Analyse des correspondances multiples
(TP Splus)

Analyse des correspondances multiples sur les données Pommes

	arb	rec	cal	coul	pour	type	form	ferm	gout
Granny-spur	1	3	2	1	1	1	2	3	3
Granny-stand.	4	3	2	1	1	1	2	3	3
Boskoop-mut.	3	3	3	4	4	1	1	2	2
Boskoop-stand.	2	3	3	4	3	1	1	2	2
Gala-red	2	2	1	3	3	3	2	2	1
Gala	3	2	1	3	2	3	2	2	1
Elstar	3	1	1	2	2	3	2	2	3
Elista	3	1	1	3	3	1	2	2	3
Jonagold	3	2	3	3	2	2	2	2	3
Jonagored	3	2	3	4	3	3	2	2	3
New-Jonagold	3	2	3	2	3	1	2	2	3
Gloster	2	3	2	3	4	1	1	3	2
Gloster-Spur	1	3	2	3	4	1	1	3	2
Averdal	3	2	2	4	4	2	1	2	1
StarKrimson	1	2	2	4	4	3	1	2	1
Erovan	2	2	2	4	4	1	1	2	1
Goldenspur	1	2	1	1	1	1	1	2	2
Quemoni	3	2	1	1	2	1	1	2	2
Golden-972	3	2	1	1	1	1	1	2	2
Sinta	2	2	1	3	2	1	3	2	2
Fiesta	3	1	1	2	3	3	2	3	2
Arlet	3	1	1	3	3	2	2	2	2
Hilrome	4	3	3	3	4	1	2	3	2
Rome-Beauty	4	3	3	3	3	3	2	3	2
Melrose-mut.	3	3	2	2	4	1	3	3	2
Melrose	3	3	3	2	3	3	3	3	2

57 variétés de pommes ont été caractérisées par 9 variables qualitatives codées.

La plupart de ces variétés sont génétiquement différentes, d'autres sont issues de mutations et représentent la part la plus importante des variétés commercialisées aujourd'hui car elles apportent un petit *plus* par rapport à la variété standard dont elles dérivent, notamment par leur coloration. Ces variétés mutantes sont :

- *Granny-spur* = mutation de *Granny-Smith* modifiant la forme de l'arbre
- *Golden-spur* = mutation de *Golden* modifiant la forme de l'arbre
- *Gloster-Spur* = mutation de *Gloster* modifiant la forme de l'arbre
- *Boskoop-mut.* = mutation de *Boskoop* plus colorée
- *Gala-red* = mutation de *Gala* plus colorée
- *Elista* = mutation de *Elstar* plus colorée
- *Jonagored, New-Jonagold* = mutations de *Jonagold* plus colorées
- *Quemoni* = mutation de *Golden* plus colorée
- *Hilrome* = mutation de *Rome-Beauty* plus colorée
- *Melrose-mut.* = mutation de *Melrose* plus colorée
- *Belrene* = mutation de *R.-des-Reinettes* plus colorée
- *Fuji-mut.* = mutation de *Fuji* plus colorée

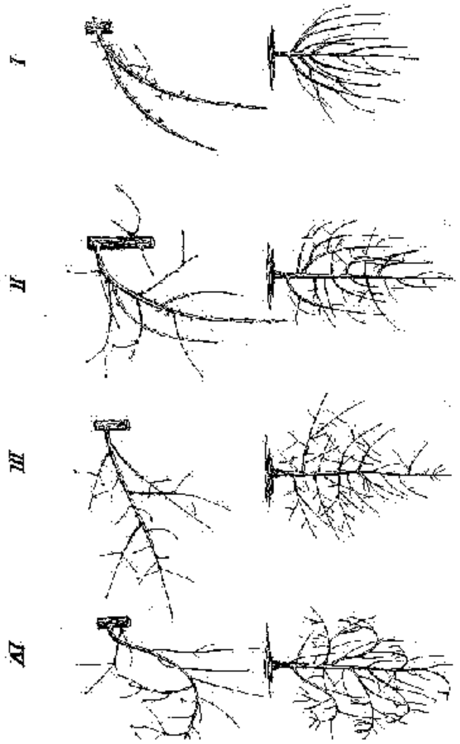
R.-des-Reinettes	2	2	1	2	3	3	2	2	2
Belrene	2	2	1	3	4	2	2	2	2
Generos	3	2	2	1	2	3	2	2	2
Merano	3	2	1	1	2	3	1	2	2
Delblush	3	2	1	1	2	1	1	2	3
Chantecler	3	3	1	1	1	1	2	2	3
Fuji	3	3	1	3	2	3	2	3	1
Fuji-mut.	3	3	1	3	4	1	2	3	1
Braeburn	3	3	1	2	3	2	2	3	3
Baujade	4	3	1	1	1	1	2	3	3
Ginger-Gold	3	1	1	1	2	1	2	2	2
Delcorf	3	1	2	3	2	3	1	2	2
Dalili	3	1	2	3	3	2	1	2	2
Delgollune	2	2	3	3	3	1	1	2	2
Pink-lady	3	3	1	3	3	1	1	3	3
Hanners	4	2	3	4	4	1	1	2	3
Canada-blanc	2	2	3	1	2	1	3	1	2
Cadel	3	2	1	3	3	3	1	2	2
Sansa	3	1	1	2	3	1	2	2	1
Akane	3	1	1	3	3	1	3	2	2
Jerseymac	3	1	1	4	4	1	3	2	2
Vista-Bella	3	1	1	3	4	2	3	1	2
Ota	2	2	1	3	2	1	2	2	2
Rubi	4	2	1	3	3	2	3	2	2
Stayman-Winesap	4	3	3	4	4	1	2	3	1
Hyb1	3	2	1	3	4	1	2	2	1
Hyb2	3	1	2	2	3	2	1	2	2
Rubinette	3	2	1	3	3	2	2	2	2
Rubin	4	1	2	2	4	3	2	2	3
Hyb3	4	2	2	4	4	1	2	3	3
Hyb4	3	2	2	1	1	1	2	1	2

Les variables concernent différents types de caractères: forme de l'arbre, date de récolte, description et qualité du fruit :

- *arb* = forme de l'arbre, caractérisée par 4 modalités (1 = colonnaire ; 2 = spur ; 3 = étalé ; 4 = très étalé)
- *rec* = date de récolte, caractérisée par 3 modalités (1 = précoce ; 2 = intermédiaire ; 3 = tardif ou très tardif)
- *cal* = calibre du fruit, caractérisé par 3 modalités (1 = petit ou moyen ; 2 = gros ; 3 = très gros)
- *coul* = couleur dominante du fruit ou intensité de la couleur rouge= additionnelle, caractérisée par 4 modalités (1 = jaune ou vert ; 2 = rouge-orangé ; 3 = rouge ; 4 = rouge-violacé)
- *pour* = proportion de coloration rouge additionnelle, caractérisée par 4 modalités (1 = absente (0-25 %) ; 2 = faible (25-50 %) ; 3 = moyenne= (50-75 %) ; 4 = forte (75-100 %))
- *type* = type de coloration du fruit, caractérisé par 3 modalités (1 = lavé ; 2 = lavé-strié ; 3 = strié)
- *form* = forme du fruit, caractérisée par 3 modalités (1 = allongé ; 2 = intermédiaire ; 3 = aplati)
- *ferm* = fermeté du fruit, caractérisée par 3 modalités (1 = peu ferme ; 2 = moyen ; 3 = très ferme)
- *gout* = rapport sucre / acidité, caractérisé par 3 modalités (1 = doux ; 2 = équilibré ; 3 = acidulé)

Remarquons que certaines de ces variables ont des modalités ordonnées (ex : *pour*).

Grille caractérisant les ports d'arbres chez le pommier
(d'après J.M. Lespinasse, 1977)



Récupération du fichier de données

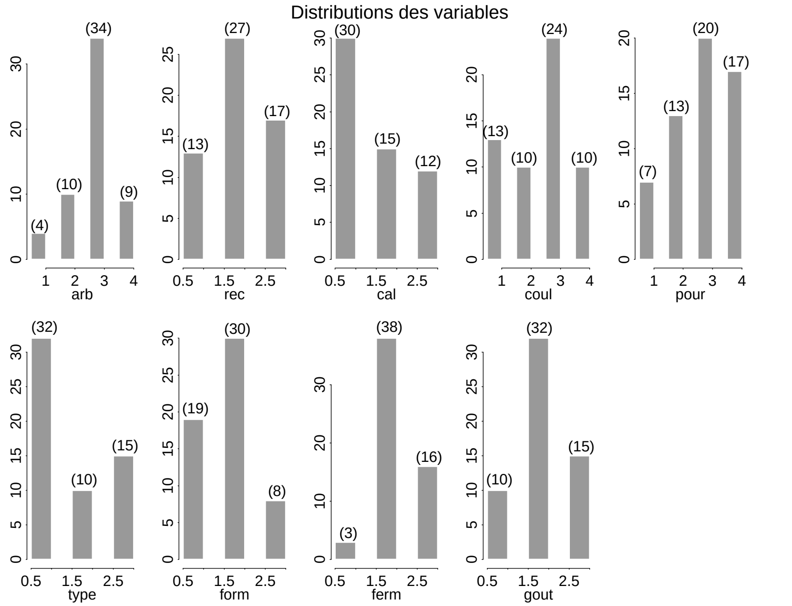
```
> Splus       # lancement de Splus

> pomd <- read.table("pommes.txt",header = T) # le fichier est
                       importé avec les noms des lignes et des
                       colonnes, il est appelé "pomd"

> pomd[1:10,] # présentation des 10 premières lignes seulement

                arb rec cal coul pour type form ferm gout
   Granny-spur   1   3   2    1    1    1    2    3    3
  Granny-stand.  4   3   2    1    1    1    2    3    3
   Boskoop-mut.  3   3   3    4    4    1    1    2    2
Boskoop-stand.   2   3   3    4    3    1    1    2    2
       Gala-red  2   2   1    3    3    3    2    2    1
           Gala  3   2   1    3    2    3    2    2    1
         Elstar  3   1   1    2    2    3    2    2    3
         Elista  3   1   1    3    3    1    2    2    3
       Jonagold  3   2   3    3    2    2    2    2    3
      Jonagored  3   2   3    4    3    3    2    2    3
```


Le but de cette analyse est d'étudier les liaisons entre ces différentes variables et la façon dont cet échantillon de variétés se structure en fonction de ces critères.

Les données sont présentées dans le fichier texte *pommes.txt*. Le fichier *pommes.txt* est récupéré dans Splus sous forme de data.frame.

Représentation graphique des données de "pomd"

```
> motif()            # ouverture de la fenêtre graphique
> par(mfrow=c(2,5)) # division de la fenêtre graphique en 10
                      parties sur 2 lignes et 5 colonnes
> for(i in 1:9){hist(pomd[,i],xlab=names(pomd)[i],cex=0.8)}
                    # boucle pour sortir les histogrammes de
                      toutes les variables
> par(mfrow=c(1,1)) # retour à la fenêtre entière
> text(locator(1),"Distributions des variables",cex=1.2)
                    # rajout d'un titre général positionné à
                      l'intérieur de la fenêtre à l'aide de
                      la fonction= locator()
```



On observe que certaines variables ont des effectifs très déséquilibrés entre leurs différentes modalités :

arb : mod 1 = 4
ferm : mod 1 = 3

Ces modalités plus rares peuvent avoir une influence sur le résultat de l'analyse.

Une règle empirique consiste à respecter pour les modalités un effectif minimum de $\sqrt{3 \times n}$, soit 13 individus pour $n = 57$.

Analyse des correspondances multiples (ACM)

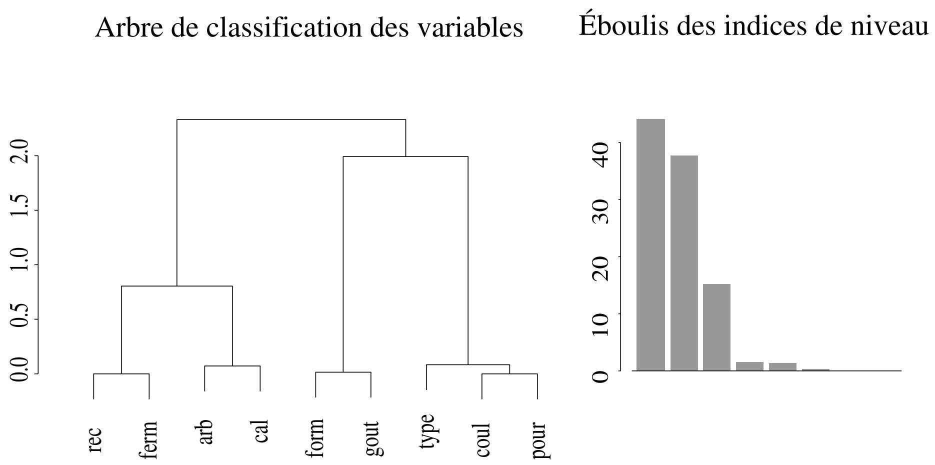
```
> library(multidim,first=T) # appel de la librairie "multidim"

> search() # vérification de la position de multidim et
           repérage de son chemin d'accès

> dyn.load("chemin d'accès/multidim/lib/cahds.o") # chargement
    dynamique de la fonction cahd() qui sera utilisée  pour la
    classification des variétés

> attach(pomd) # donne la possibilité d'accéder aux colonnes
               par leur nom
> chisq.test(table(arb,rec)) # calcul du khi2 et de la proba
                              sur la table arb x rec

Pearson's chi-square test without Yates' continuity correction

data:  table(arb, rec)
X-squared = 12.2798, df = 6, p-value = 0.056

rec   0.06                 tableau des probabilités
cal   0.07  0.17
coul  0.61  0.25  0.07
pour  0.06  0.14  0.08  0.00
type  0.74  0.25  0.90  0.08  0.03
form  0.38  0.71  0.37  0.38  0.86  0.72
ferm  0.10  0.00  0.55  0.56  0.25  0.47  0.03
gout  0.25  0.65  0.89  0.10  0.18  0.70  0.01  0.45
      arb   rec   cal   coul  pour  type  form  ferm
```


Pour avoir une première idée des liaisons entre les variables on peut examiner les tables de contingence 2 à 2 et calculer les χ^2 correspondants.

Remarquons qu'il n'y a pas beaucoup de liaisons significatives au seuil 5 % et que le tableau ci-contre est une matrice de dissimilarité.

Classification des variables

```
> resc <- classqual(pom)
> nom <- dimnames(pom)[[2]]
> plclust(resc,label=nom)
> eboulis(resc,subpl=T)
> title("Arbre de classification des variables")
```



Pour avoir une première idée des liaisons entre les variables on peut représenter un arbre de classification des variables avec l'histogramme des indices de niveau en graphe superposé.

La fonction classqual effectue la classification sur la matrice de dissimilarité des p-values.

On constate que les variables présentant un χ^2 significatif se retrouvent étroitement associées sur l'arbre.

La hauteur des barres de l'éboulis est égale à un facteur près à l'ordonnée des nœuds.

Tableau des indicatrices

```
> pomZ <- matind(pom) # création du tableau des indicatrices

> pomZ[1:10,1:10]     # affichage des 10 premières lignes pour
                        les modalités des 3 premières variables
```

	arb1	arb2	arb3	arb4	rec1	rec2	rec3	cal1	cal2	cal3
Granny-spur	1	0	0	0	0	0	1	0	1	0
Granny-stand.	0	0	0	1	0	0	1	0	1	0
Boskoop-mut.	0	0	1	0	0	0	1	0	0	1
Boskoop-stand.	0	1	0	0	0	0	1	0	0	1
Gala-red	0	1	0	0	0	1	0	1	0	0
Gala	0	0	1	0	0	1	0	1	0	0
Elstar	0	0	1	0	1	0	0	1	0	0
Elista	0	0	1	0	1	0	0	1	0	0
Jonagold	0	0	1	0	0	1	0	0	0	1
Jonagored	0	0	1	0	0	1	0	0	0	1

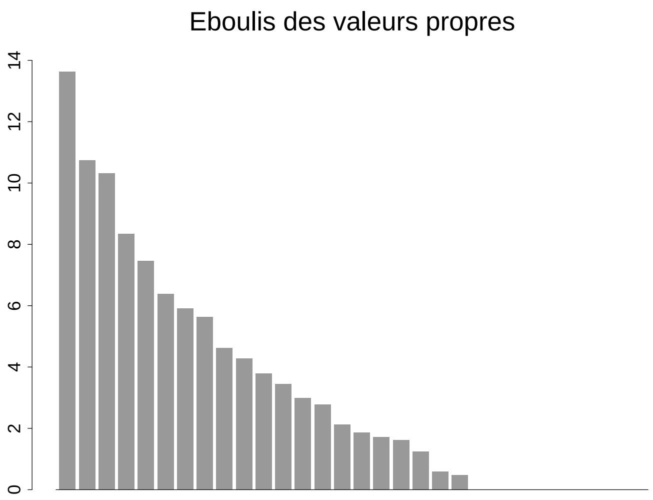
La fonction *matind()* transforme un tableau individus × variables codées en tableau des indicatrices (tableau disjonctif complet).

Le tableau d'indicatrices n'étant pas très grand, 57 lignes et 30 colonnes, son ACM ne pose pas de difficultés de calcul.

Lorsque les effectifs sont très élevés et que l'on s'intéresse surtout aux rapports entre variables, l'ACM du tableau de Burt est mieux indiquée car plus économe en calculs.

Lancement de l'ACM

```
> afc.pomZ <- afc(pomZ) # lancement de l'ACM sur pomZ

  Inertie totale: 2.33

> prexpl(afc.pomZ)

Pourcentage d'inertie expliquée

            f1 f2 f3 f4 f5 tot
inertie exp 14 11 10  8  7  51
inertie cum 14 24 35 43 51  51

> eboulis(afc.pomZ) # représentation graphique
```



$$\text{Inertie totale} = \frac{\text{Nb total de modalités - Nb total de variables}}{\text{Nb total de variables}}$$

$$\text{Inertie totale} = \frac{30-9}{9} = 2.33$$

(elle est indépendante des liaisons entre les variables)

La fonction *prexpl()* donne le pourcentage de variance expliquée par les 5 premiers facteurs et les pourcentages cumulés.

La fonction *éboulis()* donne une représentation graphique plus complète et plus parlante des pourcentages de variance expliquée par les facteurs.

Dans notre cas l'éboulis présente un premier coude après 3 facteurs (soit 35 % de la variabilité) et un second après 5 facteurs (soit 51 % d'inertie expliquée, l'ACM sur tableau Z donne toujours un faible pourcentage expliqué).

Résultats de l'ACM

```
> afc.pomZ       # affichage des résultats de l'ACM

$values:
       f1         f2          f3          f4          f5          f6
0.318305  0.2505639   0.2410143   0.1947314   0.1739748   0.1488507
       f7         f8          f9         f10         f11         f12
0.1379745  0.1315697   0.1079107   0.1000326   0.08847663  0.08041978
      f13        f14         f15         f16         f17         f18
0.06984073 0.06509006  0.04954206  0.04338147  0.04022643  0.03754731
      f19        f20         f21         f22         f23         f24
0.02899809 0.01404053  0.01084272 1.549158e-16 8.937678e-17  7.710185e-17
      f25        f26         f27         f28         f29         f30
5.698231e-17 4.434532e-17 3.436541e-17 2.861928e-17 1.900358e-18 -2.316727e-17
 ...........

$vectors:
             f1          f2         f3         f4          f5          f6
 arb1  2.5451471  1.72141196  1.4079781  1.2563343  3.691801620 -1.93475413
 arb2 -0.6526067  0.01625984  2.0480391 -0.9644125 -2.052712762 -0.99459585
 arb3 -0.6732358  0.13066885 -0.5607562  0.2633841  0.507328337  0.20543381
 arb4  2.1372773 -1.27677634 -0.7829548 -0.4818081 -1.276582479  1.18891394
 rec1 -1.4899817 -0.46586018 -1.1972765 -0.2476271  2.200980559  0.88136041
 rec2 -0.4211688  0.69351894  1.1142410  0.6283349 -0.880218093  0.37011712
 rec3  1.8083129 -0.74522524 -0.8541125 -0.8085818 -0.285109339 -1.26181456
 cal1 -0.7786283  0.51131540 -0.5645279  0.2412463 -0.249178672 -0.91715507
 cal2  1.0318630  0.02044440  0.6506880  0.3082864  2.503087150  0.15760005
 cal3  0.6567420 -1.30384399  0.5979597 -0.9884739 -2.505912257  2.09588760
 ...........

$cmpr:
                          f1           f2          f3          f4
     Granny-spur  1.3932676043  0.785665274 -0.478377283  0.131398030
   Granny-stand.  1.3479487369  0.452533241 -0.721814279 -0.061728905
    Boskoop-mut.  0.4150699645 -0.393172922  0.649637276 -0.155621309
  Boskoop-stand.  0.1881634572 -0.375847975  0.683006410 -0.264330646
        Gala-red -0.5134037456 -0.222195957  0.230017491  0.502271652
            Gala -0.5272712081  0.086243414  0.050156771  0.739311159
          Elstar -0.5476160144  0.003536646 -0.779689238  0.603546803
          Elista -0.3729481974 -0.080026176 -0.543111629  0.174369037
        Jonagold -0.3558117091  0.002803386 -0.098781397  0.191201591
       Jonagored -0.0469588776 -0.383940837  0.013392277  0.526307399
```


La fonction *afc()* appliquée au tableau des indicatrices donne les résultats numériques de l'ACM. Les sorties sont analogues à celles de la fonction *acp()* :

$values = valeurs propres

$vectors = vecteurs propres

$cmpr = coordonnées des individus sur les axes principaux

La fonction contri

La commande

```
> contri(afc.pomZ)
```

nous fournit l'ensemble des indices associés aux individus et aux variables.

La fonction *contri()* donne les aides à l'interprétation de l'ACM ; elle permet d'apprécier la qualité des approximations de chaque individu et de chaque modalité de variable sur les axes et les sous-espaces principaux :

- cosinus carrés lignes - axes principaux
- cosinus carrés lignes - sous-espaces principaux
- cosinus carrés colonnes - axes principaux
- cosinus carrés colonnes - sous-espaces principaux

et la contribution de chaque individu et de chaque variable à la variance de chaque facteur et à l'inertie totale :

- contribution des lignes aux facteurs
- contribution des colonnes aux facteurs

ORI est un indice d'originalité proportionnel au carré de la distance d'un individu ou d'une modalité à l'origine.

On peut analyser à ce stade les contributions des individus et des modalités des variables ou, de préférence, le faire en parallèle de la représentation graphique. Ici, on s'intéresse aux indices associés aux modalités.

```
> contri(afc.pomZ,var=F) # on affiche les aides à
                           l'interprétation de l'ACM

 Contribution des lignes aux facteurs

                  f1   f2   f3   f4   f5  f>5  ORI
     Granny-spur 107   43   17    2   27   30   30
   Granny-stand. 100   14   38    0    0   23   23
    Boskoop-mut.   9   11   31    2    0   15   15
  Boskoop-stand.   2   10   34    6   19   18   18
        Gala-red  15    3    4   23   30   16   16
            Gala  15    1    0   49   16   14   14
          Elstar  17    0   44   33    0   17   17
          Elista   8    0   21    3    0   11   11
        Jonagold   7    0    1    3   21   16   16
       Jonagored   0   10    0   25   28   16   16
    New-Jonagold   0    5   12    4   17   14   14
         Gloster  15    5   17   19    4   16   16
    Gloster-Spur  42    0   13    4   72   24   24
         Averdal   0   11   63   20   42   20   20
     StarKrimson  13    2   90   67   58   29   29
          Erovan   5    5  118   13    2   20   20
      Goldenspur  13  102    7    3   20   25   25
         Quemoni   4   48    5    3    2   11   11
      Golden-972   1   74    1    1    1   14   14
           Sinta  18    9   16   25   27   16   16
          Fiesta  11   12   56    1    7   16   16
           Arlet  36    3   11    1   13   13   13
         Hilrome  35   27    1   26   22   17   17
     Rome-Beauty   9   34   17    6   37   18   18
     Melrose-mut.  9   12    3   49   29   19   19
         Melrose   0   29   23   34    3   21   21
```


Informations sur les individus

– Certaines variétés sont mal représentées dans le premier plan principal (cf. données de *contri()) : Fuji, Hyb1, New Jonagold, Delgollune, Jerseymac, Jonagold, Averdal, Erovan, ...* il faut rester prudent sur l'interprétation de leur position dans le biplot iso-ligne f1 × f2. Le sous-espace des 5 premiers axes représente imparfaitement les variétés *Pink Lady, Jonagold, Hyb1, Fuji mut. Fuji, ...*

– Certaines variétés ont une forte originalité (*ORI*), correspondant à des profils rares : *Canada Blanc, Vista Bella, Granny Spur, Hyb4, Star Krimson...* elles apportent souvent une très forte contribution à un axe particulier (*Vista Bella* - axe 4, *Canada Blanc* - axe 4, *Granny Spur* - axe 1, ...), leur position est alors très excentrée.

– D'une façon générale les dispersions du nuage dans les différents plans principaux (biplots f1 × f2, f1 × f3, ...)sont bonnes, avec une densité de points plus forte au voisinage de l'origine et des points excentrés qui correspondent aux individus à forte originalité. A part quelques variétés très typées, il est difficile de distinguer des clivages nets permettant de séparer des sous-populations éventuelles ; pour approfondir cette question il est nécessaire de prolonger l'analyse par une classification des variétés. Nous procéderons également à une classification des modalités.

R.-des-Reinettes	20	2	1	5	13	15	15
Belrene	11	1	13	3	2	14	14
Generos	3	26	0	18	0	14	14
Merano	10	36	2	16	2	13	13
Delblush	0	46	0	16	5	13	13
Chantecler	17	45	30	3	2	16	16
Fuji	0	3	10	16	11	16	16
Fuji-mut.	9	13	1	0	0	14	14
Braeburn	0	17	74	0	2	18	18
Baujade	72	18	53	0	11	22	22
Ginger-Gold	8	29	6	1	0	12	12
Delcorf	16	1	1	7	25	15	15
Dalili	20	3	0	1	76	15	15
Delgollune	4	1	30	6	29	14	14
Pink-lady	4	3	12	0	0	13	13
Hanners	24	12	29	2	15	19	19
Canada-blanc	3	50	22	153	55	35	35
Cadel	20	0	1	6	0	10	10
Sansa	12	8	17	13	5	15	15
Akane	26	0	3	31	6	14	14
Jerseymac	3	3	8	19	16	17	17
Vista-Bella	27	0	0	167	38	32	32
Ota	11	8	8	0	39	12	12
Rubi	14	4	0	34	1	19	19
Stayman-Winesap	66	57	4	0	22	23	23
Hyb1	1	1	6	12	2	11	11
Hyb2	20	7	3	0	103	18	18
Rubinette	26	0	1	0	0	11	11
Rubin	2	20	20	14	16	21	21
Hyb3	53	14	2	2	0	19	19
Hyb4	6	103	0	33	5	30	30
TOT	1000	1000	1000	1000	1000	1000	1000

Cosinus carré colonnes axes principaux

	f1	f2	f3	f4	f5	f>5	TOT
arb1	156	56	36	23	179	550	1000
arb2	29	0	215	39	156	562	1000
arb3	213	6	112	20	66	582	1000
arb4	273	77	28	8	53	561	1000
rec1	209	16	102	4	249	421	1000
rec2	51	108	269	69	121	381	1000
rec3	442	59	75	54	6	364	1000
cal1	214	73	85	13	12	603	1000
cal2	121	0	36	7	389	447	1000
cal3	37	114	23	51	291	485	1000
coul1	77	761	19	0	2	140	1000
coul2	31	106	239	3	34	586	1000
coul3	120	43	3	34	11	789	1000
coul4	101	134	320	30	0	416	1000
pour1	261	409	65	0	10	255	1000
pour2	117	186	1	34	94	568	1000
pour3	177	157	94	2	2	568	1000
pour4	149	194	219	12	67	359	1000
type1	255	74	28	64	23	556	1000
type2	143	32	1	34	109	680	1000
type3	59	23	24	196	13	685	1000
form1	4	13	314	33	101	535	1000
form2	10	8	311	104	106	461	1000
form3	53	1	2	504	1	438	1000
ferm1	7	111	8	338	0	536	1000
ferm2	321	21	110	210	0	338	1000
ferm3	403	101	153	37	0	307	1000
gout1	3	75	88	188	0	645	1000
gout2	133	29	31	325	19	464	1000
gout3	129	2	207	72	19	571	1000
TOT	136	107	103	83	75	495	1000

Cosinus carré colonnes sous-espaces principaux

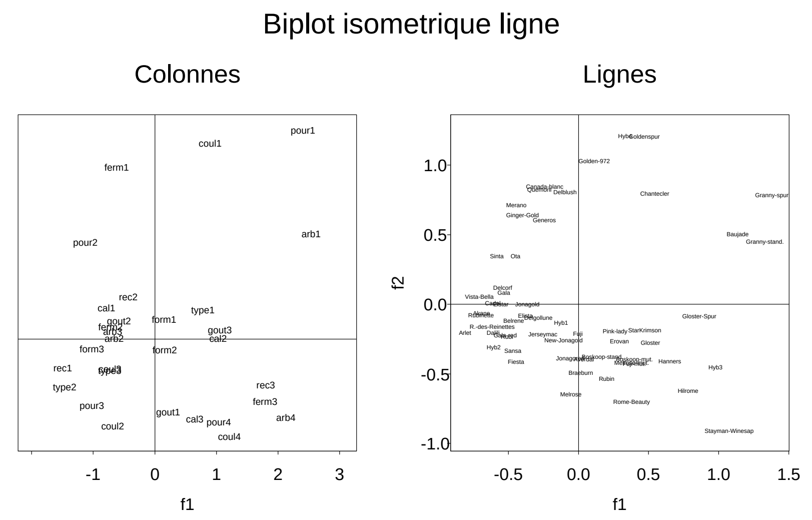
	f1	f1a2	f1a3	f1a4	f1a5	TOT
arb1	156	212	248	271	450	1000
arb2	29	29	244	282	438	1000
arb3	213	220	332	352	418	1000
arb4	273	349	377	385	439	1000
rec1	209	225	327	330	579	1000
rec2	51	159	429	498	619	1000
rec3	442	502	576	630	636	1000
cal1	214	287	373	385	397	1000
cal2	121	121	158	164	553	1000
cal3	37	150	173	224	515	1000
coul1	77	838	857	857	860	1000
coul2	31	137	376	379	414	1000
coul3	120	163	166	200	211	1000
coul4	101	235	554	584	584	1000
pour1	261	670	735	736	745	1000
pour2	117	303	304	339	432	1000
pour3	177	334	428	430	432	1000
pour4	149	343	561	574	641	1000
type1	255	329	357	420	444	1000
type2	143	176	177	211	320	1000
type3	59	82	106	302	315	1000
form1	4	18	331	364	465	1000
form2	10	18	329	433	539	1000
form3	53	54	56	560	562	1000
ferm1	7	117	125	463	464	1000
ferm2	321	342	451	661	662	1000
ferm3	403	503	657	693	693	1000
gout1	3	79	167	355	355	1000
gout2	133	162	193	517	536	1000
gout3	129	132	339	410	429	1000
TOT	136	244	347	431	505	1000

Contribution des colonnes aux facteurs

	f1	f2	f3	f4	f5	f>5	TOT	ORI
arb1	51	23	15	12	106	49	44	116
arb2	8	0	82	18	82	45	39	41
arb3	30	1	21	5	17	23	19	6
arb4	80	29	11	4	29	45	40	47
rec1	56	5	36	2	123	31	37	30
rec2	9	25	65	21	41	19	25	10
rec3	108	18	24	22	3	25	33	21
cal1	35	15	19	3	4	27	23	8
cal2	31	0	12	3	183	32	35	25
cal3	10	40	8	23	147	37	38	33
coul1	21	261	7	0	1	10	37	30
coul2	9	39	91	2	18	47	39	41
coul3	24	11	1	11	4	44	28	12
coul4	29	49	121	14	0	33	39	41
pour1	80	159	26	0	5	22	42	63
pour2	32	64	1	15	46	42	37	30
pour3	40	45	28	1	1	35	31	16
pour4	37	60	71	5	30	24	33	21
type1	39	14	6	16	7	23	21	7
type2	41	12	1	16	57	54	39	41
type3	15	7	8	83	6	49	35	25
form1	1	4	96	13	43	34	32	18
form2	2	2	68	28	32	21	23	8
form3	16	0	1	247	1	36	41	54
ferm1	2	46	3	183	0	49	45	158
ferm2	37	3	17	40	0	11	16	4
ferm3	101	32	51	15	0	21	34	23
gout1	1	28	34	89	0	51	39	41
gout2	20	6	6	81	5	20	21	7
gout3	33	1	70	30	9	40	35	25
	1000	1000	1000	1000	1000	1000	1000	1000

- La contribution globale d'une variable à un facteur est donnée par la somme des contributions de ses modalités.
  Ex : Facteur 1 variable *pour* : $80 + 32 + 40 + 37 = 189$
  form : $1 + 2 + 16 = 19$
- La contribution d'une variable à l'inertie totale (colonne *TOT*) ne dépend que du nombre de ses modalités (cf. cours AFC-ACM, p. 37 à 39)
  Ex : variable *rec* : $37 + 25 + 33 = 95$
  cal: $23 + 35 + 38 = 96$
- Un facteur s'interprète en fonction des variables qui y contribuent le plus (ici pour le facteur 1 , *pour*, *arb*, *rec* et *ferm*).

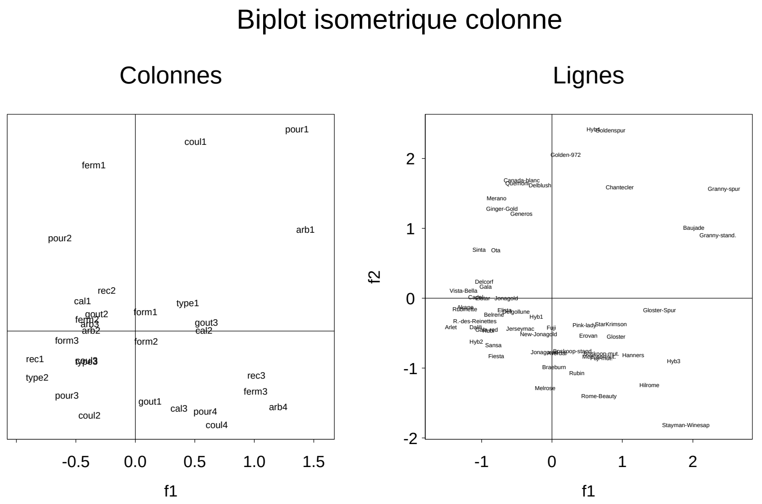
```
> biplot(afc.pomZ,titre="Biplot isometrique ligne",
        cexi=0.4,cexv=0.6) # projections sur le
                              premier plan principal
1: Biplot isometrique ligne ?
2: Biplot isometrique colonne ?
3: Biplot a_l'hollandaise ?
Selection: 1
```



On se propose d'analyser les représentations graphiques et comparer les sorties biplot isométriques ligne et colonne.

biplot isométrique ligne → représentation exacte des individus.

```
> biplot(afc.pomZ,titre="Biplot isometrique colonne",
         cexi=0.4,cexv=0.6)
1: Biplot isometrique ligne ?
2: Biplot isometrique colonne ?
3: Biplot a_l'hollandaise ?
Selection: 2
```



biplot isométrique colonne → représentation exacte des variables

Informations sur les variables (modalités)

- Représentativité médiocre sur le premier plan factoriel (cf. données de *contri()*) qui n' explique que 24 % de l'inertie totale ; certaines modalités comme *form1, form2, arb2* ... y sont très mal représentées.

  Avec 5 facteurs *coul3, type2, type3, gout1,* ... restent mal représentées ; attention donc à l' interprétation des biplots (!).

  Les modalités d'une même variable, affectés de poids proportionnels aux effectifs, ont pour barycentre le barycentre global.

- Contributions des modalités aux axes (cf. données de *contri()* et les différents biplots iso-colonnes faisant intervenir les 5 premiers axes).

  . Le facteur 1 est caractérisé par une forte contribution des modalités *rec3* (récolte tardive) et *ferm3* (très ferme) et, dans une moindre mesure *arb4* (très étalé) et *pour1* (absence de rouge).
  Les modalités *rec3, ferm3, ferm2* et *arb4* sont groupées et bien représentées sur cet axe.

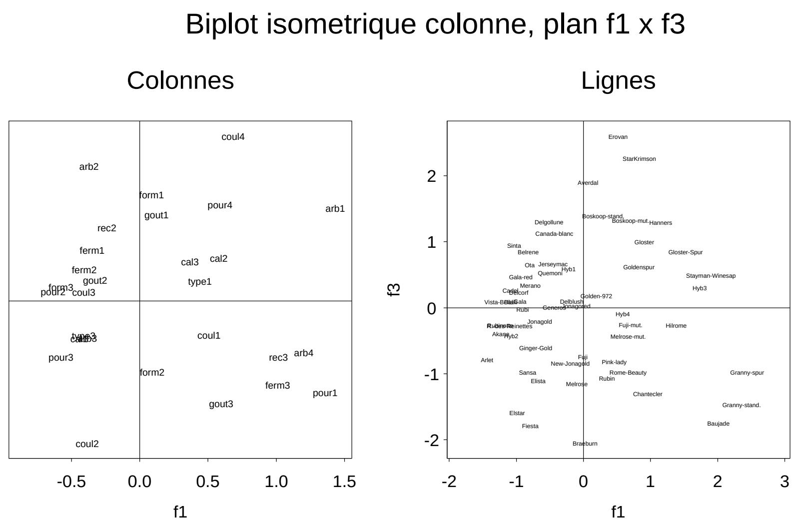
  Grossièrement ce facteur sépare les variétés précoces (*rec1*) des variétés tardives et très fermes. La situation anormale des variétés très tardives *Fuji* et *Pink Lady* près de l'origine peut s'expliquer par leur très mauvaise représentation sur cet axe (cf. contri).

  . Le facteur 2 est caractérisé d'une part par la très forte contribution de *coul1* (jaune ou vert), celle moins forte de *pour1* (absence) qui lui est proche, et celle plus modeste de *pour2* (% faible), et d'autre part par la contribution modeste de *pour4* (% fort) et*coul4* (rouge violacé), ainsi que *pour3* (% moyen) et *coul2* (rouge orangé).

  Ce facteur sépare nettement les variétés jaunes et vertes des autres. Il sépare aussi les arbres colonnaires des arbres très étalés (les modalités *arb1* et *arb4* sont bien représentées sur cet axe). On notera que quand une variété a un mutant spur, ce dernier est toujours décalé positivement par rapport à cet axe.

  Le plan factoriel principal permet d'identifier les variétés jaunes ou vertes, sans coloration additionnelle, d'arbre plutôt colonnaire, des variétés tardives, très fermes, d'arbre étalé et enfin des varétés précoces.

```
> biplot(afc.pomZ,1,3,titre="Biplot isometrique colonne,
        plan f1 x f3",cexi=0.4,cexv=0.6) # projection dans
                                              le plan  f1xf3
1: Biplot isometrique ligne ?
2: Biplot isometrique colonne ?
3: Biplot a_l'hollandaise ?
Selection: 2
```

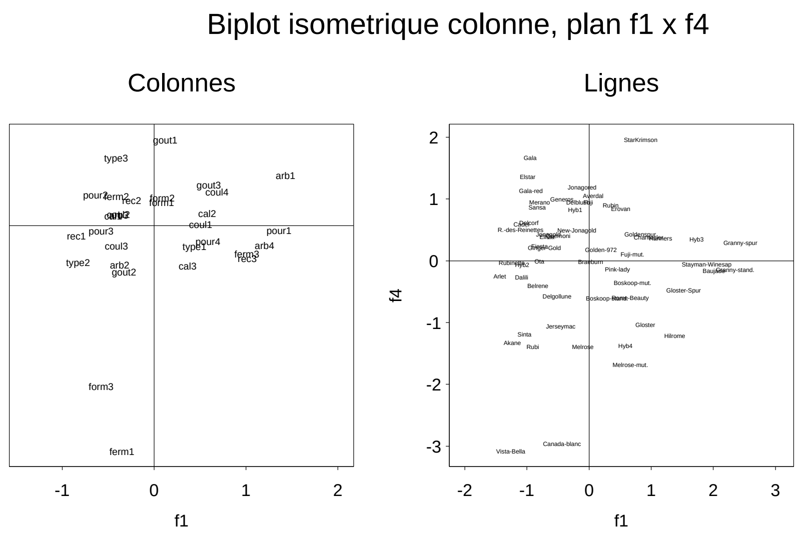


Informations sur les variables (modalités)

. Le facteur 3 est caractérisé par l'opposition entre les fortes contributions de *coul4* (rouge violacé), *form1* (fruit allongé), *arb2* (spur) et *pour4* (% fort) d'une part et d'autre part les contributions de *coul2* (rouge orangé) et *gout3* (acidulé).

Ce facteur sépare les variétés douces des variétés acides, ainsi que parmi les variétés colorées, celles qui le sont peu (rouge orangé) de celles qui le sont beaucoup (rouge violacé). On notera aussi les positions relatives des variétés avec leurs mutants colorés.

```
> biplot(afc.pomZ,1,4,titre="Biplot isometrique colonne,
       plan f1 x f4",cexi=0.4,cexv=0.6) # projection dans
                                           le plan f1xf4
1: Biplot isometrique ligne ?
2: Biplot isometrique colonne ?
3: Biplot a_l'hollandaise ?
Selection: 2
```



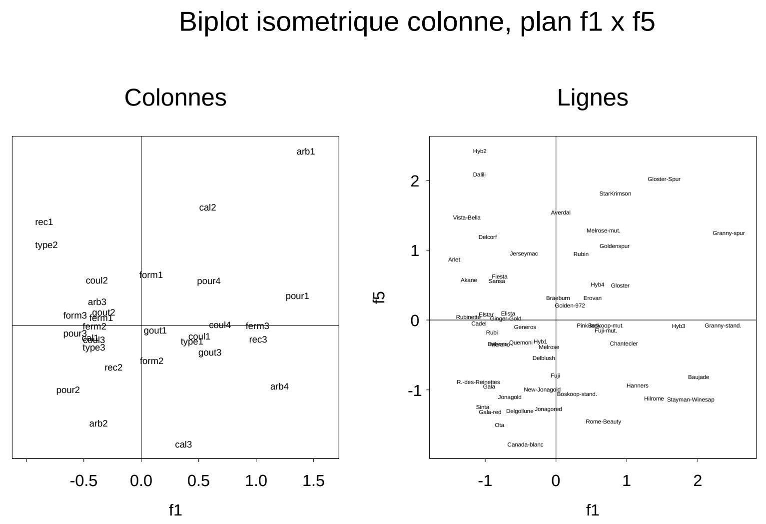
Informations sur les variables (modalités)

. Le facteur 4 est caractérisé par la très forte contribution de *form3* (fruit aplati) et *ferm1* (peu ferme) et la contribution plus faible de *gout2* (équilibré) d'une part, et d'autre part par la contribution de *gout1* (doux) et *type3* (strié).

Ce facteur isole les variétés peu fermes du reste du groupe.

Informations sur les variables (modalités)

```
> biplot(afc.pomZ,1,5,titre="Biplot isometrique colonne,
        plan f1 x f5",cexi=0.4,cexv=0.6) # projection dans
                                             le plan f1xf5
1: Biplot isometrique ligne ?
2: Biplot isometrique colonne ?
3: Biplot a_l'hollandaise ?
Selection: 2
```



Informations sur les variables (modalités)

. Le facteur 5 est caractérisé par la forte contribution de *cal2* (gros), *rec1* (précoce) et *arb1* (colonnaire), opposées à *cal3* (très gros) et *arb2* (spur).

Ce facteur différencie surtout les arbres colonnaires (*arb1*) des arbres à port très étalé (*arb4*). On notera que curieusement les mutants spur sont très éloignés de la modalité *arb2* (!).

Le critère d'originalité (ORI) met en évidence les modalités *ferm1* (peu ferme) et *arb1* (colonnaire), qui correspondent à des effectifs très faibles et pèsent sur le résultat de l'ACM au niveau des axes 4 et 5 (forte contribution et éloignement de l'origine). Il faudra nuancer leur rôle lors de l'interprétation des variables.

En conclusion, la rareté des liaisons existantes entre les variables rend ce tableau difficile à résumer. On retrouve partiellement les associations révélées par la classification : $1er$ axe (*rec, ferm, arb, cal*, $2ème$ axe (*coul, pour*), $3ème$ axe (*coul, form, . . .*), $4ème$ axe (*form, ferm, gout*).

Poursuite du TP sur l'ACM du fichier "Pommes"

```
Classification hiérarchique des modalités.

> pomB <- burt(pom) # création du tableau de Burt
> pomB[1:7,1:7]     # affichage des 7 premières lignes des 7
                      premières colonnes (variables arb et rec)

     arb1 arb2 arb3 arb4 rec1 rec2 rec3
arb1    4    0    0    0    0    2    2
arb2    0   10    0    0    0    8    2
arb3    0    0   34    0   12   14    8
arb4    0    0    0    9    1    3    5
rec1    0    0   12    1   13    0    0
rec2    2    8   14    3    0   27    0
rec3    2    2    8    5    0    0   17

> afc.pomB <- afc(pomB) # lancement de l'ACM
                          sur ``pomB''

Inertie totale:  0.4

> prexpl(afc.pomB)

Pourcentage de inertie expliquee
            f1 f2 f3 f4 f5 tot
inertie exp 25 16 14  9  8  72
inertie cum 25 41 55 65 72  72

> eboulis(afc.pomB)       # facultatif
> contri(afc.pomB)        # facultatif
> biplot(afc.pomB)        # facultatif
```


Pour détecter les proximités entre modalités des différentes variables on peut réaliser une classification à partir de leurs coordonnées dans l'espace des facteurs principaux de l'ACM: tous les facteurs sont pris en considération, la métrique est celle du χ^2, on peut donc travailler sur la matrice des distances euclidiennes entre modalités dans cet espace.
On peut utiliser les coordonnées des modalités à partir de l'ACM des indicatrices ou bien à partir de l'ACM du tableau de Burt. Nous choisirons cette deuxième option ce qui nous permettra de comparer les résultats des deux analyses (facultatif).

On notera que par rapport à la première analyse, l'inertie totale est plus faible mais les pourcentages d'inertie expliqué par les 5 premiers facteurs sont plus forts. Les contributions des modalités aux axes sont identiques ainsi que leurs projections, seules les unités changent.

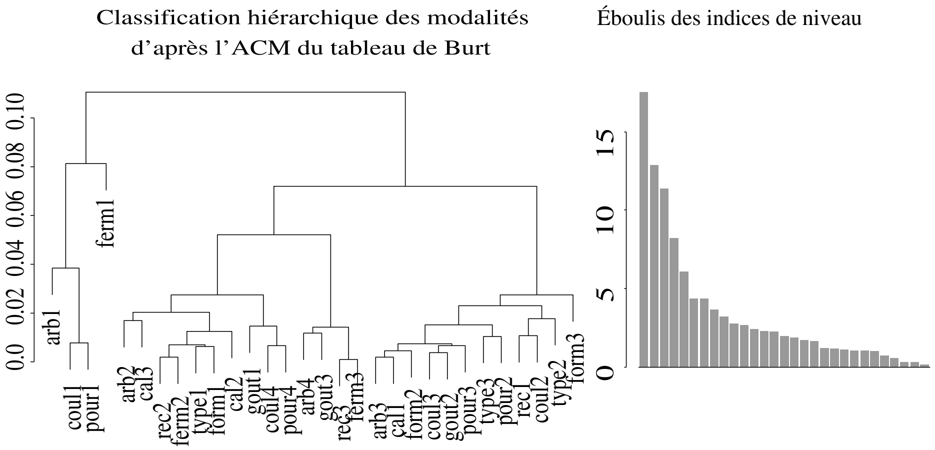
Distances euclidiennes

```
> d.afc.pomB <- dist(afc.pomB$cmpr) # calcul des distances
                                      euclidiennes entre
                                      modalités, dans l'espace
                                      des facteurs principaux

> d.afc.pomB[1:10] # affichage des 10 premières distances

 [1] 1.773100 1.611561 1.703993 1.859797 1.508162 1.409438
 [7] 1.615740 1.184303 1.763603 1.391986

> ch.afc.pomB <- cahd(d.afc.pomB) # classification hiérarchique
                                    ascendante
```


La fonction *cahd()* exécute une classification hiérarchique ascendante en utilisant le critère du saut de Ward qui cherche à minimiser la perte d'inertie inter à chaque regroupement.

Remarquons qu'on classe ici les modalités et non les variables.

```
> clust.afc.pomB <- plclust(ch.afc.pomB,label=F) # tracé de
                                     l'arbre hiérarchique
> labclust(clust.afc.pomB,label=dimnames(pomB)[[1]]) # rajout
                                             des étiquettes
> eboulis(ch.afc.pomB,subpl=T) # rajout de l'histogramme des
                               indices de niveau
> text(locator(1),"Classification hierarchique des modalites \n
                 d'apres l'ACM du tableau de Burt") # rajout
                 d'un titre positionné à l'aide de la
                 fonction locator()
```



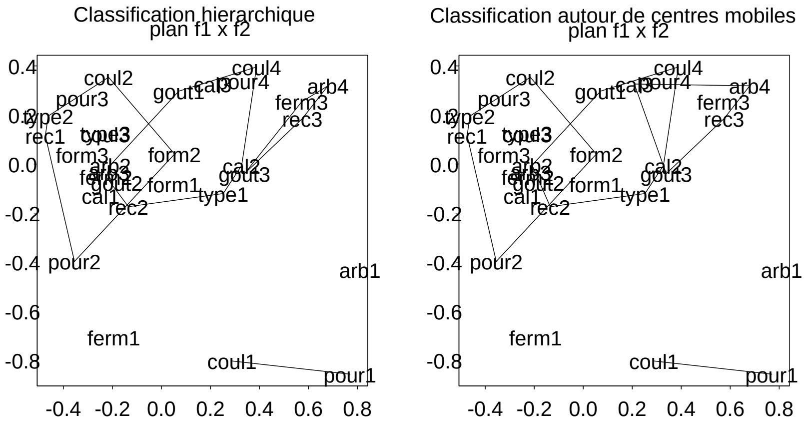
L'examen de l'arbre de classification et de l'éboulis des indices de niveau indique les possibilités de partition en 4 ou 6 classes. On observe que les modalités *ferm1* (peu ferme) et *arb1* (colonnaire) sont nettement séparées des autres, elles correspondent aux deux modalités les plus rares. On trouve *coul1* et *pour1* étroitement associées, cela correspond bien à l' originalité des quelques variétés vertes et jaunes repérables dans le plan principal. Procédons à une répartition en 6 classes.

```
> ph6.pomB <- cutree(ch.afc.pomB,6) # découpage en 6 classes

> ph6.pomB       # affichage de la partition et des niveaux

 [1] 5 2 1 3 1 2 3 1 2 2 4 1 1 2 4 1 1 2 2 1 1 2 1 1 6 2 3 2 1 3
attr(, "height"):
[1] 0.027474793 0.027451970 0.014256261 0.007759315 0.000000000
[6} 0.000000000

> cm6.pomB <- kmeans(afc.pomB$cmpr,gravc(afc.pomB$cmpr,ph6.pomB))
# nouvelle classification par l'algorithme des centres mobiles à
  partir des 6 centres de gravité de la partition hiérarchique

> pcm6.pomB <- cm6.pomB$cluster # nouvelle répartition en 6
                                  classes

> pcm6.pomB                     # affichage de la partition

 [1] 5 2 1 3 1 2 3 1 2 3 4 1 1 2 4 1 1 2 2 1 1 2 1 1 6 1 3 2 1 3
```


La fonction *cutree()* découpe un arbre hiérarchique en un nombre de classes donné.

Pour améliorer la classification obtenue, on procède à une nouvelle classification par l'algorithme des centres mobiles à partir de 6 pôles qui sont les centres de gravité des classes obtenues dans la partition précédente.

```
> par(mfrow=c(1,2))   # division en 2 de la fenêtre graphique

> plotcl(afc.pomB$cmpr[,1:2],ph6.pomB,cl=F) # représentation des 6 classes
                    de modalités  avec leur contour dans le premier plan
                    factoriel

> plotcl(afc.pomB$cmpr[,1:2],pcm6.pomB,cl=F) # idem avec la classification
                    améliorée après l'utilisation de l'algorithme des
                    centres mobiles

> par(mfrow=c(1,1)) # recadrage de la fenêtre graphique en pleine page

> text(locator(1),"Classification hierarchique \n plan f1 x f2") # rajout
                    des titres des graphes

> text(locator(1),"Classification autour de centres mobiles \n plan f1 x f2")
```



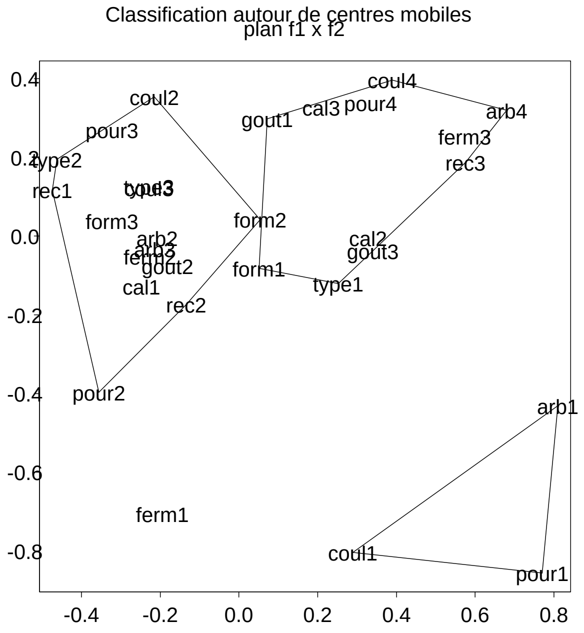
Par rapport à l'AFC sur tableau Z, l'axe 2 est ici inversé. Rappelons qu'un facteur est défini au signe près.

L'arbre de classification comme la représentation dans le plan principal font ressortir l'originalité des profils des modalités *ferm1* et *arb1*, ainsi que la proximité et l'originalité du couple coul1-pour1 qui caractérise les variétés sans coloration rouge.

On remarque également la proximité des profils de *coul4* et *pour4* (variétés très colorées), de *rec3*, *ferm3* et *arb4* …associations qui déterminent certains des axes principaux. À part les 3 classes constituées par *ferm1*, *arb1* et le couple *coul1-pour1*, les autres s'identifient mal dans un ensemble de modalités peu structuré.

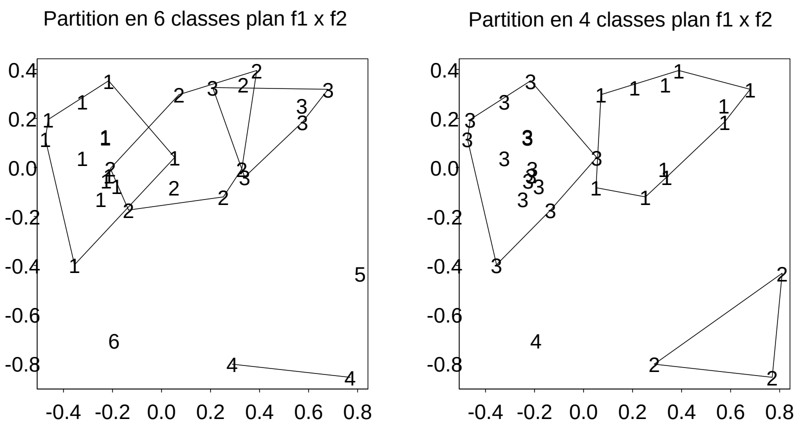
On peut aussi voir ce que donne la répartition en 4 classes (facultatif).

```
> ph4.pomB <- cutree(ch.afc.pomB,4)

> cm4.pomB <- kmeans(afc.pomB$cmpr,gravc(afc.pomB$cmpr,ph4.pomB))

> pcm4.pomB <- cm4.pomB$cluster

> plotcl(afc.pomB$cmpr[,1:2],pcm4.pomB,cl=F)

> text(locator(1),"Classification autour de centres mobiles \n
                  plan f1 x f2")
```



Et comparer les deux partitions (facultatif).

```
> par(mfrow=c(1,2))

> plotcl(afc.pomB$cmpr[,1:2],pcm6.pomB)

> plotcl(afc.pomB$cmpr[,1:2],pcm4.pomB)

> par(mfrow=c(1,1))

> text(locator(1),"Partition en 6 classes plan f1 x f2")

> text(locator(1),"Partition en 4 classes plan f1 x f2")
```

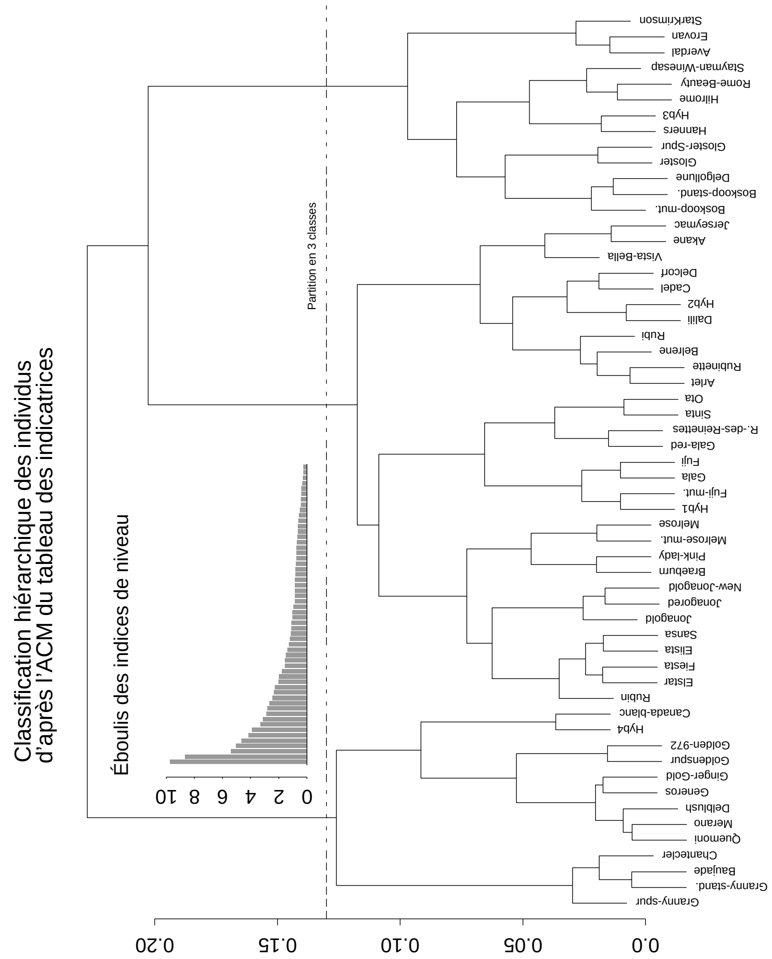


Classification hiérarchique des individus (variétés)

```
> d.afc.pomZ <- dist(afc.pomZ$cmpr) # calcul des distances
                       euclidiennes entre individus dans
                       l'espace des facteurs  principaux
                       (métrique du  Chi-deux)

> ch.afc.pomZ <- cahd(d.afc.pomZ) # classification hiérarchique
                                  ascendante

> clust.afc.pomZ <- plclust(ch.afc.pomZ,label=F) # tracé de
                                                 l'arbre

> labclust(clust.afc.pomZ,label=dimnames(pomZ)[[1]],cex=0.6)
                                    # ajout des étiquettes

> eboulis(ch.afc.pomZ,subpl=T) # ajout de l'histogramme des
                               indices de niveau

> abline(h=0.13,lty=3)
# représentation de la troncature en 3 classes

> text(locator(1),"Partition en 3 classes",cex=0.6)

> text(locator(1),"Classification hierarchique des individus \n
                  d'apres l'ACM du tableau des indicatrices",
                  cex=1.2)
```


On procède comme pour les modalités mais en reprenant le résultat de l'ACM sur le tableau des indicatrices pour être renseigné sur les individus.

Classification hiérarchique ascendante par la méthode du saut de Ward : fonction *cahd()*.



Le *coude* de l'éboulis des indices de niveau oriente le choix vers une partition en 3 classes (on pourrait aussi envisager 8 classes).

```
> ph3.pomZ <- cutree(ch.afc.pomZ,3) # partition en 3 classes

> cm3.pomZ <- kmeans(afc.pomZ$cmpr,gravc(afc.pomZ$cmpr,ph3.pomZ))
                                    # classification autour de
                                      centres mobiles

> pcm3.pomZ <- cm3.pomZ$cluster # nouvelle partition en 3 classe

> pcm3.pomZ # affichage de la partition

[1] 1 1 3 3 2 2 2 2 2 2 2 3 3 3 3 3 1 1 1 2 2 2 3 2 3 2 2 2 1
[30] 1 1 1 2 3 2 1 1 2 2 2 2 3 1 2 2 2 3 2 2 2 3 2 2 2 2 3 1
```

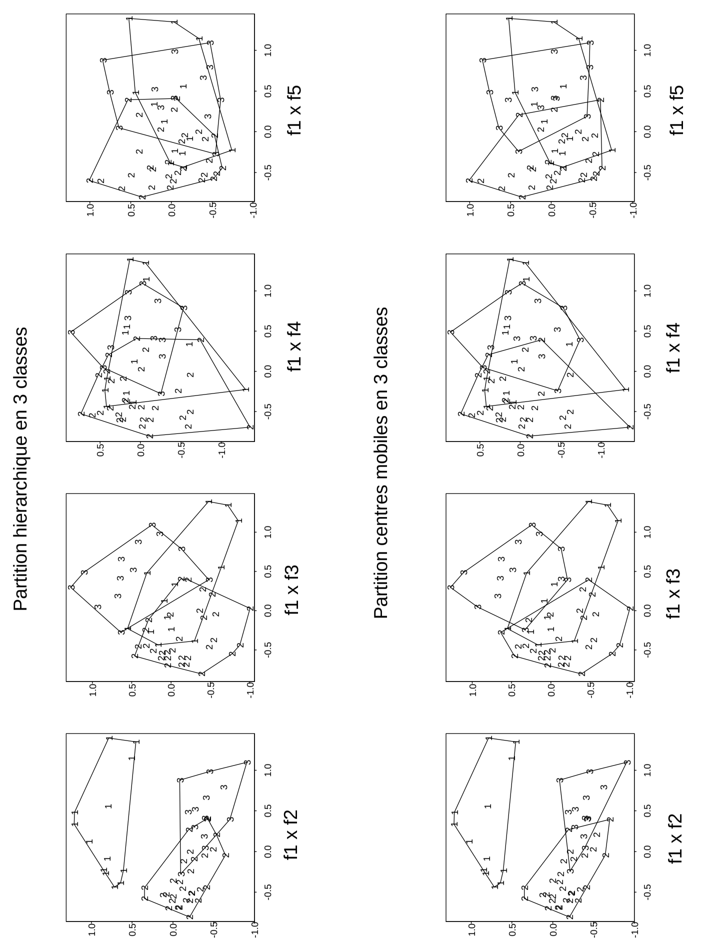

Amélioration de la partition en 3 classes par réallocation itérative des individus à partir des noyaux constitués par les centres de gravité des classes précédentes ;

fonction *kmeans()* : algorithme des centres mobiles optimisant le gain d'inertie inter

fonction *gravc()* : détermination des centres de gravité des classes d'une partition.

Représentation des classes (projections)

```
> par(mfrow=c(2,4)) # division de la fenêtre graphique en
                      matrice 2 lignes x 4 colonnes

> for(i in 2:5){plotcl(afc.pomZ$cmpr[,c(1,i)],ph3.pomZ,nmax=60)}
  # boucle pour sortir les graphes des 3 classes hiérarchiques
    avec leur contour dans les plans principaux f1xf2, f1xf3,
    f1xf4 et f1xf5

> for(i in 2:5){plotcl(afc.pomZ$cmpr[,c(1,i)],pcm3.pomZ,nmax=60)}
   # idem avec les classes modifiées par la méthode des centres
     mobiles

> par(mfrow=c(1,1)) # retour à la fenêtre entière

> text(locator(1),"Partition hierarchique en 3 classes")
     # positionnement du premier titre

> text(locator(1),"Partition centres mobiles en 3 classes")
     # second titre

> text(locator(2),"f1 x f2") # indication du plan de projection
                               pour les deux séries de graphes
> text(locator(2),"f1 x f3")

> text(locator(2),"f1 x f4")
> text(locator(2),"f1 x f5")
```



On remarque que l'amélioration de la partition par la méthode des centres mobiles ne permet pas de discriminer totalement entre les classes 2 et 3 dans les deux premiers plans factoriels.

La classe 1, par contre, s'individualise très bien dans le plan f1 $\times$ f2.

Représentation des classes (distributions)

```
>  par(mfrow=c(2,5)) # division de la fenêtre en matrice
                       graphique

> for(i in 1:5){boxplot(split(afc.pomZ$cmpr[,i],ph3.pomZ),
               main=dimnames(afc.pomZ$cmpr)[[2]][i])}
  # boucle pour sortir les boxplots des distributions des
    3 classes pour les 5 premiers facteurs (partition
    hiérarchique)

> for(i in 1:5){boxplot(split(afc.pomZ$cmpr[,i],pcm3.pomZ),
               main=dimnames(afc.pomZ$cmpr)[[2]][i])} # idem
                           pour la partition centres mobiles

> par(mfrow=c(1,1)) # retour à la fenêtre entière

> text(locator(1),"Partition hierarchique en 3 classes")
  # positionnement du premier titre

> text(locator(1),"Partition centres mobiles en 3 classes")
  # second titre
```

Séparation des classes

```
> ph3 <- ph3.pomZ
> pcm3 <- pcm3.pomZ # raccourcissement des noms des partitions

> pomh3 <- cbind(ph3,pom) # on ajoute à la matrice "pom" une
                            colonne avec le code de la classe
                            des individus (partition
                            hiérarchique)

> pomh3 <- as.data.frame(pomh3) # transformation de la matrice
                                  en data.frame

> cl1h3 <- pomh3[pomh3$ph3=1,] # sélection des individus ayant
                                 la valeur 1 pour la colonne ph3
                                 (= classe1)

> cl2h3 <- pomh3[pomh3$ph3=2,] # classe 2
> cl3h3 <- pomh3[pomh3$ph3=3,] # classe 3
> pomcm3 <- cbind(pcm3,pom)
# on ajoute à la matrice "pom" une colonne avec le code de
# la classe des individus (partition centres mobiles)

> pomcm3 <- as.data.frame(pomcm3)
> cl1pcm3 <- pomcm3[pomcm3$pcm3=1,] # classe 1
> cl2pcm3 <- pomcm3[pomcm3$pcm3=2,] # classe 2
> cl3pcm3 <- pomcm3[pomcm3$pcm3=3,] # classe 3
```


On coupe en 3 le fichier de départ selon les classes issues de la partition hiérarchique puis selon celles issues de la partition centres mobiles.

```
> cl1pcm3       # on affiche la classe 1

              pcm3 arb rec cal coul pour type form ferm gout
  Granny-spur    1   1   3   2    1    1    1    2    3    3
Granny-stand.    1   4   3   2    1    1    1    2    3    3
   Goldenspur    1   1   2   1    1    1    1    1    2    2
      Quemoni    1   3   2   1    1    2    1    1    2    2
   Golden-972    1   3   2   1    1    1    1    1    2    2
      Generos    1   3   2   2    1    2    3    2    2    2
       Merano    1   3   2   1    1    2    3    1    2    2
     Delblush    1   3   2   1    1    2    1    1    2    3
   Chantecler    1   3   3   1    1    1    1    2    2    3
      Baujade    1   4   3   1    1    1    1    2    3    3
  Ginger-Gold    1   3   1   1    1    2    1    2    2    2
 Canada-blanc    1   2   2   3    1    2    1    3    1    2
         Hyb4    1   3   2   2    1    1    1    2    1    2
```


L'examen de la composition des classes permet de les interpréter en fonction des variables qui contribuent le plus aux facteurs responsables de ces partitions (cf. les *plotcl*).

Dans la classe 1 on trouve toutes les variétés *coul1* (jaune ou vert), majoritairement de type1 (lavé), partagées entre *pour1* (absence) et *pour2* (faible), ce qui correspond bien à leur position sur le biplot f1 $\times$ f2. C'est l'axe 2 qui discrimine bien cette classe puisqu'il est déterminé essentiellement par les contributions des modalités *coul1* et *pour1*.

```
> cl2pcm3       # classe 2
```

	pcm3	arb	rec	cal	coul	pour	type	form	ferm	gout
Gala-red	2	2	2	1	3	3	3	2	2	1
Gala	2	3	2	1	3	2	3	2	2	1
Estar	2	3	1	1	2	2	3	2	2	3
Elista	2	3	1	1	3	3	1	2	2	3
Jonagold	2	3	2	3	3	2	2	2	2	3
Jonagored	2	3	2	3	4	3	3	2	2	3
New-Jonagold	2	3	2	3	2	3	1	2	2	3
Sinta	2	2	2	1	3	2	1	3	2	2
Fiesta	2	3	1	1	2	3	3	2	3	2
Arlet	2	3	1	1	3	3	2	2	2	2
Rome-Beauty	2	4	3	3	3	3	3	2	3	2
Melrose	2	3	3	3	2	3	3	3	3	2
R.des-Reinettes	2	2	2	1	2	3	3	2	2	2
Belrene	2	2	2	1	3	4	2	2	2	2
Fuji	2	3	3	1	3	2	3	2	3	1
Braeburn	2	3	3	1	2	3	2	2	3	3
Delcorf	2	3	1	2	3	2	3	1	2	2
Dalili	2	3	1	2	3	3	2	1	2	2
Delgollune	2	2	2	3	3	3	1	1	2	2
Pink-lady	2	3	3	1	3	3	1	1	3	3
Cadel	2	3	2	1	3	3	3	1	2	2
Sansa	2	3	1	1	2	3	1	2	2	1
Akane	2	3	1	1	3	3	1	3	2	2
Vista-Bella	2	3	1	1	3	4	2	3	1	2
Ota	2	2	2	1	3	2	1	2	2	2
Rubi	2	4	2	1	3	3	2	3	2	2
Hyb1	2	3	2	1	3	4	1	2	2	1
Hyb2	2	3	1	2	2	3	2	1	2	2
Rubinette	2	3	2	1	3	3	2	2	2	2
Rubin	2	4	1	2	2	4	3	2	2	3

Dans la classe 2 on trouve des variétés *coul2* (rouge orangé) et *coul3* (rouge), plutôt précoces *rec1* et *rec2*, et de *gout2* (équilibré). Cette classe se distingue bien de la classe 1 sur l'axe 2 (coloration), et un peu moins bien de la classe 3 dans le plan f1 × f3 (gradients précocité et goût).

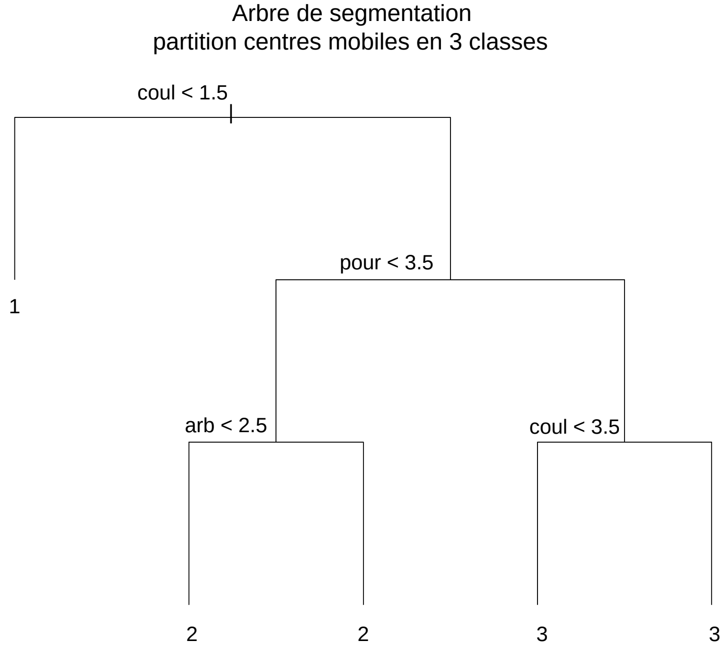
```
> cl3pcm3       # classe 3

                 pcm3 arb rec cal coul pour type form ferm gout
    Boskoop-mut.    3   3   3   3    4    4    1    1    2    2
  Boskoop-stand.    3   2   3   3    4    3    1    1    2    2
         Gloster    3   2   3   2    3    4    1    1    3    2
    Gloster-Spur    3   1   3   2    3    4    1    1    3    2
         Averdal    3   3   2   2    4    4    2    1    2    1
     StarKrimson    3   1   2   2    4    4    3    1    2    1
          Erovan    3   2   2   2    4    4    1    1    2    1
         Hilrome    3   4   3   3    3    4    1    2    3    2
    Melrose-mut.    3   3   3   2    2    4    1    3    3    2
       Fuji-mut.    3   3   3   1    3    4    1    2    3    1
         Hanners    3   4   2   3    4    4    1    1    2    3
       Jerseymac    3   3   1   1    4    4    1    3    2    2
 Stayman-Winesap    3   4   3   3    4    4    1    2    3    1
            Hyb3    3   4   2   2    4    4    1    2    3    3
```


Dans la classe 3 on trouve quasiment toutes les variétés *coul4* (rouge violacé) sauf une, et quelques *coul3*, la coloration est toujours forte (*pour4*), la récolte plutôt tardive (*rec3*), et de goût doux à équilibré (*gout1* et *gout2*). Elles ont en général un calibre supérieur à celui des deux classes précédentes.

En résumé, les trois classes peuvent être discriminées par une combinaison des trois premiers facteurs. Le facteur 2 sépare nettement la classe 1 des deux autres (caractère absence de coloration rouge) ; les classes 2 et 3 ne peuvent être séparées que dans le plan f1 $\times$ f3 (précocité de la récolte et intensité de la coloration rouge, voire le gout plutôt doux ou acidulé). On constate l'importance de la coloration rouge additionnelle dans la structuration de cet échantillon de variétés de pommes, ce phénomène est à rapprocher du fait que la coloration est caractérisée par deux variables (*coul* et *pour*) très étroitement corrélées (par nature ?), représentées chacune par 4 modalités soit 8 sur un total de 30 ! Il conviendrait peut-être de recommencer cette analyse en écartant la variable pourcentage, redondante avec coloration, qui la déséquilibre.

Arbre de segmentation

```
> attach(pomcm3)
> pcm3 <- factor(pcm3)
> tree3 <- tree(pomcm3,pcm3)
> plot(tree3,type="u")
> text(tree3)
> title("Arbre de segmentation \n
       partition centres mobiles en 3 classes",
       cex=0.8)
```



L'arbre de segmentation confirme ces résultats :

- la classe 1 est discriminée par la variable couleur ($coul = 1$) ; ce sont les variétés vertes ou jaunes ;
- la classe 2 est séparée de la classe 3 par la variable proportion de coloration rouge ($pour = 3$ et quelques $pour = 2$) ; ce sont les variétés les moins colorées des variétés rouges ;
- la classe 3 regroupe les variétés les plus fortement colorées ($pour = 4$ et aussi $coul = 4$ majoritairement).

On remarque que les variétés mutantes restent dans la même classe que leur variété d'origine, y compris les mutants plus colorés (à l'exception cependant des variétés *Hilrome, Melrose mut.* et *Fuji mut.* qui passent de la classe 2 à la classe 3).

La segmentation est une méthode qui cherche à expliquer une variable qualitative (ici la classe) par d'autres variables.
À chaque pas elle fait une dichotomie de la population en utilisant la variable la plus discriminante (ici *coul*) de façon à générer des ensembles les plus homogènes relativement à la variable à expliquer (cf. Statistical Models in S, Chambers & Hastie 1990).