Analyse des correspondances multiples (TP SAS)

©1998 - Formation statistique - Module 7/8 (ACM TP SAS)-1

Exemple Pommes

Tableau croisant les modalités de 9 variables qualitatives.

Les variables sont mesurées sur 57 variétés de pommes.

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.

Les variables concernent différents types de caractères : forme de l'arbre, date de récolte, description et qualité du fruit. Certaines de ces variables ont des modalités ordonnées et d'autres non.

arb forme de l'arbre (4 modalités)

rec date de la récolte (3 modalités)

cal calibre du fruit (3 modalités)

coul couleur dominante du fruit ou intensité de la couleur rouge additionnelle (3 modalités)

pour proportion de coloration rouge additionnelle (4 modalités)

type type de coloration du fruit (3 modalités)

form forme du fruit (3 modalités)

ferm fermeté du fruit (3 modalités)

gout rapport sucre/acidité (3 modalités)

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 la variété standard dont elles dérivent, notamment par leur coloration.

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

arb forme de l'arbre (4 modalités)

rec date de la récolte (3 modalités)

cal calibre du fruit (3 modalités)

coul couleur dominante du fruit ou intensité de la couleur rouge additionnelle (3 modalités)

pour proportion de coloration rouge additionnelle (4 modalités)

type type de coloration du fruit (3 modalités)

form forme du fruit (3 modalités)

ferm fermeté du fruit (3 modalités)

gout rapport sucre/acidité (3 modalités)

Les différentes modalités des variables sont codées de 1 à 3 (4) selon le tableau ci-après :

Variable	Modalités			
	1	2	3	4
arb	spur	1/2 spur	étalé	très étalé
rec	précoce	intermédiaire	tardif ou très tardif	
cal	petit ou moyen	gros	très gros	
coul	jaune ou vert	rouge-orangé	rouge	rouge-violacé
pour	0-25%	25-50%	50-75%	> 75%
type	lavé	lavé-strié	strié	
form	allongé	intermédiaire	aplati	
ferm	peu ferme	moyen	très ferme	
gout	doux	équilibré	acidulé	

Fichiers de Données : Pommes.txt

Tableau croisant les modalités de 9 variables qualitatives mesurées sur 57 variétes de 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
...	...	...	...	...	...	...	...	...	...
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

À l'intersection d'une ligne et d'une colonne, on trouve le numéro de la modalité.

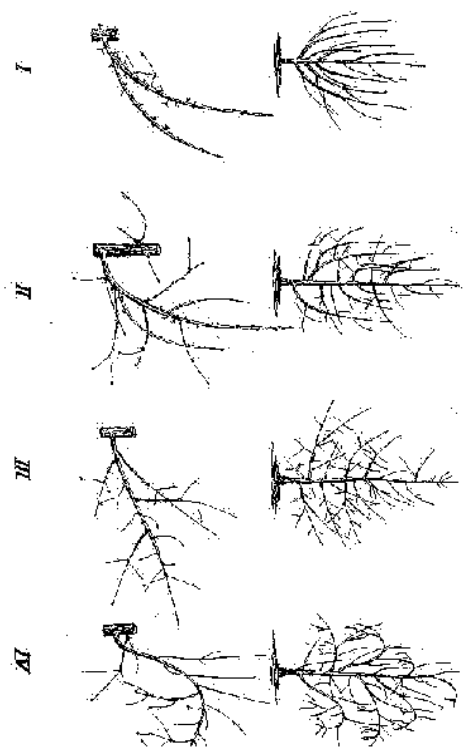
57 variétés de pommes ont été caractérisées par 9 variables qualitatives codées.

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

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

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



Création du tableau SAS

```
Data Pom ;
 infile 'pommes.txt' firstobs=2;
 input Nom :$18. arb$ rec$ cal$ coul$
       pour$ type$ form$ ferm$ gout$ ;
  id=_n_;
 length labl $8.;
 array tmp(*) arb--gout;
  do i=1 to dim(tmp);
     call vname(tmp{i},labl);
     tmp{i}=compress(labl||tmp{i});
  end;
drop i labl;
run;

Proc print data=Pom(obs=10);
run;
```

OBS	NOM	ARB	REC	CAL	COUL	POUR	TYPE	FORM	FERM	GOUT	ID
1	Granny-spur	ARB1	REC3	CAL2	COUL1	POUR1	TYPE1	FORM2	FERM3	GOUT3	1
2	Granny-stand.	ARB4	REC3	CAL2	COUL1	POUR1	TYPE1	FORM2	FERM3	GOUT3	2
3	Boskoop-mut.	ARB3	REC3	CAL3	COUL4	POUR4	TYPE1	FORM1	FERM2	GOUT2	3
4	Boskoop-stand.	ARB2	REC3	CAL3	COUL4	POUR3	TYPE1	FORM1	FERM2	GOUT2	4
5	Gala-red	ARB2	REC2	CAL1	COUL3	POUR3	TYPE3	FORM2	FERM2	GOUT1	5
6	Gala	ARB3	REC2	CAL1	COUL3	POUR2	TYPE3	FORM2	FERM2	GOUT1	6
7	Elstar	ARB3	REC1	CAL1	COUL2	POUR2	TYPE3	FORM2	FERM2	GOUT3	7
8	Elista	ARB3	REC1	CAL1	COUL3	POUR3	TYPE1	FORM2	FERM2	GOUT3	8
9	Jonagold	ARB3	REC2	CAL3	COUL3	POUR2	TYPE2	FORM2	FERM2	GOUT3	9
10	Jonagored	ARB3	REC2	CAL3	COUL4	POUR3	TYPE3	FORM2	FERM2	GOUT3	10

Lecture du fichier : pommes.txt

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.

Création d'un tableau SAS que l'on nomme *Pom* à partir du fichier externe *pommes.txt*. Les tableaux créés par SAS sont du type lignes × colonnes avec en lignes des *individus* et en colonnes des *variables*.

L'option *firstobs=2* impose la lecture du fichier à partir de la $2ème$ ligne.

Le tableau *Pom* contient 10 colonnes ou variables ayant les intitulés : *Nom arb rec cal coul pour type form ferm gout*.

Chaque variable correspond à une des colonnes du fichier *pommes.txt*.

Toutes les variables sont lues avec un type alphanumérique. On les fait suivre par (*$*). Pour la première variable (*Nom*) on ajoute en plus un format de lecture (:18) indiquant que la longueur maximales des valeurs de cette variable est de 18 caractères.

On crée une nouvelle variable qui servira plus tard que l'on appelle *id*.

Le reste du programme permet de remplacer les valeurs {1,2,3,4} des modalités par un label reprenant le nom de la variable et le numéro de la modalité.
Par exemple la modalité 1 de la variable *arb* notée 1 dans le fichier *pommes.txt* devient *arb1*.

Facultatif : explication code

Pour cela on va utiliser une boucle après avoir créé un vecteur.

array tmp (*) arb- -gout;	on range les variables *arb, rec, cal, coul, pour, type, form, ferm, gout* dans un vecteur *tmp*, avec *tmp(1)=arb, tmp(2)=rec, ..., tmp(9)=gout.*
do i=1 to dim(tmp);	on commence une boucle de *i=1* à *i=9*, dimension du vecteur *tmp* (= nombre de variables)
call vname(tmp(i),labl);	on affecte à la nouvelle variable *labl* le label de *tmp(i)*, c'est-à-dire la valeur *arb* pour *tmp(1)*, la valeur *rec* pour *tmp(2)*, ...
tmp(i)=compress(labl‖tmp(i));	on affecte à *tmp(i)* le résultat de la concaténation de la nouvelle variable *labl* et du contenu de *tmp(i)*
end;	fin de la boucle
drop i labl;	on enlève du tableau *Pom* ces 2 variables

Exploration graphique des données

```
Goptions nodisplay;

Proc Datasets library=work;
  delete graphe/ memtype=catalog;
run;
quit;

Proc Gchart data=Pom gout=graphe;
 vbar arb rec cal coul pour type form ferm gout
      /autoref clipref frame;
run;
quit;

Goptions display;

%inc 'panels.sas';

%panels(rows=3,cols=3,plots=9,order=DOWN,equate=Y,
        gin=graphe,gout=graphe);
```


Facultatif : explication code

Goptions : options graphiques permettant de sortir les graphiques,
nodisplay n'affiche pas sur l'écran les graphiques produits
display rétablit l'affichage écran

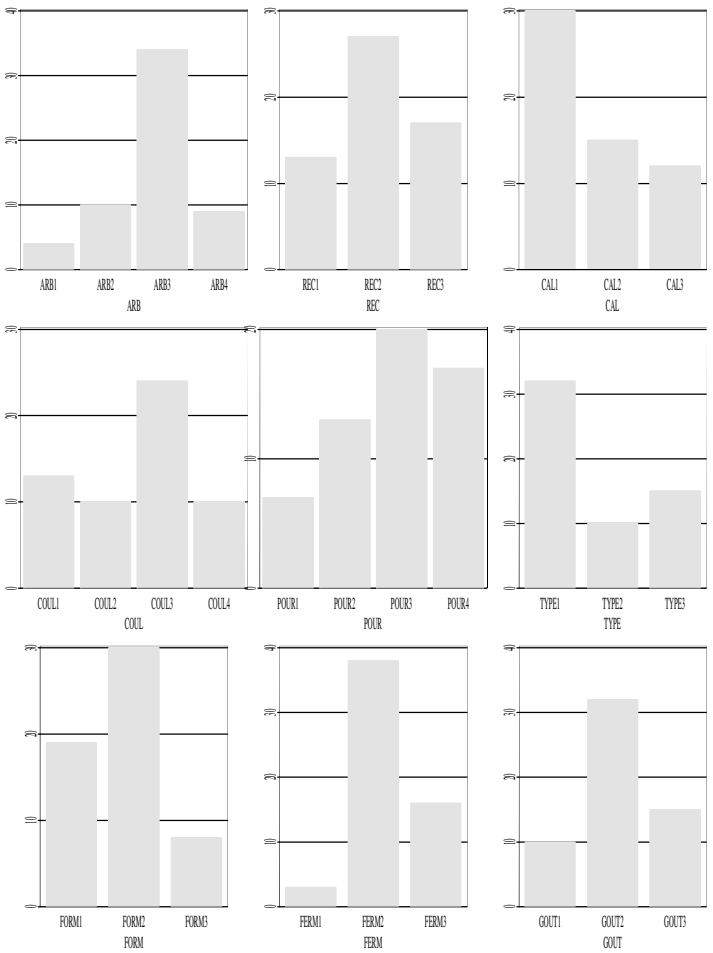
delete : détruit le catalogue dans le répertoire de travail. Le catalogue (*graphe*) doit contenir les graphiques réalisés par le reste du programme. Cette opération n'est pas utile la première fois que l'on lance le programme. Par contre elle est nécessaire si l'on lance plusieurs fois le programme.

Proc Gchart : tracé des graphiques avec la procédure *Gchart*. Tracé d'histogrammes avec barres verticales. Les graphiques sont stockés dans le catalogue *graphe*.

%inc : chargement d'une macro SAS *panels.sas* permettant de mettre plusieurs graphiques sur une même page.

%panels : utilisation de la macro qui trace 9 graphiques dans une grille de 3 lignes et 3 colonnes.

Histogrammes des modalités des différentes variables



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.

Facultatif : explication code

Les graphiques ont été obtenus avec le code suivant :

```
Goptions reset=all device=xcolor nodisplay target=lpmh
ftext=hwpsl005 htext=5;

pattern v=solid c=yellow;
axis1 label=none value=(a=90 h=5) minor=none;

Proc Gchart data=Pom gout=graphe;
 vbar arb rec cal coul pour type form ferm gout
      /autoref clipref frame raxis=axis1 width=14;
run;
quit;
```

ACM du tableau de Burt

```
* ACM du tableau de Burt: Proc Corresp avec l'option MCA;

Proc corresp data=Pom mca outc=graphem dimens=5;
 table arb rec cal coul pour type form ferm gout;
run;

The Correspondence Analysis Procedure

Inertia and Chi-Square Decomposition

Singular  Principal Chi-
Values    Inertias  Squares Percents    3    6    9   12   15
                                      ----+----+----+----+----+--
0.5641   0.31830   253.484  13.64% **********************
0.5005   0.25056   199.538  10.74% ******************
0.4909   0.24101   191.933  10.33% *****************
0.4412   0.19473   155.075   8.35% **************
0.4171   0.17397   138.546   7.46% ************
0.3858   0.14885   118.538   6.38% ***********
0.3714   0.13797   109.877   5.91% **********
0.3627   0.13157   104.776   5.64% *********
0.3285   0.10791    85.935   4.62% ********
0.3162   0.10003    79.661   4.29% *******
0.2974   0.08848    70.459   3.79% ******
0.2835   0.08042    64.043   3.45% ******
0.2642   0.06984    55.618   2.99% *****
...
0.1937   0.03755    29.901   1.61% ***
0.1702   0.02900    23.093   1.24% **
0.1184   0.01404    11.181   0.60% *
0.1041   0.01084     8.635   0.46% *
         -------   -------
         2.33333   1858.16 (Degrees of Freedom = 841)
```


L'ACM peut se faire au travers de l'analyse d'un tableau de Burt ou au travers d'un tableau d'indicatrices. Dans le premier cas, ne sont accessibles que les colonnes; dans le second cas, on a l'information sur les lignes et les colonnes.

Lorsque les effectifs sont très élevés, l'ACM du tableau de Burt est préférable car économe en calcul.

L'analyse factorielle des correspondances multiples se fait avec la procédure *corresp*. Les options utilisés sont :

data=	non du tableau utilisé.
outc=	nom du tableau dans lequel sont stockés un certain nombre d'informations permettant entre autre de tracer les graphes.
mca	indique que l'on analyse un tableau de Burt
dim=	le nombre d'axes retenus (ici 5).

La procédure produit une sortie très *verbeuse*.

En colonne on a :

Singular Values les valeurs singulières

Principal Inertias les valeurs propres (égales au carré des valeurs singulières) associées à chaque axe (inertie) et l'inertie totale.

Chi-Squares une décomposition du χ^2, sa valeur totale et le nombre de degrés de liberté.

Percents le pourcentage d'inertie associé à chaque axe.

SAS fournit aussi un éboulis des pourcentages d'inertie.

L'inertie associée au tableau est égale au (nombre de modalités - nombre de variables) / nombre de variables: $(30 - 9)/9 = 2.33$

```
Column Coordinates
             Dim1          Dim2          Dim3      ..     Dim5

ARB1      1.43593       0.86168       0.69122      ..  1.53986
ARB2     -0.36819       0.00814       1.00545      .. -0.85619
ARB3     -0.37983       0.06541      -0.27529      ..  0.21161
ARB4      1.20582      -0.63911      -0.38438      .. -0.53247
REC1     -0.84063      -0.23319      -0.58778      ..  0.91804
REC2     -0.23762       0.34715       0.54702      .. -0.36714

Summary Statistics for the Column Points
           Quality          Mass       Inertia

ARB1      0.449865      0.007797      0.044277
ARB2      0.438455      0.019493      0.039265
ARB3      0.417789      0.066277      0.019215
ARB4      0.438550      0.017544      0.040100
REC1      0.579459      0.025341      0.036759
REC2      0.619089      0.052632      0.025063

Partial Contributions to Inertia for the Column Points
              Dim1          Dim2          Dim3      ..       Dim5

ARB1      0.050509      0.023105      0.015457      ..   0.106272
ARB2      0.008302      0.000005      0.081763      ..   0.082137
ARB3      0.030040      0.001132      0.020841      ..   0.017058
ARB4      0.080140      0.028599      0.010755      ..   0.028591
REC1      0.056258      0.005500      0.036326      ..   0.122760
REC2      0.009336      0.025314      0.065344      ..   0.040778

Squared Cosines for the Column Points
               Dim1          Dim2          Dim3      ..       Dim5

ARB1      0.155616      0.056037      0.036059      ..   0.178956
ARB2      0.028844      0.000014      0.215091      ..   0.155971
ARB3      0.213270      0.006324      0.112032      ..   0.066194
ARB4      0.272625      0.076586      0.027702      ..   0.053160
REC1      0.208783      0.016066      0.102076      ..   0.249006
REC2      0.050816      0.108462      0.269304      ..   0.121314
```


Column Coordinates:

Coordonnées des colonnes sur les différents axes retenus (*Dim1*, ..., *Dim5*).

Summary Statistics for the Column Points:

Quality = Qualité de la représentation des colonnes dans le sous-espace de dimension 5. La valeur correspond à la somme des cosinus carrés (*Squared cosines*) de la modalité sur chacun des axes.

Mass = masse affectée au point-colonne.

Inertia = contribution du point-colonne à l'inertie totale.

Partial Contributions to Inertia for the Column Points:

Contribution du point-colonne à l'inertie de l'axe.

Squared Cosines for the Column Points:

Qualité de la représentation des points-colonnes par les différents axes.

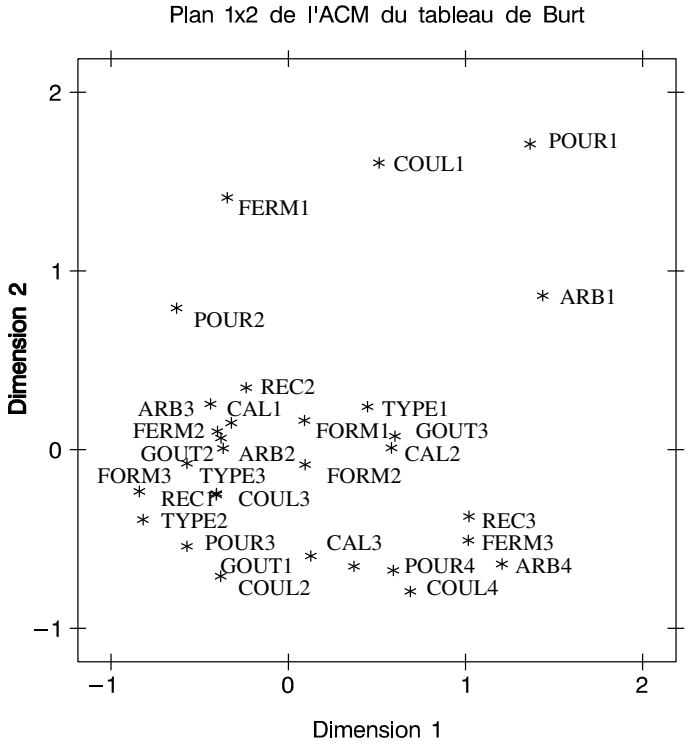
On peut analyser à ce stade les contributions des modalités des variables ou, de préférence, le faire en parallèle de la représentation graphique

Plan 1× 2 de l'ACM du tableau de BURT

```
%inc 'plotit.sas';

title "Plan 1x2 de l'ACM du tableau de Burt";

%plotit(data=graphem,plotvars=dim2 dim1,labelvar=_name_,
       datatype=mca,method=print,post=tmp.eps,options=square,
       labfont=HWpsl005,monochro=black,
       gopprint=dev=psepsf gsfmode=replace gaccess=gsasfile);
```



On utilise une macro SAS pour tracer le plan 1× 2 de l'ACM.

Seules les colonnes (modalités des variables) sont représentées.

- Représentativité médiocre sur le premier plan factoriel (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 !
- Contributions des modalités aux 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 *prec3, 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.
  - 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 spur des arbres très étalés (les modalités *arb1* et *arb4* sont bien représentées sur cet axe). On notera que quand une variété a un mutant 1/2 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.

pour les autre axes, non représentés, on a :

- 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.
- 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.
- 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 spur (*arb1*) des arbres à port très étalé (*arb4*). On notera que les mutants 1/2 spur sont curieusement très éloignés de la modalité *arb2* !

Création du tableau des indicatrices : Ind

```
Proc glmmod data=Pom outdesign=ind noprint;
 class arb rec cal coul pour type form ferm gout;
 model id=arb rec cal coul pour type form ferm gout /noint;
run;

Data Ind;
 merge Pom(keep= Nom) Ind(drop=id);
run;

Proc print data=ind(obs=10);run;
```

		C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C
		C	C	C	C	C	C	C	C	C	O	O	O	O	O	O	O	O	O
O	N	O	O	O	O	O	O	O	O	O	L	L	L	L	L	L	L	L	L
B	O	L	L	L	L	L	L	L	L	L	1	1	1	1	1	1	1	1	1
S	M	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8
1	Granny-spur	1	0	0	0	0	0	1	0	1	0	1	0	0	0	1	0	0	0
2	Granny-stand.	0	0	0	1	0	0	1	0	1	0	1	0	0	0	1	0	0	0
3	Boskoop-mut.	0	0	1	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1
4	Boskoop-stand.	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0	0	1	0
5	Gala-red	0	1	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0
6	Gala	0	0	1	0	0	1	0	1	0	0	0	0	1	0	0	1	0	0
7	Elstar	0	0	1	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0
8	Elista	0	0	1	0	1	0	0	1	0	0	0	0	1	0	0	0	1	0
9	Jonagold	0	0	1	0	0	1	0	0	0	1	0	0	1	0	0	1	0	0
10	Jonagored	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0

Transformation du Tableau d'origine *Pom* en Tableau d'indicatrices *ind*.

Pour cela on utilise la Proc *glmmod*. Cette procédure permet de créer une matrice d'indicatrices que l'on place dans un fichier *Ind*.

Ce tableau compte 30 colonnes, correspondant à l'ensemble des modalités des 9 variables de départ et 57 lignes correspondant aux variétés.

On concatène ce tableau avec le tableau de départ de façon à récupérer les nom des variétes.

Facultatif : explication code

L'option *outdesign=ind* indique le nom de la matrice d'indicatrices, et l'option *noprint* évite l'affichage de la matrice dans la fenêtre de sortie.

L'instruction *class* indique les variables pour lesquelles on veut construire des indicatrices.

On specifie le modèle tel qu'on pourrait l'écrire dans un ML avec l'instruction *model*. Avant le signe = on met la variable dépendante (ici la variable *id* qui correspond au numéro de chaque individu-variété) et après le signe =, on met les variables indépendantes (ici *arb rec cal coul pour type form ferm gout*). L'option *noint* évite d'avoir une première colonne à 1.

```
* AFC du tableau des Indicatrices;

Proc Corresp data=Ind outc=graphe dimens=5;
 var col1-col30;
 id nom;
run;

The Correspondence Analysis Procedure

Inertia and Chi-Square Decomposition

Singular  Principal Chi-
Values    Inertias  Squares Percents    3    6    9   12   15
                                     ----+----+----+----+----+---
0.5641   0.31830   163.290  13.64% **********************
0.5005   0.25056   128.539  10.74% ****************
0.4909   0.24101   123.640  10.33% ****************
0.4412   0.19473    99.897   8.35% *************
0.4171   0.17397    89.249   7.46% ************
0.3858   0.14885    76.360   6.38% ***********
0.3714   0.13797    70.781   5.91% **********
0.3627   0.13157    67.495   5.64% *********
0.3285   0.10791    55.358   4.62% ********
0.3162   0.10003    51.317   4.29% *******
0.2974   0.08848    45.389   3.79% ******
0.2835   0.08042    41.255   3.45% ******
0.2642   0.06984    35.828   2.99% *****
...
0.1937   0.03755    19.262   1.61% ***
0.1702   0.02900    14.876   1.24% **
0.1184   0.01404     7.203   0.60% *
0.1041   0.01084     5.562   0.46% *
         -------   -------
         2.33333      1197 (Degrees of Freedom = 1624)
```


L'ACM est faite ici par l'AFC du tableau des indicatrices (*Ind*). Dans ce second cas, on obtient l'information sur les lignes et les colonnes.

On utilise la procédure *Corresp*. Les options utilisés sont :

data=	non du tableau utilisé.
outc=	nom du tableau dans lequel sont stockés un certain nombre d'informations permettant entre autre de tracer les graphes.
dim=	le nombre d'axes retenus (ici 5).

La procédure produit une sortie très «verbeuse».

En colonne on a :

Singular Values les valeurs singulières

Principal Inertias les valeurs propres (égales au carré des valeurs singulières) associées à chaque axe (inertie) et l'inertie totale.

Chi-Squares une décomposition du χ^2, sa valeur totale et le nombre de degrés de liberté.

Percents le pourcentage d'inertie associé à chaque axe.

SAS fournit aussi un éboulis des pourcentages d'inertie.

L'inertie associée au tableau *ind* est la même que celle associée à l'analyse du tableau *Pom* : 2.33.

```
Row Coordinates

                    Dim1        Dim2        Dim3     ...      Dim5

Granny-spur      1.39327     0.78567    -0.47838     ...   0.52196
Granny-stand.    1.34795     0.45253    -0.72181     ...  -0.03008
Boskoop-mut.     0.41507    -0.39317     0.64964     ...  -0.03076
Boskoop-stand.   0.18816    -0.37585     0.68301     ...  -0.43833
Gala-red        -0.51340    -0.22220     0.23002     ...  -0.54687

Summary Statistics for the Row Points

                   Quality          Mass       Inertia

Granny-spur       0.769807      0.017544      0.030054
Granny-stand.     0.817156      0.017544      0.023440
Boskoop-mut.      0.389304      0.017544      0.014950
Boskoop-stand.    0.380406      0.017544      0.017891
Gala-red          0.438244      0.017544      0.015736

Partial Contributions to Inertia for the Row Points

                    Dim1        Dim2        Dim3     ...      Dim5

Granny-spur      0.106992    0.043220    0.016658     ...  0.027474
Granny-stand.    0.100145    0.014339    0.037926     ...  0.000091
Boskoop-mut.     0.009496    0.010824    0.030720     ...  0.000095
Boskoop-stand.   0.001951    0.009891    0.033957     ...  0.019375
Gala-red         0.014528    0.003457    0.003851     ...  0.030158

Squared Cosines for the Row Points

                    Dim1        Dim2        Dim3     ...      Dim5

Granny-spur      0.485647    0.154428    0.057252     ...  0.068160
Granny-stand.    0.582828    0.065689    0.167126     ...  0.000290
Boskoop-mut.     0.086648    0.077747    0.212254     ...  0.000476
Boskoop-stand.   0.014879    0.059366    0.196049     ...  0.080747
Gala-red         0.125942    0.023590    0.025280     ...  0.142894
```

Row Coordinates :

Coordonnées des lignes sur les différents axes retenus (*Dim1* ... *Dim5*).

Summary Statistics for the Row Points :

Quality = Qualité de la représentation des lignes dans le sous-espace de dimension 5. La valeur correspond à la somme des cosinus carrés (*Squared cosines*) de la modalité sur chacun des axes.

Mass = masse affectée au point-ligne.

Inertia = contribution du point-ligne à l'inertie totale.

Partial Contributions to Inertia for the Row Points :

Contribution du point-ligne à l'inertie de l'axe.

Squared Cosines for the Row Points :

Qualité de la représentation des points-lignes par les différents axes.

Column Coordinates

	Dim1	Dim2	Dim3	...	Dim5
COL1	1.43593	0.86168	0.69122	...	1.53986
COL2	-0.36819	0.00814	1.00545	...	-0.85619
COL3	-0.37983	0.06541	-0.27529	...	0.21161
COL4	1.20582	-0.63911	-0.38438	...	-0.53247
COL5	-0.84063	-0.23319	-0.58778	...	0.91804
COL6	-0.23762	0.34715	0.54702	...	-0.36714

Summary Statistics for the Column Points

	Quality	Mass	Inertia
COL1	0.449865	0.007797	0.044277
COL2	0.438455	0.019493	0.039265
COL3	0.417789	0.066277	0.019215
COL4	0.438550	0.017544	0.040100
COL5	0.579459	0.025341	0.036759
COL6	0.619089	0.052632	0.025063

Partial Contributions to Inertia for the Column Points

	Dim1	Dim2	Dim3	...	Dim5
COL1	0.050509	0.023105	0.015457	...	0.106272
COL2	0.008302	0.000005	0.081763	...	0.082137
COL3	0.030040	0.001132	0.020841	...	0.017058
COL4	0.080140	0.028599	0.010755	...	0.028591
COL5	0.056258	0.005500	0.036326	...	0.122760
COL6	0.009336	0.025314	0.065344	...	0.040778

Squared Cosines for the Column Points

	Dim1	Dim2	Dim3	...	Dim5
COL1	0.155616	0.056037	0.036059	...	0.178956
COL2	0.028844	0.000014	0.21509	...	0.155971
COL3	0.213270	0.006324	0.112032	...	0.066194
COL4	0.272625	0.076586	0.027702	...	0.053160
COL5	0.208783	0.016066	0.102076	...	0.249006
COL6	0.050816	0.108462	0.269304	...	0.121314

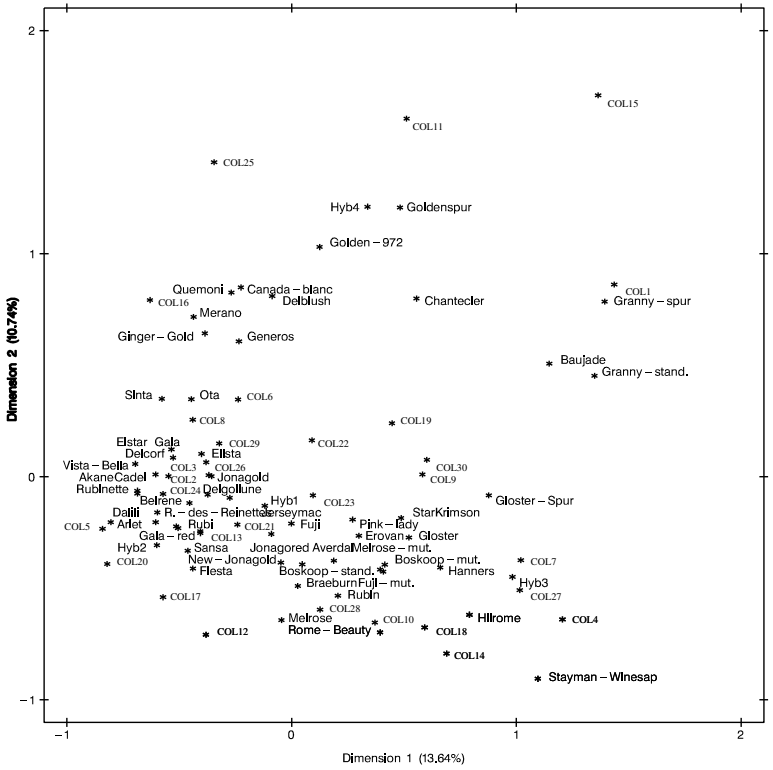
Mêmes explications que pour les lignes. Les informations données pour les colonnes sont les mêmes que celles fournies par l'ACM du tableau de Burt avec *COL1 = ARB1*, *COL2 = ARB2*, ...

Plan 1× 2 de l'AFC du tableau d'indicatrices

```
title "Plan 1x2: representation simultanee";

%plotit(data=graphe,plotvars=dim2 dim1,labelvar=nom,
        datatype=corresp,method=print,post=AFC12.eps,
        options=square,labfont=HWpsl005,monochro=black,
        gopprint=dev=psepsf gsfmode=replace gaccess=gsasfile);
```

Plan 1× 2 : représentation simultanée



On utilise une macro SAS pour tracer le plan 1× 2 de l'AFC.

Seules les colonnes sont représentées.

L'information relative aux modalités des variables est la même que celle obtenue avec l'ACM du tableau de Burt.

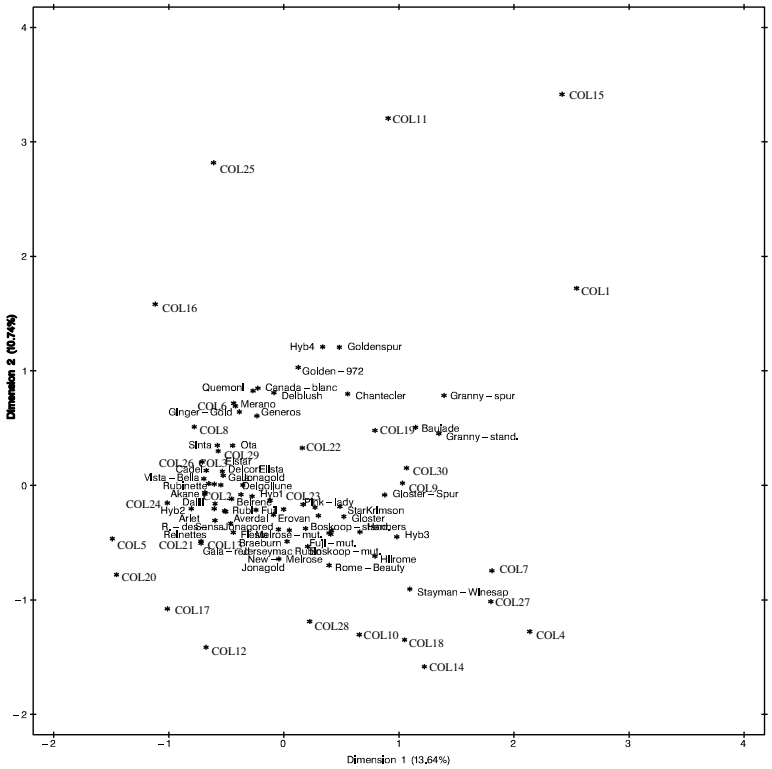
Pour les lignes (variétés), on a :

- Certaines variétés sont mal représentées dans le premier plan principal (*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.* et *Fuji.*
- 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. À 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.

Autres représentations

```
Proc Corresp data=Ind outc=graphe noprint profile=row;
 var col1-col30;
 id nom;
run;

%plotit(data=graphe,plotvars=dim2 dim1,labelvar=nom,
        datatype=corresp,options=square,
        method=gplot,gopplot=device=xcolor);
```



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

Rappel :

- représentation isométrique ligne → représentation exacte des individus
- représentation isométrique colonne → représentation exacte des variables
- représentation simultanée → représentations superposables

Globalement il y a peu de différences de représentation mais les échelles sont différentes.

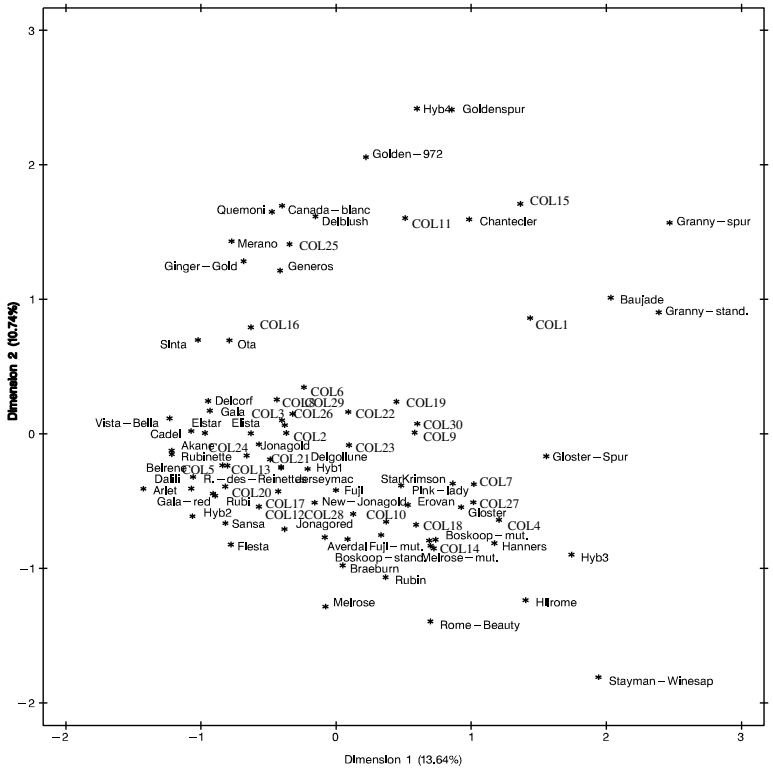
Facultatif : explication code

Pour l'impression :

```
%plotit(data=graphe,plotvars=dim2 dim1,labelvar=nom,
        datatype=corresp,method=print,post=AFCl12.eps,
        options=square,labfont=HWpsl005,gopprint=dev
        =psepsf gsfmode=replace gaccess=gsasfile,
        monochro=black,labsize=1.5,tsize=1.2);
```

```
Proc Corresp data=Ind outc=graphe noprint profile=column;
 var col1-col30;
 id nom;
run;

%plotit(data=graphe,plotvars=dim2 dim1,labelvar=nom,
        datatype=corresp,options=square,
        method=gplot,gopplot=device=xcolor);
```



Facultatif : explication code

Pour l'impression :

```
%plotit(data=graphe,plotvars=dim2 dim1,labelvar=nom,
        datatype=corresp,method=print,post=AFCc12.eps,
        options=square,labfont=HWpsl005,gopprint=dev
        =psepsf gsfmode=replace gaccess=gsasfile,
        monochro=black,labsize=1.2,tsize=1);
```

Pour aller plus loin ...

Pour détecter les affinités 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.

On peut procéder de la même façon pour les individus-variétés en reprenant le résultat de l'AFC sur le tableau des indicatrices pour être renseigné sur les individus.

Classification hiérarchique des modalités

```
Proc corresp data=Pom dimens=21 noprint
     outc=graphem(keep=Dim1-Dim21 _name_ _type_);
 table nom, arb rec cal coul pour type form ferm gout;
run;

Proc print data=Graphem(where=(_type_='VAR'));
run;
```

OBS	_TYPE_	_NAME_	DIM1	DIM2	DIM3	DIM4
59	VAR	ARB1	1.43593	0.86168	0.69122	-0.55440
60	VAR	ARB2	-0.36819	0.00814	1.00545	0.42558
61	VAR	ARB3	-0.37983	0.06541	-0.27529	-0.11623
62	VAR	ARB4	1.20582	-0.63911	-0.38438	0.21261
63	VAR	REC1	-0.84063	-0.23319	-0.58778	0.10927
64	VAR	REC2	-0.23762	0.34715	0.54702	-0.27727
65	VAR	REC3	1.02022	-0.37303	-0.41931	0.35681
66	VAR	CAL1	-0.43929	0.25595	-0.27714	-0.10646

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 de χ^2, on peut donc travailler sur la matrice des distances euclidiennes entre modalités dans cet espace.

On peut utiliser les coordonnées des modalités obtenues à partir de l'ACM des indicatrices ou bien à partir de l'ACM du tableau de Burt.

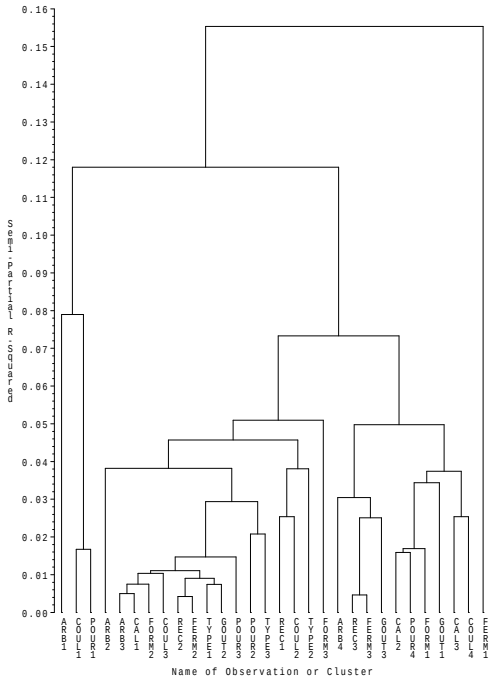
Ici on a choisi de faire l'ACM à partir des indicatrices afin d'avoir les informations sur les individus et les variables.

Dans ce qui suit on ne traitera que les variables mais les raisonnements sont identiques pour les individus.

Classification hiérarchique des modalités

```
Proc cluster data=Graphem(where=(_type_='VAR'))
     outtree=arbre method=ward ;
 var Dim1-dim21;
 id _name_;
run;

Proc tree data=arbre n=5 out=tree graphics;
 id _name_;
 copy dim1-dim21;
run;
```



La procédure *Cluster* effectue une classification ascendante hiérarchique en utilisant le critère du saut de WARD qui cherche à minimiser la perte d'inertie inter à chaque regroupement.

On ne garde que les observations qui correspondent aux variables (*_type_='VAR'*).

La procédure *Tree* permet de tracer le dendrogramme (*graphics*). L'examen du graphique permet de déterminer le nombre de classes à retenir.

Ici, on peut couper l'arbre en 3 ou 5 classes.

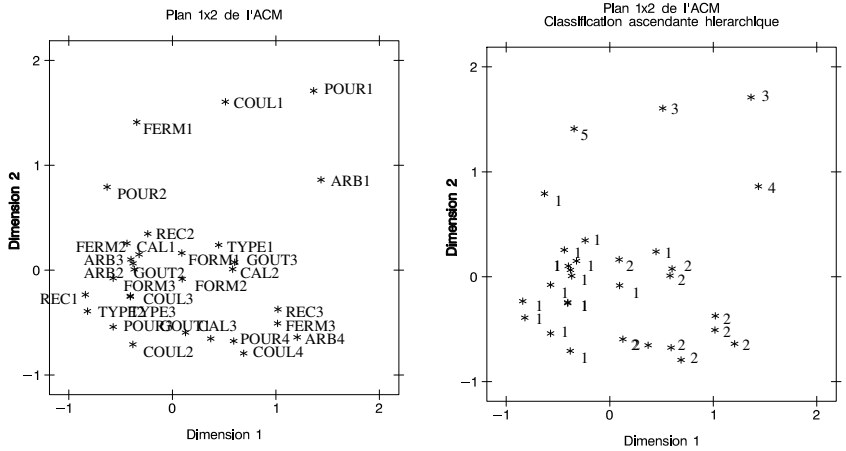
On observe que la modalité *ferm1* est très différentes des autres modalités. Elle forme un groupe à elle seule. La modalité *arb1* a aussi un comportement différent. Ces 2 modalités correspondent aux modalités les plus rares.

On indique qu'on veut 5 classes en utilisant l'option *n=5* et on met le résultat dans un fichier *out=tree*.

Représentation de la classification sur le plan 1×2 de l'ACM

```
title "Plan 1x2 de l'ACM du tableau de Burt";

%plotit(data=tree,plotvars=dim2 dim1,labelvar=_name_,
       datatype=mca,method=gplot,gopplot=device=xcolor,
       options=square);
%plotit(data=tree,plotvars=dim2 dim1,labelvar=cluster,
       datatype=mca,method=gplot,gopplot=device=xcolor,
       options=square);
```

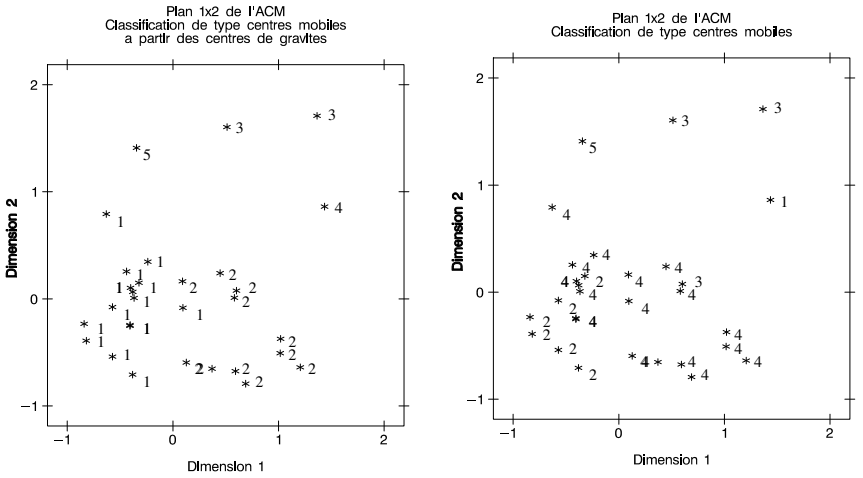


L'arbre de classification comme la représentation dans le plan principal font ressortir l'originalité des modalités *ferm1* et *arb1*, ainsi que la proximité et l'originalité du couple (*pour1, coul1*) qui caractérisent les variétés sans coloration rouge.

À part les classes formées par *arb1*, *ferm1* et le couple (*pour1*, *coul1*), les autres classes s'identifient mal dans un ensemble de modalités peu structuré.

Classification par les centres mobiles

```
Proc means data=Tree noprint;
 class cluster;
 var dim1-dim21;
 output out=moy(where=(_type_=1)) mean=;
run;
Proc fastclus data=Graphem(where=(_type_='VAR'))
              seed=moy mean=moy2 out=mixte maxc=6 maxiter=7;
 var dim1-dim21;
run;
```



Pour améliorer la classification obtenue, on procède à une nouvelle classification par l'algorithme des centres mobiles à partir des 5 pôles qui sont les centres de gravité des classes obtenues par la classification précédente.

On n'observe aucune amélioration de la partition.

On peut aussi effectuer une classification directement à partir des coordonnées sans indiquer de centres de gravité. On retouve les 3 classes individualisées précédemment : *ferm1*, *arb1* et le couple (*pour1*, *coul1*).

Mais les 2 autres classes sont modifiées.

On peut aussi voir ce que donne la répartition en 3 classes.

On peut aussi faire le même genre d'opération sur les individus.