Mise en œuvre de la classification (TP Splus)

Classification automatique sur un tableau présence absence

```
library(multidim,first=T)
```

Acquisition et transformation des données

```
       nasa1a_read.table("beer301.prn", header = T)
       nasa1b_read.table("beer302.prn", header = T)
       nasa1_cbind(nasa1a,nasa1b)
       etiq_c(scan("souch",what=character()))

carac <- names(nasa1)
nasa <- as.matrix(nasa1)
cat("\n matrices initiales :\n")
print(nasa[1:15,1:10])
```

```
 matrices initiales :
    C4 C5 C6 C10 C12 C13 C18 C21 C22 C23
 B2  0  0  0   1   1   1   0   1   0   0
 B4  0  0  0   1   1   0   1   1   0   0
 B5  0  0  0   1   1   0   0   0   0   0
 B7  0  0  0   1   1   0   0   1   1   0
 B8  0  0  0   1   1   0   0   0   0   1
B10  0  0  0   1   1   0   0   1   0   0
B11  0  0  0   1   1   0   0   0   0   0
B12  0  0  0   1   1   0   0   1   0   0
B14  0  0  0   1   1   0   0   1   0   0
B15  0  0  0   1   1   0   0   0   0   0
B16  0  0  0   1   1   0   0   0   0   0
B17  1  0  0   1   1   0   0   0   0   0
B18  0  0  0   1   1   0   0   1   0   0
B21  0  0  0   1   1   0   0   1   0   0
B25  1  0  0   1   1   0   0   1   0   0
```

```
        long_length(nasa[,1])
        larg_length(nasa[1,])
        tota_long*larg
cat("\n nombre de souches :\t")
print(long)
print(etiq)
```

```
 nombre de souches : [1] 72
 [1] "B2"  "B4"  "B5"  "B7"  "B8"  "B10" "B11" "B12" "B14" "B15" "B16" "B17"
[13] "B18" "B21" "B25" "B26" "B27" "B29" "T32" "T33" "T34" "T35" "T39" "B9"
[25] "B19" "T37" "T42" "T43" "T44" "T45" "T46" "T47" "T69" "T70" "R50" "R51"
[37] "R53" "R56" "R58" "B13" "T41" "R52" "R55" "H1"  "H2"  "H3"  "H4"  "H5"
[49] "H6"  "H7"  "H8"  "H9"  "H10" "B20" "B30" "T40" "T38" "R59" "B60" "B61"
[61] "B62" "B64" "B65" "R54" "B63" "B67" "T36" "B72" "B75" "B76" "B77" "B78"
```

```
cat("\n nombre de caracteres :\t")
print(larg)
print(carac)
```

```
 nombre de caracteres : [1] 31
 [1] "C4"  "C5"  "C6"  "C10" "C12" "C13" "C18" "C21" "C22" "C23" "C24" "C25"
[13] "C26" "C27" "C28" "C30" "C31" "C32" "C33" "C34" "C35" "C36" "C37" "C39"
[25] "C40" "E3"  "E4"  "E7"  "E8"  "E10" "E11"
```

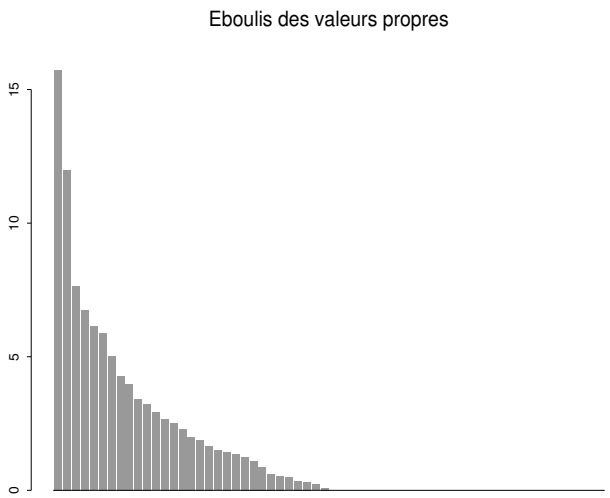

Décomposition en matrice indicatrice

```
        nasa.Z_matind((nasa+1))
cat("\n matrice des indicatrices :\n")
        print(nasa.Z[1:15,1:10])
```

```
 matrice des indicatrices :
    C41 C42 C51 C52 C61 C62 C101 C102 C121 C122
 B2   1   0   1   0   1   0    0    1    0    1
 B4   1   0   1   0   1   0    0    1    0    1
 B5   1   0   1   0   1   0    0    1    0    1
 B7   1   0   1   0   1   0    0    1    0    1
 B8   1   0   1   0   1   0    0    1    0    1
B10   1   0   1   0   1   0    0    1    0    1
B11   1   0   1   0   1   0    0    1    0    1
B12   1   0   1   0   1   0    0    1    0    1
B14   1   0   1   0   1   0    0    1    0    1
B15   1   0   1   0   1   0    0    1    0    1
B16   1   0   1   0   1   0    0    1    0    1
B17   0   1   1   0   1   0    0    1    0    1
B18   1   0   1   0   1   0    0    1    0    1
B21   1   0   1   0   1   0    0    1    0    1
B25   0   1   1   0   1   0    0    1    0    1
```

AFC sur tableau d'indicatrice (.z)

```
res_afc(nasa.Z)      Inertie totale:  1
eboulis(res)
```

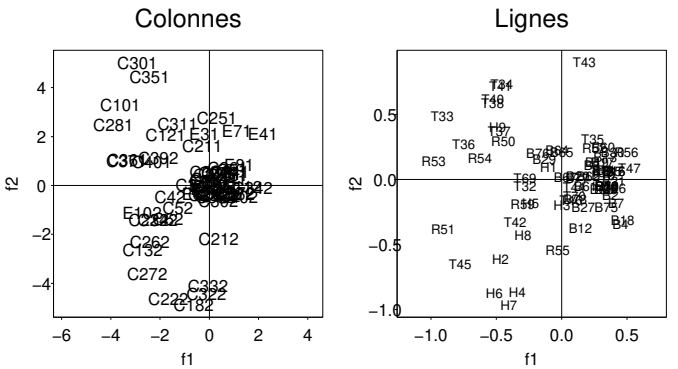



```
print(prexpl(res))
```

```
Pourcentage de inertie expliquee
            f1 f2 f3 f4 f5 tot
inertie exp 16 12  8  7  6  48
inertie cum 16 28 35 42 48  48
```



```
biplot(res)
```



```
print(res$cmpr[1:20,1:6])

            f1          f2           f3           f4          f5           f6
 B2  0.3719532 -0.11672491  0.039806767 -0.158104646  0.26919734 -0.055523139
 B4  0.4494255 -0.34541984  0.163194527 -0.049216190  0.23904087 -0.181740623
 B5  0.3989282  0.06801024 -0.044306776 -0.182097090 -0.03402045 -0.052807389
 B7  0.4164522 -0.18013446  0.152169289 -0.167273243  0.15307922 -0.082554744
 B8  0.2314296  0.11262520  0.101350567 -0.003533524 -0.25980485  0.117267771
B10  0.3191939  0.03526559  0.438786824  0.256636876  0.04120525  0.153377355
B11  0.3243586  0.07124845  0.028772398 -0.113537300 -0.15596669  0.015230428
B12  0.1454707 -0.37359513  0.306722518 -0.128725052 -0.03955731 -0.061947611
B14  0.3454898 -0.05197019  0.091356617 -0.153007402 -0.10553505 -0.007237367
B15  0.3989282  0.06801024 -0.044306776 -0.182097090 -0.03402045 -0.052807389
B16  0.3989282  0.06801024 -0.044306776 -0.182097090 -0.03402045 -0.052807389
B17  0.3164525  0.13499117 -0.323797534 -0.247804083 -0.19397572 -0.638808794
B18  0.4689952 -0.31600255  0.053055486 -0.326760346  0.53644328 -0.167163106
B21  0.4004325 -0.01362730  0.128245116 -0.071975772 -0.02521915  0.015022112
B25  0.3397808 -0.07773270 -0.142814488 -0.166186588  0.02288675 -0.125018918
B26  0.4040930 -0.06621840 -0.045517174 -0.043535963  0.11771361 -0.041403807
B27  0.1672833 -0.21049915  0.004524961 -0.038301738 -0.11873491 -0.020707238
B29 -0.1301485  0.15592718 -0.658033385  0.154908298 -0.25581887 -0.252929735
T32 -0.2780770 -0.05022512 -0.318059647 -0.119044303  0.01433910  0.005798797
T33 -0.9117838  0.48688779  0.444013037 -0.093686108  0.35583654 -0.216898218
```


Classification

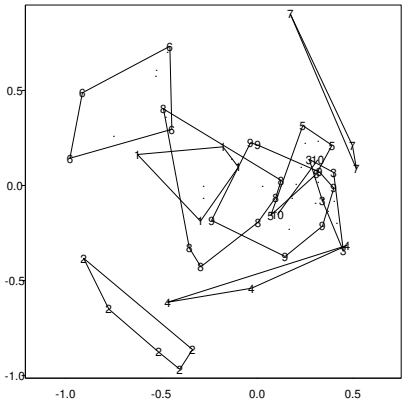
La classification automatique sera abordée de plusieurs façons :

1. À partir des composantes des individus obtenues après l'AFC sur le tableau des indicatrices,
   - classification type nuées dynamiques,
   - classification hiérarchique (critère de Ward).

   Ensuite possibilité d'effectuer une classification mixte,
   - classification hiérarchique puis nuées dynamiques.
2. À partir d'un tableau de distance entre individus. Ici, cette distance n'est pas euclidienne, elle s'obtient comme complément à 1 de l'indice de similarité de Jaccard.

Classification non hiérarchique, nuées dynamiques

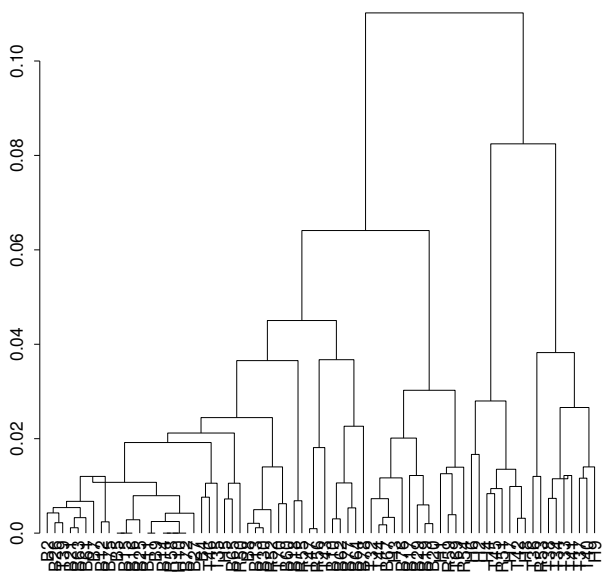
```
#    k classes demandees :
     k <- 10     #par exemple
#  res3_rkmeans(res$cmpr,k)                # au hasard
   res3_kmeans(res$cmpr,centers=rbind(res$cmpr[1:5,],
               res$cmpr[21:25,]))
   p3_res3$cluster
   plotcl(res$cmpr[,1:2],p3)
```



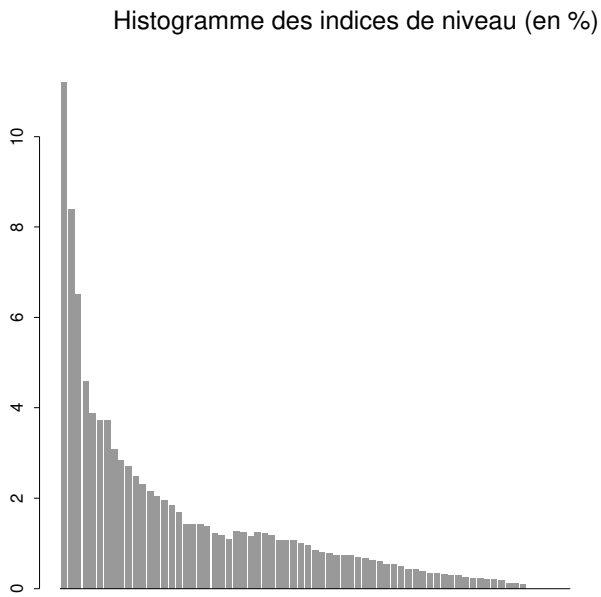

```
     cat("\n repartion des souches dans les groupes : \n")
     print(p3)

 repartion des souches dans les groupes :
 [1] 3  3  3  3  5  9  9  9  9  3  3 10  4  9  3  3  9  8  1  6  6  5  5  9  3
[26] 8  8  7  8  2  5  7  1  5  6  2  6  7  9  5  6  5  4  1  4  8  2  9  2  2
[51] 8  8  9  8  5  6  6  1  5  9  9  9  9  1  9  8  6  9  9  1  9 10

     cat("\n inertie inter apres classification non hierarchique, \n")
     cat("              noyaux tires au sort :  \t")
     print(interx(res$cmpr,p3))

 inertie inter apres classification non hierarchique,
              noyaux tires au sort :   [1] 0.4699885
```


Classification hiérarchique ascendante après une AFC

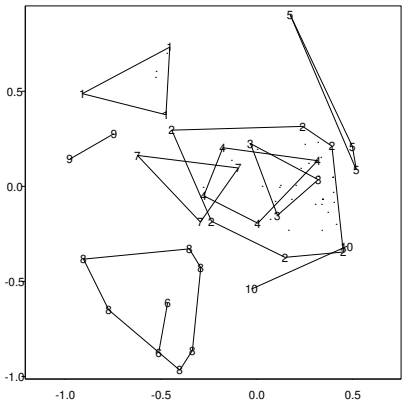
```
dista_dist(res$cmpr) # distance euclidienne arpenteur
hcz_cahd(dista)      # sur le resultat d une AFC
                     # critere du saut de ward

#     construction de l arbre :

      graph_plclust(hcz, hang=-1,plot = F)
      plclust(hcz, hang=-1, label = F)
      labclust(graph, label = etiq)
```



```
scan()
eboulis(hcz)
```




```
scan()
cat("\n repartition en k groupes et inertie inter \n")
po.hcz_cutree(hcz,k)
cat("\n inertie inter apres cah / ward apres AFC :\t")
print(interx(res$cmpr,po.hcz))

inertie inter apres cah / ward apres AFC : [1] 0.4713956

plotcl(res$cmpr[,1:2],po.hcz)
```



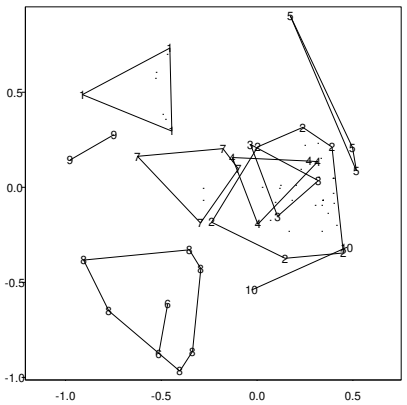

```
scan()
cat("\n repartion des souches dans les groupes : \n")
print(po.hcz)

 repartion des souches dans les groupes :
 [1] 2  2  2  2  2  3  2  2  2  2  2  4 10  2  2  2  2  4  4  1  1  2  2  2  4
[26] 1  8  5  4  8  2  5  7  2  2  8  9  5  2  2  1  2 10  7  6  4  8  2  6  8
[51] 8  1  2  4  2  1  1  7  2  2  3  3  2  7  2  4  9  2  2  4  3  3

attr(, "height"):
 [1] 0.026579937 0.024464628 0.022651687 0.020092169 0.018102579 0.016661802
 [7] 0.013909425 0.013473909 0.011987222 0.006799995
```

Classification mixte hiérarchique ascendante puis nuées dynamiques avec les pôles initiaux les centres de gravité obtenus en CAH

```
cntr1 <- gravc(res$cmpr, po.hcz)
res1_kmeans(res$cmpr,cntr1)
p1.hcz_res1$cluster

cat("\n inertie inter apres mixte : kmeans / AFC :\t")
print(interx(res$cmpr,p1.hcz))
```

```
 inertie inter apres mixte : kmeans / AFC : [1] 0.4789649
```

```
plotcl(res$cmpr[,1:2],p1.hcz)
```

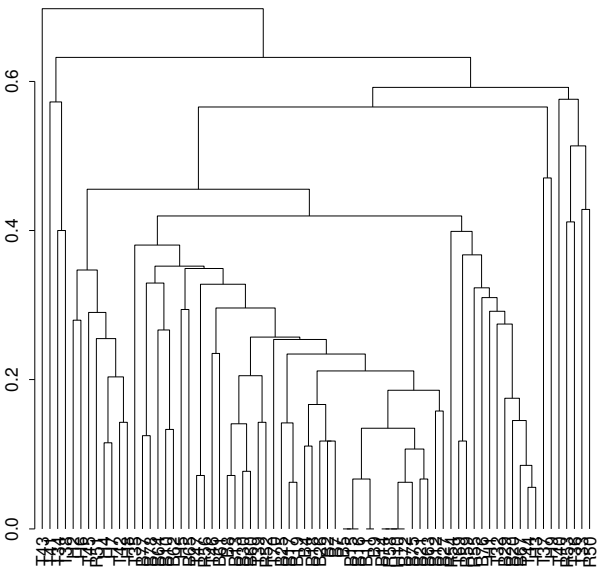



```
scan()
cat("\n repartion des souches dans les groupes : \n")
print(p1.hcz)
```

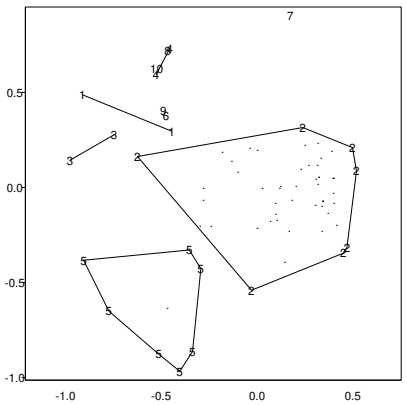
```
 repartion des souches dans les groupes :
 [1] 2  2  2  2  2  3  2  2  2  2  2  4 10  2  2  2  2  4  7  1  1  2  2  2  4
[26] 1  8  5  4  8  2  5  7  2  1  8  9  5  2  2  1  2 10  7  6  4  8  2  6  8
[51] 8  1  2  4  2  1  1  7  2  2  3  3  2  7  2  4  9  2  2  7  3  3
```


Méthode des bactériologistes, indice de Jaccard

```
disnasa <- dist(nasa, metric = "binary")        # 1-indice de jaccard
arbnasa <- hclust(disnasa, method = "average")  # saut moyen

#     construction de l arbre :

graph <- plclust(arbnasa, hang=-1,plot = F)
plclust(arbnasa, hang=-1, label = F)
labclust(graph, label = etiq)
```



```
     scan()
#    repartition en k groupes :
     po_cutree(arbnasa,k)
     cat("\n inertie inter avec la methodes des bacteriologistes :\t")
     print(interx(res$cmpr,po))

 inertie inter avec la methodes des bacteriologistes : [1] 0.3444748

     plotcl(res$cmpr[,1:2],po)
```




```
     scan()
     cat("\n repartion des souches dans les groupes : \n ")
     print(po)

 repartion des souches dans les groupes :
 [1] 2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  1  4  2  2  2  2
[26] 6  5  7  2  5  2  2  2  2  1  5  3  2  2  2  8  2  2  2  5  2  5  2  5  5
[51] 5  9  2  2  2 10  4  2  2  2  2  2  2  2  2  2  3  2  2  2  2  2

attr(, "height"):

 [1] 0.4285714 0.4197177 0.4117647 0.4000000 0.3469404 0.0000000 0.0000000
 [8] 0.0000000 0.0000000 0.0000000
```