(V[3,3],Bav[3,3],Vav[3,3])
Eap <- c(V[3,3],Bap[3,3],Vap[3,3])

M <- as.matrix(cbind(Eav,Eap))
colnames(M) <- c("E avant","E après")
rownames(M) <- c("Variance totale","Variance interclasse","Variance intraclasse")

barplot(t(M),
        beside = TRUE,
        legend.text = TRUE,
        ylab = "Variance",
        xlab = "Type de variance",
        main = "Variance en fct de avant/après ouverture du site de E",
        col = c(2,4))
```




```
#Le reste des variables est calculées de la même façon
```



La différence est flagrante dans ce cas:

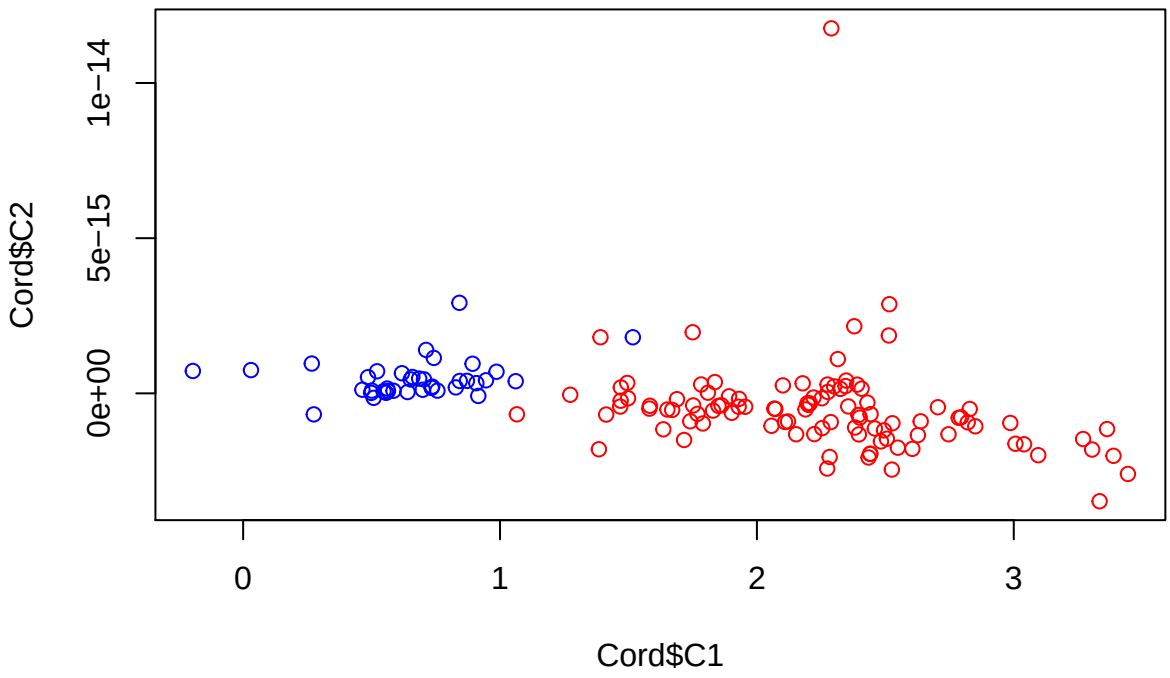
- Avant ouverture du site, on enregistrait des valeurs très proches des différentes molécules contrôlées (variance intraclasse très faible par rapport à la variance interclasse).
- Après ouverture du site, on a obtenu des valeurs très importantes par rapport à ce qu'on avait avant, le fait qui a boosté la variance intraclasse contrairement à celle interclasse restée relativement faible cette fois.

9) Afin d'évaluer la contribution d'un site au niveau de la qualité de l'air ambiant, on vous propose de mettre en oeuvre une AFD sur la variable qualitative période (CA et BF) et donner les résultats de l'AFD (qualité de la réduction, fonction linaire discriminante, critère de et votre interprétation en terme de variables contribuant le plus à la discrimination).

Comme réalisé avant pour la variable SAISON, le code suivant effectue une AFD sur la variable période (CA et BF) :

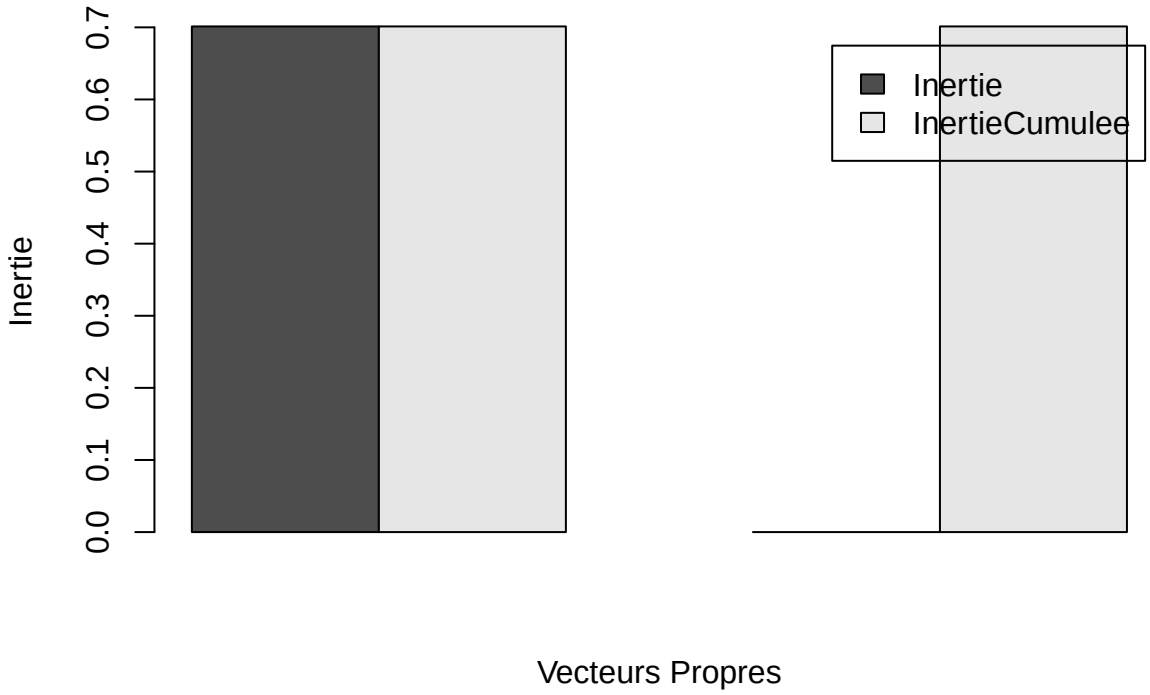
```
# Calcul de B
B <- (1/n)*(38*(moyav-moy)%*%t(moyav-moy)+102*(moyap-moy)%*%t(moyap-moy))

# Calcul de W
n1 <- 38
Wav <- (quantitativeav-Meanav)
W1 <- (1/n1)*t(Wav)%*%Wav
```

```r
#variance intraclasse pour la modalité hiver


n2 <- 102
Wap <- (quantitativeap-Meanap)
W2 <- (1/n2)*t(Wap)%*%Wap
#variance intraclasse pour la modalité été


W <- (1/n)*(n1*W1+n2*W2)
# variance intra classe totale

#V-B-W
#vérification de la formule : c bien vérifié

#Décomposition pour assurer la diogonalisibité
C <- matrix(0,nrow=14,ncol=2)

C[,1] <- sqrt(n1/n)*(moyav-moy)
C[,2] <- sqrt(n2/n)*(moyap-moy)

#Matrice à diagonaliser
A <- t(C) %*% solve(V) %*% C

decomp1 <- eigen(A)
values1 <- decomp1$values
vectors1 <- decomp1$vectors

vectors1 <- solve(V) %*% C %*% vectors1

#Les nouvelles coordonnées des individus sur le nouveau plan factoriel
Cord <- matrix(0,nrow=140,ncol=4)
colnames(Cord) <- c("C1","C2","PERIODE","binary")

Cord <- as.data.frame(Cord)

Cord[,1:2] <- quantitative %*% vectors1

Cord[,3] <- data$Campagne

#La colonne binaire pour différencier la couleur
Cord[which(substr(Cord$PERIODE, 1, 2) == "BF"),4] <- 4
Cord[which(substr(Cord$PERIODE, 1, 2) == "CA"),4] <- 2

plot(Cord$C1,Cord$C2,col=Cord$binary)
legend(1, y=-8e-15, legend=c("BF", "CA"),
       col=c(4, 2), lty=1, cex=0.8)
```



L'histogramme d'inertie est le suivant :

```
#Inertie
Inertie<-values1
InertieCumulee<-rep(0,length(values1))
for (i in 1:length(values1)) {
  InertieCumulee[i]<-sum(values1[1:i])
}

barplot(t(as.matrix(cbind(Inertie,InertieCumulee))),
        beside = TRUE,
        legend.text = TRUE,
        ylab = "Inertie",
        xlab = "Vecteurs Propres")
```



Toujours la même remarque, notre premier axe principale contient presque toute l'information, confirmons cela par le calcul de la qualité de réduction :

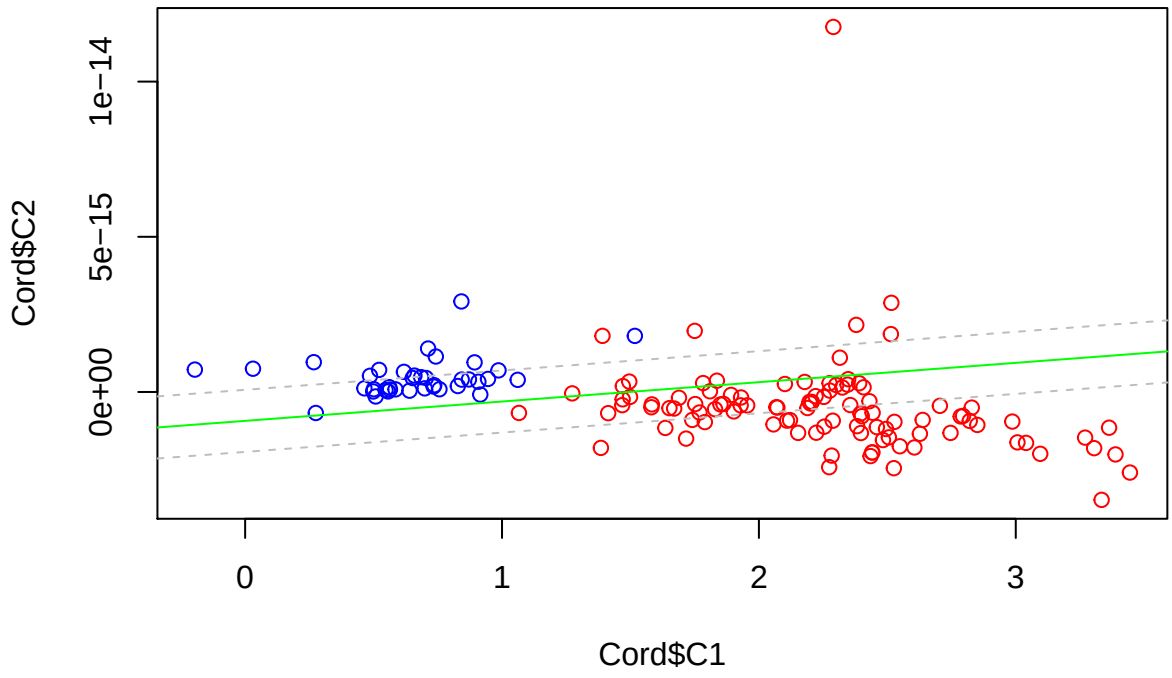
```
Qual<- function(values,k){
  A<-sum(values)
  return(values[k]/A)
}

Quality <- rep(1,2)

Quality[1] <- Qual(values1,1)
Quality[2] <- Qual(values1,2)

print(Quality)
```

```
## [1] 1.000e+00 3.958e-16
```

Traçons maintenant la droite discriminante entre les deux périodes :

```
centroide1 <- colMeans(Cord[which(Cord$binary == 4),1:2])
centroide2 <- colMeans(Cord[which(Cord$binary == 2),1:2])

m  <- (centroide1[2]-centroide2[2])/(centroide2[1]-centroide1[1])
p <- (centroide2[2]*centroide2[1]-centroide1[2]*centroide1[1])/(centroide2[1]-centroide1[1])
```

```r
plot(Cord$C1,Cord$C2,col=Cord$binary)
abline(a = p,b = m, lty=1, col = "green")
abline(a = p+1e-15,b = m, lty=2, col = "gray")
abline(a = p-1e-15,b = m, lty=2, col = "gray")
legend(1, y=-8e-15, legend=c("BF", "CA"),
       col=c(4, 2), lty=1, cex=0.8)
```



Notre droite discriminante est un peu biaisé à cause des outliers, d'où l'imperfection dans la séparation.

Calculons maintenant le critère de réduction pour cette AFD :

```r
rapp = rep(1, 2)

for (i in 1:2) {
  v <- vectors1[,i]

  SCT <- t(v) %*% V %*% v
  SCR <- t(v) %*% W %*% v
  SCE <- t(v) %*% B %*% v

  rapp[i] <- SCE/SCT
}

recap <- matrix(0,nrow=2,ncol=2)
colnames(recap) <- c("Axe1","Axe2")
rownames(recap) <- c("Valeur propre","critère éta")
```

```
recap[1,] <- values1
recap[2,] <- rapp

print(recap)
```

```
##                     Axe1         Axe2
## Valeur propre 0.7012525 2.775558e-16
## critère éta   0.7012525 9.076506e-02
```

10) Vous avez alors deux résultats ACP et AFD sur les données : les deux réductions ne sont pas faites selon le même critère mais pouvez-vous conclure sur l'effet sur de l'activité du site sur l'environnement ou non ?

On a remarqué pour les deux méthodes une obtention de deux classes(avant/après l'ouverture) séparées en projetant sur le plan factoriel.

Ceci nous a permis de dire qu'il y a une différence flagrante de concentration d'éléments chimiques entre les deux périodes, et qui nous pousse par conséquent à dire qu'une pollution à pu avoir lieu (sachant qu'on a pas de seuil de pollution qui pourra nous permettre de trancher).

Etape :4éme

Il existe une généralisation de l'AFDM qui intègre ACP et AFC pour plusieurs var. quantitatives et qualitatives, appelée Analyse factorielle des données mixtes.

11) Préparer les données au format attenus par le package FactoMineR

Afin de réaliser l'AFDM sur toutes les données, on va enlever l'observation dont le TYPE est manquant ('?'), et mettre en data frame :

```
index <- which(data$TYPE == "?")
data <- data[-index,]
```

12) Mettre en oeuvre cette méthode à partir des packages disponibles pour avoir une réduction de dimension sur l'espace de 14 var. quantitatives et des 4 variable qualitatives

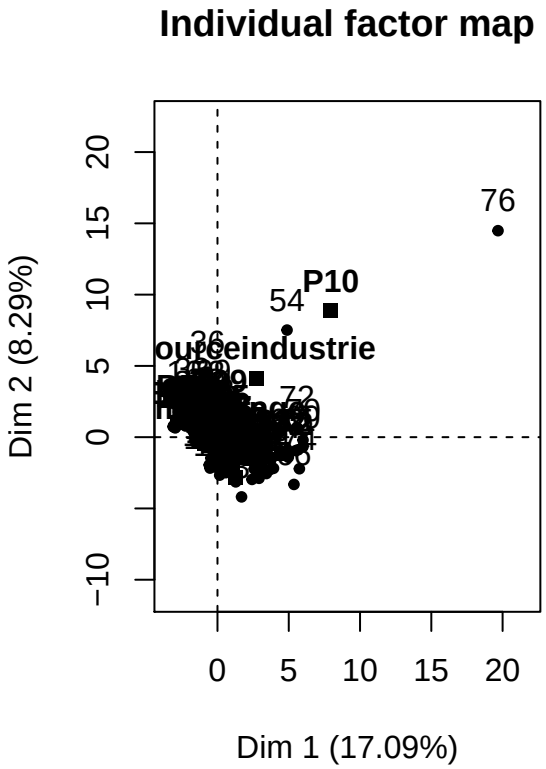
L'aFDM mise en oeuvre :

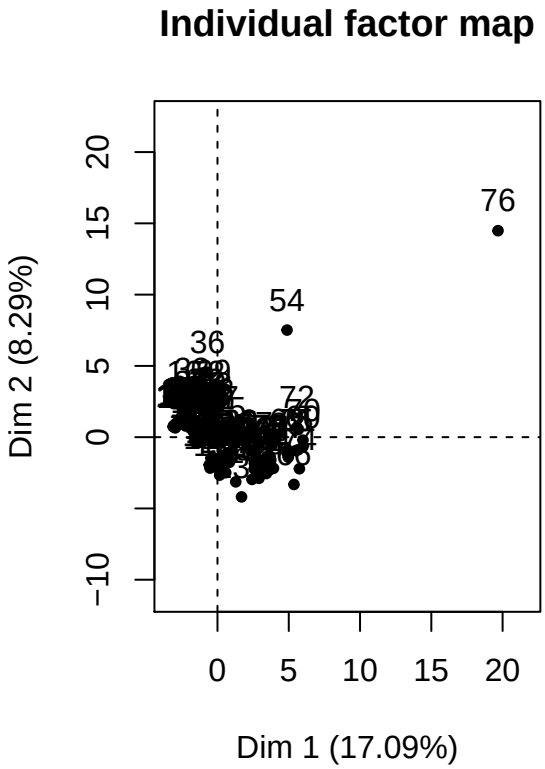
```
library("FactoMineR")
library("factoextra")
```

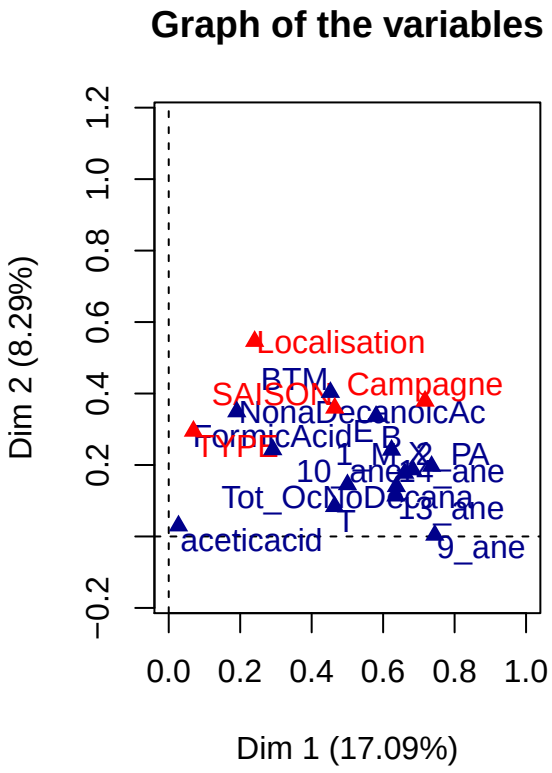
```
## Loading required package: ggplot2
```

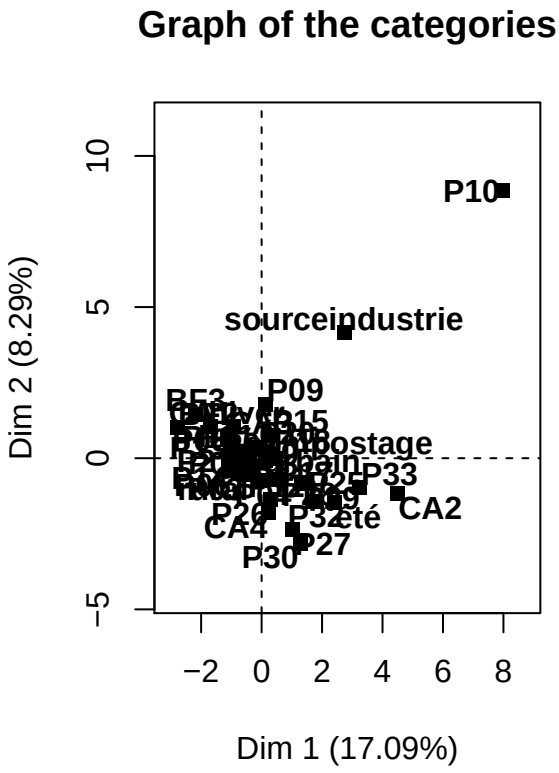
```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```
par(mfrow=c(1,2))
AFDM <- FAMD (data, ncp = 2, sup.var = NULL, ind.sup = NULL, graph = TRUE)
```

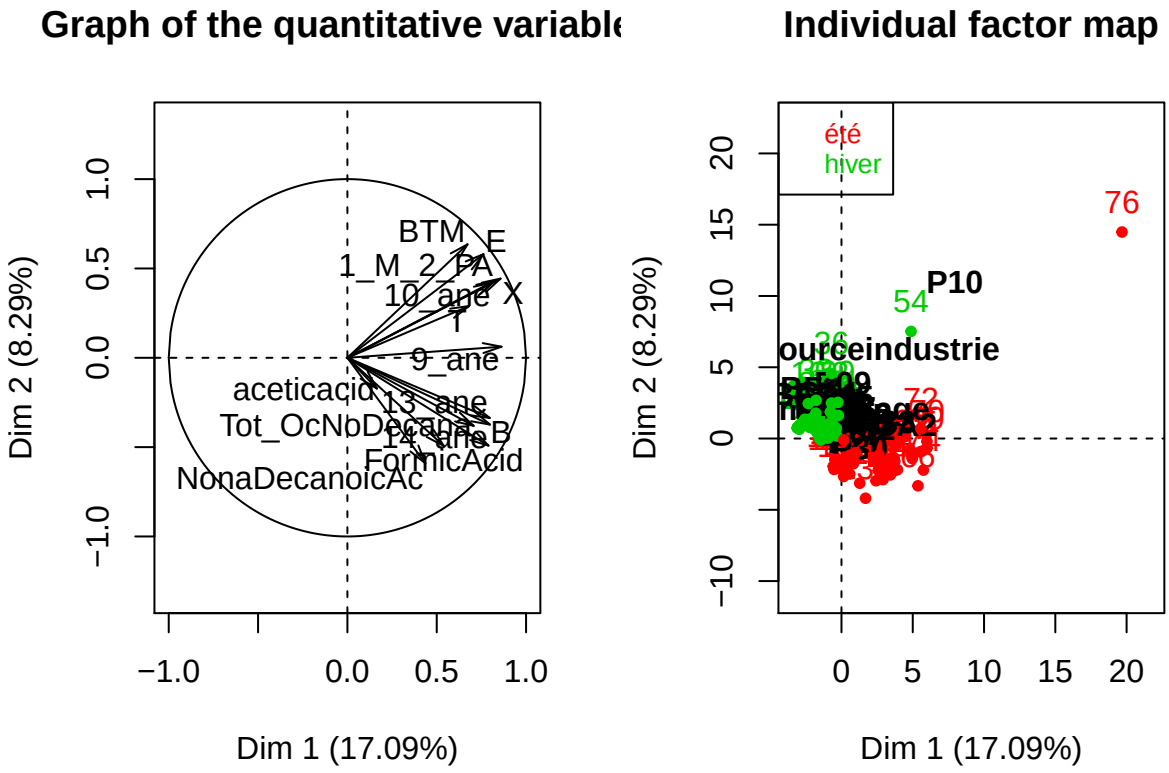




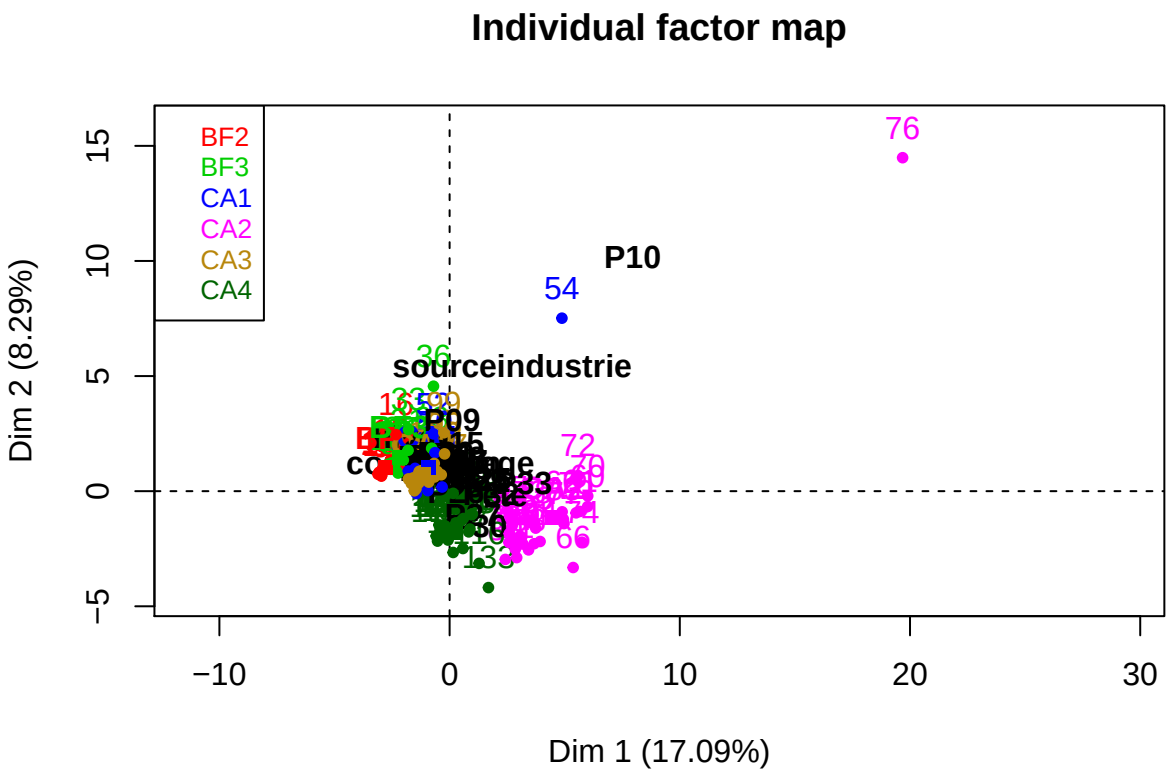





```
#Graphe des individus en fonction de la saison
plot(AFDM,choix ="ind",habillage = 16)
```




```
par(mfrow=c(1,1))
#Graphe des individus en fonction de la Campagne
plot(AFDM,choix ="ind",habillage = 17)
```



13) Interprétation des résultats obtenus intégrant ces 18 var. au total

Différentes conclusions peuvent être tirées depuis les graphes obtenus :

- Depuis le graphe des catégories, Le nuage n'est pas assez clair pour déduire des correspondances entre les différentes catégories des différentes variables, mais on peut voir que la modalité "été" est assez proche de la Campagne "CA2" (deuxième campagne après ouverture site), et de même hiver est proche des Campagnes BF2, BF3 (avant ouverture du site) ce qui est assez logique vu les traitement statistiques réalisés dans les étapes précédentes.
- La graphe des variables par contre est assez parlant, et permet de confirmer nos études préalables (ACP et AFD). Ce graphe represente la contribution des différentes variables aux deux axes principaux, on peut voir que l'aceticacid est assez proche de l'origine ainsi elle ne contribue presque pas aux axes factoriels résultat retrouvé en étape 2. La variable Campagne est assez éliognée sur la première bissectrice donc elle a une bonne contribution aux deux axes ce qui est assez significatifs pour nous permettre de conclure sur l'influence du site industriel sur l'environnement.
- Le cercle de corrélation des variables quantitatives ne nous donnent pas plus d'informations de ce qu'on a déja, il confirme juste nos constats jusque là.
- Enfin les deux plots qu'on a ajouté, le plot des individus en fonction de la saison permet de voir que le nuage de points obtenu est séparé significativement en fonction des modalités de la variable saison, ce qui est en harmonie avec bos résultats de l'AFD (étape 3)
- Pareil que pour le plot des individus en fonction des campagnes (vu aussi en fonction des périodes avant/après l'ouverture du site), il est assez séparateur en terme de la période ce qui va nous permettre de conclure sur l'influence du site industriel.

14) Cela vous apporte-t-il des éléments complémentaires à la première ACP ?

Par rapport à l'ACP réalisée sur les variables quantitatives, l'AFDM nous a apporté en supplémentaire tout ce qui est en relation avec les variables qualitatives :

- La contribution des variables qualitatives aux axes factoriels, ce qui est assez important pour juger la variabilité apportée par les différentes modalités de ces variables.
- Le graphe des catégories qui nous permet de conclure sur les correspondances qui existent entre les différentes modalités des différentes variables.

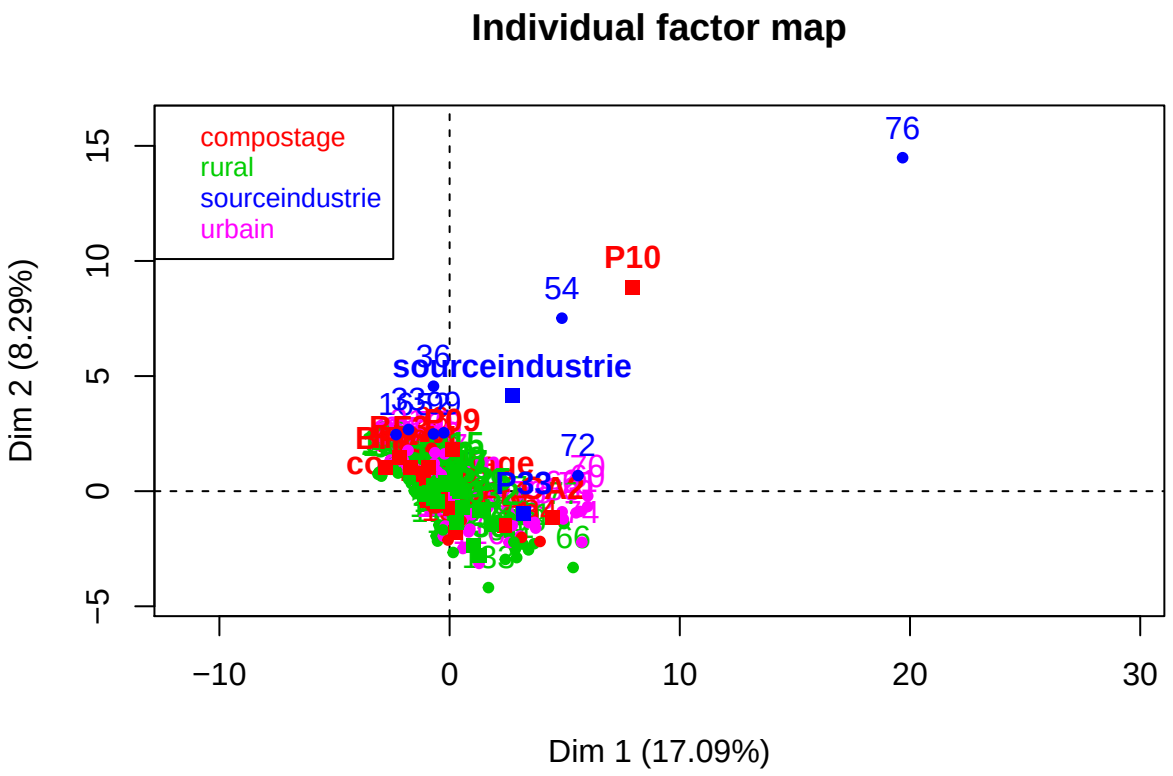
15) Que pourriez-vous suggérer pour établir les éléments de comparaison entre groupes campagne hiver/été, groupe avant et après installation industrielle, groupe en fonction de la localisation du point de mesure (urbain, rural, site industriel, sur site de compostage).

Le graphe des individus en fonction des modalités de la variable qualitative qu'on veut étudier est le meileur moyen de comparer entre les différents groupes.

Pour la variable SAISON et CAMPAGNE (vu aussi come période avant/après ouverture du site) ce graphe a déjà était fait dans la question 12.

Pour la variable LOCALISATION voila le graphe correspondant :

```
#Graphe des individus en fonction de la localisation
plot(AFDM,choix ="ind",habillage = 15)
```



16) Vous avez alors deux résultats ACP et AFDM sur les données : les deux réductions ne sont pas faites selon le même critère mais pouvez-vous conclure sur l'effet sur de l'activité du site sur l'environnement ou non ?

Sans prendre en compte le critère considéré pour les deux analyses, les deux nous ont ammené à une séparation plutôt précise et radicale du nuage du points en fonction de la période (avant/après ouverture du site industriel) et donc cela est suffisant pour pouvoir conclure sur l'impact du site industriel sur l'environement vu que les concentrations des molécules étudiées ont significativement changé.

17) Questions complémentaires (compter en plus si réaliser)

On s'intéresse maintenant à la signature de profils i de concentration de chaque individu (i point échantillonné, parmi les n) : une première estimation faite par les chimistes est d'attribuer un type (rural, urbain, compostage, site industriel) à chaque individu : pouvez-vous à partir d'une statistique de type AFD sur cette variable qualitative 'type' proposer une réduction et une analyse de la discrimination des groupes :

pensez- vous que ce regroupement empirique initiale est cohérente avec la localisation effective du point de mesure dans son environnement immédiat ?

Afin de ne pas perdre le temps à reécrire l'AFD qu'on a fait dans l'étape 3, on va utiliser une fonction prédifinie de R.

pour ajouter l'impact des localisations effectives des individus on va rajouter aussi la variable 'Localisation' dans l'explication de la variable 'TYPE' dans notre AFD, c'est seulement comme cela qu'on pourrait voir s'il existe un regroupement d'individus selon les modalités de la variable 'TYPE' qui est cohérent avec les localisations effectives des points de mesure.

En voila le code correspondant

```
#Imporatations nécessaires
library(tidyverse)
library(caret)
library(MASS)
```

```
theme_set(theme_classic())

set.seed(123)
model <- lda(TYPE~B+T+E+X+`9_ane`+`10_ane`+`13_ane`+`14_ane`+`1_M_2_PA`+BTM+FormicAcid+
               aceticacid+NonaDecanoicAc+`Tot_OcNoDecana`+Localisation, data = data)
```

```
## Warning in lda.default(x, grouping, ...): variables are collinear
```

```
model
```

```
## Call:
## lda(TYPE ~ B + T + E + X + `9_ane` + `10_ane` + `13_ane` + `14_ane` +
##     `1_M_2_PA` + BTM + FormicAcid + aceticacid + NonaDecanoicAc +
##     Tot_OcNoDecana + Localisation, data = data)
##
## Prior probabilities of groups:
##      compostage           rural sourceindustrie          urbain
```

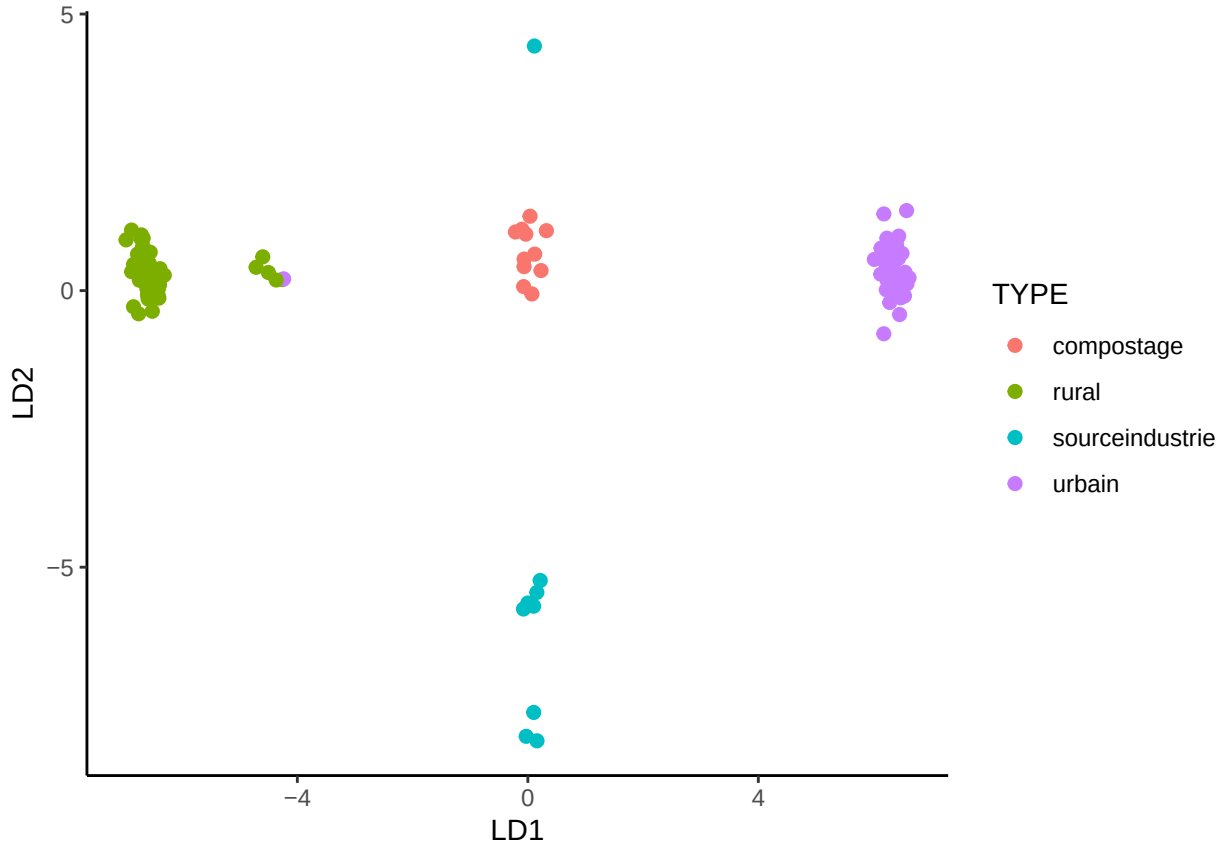
```
##      0.07971014      0.42028986      0.06521739      0.43478261
##
## Group means:
##                        B        T        E         X   `9_ane` `10_ane`
## compostage      51.55895 208.3958 158.7788  634.4140 109.25457 244.3136
## rural           56.07831 171.4796 155.2816  624.4715  84.30452 188.8085
## sourceindustrie 53.01937 302.9093 861.5854 2496.8124 175.19973 770.5307
## urbain          57.06639 230.3482 230.3248  937.0685 112.76325 274.2752
##                  `13_ane` `14_ane` `1_M_2_PA`       BTM FormicAcid
## compostage      1041.6656 1669.487   261.4287  336.0570   537.5859
## rural            946.1124 1583.783   211.0149  300.1735   493.2734
## sourceindustrie 1027.7899 2138.873  1284.5329 3603.1616   404.1339
## urbain          1052.6471 2149.304   277.3798  441.7506   468.9639
##                 aceticacid NonaDecanoicAc Tot_OcNoDecana LocalisationP02
## compostage        527.7090       695.2390       414.7427       0.5454545
## rural             355.7253       823.1005       718.8003       0.0000000
## sourceindustrie   223.9699       459.2908       528.7826       0.0000000
## urbain            326.2676       716.4428       641.6696       0.0000000
##                 LocalisationP03 LocalisationP04 LocalisationP05
## compostage            0.0000000      0.00000000             0.0
## rural                 0.1034483      0.06896552             0.0
## sourceindustrie       0.0000000      0.00000000             0.0
## urbain                0.0000000      0.00000000             0.1
##                 LocalisationP06 LocalisationP07 LocalisationP08
## compostage            0.0000000       0.0000000       0.0000000
## rural                 0.1034483       0.1034483       0.1034483
## sourceindustrie       0.0000000       0.0000000       0.0000000
## urbain                0.0000000       0.0000000       0.0000000
##                 LocalisationP09 LocalisationP10 LocalisationP11
## compostage            0.0000000       0.0000000      0.00000000
## rural                 0.0000000       0.0000000      0.00000000
## sourceindustrie       0.6666667       0.3333333      0.00000000
## urbain                0.0000000       0.0000000      0.08333333
##                 LocalisationP12 LocalisationP13 LocalisationP14
## compostage            0.0000000       0.0000000      0.00000000
## rural                 0.0862069       0.1034483      0.08620690
## sourceindustrie       0.0000000       0.0000000      0.00000000
## urbain                0.0000000       0.0000000      0.01666667
##                 LocalisationP15 LocalisationP16 LocalisationP17
## compostage                  0.0             0.0      0.00000000
## rural                       0.0             0.0      0.00000000
## sourceindustrie             0.0             0.0      0.00000000
## urbain                      0.1             0.1      0.08333333
##                 LocalisationP18 LocalisationP19 LocalisationP20
## compostage           0.00000000       0.0000000             0.0
## rural                0.00000000       0.1034483             0.0
## sourceindustrie      0.00000000       0.0000000             0.0
## urbain               0.08333333       0.0000000             0.1
##                 LocalisationP21 LocalisationP25 LocalisationP26
## compostage                  0.0            0.00            0.00
## rural                       0.0            0.00            0.00
## sourceindustrie             0.0            0.00            0.00
## urbain                      0.1            0.05            0.05
##                 LocalisationP27 LocalisationP28 LocalisationP29
```

```
## compostage           0.00000000            0.00      0.00000000
## rural                0.05172414            0.00      0.05172414
## sourceindustrie      0.00000000            0.00      0.00000000
## urbain               0.00000000            0.05      0.00000000
##                 LocalisationP30 LocalisationP32 LocalisationP33
## compostage           0.00000000            0.00      0.00000000
## rural                0.03448276            0.00      0.00000000
## sourceindustrie      0.00000000            0.00      0.00000000
## urbain               0.00000000            0.05      0.03333333
##
## Coefficients of linear discriminants:
##                            LD1           LD2           LD3
## B                -7.223791e-03 -4.797799e-03 -1.451327e-03
## T                -3.086710e-04 -1.877066e-03  7.489588e-04
## E                 3.925349e-04  1.747666e-03 -4.985434e-04
## X                 1.198529e-04 -5.284691e-05  4.599457e-04
## `9_ane`          -3.696944e-04 -1.721305e-03 -2.207416e-03
## `10_ane`          4.453801e-05  5.404227e-03 -5.992366e-04
## `13_ane`          7.801671e-05  1.685855e-04 -2.260251e-04
## `14_ane`         -1.759947e-05 -2.046678e-04  7.999962e-05
## `1_M_2_PA`        1.104991e-04 -1.054882e-03 -1.097480e-03
## BTM              -1.586680e-04 -1.924258e-03  2.190483e-04
## FormicAcid        4.049800e-05 -2.887109e-05  1.101975e-04
## aceticacid       -8.705610e-06  1.721111e-04 -3.400574e-05
## NonaDecanoicAc   -7.411381e-06  2.459779e-04  6.456527e-06
## Tot_OcNoDecana    1.610557e-06  2.407733e-04  6.148505e-04
## LocalisationP02 -1.622480e-02  5.607190e-01 -6.749143e+00
## LocalisationP03 -6.618880e+00 -3.036439e-01 -2.489277e-01
## LocalisationP04 -6.618168e+00 -1.570010e-01 -1.283965e-01
## LocalisationP05  6.345942e+00  8.901714e-02  8.015923e-02
## LocalisationP06 -6.673313e+00 -1.140588e-01  1.464415e-01
## LocalisationP07 -6.601635e+00 -1.093213e-01 -5.575778e-02
## LocalisationP08 -6.699850e+00 -1.307937e-02  2.398861e-01
## LocalisationP09  3.149681e-02 -5.302399e+00 -5.103564e-01
## LocalisationP10 -3.149681e-02  5.302399e+00  5.103564e-01
## LocalisationP11  6.310565e+00 -1.085580e-01  3.454682e-02
## LocalisationP12 -6.586775e+00 -3.372673e-02 -1.440283e-01
## LocalisationP13 -6.657771e+00  4.892850e-02  2.001083e-02
## LocalisationP14 -4.494801e+00 -1.957962e-02  1.776266e-02
## LocalisationP15  6.246625e+00  2.331523e-01 -2.041882e-01
## LocalisationP16  6.316507e+00 -1.353766e-01  8.655912e-02
## LocalisationP17  6.354944e+00  6.228803e-02 -2.045133e-02
## LocalisationP18  6.317882e+00  1.693697e-01  3.448150e-01
## LocalisationP19 -6.572331e+00  1.080776e-01 -2.891087e-02
## LocalisationP20  6.360266e+00  1.276475e-01  5.979004e-02
## LocalisationP21  6.369212e+00 -1.495520e-01 -1.448340e-02
## LocalisationP25  6.408705e+00  6.178143e-02  1.471085e-01
## LocalisationP26  6.484373e+00  3.825479e-02  4.979168e-02
## LocalisationP27 -6.531136e+00 -7.151853e-01 -2.216663e-01
## LocalisationP28  6.431800e+00  4.969299e-01  5.489279e-02
## LocalisationP29 -6.550006e+00  5.395171e-01  2.295371e-01
## LocalisationP30 -6.550505e+00  5.405249e-01 -2.179595e-01
## LocalisationP32  6.492126e+00 -1.199693e-01 -3.949375e-01
## LocalisationP33  6.388789e+00 -2.326401e-01 -1.213274e-01
```

```
##
## Proportion of trace:
##    LD1    LD2    LD3
## 0.9171 0.0519 0.0310
```

On peut voir que l'inertie des deux premiers axes est très proche de 1 donc on prévoit une bonne séparation par la suite, vérifions cela :

```
lda.data <- cbind(data, predict(model)$x)
ggplot(lda.data, aes(LD1, LD2)) +
  geom_point(aes(color = TYPE), size = 2)
```



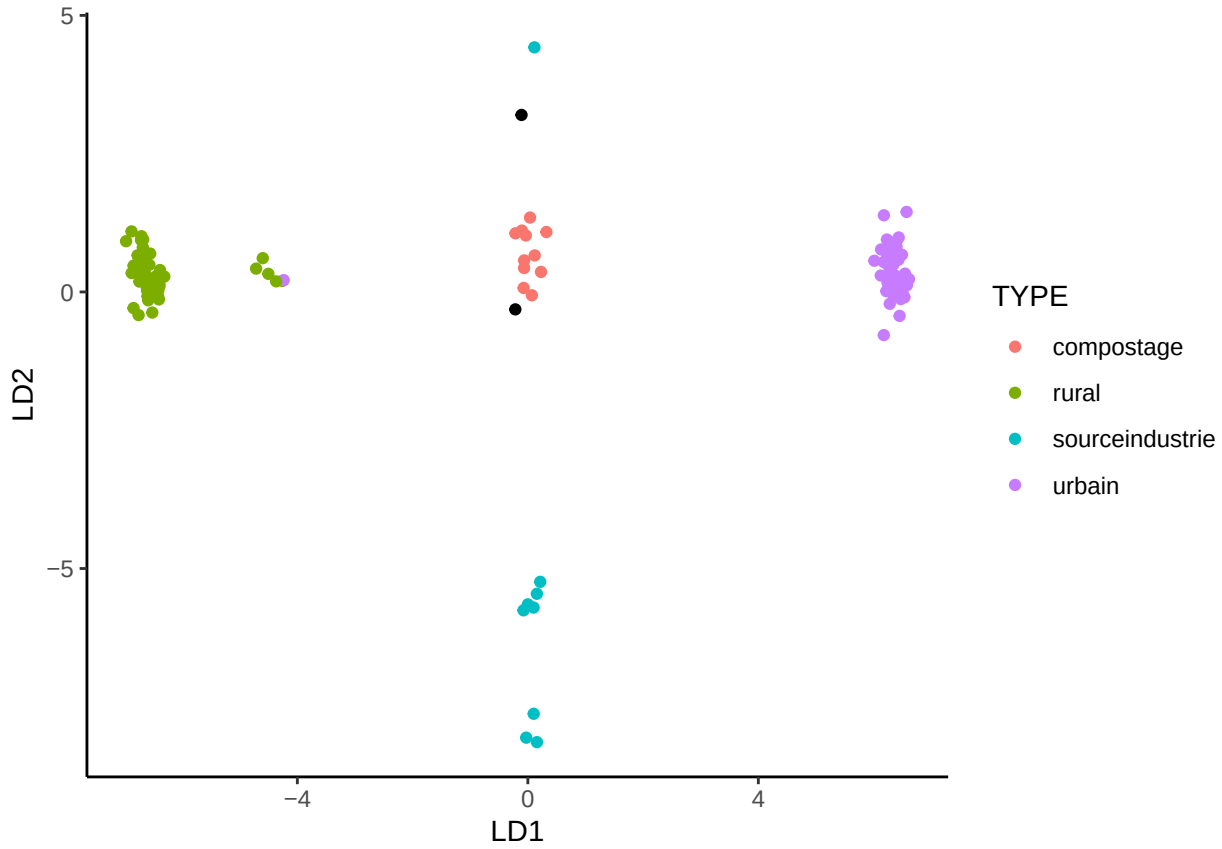
Ainsi on obtient un regroupement assez clair des individus selon les 4 modalités de la variable TYPE, ce qui va nous permettre par la suite de prédire l'emplacement des individus non typés.

Cette explication de la variable 'TYPE' avec les variables quantitatives et la variable 'Localisation' nous permet de dire que le regroupement obtenu est cohérent avec la localisation effective des points de mesure dans leur environnement immédiat.

18) Enfin un individu n'est pas typé par son environnement (?) pouvez l'extraire et refaire l'AFD et faire la prévision d'appartenance à sa classe en utilisant AFD en mode prédictif ?

On a remarqué que deux individus et non pas un seul n'ont pas été typés, et afin de les placer dans le graphe précédent et conclure sur leur appartenance on va calculer leurs coordonnées dans le plan factoriel obtenu par l'AFD qu'on vient de réaliser :

```r
#On va reimporter nos données pour remettre tous les individus dedant

data <- read_excel("TP4_covC1234_DS19_20.xlsx")

data <- data[,2:19]

moy <- colMeans(data[,1:14])
for (i in 1:14) {
  data[which(data[,i] == 0),i] <- moy[i]
}

# Maintenant on va éxtraire les individus concernés par la classification

index <- which(data$TYPE == "?")

individu1 <- as.matrix(data[index[1],1:14])
individu2 <- as.matrix(data[index[2],1:14])

#Puisque on a utilisé la variable Localisation aussi on va rajouter des zéros dans les indices correspo

Local1 <- rep(0,28)
Local1[9] <- 1

#le deuxième individu est seul dans la localisation P22 ainsi cettt modalités n'est pas considéré par l
Local2 <- rep(0,28)

Cord1 <- c(individu1,Local1)
Cord2 <- c(individu2,Local2)

#Afin d'extraire les axes principaux
Axes <- model$scaling

#Calcul des coordonnées des deux points
New <- rbind(t(Cord1),t(Cord2)) %*% cbind(Axes[,1],Axes[,2])

p = ggplot() +
  geom_point(data = lda.data, aes(x = LD1, y = LD2,color = TYPE))+
  geom_point(data = as.data.frame(New), aes(x = New[,1], y = New[,2]))

print(p)
```



Le premier individu avec un '?' est localisé à P10, qui correspond à la 'sourceindustrie' d'après les autres individus situés dans le même endroit, est donc c'est lui qui est décalé en haut dans notre graphe.

Quant au deuxième, il est localisé à P22 où il est le seul, ainsi on va prédire sa classe à partir de son emplacement. On peut voir qu'il est situé en plein milieu du nuage des 'compostages' et donc on va qualifié son type de 'Compostage'.

Pour vérifier ce résultat on a remarqué dans l'image des localisations (voir l'image en dessous) que les endroits typés de 'Compostage' étaient P01 et P02, et il y a un troisième spot non nommé collé à ces deux spots de compostage ainsi logiquement il serait typé de 'Compostage' aussi, ainsi c'est l'individu cherché et donc ce troisième spot en compostage est P22.

Figure 1: Carte