Analyse de Données

Extrait du cours à l'ISS

Prof. KIZUNGU Vumilia Roger

Introduction

Beaucoup d'entreprises, beaucoup de chercheurs disposent de beaucoup de données dont les analyses restent souvent rudimentaires. Il suffit de penser aux données accumulées par les programmes de sélection et d'amélioration des cultures de l'INERA et des universités. Il suffit d'observer les registres de la Direction Générale des Migrations (DGM) remplies à chacun de nos passages à l'aéroport. Il suffit de considérer les données saisies chaque jour par les services des urgences des hôpitaux, les morgues, les maternités. Il suffit d'imaginer les données issues des milliers d'enquêtes organisées par les services spécialisés de l'Etat, par les entreprises de téléphonie etc.

Une enquête sur les rapports et les travaux de recherche issues de ces données montre que la description populaire des caractères qualitatifs un à un, reste encore aujourd'hui la présentation des tableaux de distribution de fréquence qui donnent des probabilités et des pourcentages. La même enquête montre que les chercheurs qui analysent les mêmes types des caractères deux à deux, s'intéressent à l'indépendance ou non de deux caractères par le tableau de contingence. En ce qui concerne les données quantitatives, les études se limitent souvent à la corrélation

Cette lecture va vous montrer comment décrire simultanément plusieurs caractères. L'Analyse de Données est le nom dédié aux analyses descriptives multivariées. C'est tout simplement le prolongement de la Statistique Descriptive Classique.

A la fin de la lecture, vous devriez être capable, à l'aide de l'Analyse en Composantes Principales, premièrement, de grouper les produits, les attitudes des clients, les accessions suivant leurs ressemblances et leurs dissemblances. Deuxièmement, de juger si les caractères mesurés sur ces produits sont corrélés ou pas. Troisièmement, de mettre en évidence la structuration des réponses en montrant le regroupement des individus selon des combinaisons de réponses aux questions prises en compte.

Vous apprendrez à élaborer une typologie ou une Classification Hiérarchique. Vous apprendrez donc à répartir la population d'une enquête en un nombre défini de sous-groupes aussi différents que possibles les uns des autres et dans lesquels les individus sont aussi semblables que possible entre eux. Vous apprendrez la méthode des moyennes mobiles (K-means) et la méthode des nuées dynamiques. La première méthode permet d'effectuer plusieurs combinaisons des individus. Elle calcule chaque combinaison la variance entre groupes et la variance dans les groupes. Elle retient la combinaison correspondant à une variance dans les groupes minimale et la variance entre groupes maximale. Les deux algorithmes sont proches à la différence que le deuxième part d'une sélection d'un noyau d'individus au lieu de prendre des individus isolés pour constituer les partitions de démarrage (ce qui est censé donner de meilleurs résultats). En final, les calculs itératifs des analyses typologiques aboutissent au classement des individus dans le nombre de groupes défini initialement. L'effectif de ces groupes peut être très différent.

Comme la typologie, la classification est une méthode de regroupement des individus selon leurs ressemblances. La différence est que le nombre de groupes n'est pas à fixer a priori et que le résultat est représenté sous la forme d'un arbre de classification. L'élaboration de cet arbre peut être ascendante (méthode la plus fréquemment utilisée), par regroupements successifs des individus ou descendante, par divisions successives.

L'arbre de classification relie un individu à un autre ou à un sous-groupe d'individu issus eux-mêmes de regroupements. Lorsque l'on coupe l'arbre au niveau du dernier regroupement, on obtient deux groupes d'individus. Si la division est effectuée au niveau de l'avant-dernier regroupement, on obtient trois groupes.

La classification s'opère en trois étapes. La première étape consiste en la sélection et préparation des variables. La deuxième consiste au choix de la mesure des distances et la troisième au choix de la mesure d'agrégation.

Dans le cas des mesures, souvent puisque les échelles sont différentes, on les centre et on utilise la distance euclidienne. Deux procédures d'agrégation sont considérées : la procédure hiérarchique ou la procédure de Ward et la procédure non hiérarchique ou la procédure K-means

Le lecteur apprendra l'analyse factorielle des correspondances (Benzecri 1973) qui s'applique à deux variables qualitatives (nominales). Elle permet de positionner sur un plan les modalités de réponses des deux questions. L'analyse des correspondances Multiples (ACM) généralise l'AFC à un nombre quelconque de variables et permet donc de représenter sur le même plan les modalités de réponses de plus de deux variables.

Comme pour l'ACP, le but de ces analyses est de dégager des dimensions cachées contenues dans les réponses aux variables sélectionnées, pour faciliter l'interprétation de tableaux pas toujours lisibles au départ.

Nous allons analyser 6 cas. Le premier est celui de la segmentation dans les marchés industriels au stade de la maturité : une application aux tubes en plastique de l'arbitrage entre le prix et le service (Chéron E. et al, 1996). Le deuxième cas sera celui du choix d'un noyau d'accessions de riz de départ pour un programme de sélection du riz à faible teneur en eau commandée par les Tetela du Nord de la province du Kasaï Oriental. Le troisième cas analyse la notoriété des réseaux téléphoniques, cas des réseaux imaginaires à Kinshasa. Le quatrième cas vient d'une enquête où l'on s'est posé la question de savoir quels marchés fréquentent les sénateurs, les députés, les fonctionnaires etc. Le cinquième cas concerne l'étude de l'eau de puits par un laboratoire d'éco-toxicologie (mémoire Ighera, FACAGRO, UNIKIN).

Les analyses seront effectuées avec le logiciel  qui peut être détenu par tout le monde. Chacun pourra refaire la même chose avec le logiciel dont il dispose et comparer les résultats.

Problème 1 : La segmentation dans les marchés industriels au stade de la maturité : une application aux tubes en plastique de l'arbitrage entre le prix et le service (Chéron E. et al, 1996)



1. Contexte

Cette lecture applique l'Analyse de Données en marketing et en agronomie. En effet, les chefs d'entreprises se posent comment classer et classifier les produits, comment segmenter les marchés en vue de mieux les cibler et de mieux s'y positionner, comment analyser l'environnement d'un marché, comment comprendre la consommation et le comportement d'achat chez les clients, comment analyser la concurrence.

Les entrepreneurs font toujours l'évaluation des besoins en ressources humaines et financières à affectées à la publicité sur leur produit. Ils étudient ou mieux segmentent leur clientèle selon leur importance ou selon les bénéfices.

La littérature renseigne que la segmentation par avantages recherchés par la clientèle est plus parlante que la segmentation par secteur d'activité (Wind, 1978, Cardozo, 1980, Moriarty et al 1986). Les avantages recherchés sont, pour un produit en maturité, le prix et le service relatif au prix (publicité, visites etc).

Il arrive que le consommateur affiche un comportement en termes d'arbitrage entre le service et le prix. Pour un marché en maturité Ragan *et al* ont proposé un modèle qui permet de le segmenter en fonction du prix et de service y relatif.

2. Les individus ou les sujets d'étude

Dans ce problème, un bureau d'études en marketing a été consulté pour segmenter le marché des tubes en plastiques sur une population de 96 clients. Les données qui

composent l'étude sont les données secondaires internes, les données issues d'un questionnaire administré aux représentants et aux gestionnaires de l'entreprise.

3. Les variables ou les caractères des individus

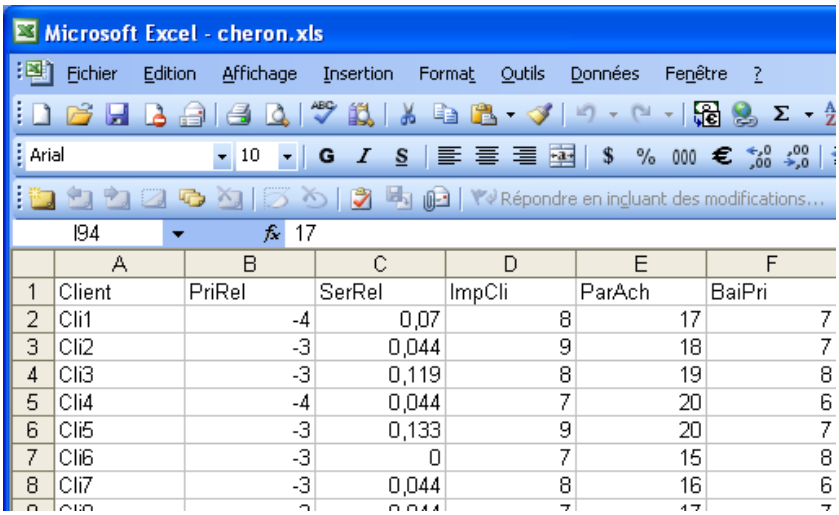
Après dépouillement des données, 12 variables ont été retenues :

1. Prix relatif en pourcentage (PriRel) : il mesure le prix qu'un client a payé en relation avec le prix minimum et le prix maximum que les autres clients de la division ont payés.
2. Service relatif de 0 à 1 (SerRel) : c'est la mesure du service obtenu par un client comparativement aux autres. Ce chiffre est obtenu après fusion de l'information sur le nombre de visites par client, le nombre des plaintes enregistrés durant l'année par le client et les dépenses de promotion durant l'année.
3. Importance du client en milliers de dollars : c'est le niveau des achats d'un client.
4. Part d'achats en pourcentage : le pourcentage des ventes qu'un fournisseur détient comparativement aux achats totaux du client pour un groupe de produits donné. Il s'agit d'une mesure de la réputation du fournisseur et de la stratégie d'achat choisie par le client en termes de nombre de fournisseurs différents qui seront retenus et du pourcentage des achats qui sera attribué à chacun d'eux.
5. Impact d'une baisse de prix en pourcentage. C'est la réponse à la question suivante au fournisseur : " Si vous pouviez baisser votre prix de 5%, quelle serait, selon votre évaluation, la hausse (baisse) du volume d'achat de vos clients? "
6. Impact d'une hausse de prix en pourcentage
7. Impact d'une baisse de service en pourcentage
8. Impact d'une hausse de service en pourcentage
9. Importance du produit de 1 à 5 : Les représentants devaient évaluer l'importance du produit pour le client sur une échelle de 1 à 5, où : 1 = Pas important; 2 = Peu important; 3 = Importance moyenne; 4 = Assez important et 5 = Très important.
10. Infidélité par rapport au fournisseur de 1 à 5 : la possibilité de changer de fournisseur fut évaluée par les représentants sur une échelle de fidélité de 1 à 5, où : 1 = Pas fidèle; 2 = Peu fidèle; 3 = Moyennement fidèle; 4 = Assez fidèle; et 5 = Très fidèle.
11. Connaissance du marché 1 à 5 : dans ce contexte, les représentants devaient évaluer la connaissance du marché par le client sur une échelle de 1 à 5 où : 1 = Pas de connaissance; 2 = Peu de connaissance; 3 = Connaissance moyennement; 4 = Bonne connaissance; et 5 = Très bonne connaissance.
12. Processus décisionnel d'achat de 1 à 5 : Les représentants devaient évaluer le processus décisionnel d'achat sur une échelle de 1 à 5 où : 1 = Pas complexe; 2 = Peu complexe; 3 = Moyennement complexe; 4 = Assez complexe; et 5 = Très complexe.

4. Tableau de données

Le tableau de données est saisi en Excel et porte le nom de cheron.xls. Il est sur le disque dur (c:/), dans un répertoire appelé « DATAR » et dans un sous répertoire « MRA ». Il peut se trouver sur n'importe quel support de stockage de données. Pour connaître le chemin qui va jusqu'à votre fichier, il suffit de cliquer avec le bouton droit de la souris sur le nom du fichier et le chemin s'affiche dans la fenêtre des propriétés. Pour prendre en charge ce tableau de données avec le logiciel R, on procède en plusieurs étapes.

Etape 1. Ouvrir le fichier avec Excel. Son extrait est donné par la figure 1.



	A	B	C	D	E	F
1	Client	PriRel	SerRel	ImpCli	ParAch	BaiPri
2	Cli1	-4	0,07	8	17	7
3	Cli2	-3	0,044	9	18	7
4	Cli3	-3	0,119	8	19	8
5	Cli4	-4	0,044	7	20	6
6	Cli5	-3	0,133	9	20	7
7	Cli6	-3	0	7	15	8
8	Cli7	-3	0,044	8	16	6

Figure 1. Extrait du tableau de données cheron.xlsx

Les données sont saisies sous forme individus en lignes et variables en colonnes. Les individus en étude sont Cli1, Cli2, …,Cli96. Les variables ou les caractères mesurés sur ces clients sont PriRel, SerRel… La première ligne du tableau contient les noms des variables ou les entêtes. Les observations commencent à la deuxième ligne.

Les logiciels statistiques demandent toujours à l'utilisateur de spécifier si la première ligne est celle des entêtes ou pas. Si oui, alors le logiciel va lire les données à partir de la deuxième ligne. Si non, il lit les données à partir de la première ligne.

Etape 2. Enregistrer le fichier Excel sous le type fichier ***texte (séparateur tabulation)*.txt*** (Fichier/Enregistrer sous…) (fig. 2 et 3).

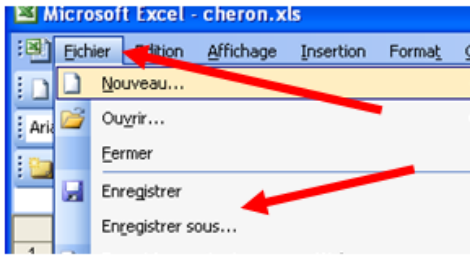


Figure 2. Processus de sauvegarde

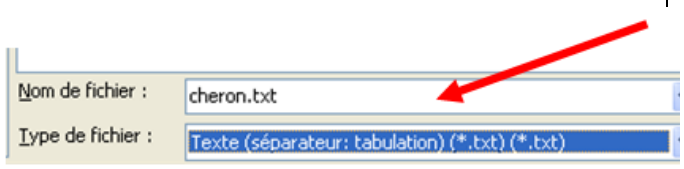


Figure 3. Choix du type de fichier

Le type texte est le type le plus reconnu par plusieurs logiciels statistiques. C'est le plus simple. On le dit aussi type codé en ASCII. Le logiciel R peut alors récupérer ce nouveau fichier. Le fichier Exel reste intacte à son emplacement. Au même emplacement on a un autre fichier qui porte le même nom que celui de Excel mais dont l'icône n'est pas celui de excel et l'extension devient « .txt ».

5. Récupération des données comme objet de R

Le logiciel est, avant d'être un auxiliaire d'analyse de données, un langage de programmation orienté objets. En d'autres termes, dans ce langage, tout est objet. Un vecteur, une matrice, un tableau de données, une moyenne d'une variable, le résultat d'une analyse, tous sont des objets.

Nous allons créer un objet appelé toto. Cet objet de R contiendra les données importées du fichier dont le chemin complet est "c:/DATAR/cheron.txt". La commande pour importer le tableau est dans le jargon de R, « read.data() ». Entre les deux parenthèses, nous donnons les arguments de la fonction. Le premier argument est le nom complet du fichier à importer entre guillemets. Ensuite, l'information selon laquelle la première ligne est celle des entêtes (« header=TRUE ». Nous indiquons aussi que le tableau a comme séparateur des colonnes, le caractère tabulation (« \t ». Le tableau des données a été saisi avec des virgules comme séparateurs des décimaux. Il faut le préciser à R car lui fonctionne par défaut avec le point comme séparateur des décimaux. Enfin nous mentionnons que la première colonne doit être considérée comme la colonne des noms des individus (row.names=1)

```
toto=read.table("c:/DATAR/cheron.txt",h=TRUE,sep="\t",dec=",",
row.names=1)
```

Le tableau dans R s'organise en ligne et en colonnes indicées. L'élément toto[i,j] donne la réponse du client i à la question j. Suivant cette logique les éléments des lignes 1, 2, …,5 désignées par le vecteur 1 :5 et des colonnes 1, 2, …, 8 désignées par le vecteur 1 :8 sont repris dans le tableau 1

Le tableau 1 donne les 2 premières observations sur les 8 premières variables.

Tableau 1. extrait du tableau cheron

```
toto[1:2,1:8]
  PriRel SerRel ImpCli ParAch BaiPri HauPri BaiSer HauSer
1     -4  0.070      8     17      7    -10     -2      0
2     -3  0.044      9     18      7    -12     -2      0
```

L'objet toto est un tableau appelé « Data Frame » dans le jargon de R. C'est une juxtaposition des vecteurs qui ont chacun une entête. La dimension de ce tableau est donnée par la commande dim.

```
dim(toto)
[1] 96 12
```

La dimension est de 96 lignes et 12 colonnes. L'enquête a consisté en 96 fiches de 12 questions chacune. On a interrogé 96 clients à qui on a posé 12 questions.

6. Analyses préliminaires à une dimension

Le fait de passer à l'analyse de données suppose que le tableau a été au préalable été nettoyé. Une partie importante de ce nettoyage se fait à l'aide du logiciel Microsoft Excel. Le logiciel R offre les commandes qui permettent de faire les analyses élémentaires à une dimension.

6.1. Le résumé des données

La commande summary() donne simultanément les statistiques suivantes : le minimum, le premier quartile, la médiane, la moyenne, le troisième quartile et le maximum.

```
summary(toto)
     PriRel           SerRel           ImpCli
 Min.   :-4.000   Min.   :0.00000   Min.   : 7.00
 1st Qu.:-2.000   1st Qu.:0.09125   1st Qu.: 9.00
 Median : 4.000   Median :0.19600   Median :10.00
 Mean   : 2.458   Mean   :0.22393   Mean   :19.20
 3rd Qu.: 5.000   3rd Qu.:0.33700   3rd Qu.:18.00
 Max.   : 8.000   Max.   :0.85900   Max.   :67.00
```

Cette étape est très importante non seulement pour l'interprétation mais aussi pour la vérification du bon nettoyage des données. En effet, une variable quantitative aura un résumé par les cinq paramètres cités. Une variable qualitative aura pour résumé une table qui donne la distribution des fréquences des modalités respectives.

Pour le tableau que nous étudions, la variable prix relatif (PriRel) varie de -4 % à 8 %. 25 % des clients ont un prix relatif inférieur ou égal à –2% (1st Qu.). 50 % des clients ont un prix relatif inférieur ou égal à 4%. 75 % des clients ont un prix relatif inférieur ou égal à 5%. (3rd Qu). Le prix relatif moyen est égal à 2.458%.

Concernant l'importance des clients, le moins important achète pour 7000 $. Le plus important achète pour 67000 $.

6.2. La moyenne et l'écart-type

La moyenne et l'écart-type peuvent aussi être calculés à part respectivement par les commandes moyenne() et sd(). Les résultats sont mis sous forme d'un tableau par la commande data.frame().

```
Moyenne=mean(toto)
Ecartype=sd(toto)
parametres=
data.frame(Moyenne,Ecartype)
parametres
```

```
        Moyenne Ecartype
PriRel    2.458    3.221
SerRel    0.224    0.158
ImpCli   19.198   17.937
ParAch   46.354   20.670
BaiPri   20.188    5.599
```

```
HauPri -36.052   10.608
BaiSer -26.146   10.520
HauSer   5.885    5.784
ImpPro   3.125    0.441
FidFou   2.938    0.949
ConMar   3.448    0.738
ProDec   2.469    0.522
```

6.3. Analyses préliminaires à deux dimensions

Le tableau ci-dessous obtenu avec la commande cor() donne les coefficients de corrélation entre les variables prises deux à deux. Pour représenter les données avec deux chiffres après la virgule, il suffit d'encadrer la commande cor() par la commande round().

```
round(cor(toto),2)
```

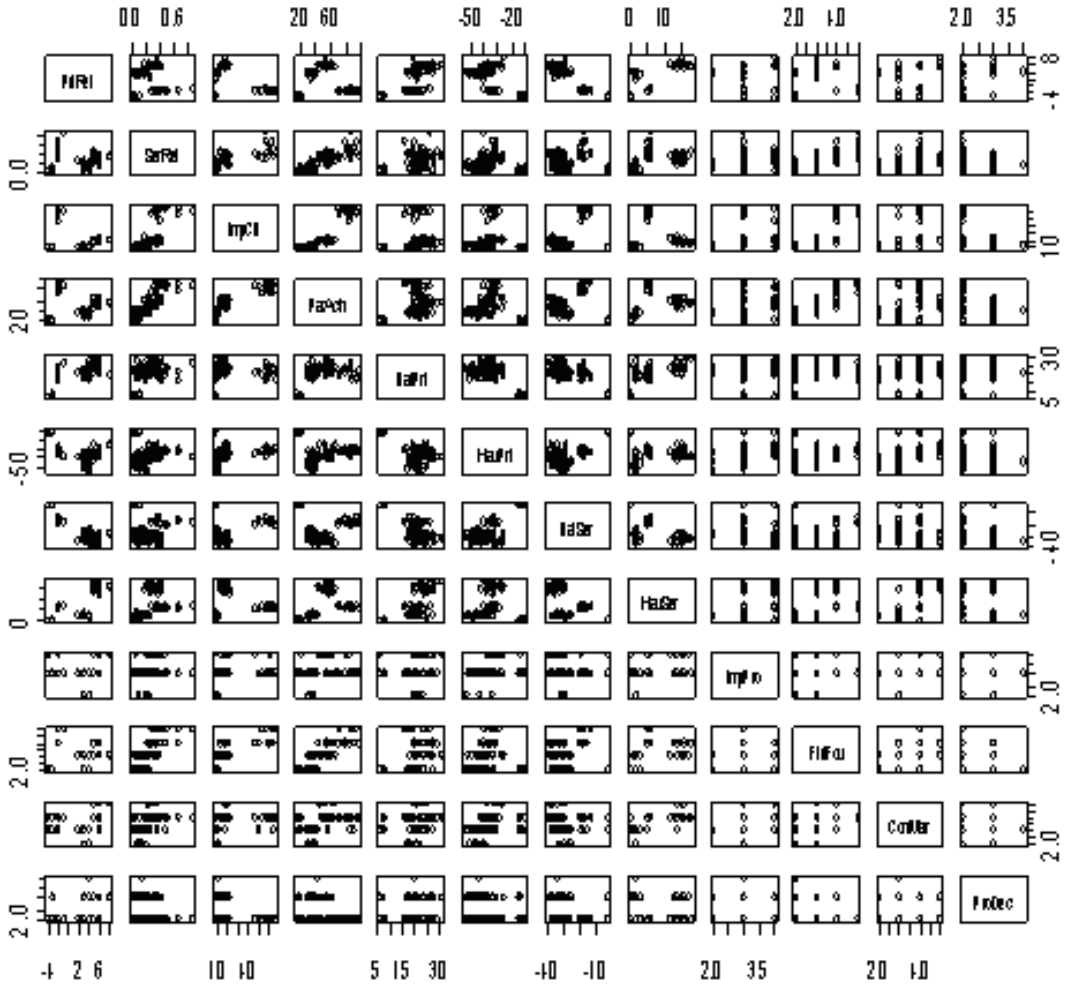
```
        PriRel SerRel ImpCli ParAch BaiPri HauPri BaiSer HauSer
PriRel    1.00  -0.03  -0.46  -0.09    0.69  -0.59  -0.90   0.54
SerRel   -0.03   1.00   0.74   0.83    0.44  -0.01  -0.09   0.54
ImpCli   -0.46   0.74   1.00   0.89    0.17   0.12   0.24   0.17
ParAch   -0.09   0.83   0.89   1.00    0.48  -0.08  -0.11   0.48
BaiPri    0.69   0.44   0.17   0.48    1.00  -0.65  -0.79   0.60
HauPri   -0.59  -0.01   0.12  -0.08   -0.65   1.00   0.74   0.02
BaiSer   -0.90  -0.09   0.24  -0.11   -0.79   0.74   1.00  -0.45
HauSer    0.54   0.54   0.17   0.48    0.60   0.02  -0.45   1.00
ImpPro    0.18   0.26   0.10   0.22    0.14   0.23  -0.09   0.52
FidFou   -0.11   0.77   0.78   0.84    0.38   0.06  -0.02   0.50
ConMar    0.07   0.52   0.40   0.50    0.25   0.37  -0.02   0.72
ProDec    0.44  -0.29  -0.38  -0.27    0.15  -0.31  -0.37   0.00

        ImpPro FidFou ConMar ProDec
PriRel    0.18  -0.11   0.07   0.44
SerRel    0.26   0.77   0.52  -0.29
ImpCli    0.10   0.78   0.40  -0.38
ParAch    0.22   0.84   0.50  -0.27
BaiPri    0.14   0.38   0.25   0.15
HauPri    0.23   0.06   0.37  -0.31
BaiSer   -0.09  -0.02  -0.02  -0.37
HauSer    0.52   0.50   0.72   0.00
ImpPro    1.00   0.24   0.47  -0.07
FidFou    0.24   1.00   0.54  -0.24
ConMar    0.47   0.54   1.00  -0.14
ProDec   -0.07  -0.24  -0.14   1.00
```

Les prix relatifs semblent liés à l'impact de la baisse de service et de la baisse des prix sur les achats.

Exercice : commenter les liaisons.

Una autre façon de voir ces corrélations est la représentation graphique des nuages des points des variables prises deux a deux.

```
pairs(toto)
```



Calcul de l'analyse en composantes principales

L'ACP se calcule avec la commandes princomp(), prcomp() et dudi.pca() de la bibliothèque ade4. Calculons l'ACP avec la commande prcomp.

```
totoacp=prcomp(toto,scale=TRUE)
summary(totoacp)
```

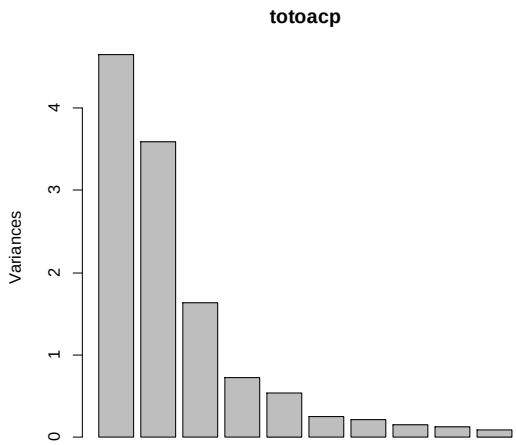
```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5     PC6     PC7
Standard deviation     2.1562 1.8943 1.2783 0.8486 0.7332 0.49278 0.46398
Proportion of Variance 0.3874 0.2990 0.1362 0.0600 0.0448 0.02024 0.01794
Cumulative Proportion  0.3874 0.6865 0.8226 0.8826 0.9274 0.94767 0.96561
```

Calcul de l'inertie expliquée par chaque axe ainsi que le cumul

Le premier axe PC1 explique 38.74% d'information. Le deuxième PC2 en explique 29.90%. Les deux premiers axes cumulés expliquent 68.65% d'information.

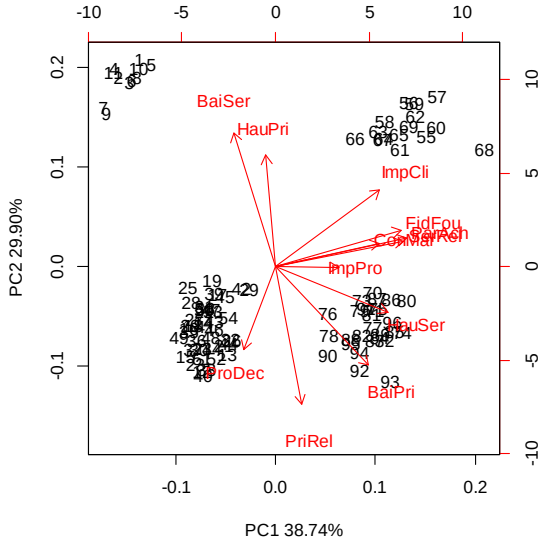
```
plot(totoacp)
```

On le voit sur le graphique.



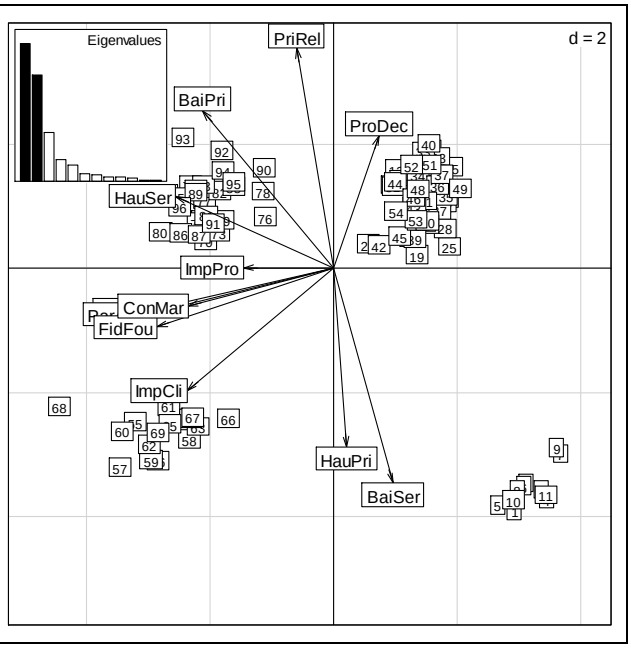
On peut donc projeter les points de l'espace nxp et symétriquement ceux de l'espace pxn sur l'espace à ceux dimensions (PC1,PC2)

```
biplot(totoacp,        xlab="PC1
38.74%", ylab="PC2 29.90%")
```



Le même résultat est obtenu par la commande dudi.pca de la bibliothèque ade4 qu'il faut au préalable télécharger et charger dans R.

```
toto.dudi <- dudi.pca(toto,,scale = TRUE, scan = FALSE)
scatter(cheron.dudi)
```



Contribution des variables aux axes

La commande summary(totoacp) donne à la fois les valeurs propres et les vecters propres de la matrice de correlation. Pour avoir seulement les vecteurs propres qui donnent les contributions des anciens axes aux nouveaux par totoacp$rotation.

Imprimons le vecteur propre correspondant à PC1 (trié (sort), en valeur absolue (abs et arrondi à deux chiffres (round))

```
sort(abs(round(totoacp$rotation[,1],2)))
```

HauPri	PriRel	ProDec	BaiSer	ImpPro	BaiPri	ImpCli	ConMar	HauSer	SerRel	FidFou	ParAch
0.03	0.08	0.10	0.13	0.20	0.29	0.33	0.33	0.35	0.40	0.40	0.42

Les trois variables qui contribuent plus à l'axe PC1 sont ParAch, FidFou et SerRel.

Pour le deuxième axe PC2 :

```
sort(abs(round(totoacp$rotation[,2],2)))
```

ImpPro	ConMar	SerRel	ParAch	FidFou	HauSer	ImpCli	ProDec	BaiPri	HauPri	BaiSer	PriRel
0.00	0.08	0.09	0.10	0.13	0.16	0.27	0.30	0.36	0.40	0.48	0.50

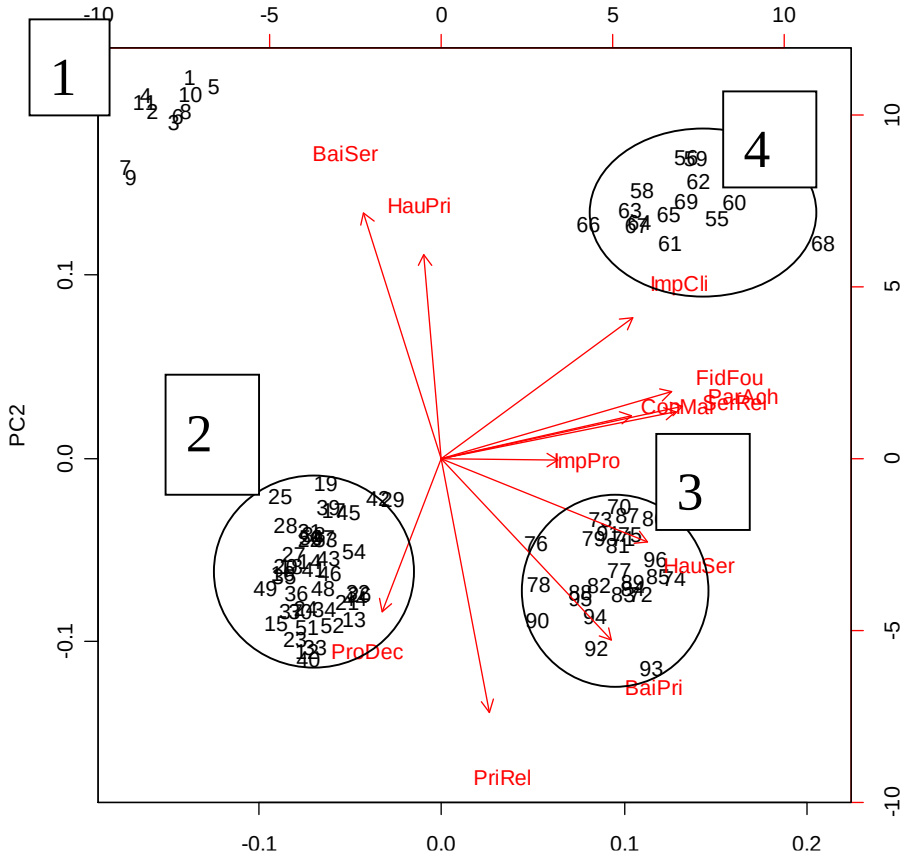
Les trois variables qui contribuent plus au deuxième axe sont PriRel, BaiSer, Haupri.

Par rapport à l'objectif de l'étude, le premier axe PC1 est celui qui est formé par le Service relatif (voir encadrée, la plus grande valeur). Le deuxième axe est celui du Prix relatif.

Les variables qui contribuent plus à la formation des axes principaux sont celles qui donnent le nom aux axes

Si on divise le graphique en 4 quadrants, les individus qui se retrouveront plus à l'Est sont ceux qui seront caractérisés par une grande valeur de ParAch, FirFou et SerRel. Ceux qui seront plus à l'ouest auront une petite valeur de ces variables.

De même, les individus qui seront au Nord auront une grande valeur de PriRel BaiSer et HauPri. Ceux qui seront au Sud seront caractérisés par une petite valeur de ces variables.



Il y a apparition de 4 segments des clients :

1. Groupe 1 : clientèle accommodante individu 1, 2, …,11 : ce sont des clients qui n'exigent pas de service et acceptent d'acheter à un prix bas
2. Groupe 2 : clientèle à potentiel limité : individus 12, 13, …, 54 : ce sont des clients qui n'exigent pas de service mais sont capable d'acheter à un prix élevé.
3. Groupe 3 : clientèle relationnelle 70, 71, …96 : ce sont des clients qui valorisent le service. Pour un service rendu, ils payent le prix.
4. Groupe 4 : clientèle à fort potentiel 55, 51, …, 69 : ce sont des clients exigeants en service mais payent le service en négociant le prix.

Souvent, comme c'est le cas, les chiffres se chévauchent et on ne sait pas dire quels sont les individus de différents groupes. Ce n'est pas cela l'objectif premier de l'ACP. Voyons maintenant comment par l'algorithme des moyennes mobiles (kmeans) comment regrouper les individus en 4 groupes

Grouper les clients en 4 groupes par la méthode des moyennes mobiles

```
mobile=kmeans(toto,4)> mobile
Clustering vector:
```

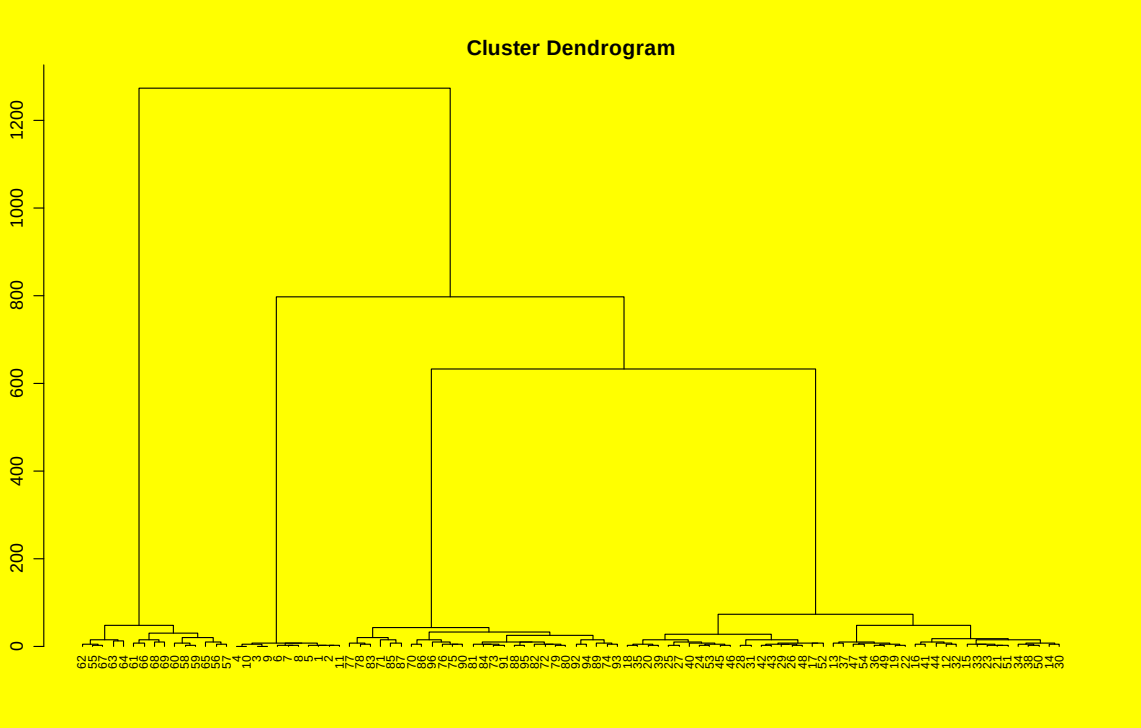
```
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 4  4  4  4  4  4  4  4  4  4  4  1  1  1  1  1  1  1  1  1
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1  1
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
 1  1  1  1  1  1  1  1  1  1  1  1  1  1  3  3  3  3  3  3
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
 3  3  3  3  3  3  3  3  3  2  2  2  2  2  2  2  2  2  2  2
81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96
 2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
```

La méthode des moyennes mobiles a trouvé les 4 meilleurs groupes respectivement de taille 43, 27, 15 et 11. La ligne supérieure donne le numéro du client et la ligne inférieure donne le groupe dans lequel il est classé.

Notons que le nombre 4 a été suggéré par l'ACP précédente.

Exécuter la classification hiérarchique

```
hc1 <- hclust(dist(toto), method = "ward")
```



Couper le dendrogramme à la hauteur correspondant à 4 classes

Quels sont les individus qui sont dans quelle classe ?

```
cutree(hc1,4)
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
 1  1  1  1  1  1  1  1  1  1  1  2  2  2  2  2  2  2  2  2
21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40
 2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2  2
41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60
 2  2  2  2  2  2  2  2  2  2  2  2  2  2  3  3  3  3  3  3
61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80
 3  3  3  3  3  3  3  3  3  4  4  4  4  4  4  4  4  4  4  4
81 82 83 84 85 86 87 88 89 90 91 92 93 94 95 96
 4  4  4  4  4  4  4  4  4  4  4  4  4  4  4  4
```

Il y a lieu de comparer les classes et les individus appartenant à ces classes et voir si les méthodes convergent.

Problème 2. Le segment du riz à faible teneur en eau commandée par les Tetela du Nord de la province du Kasaï Oriental

1. Contexte

Une enquête révèle qu'en République Démocratique du Congo, les habitants du Nord de la Province du Kasaï Oriental sont très regardants sur la qualité du riz. Ils préfèrent le riz qui ne colle pas. Ils aiment le riz à faible teneur en eau.

Vous êtes mis dans la position d'un bureau d'étude à qui le Gouverneur de la province demande d'analyser l'offre sur ce segment afin de décider quelle variété proposer aux paysans. A cet effet, vous disposez de différentes données sur le riz issues du programme de gestion des ressources phytogénétiques de l'Institut National pour l'Étude et la Recherche Agronomiqus (INERA).

2. Les individus ou les sujets d'étude

Un sélectionneur a dans sa collection 39 variétés notées de 1 à 39.

3. Les variables ou les caractères des individus

Il détient les caractères suivants sur les 39 variétés.

- HEP : hauteur de l'épi dans la tige
- DEP date d'épiaison ou de la formation de l'épi (en nombre des jours)
- QTE : quantité d'eau contenue dans les graines (décigramme)
- H2O : teneur en eau du grain (en %)
- PMS : poids de matière sèche du grain (en décigramme)
- EM2 : nombre d'épis par m2

- CTE : Coefficient de tallage (nbre d'épis par m2/nbre plants /m2)
- PMG : poids de 1000 grains à 16% d'humidité
- GE : nbre de grains par épi
- GM2 : nbre de grains par m2

4. Tableau de données

Le tableau de données est sous le format Excel. Le récupérer sous format texte avec séparateur tabulation.

5. Récupération des données comme objet de R

```
options(digits=3)
riz=read.table("c:/DATAR/ITCFRiz.txt",h=T,sep="\t",dec=",",row.names=1
)
riz[1:5,]
```

```
    HEP DEP QTE  H2O PMS EMS CTE   GE   GM2  PMG
A1 17.6  25 350 47.9 380 596 3.6 28.6 17046 38.9
A2 20.2  31 340 50.7 330 522 2.9 34.8 18192 38.3
A3 47.1  21 290 45.3 350 502   3 28.4 16854 36.3
A4 34.3  21 300 46.1 350 515 3.2 30.4 15686 33.9
A5 28.8  31 300 50.8 290 533 2.7 32.4 17274   38
```

6. Analyses préliminaires à une dimension

Moyenne et écart-type

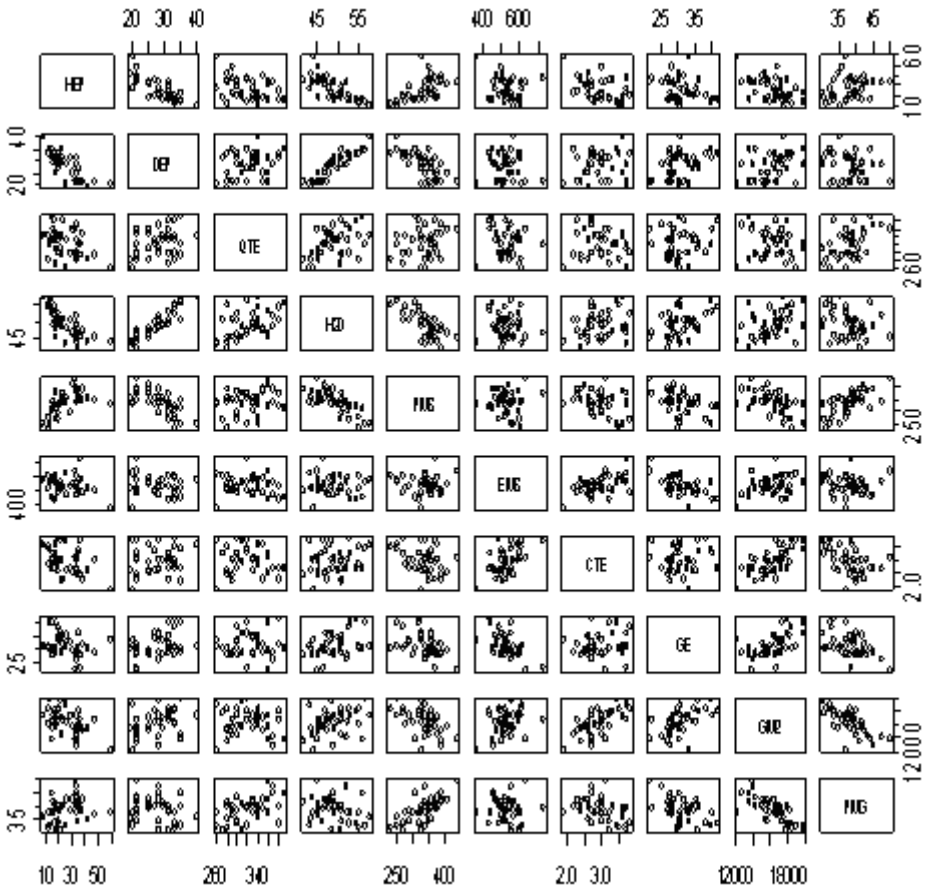
```
parametres= data.frame(Moyenne,Ecartype)
parametres
```

```
     Moyenne Ecartype
HEP    26.24   10.959
DEP    27.56    5.389
QTE   321.54   35.059
H2O    49.08    3.985
PMS   336.67   45.960
EMS   531.41   65.844
CTE     2.94    0.467
GE     30.96    4.362
GM2 16318.08 1833.357
PMG    38.36    4.385
```

Coefficients de corrélation

```
> round(cor(riz),2)
```

```
      HEP   DEP   QTE   H2O   PMS   EMS   CTE    GE   GM2   PMG
HEP  1.00 -0.77 -0.32 -0.74  0.55 -0.19 -0.42 -0.31 -0.52  0.26
DEP -0.77  1.00  0.37  0.92 -0.68 -0.06  0.19  0.40  0.38 -0.22
QTE -0.32  0.37  1.00  0.45  0.23 -0.13 -0.08  0.10 -0.09  0.51
H2O -0.74  0.92  0.45  1.00 -0.74 -0.06  0.23  0.42  0.38 -0.24
PMS  0.55 -0.68  0.23 -0.74  1.00  0.00 -0.35 -0.44 -0.53  0.68
EMS -0.19 -0.06 -0.13 -0.06  0.00  1.00  0.55 -0.57  0.32 -0.08
CTE -0.42  0.19 -0.08  0.23 -0.35  0.55  1.00  0.10  0.68 -0.45
GE  -0.31  0.40  0.10  0.42 -0.44 -0.57  0.10  1.00  0.55 -0.50
GM2 -0.52  0.38 -0.09  0.38 -0.53  0.32  0.68  0.55  1.00 -0.72
PMG  0.26 -0.22  0.51 -0.24  0.68 -0.08 -0.45 -0.50 -0.72  1.00
>
```



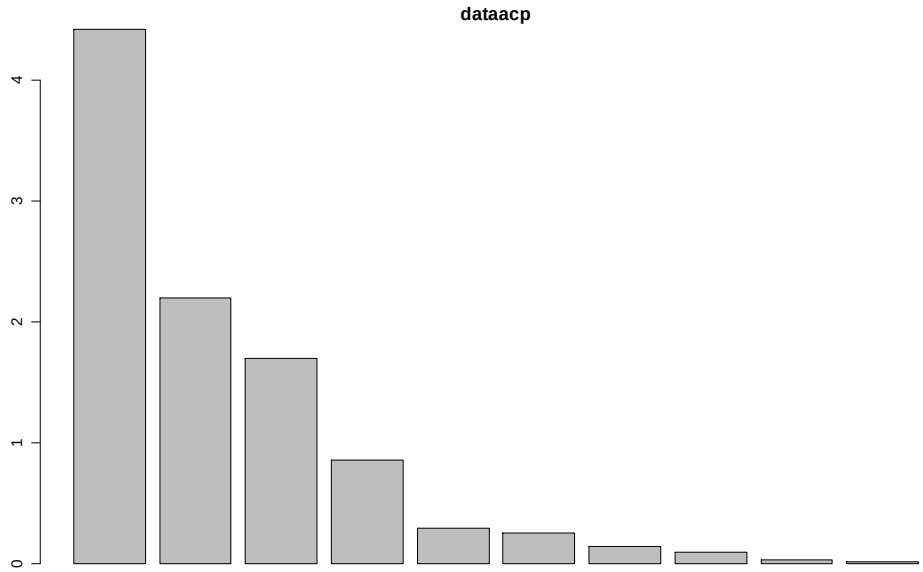
Calcul de l'analyse en composantes principales

```
>
> dataacp=prcomp(data,scale=T)
>
```

Calcul de l'inertie expliquée par chaque axe ainsi que le cumul

```
> Variance= (dataacp$sdev)^2
>
> Inertie=dataacp$sdev^2*100/sum(dataacp$sdev^2)
> cumul=rep(0,length(Inertie))
> cumul[1]=Inertie[1]
> for (i in 2:length(Inertie))
+ {cumul[i]=Inertie[i]+cumul[i-1]}
>
> data.frame(Variance,Inertie,cumul)
   Variance Inertie cumul
1    4.4240  44.240  44.2
2    2.2012  22.012  66.3
3    1.6945  16.945  83.2
4    0.8555   8.555  91.8
5    0.2911   2.911  94.7
6    0.2564   2.564  97.2
7    0.1394   1.394  98.6
8    0.0960   0.960  99.6
9    0.0280   0.280  99.9
```

```
10    0.0138    0.138 100.0
```



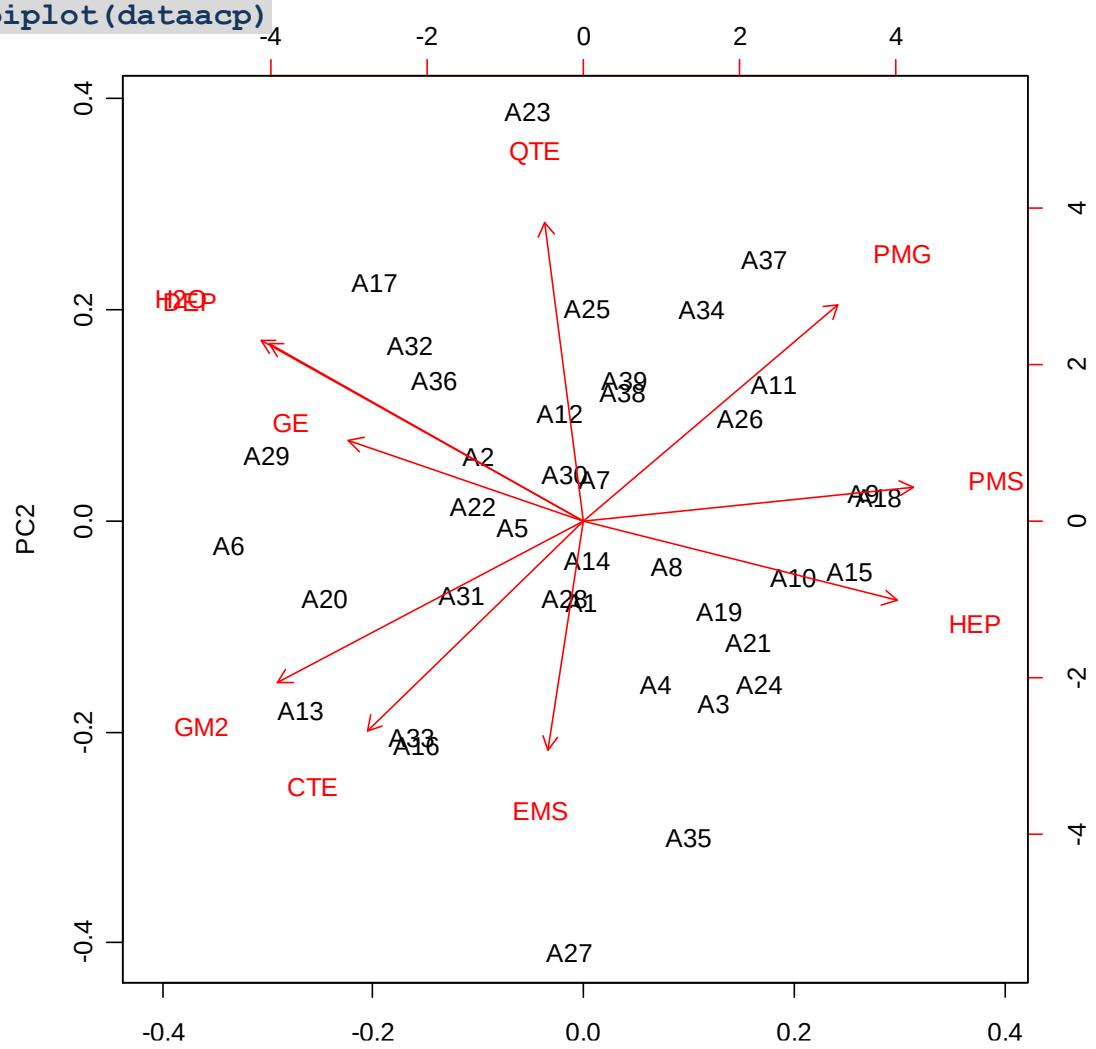
Contribution des variables aux axes

```
> dataacp$rotation
        PC1     PC2
HEP  0.3815 -0.1383
DEP -0.3837  0.3047
QTE -0.0478  0.5144
H2O -0.3938  0.3100
PMS  0.4009  0.0589
EMS -0.0434 -0.3956
CTE -0.2642 -0.3617
GE  -0.2868  0.1386
GM2 -0.3731 -0.2793
PMG  0.3098  0.3731
```

Le premier axe est formé par PMS. C'est l'axe des rendements. Le second axe est formé par QTE. C'est l'axe qui explique mieux les quantités d'eau dans la graine.

```
op <- par(mar = c(2,2,2,2))
```

```
biplot(dataacp)
```



Les variables qui forment un angle aigu sont corrélées positivement. Les variables qui forment un angle obtus sont corrélées négativement.

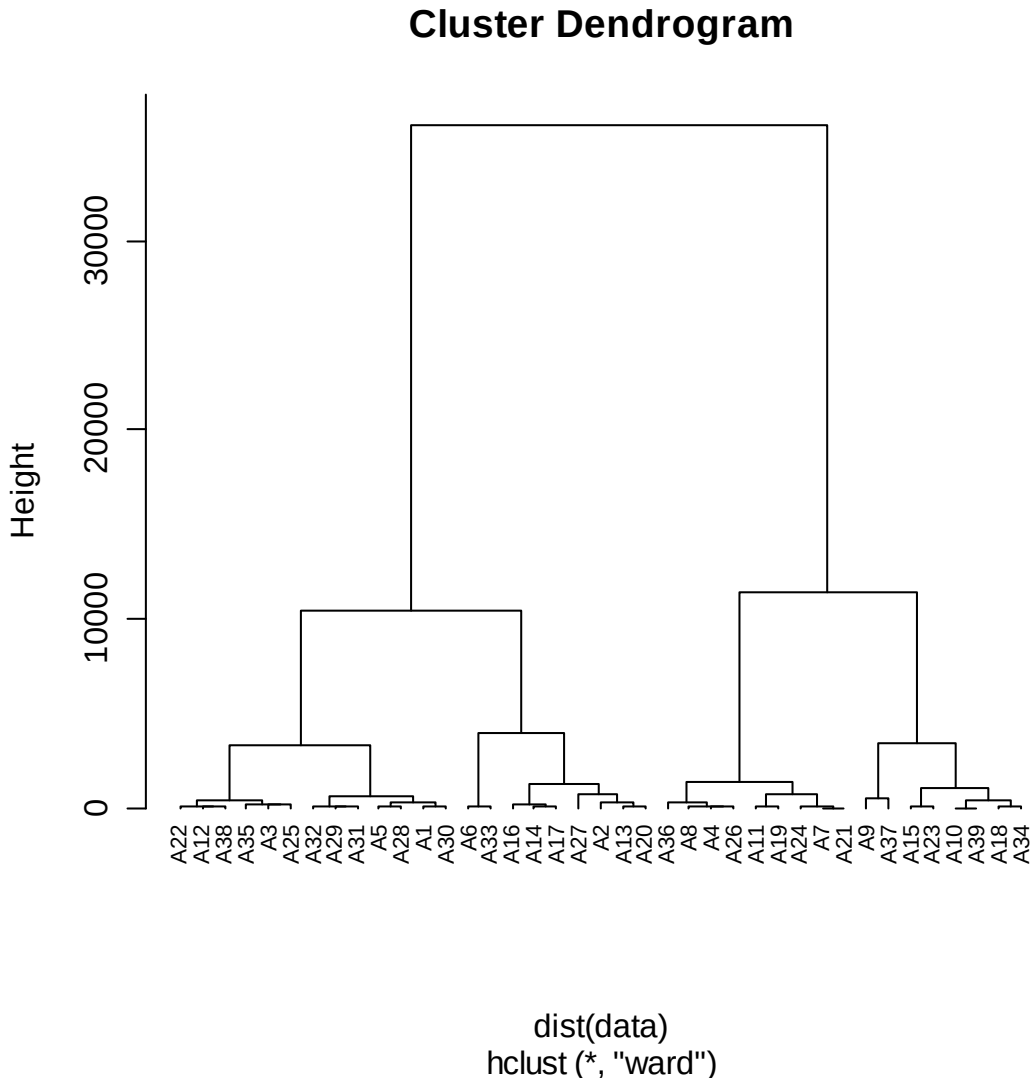
Si on doit chercher les variétés qui ont un grand nombre d'épis par m2, et qui ont moins d'eau dans les épis, alors on prend le 35 et le 27.

Grouper les variétés en 4 groupes

```
> mobile=kmeans(data,4)
> mobile
K-means clustering with 4 clusters of sizes 9, 8, 8, 14


Clustering vector:
 A1  A2  A3  A4  A5  A6  A7  A8  A9 A10 A11 A12 A13
  3   3   3   1   3   4   1   1   2   2   1   1   3
A14 A15 A16 A17 A18 A19 A20 A21 A22 A23 A24 A25 A26
  3   2   3   3   2   1   3   1   1   2   1   1   1
A27 A28 A29 A30 A31 A32 A33 A34 A35 A36 A37 A38 A39
  4   3   3   3   3   3   4   2   1   1   2   1   2
```

Exécuter la classification hiérarchique

```
plot(hc1,hang=-1,cex=0.7)
```



Couper le dendrogramme à la hauteur correspondant à 4 classes

```
> cutree(hc1,4)
 A1  A2  A3  A4  A5  A6  A7  A8  A9 A10 A11 A12 A13 A14
  1   2   1   3   1   2   3   3   4   4   3   1   2   2
A15 A16 A17 A18 A19 A20 A21 A22 A23 A24 A25 A26 A27 A28
  4   2   2   4   3   2   3   1   4   3   1   3   2   1
A29 A30 A31 A32 A33 A34 A35 A36 A37 A38 A39
  1   1   1   1   2   4   1   3   4   1   4
```

Problème 3. Notoriété des réseaux téléphoniques, cas des réseaux imaginaires à Kinshasa

1. Contexte

Un bureau de Marketing est consulté pour segmenter le marché de la téléphonie à Kinshasa. Il lui est demandé de voir la notoriété des maisons de distribution de la télécommunication dans les différentes communes, en considérant le niveau d'étude et la catégorie socio professionnelle des clients.

2. Les individus ou les sujets d'étude et les variables ou les caractères des individus

L'enquête concerne 1246 personnes dont 592 femmes et 654 hommes. Les habitants des communes suivantes ont été enquêtés : Bandalungwa (band 31 personnes), Barumbu (baru 39 personnes), Gombe (gom 15 personnes), Kasavubu (kasav 3 personnes), Kingasani (kinga 5 personnes), Kingabwa (kingab 25 personnes), Kinshasa (kinsh 30 personnes), Kintambo (kint 11 personnes), Masina (masi 20 personnes), Ndjili (ndjil 430 personnes), Ngiri-Ngiri (ngiri 601 personnes), Selembao (sele 36 personnes).

Il a été demandé à l'enquêté de mentionner le réseau qu'il utilise. Les réseaux suivants ont été spécifiés comme réponse possible : Afric, Ccc, Celta, Oaci, Smarc et Vadac[1]. Il a été aussi demandé le niveau d'études. Quatre niveaux ont été proposés : post-universitaire (post), universitaire (univ), secondaire (sec), primaire (prim) et sans niveau (sans). La catégorie socio-professionnelle a été aussi demandée. Les réponses possible étaient : Homme d'affaire (Haff), Privé (Priv), Libéral (liberal) Fonctionnaire (Fonc), Débrouillard (debr) et sans profession (sans).

3. Tableau de données

Le tableau de données est sous le format Excel. Le récupérer sous format texte avec séparateur tabulation.

4. Récupération des données comme objet de R

```
reseau=read.table("c:/DATAR/MRA/notorietereseau.txt",h=TRUE,sep="\t")
reseau[1:5,]
```

```
  SEXE COMMUNE RESEAU ETUDE  CSP
1    F   ndjil  Vadac  prim debr
2    F   ndjil  Vadac  prim debr
3    M   ndjil  Vadac  prim debr
4    F   ndjil  Afric  prim debr
```

[1] Les ressemblances des noms avec ceux des vrais réseaux n'est que fortuite. Ces noms sont pris pour des besoins purement pédagogiques et n'ont rien à avoir avec une quelconque réalité.

```
5    M   ndjil  Afric  prim debr
```

5. Analyses préliminaires à une dimension

```
summary(reseau,maxsum=20)
```

```
 SEXE       COMMUNE       RESEAU      ETUDE           CSP
 F:592   band  : 31   Afric: 23   post:113   debr   : 55
 M:654   baru  : 39   Ccc  : 52   prim:113   Debr   :142
         gom   : 15   Celta:207   sans:803   Fonc   : 72
         kasav :  3   Oaci :150   sec : 41   Fonct  :144
         kinga :  5   Smarc: 30   univ:176   Haff   :137
         kingab: 25   Vadac:784              liberal: 81
         kinsh : 30                          Liberal:177
         kint  : 11                          Priv   :100
         masi  : 20                          Retr   :238
         ndjil :430                          sans   :100
         ngiri :601
         sele  : 36
```

6. Analyses préliminaires à deux dimensions

6.1. Tableau de contingence COMMUNE, RESEAU

```
table(COMMUNE,RESEAU)
```

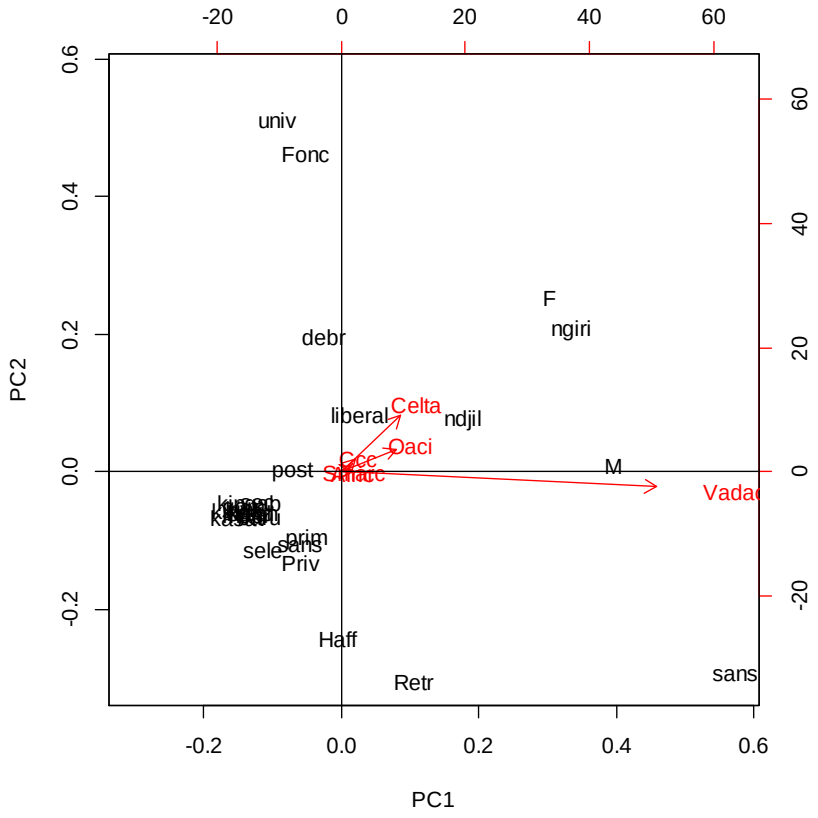
```
        RESEAU
COMMUNE  Afric Ccc Celta Oaci Smarc Vadac
  band       0   5     3    3     2    18
  baru       1   2     6    1     1    28
  gom        0   1     3    1     1     9
  kasav      0   0     0    0     0     3
  kinga      0   1     1    0     0     3
  kingab     0   2     5    3     0    15
  kinsh      1   1     5    2     1    20
  kint       0   0     3    1     0     7
  masi       2   2     3    0     0    13
  ndjil     12  27    70   43    21   257
  ngiri      6  11   108   96     3   377
  sele       1   0     0    0     1    34
```

6.2. Tableau de contingence ETUDE, RESEAU

```
        ETUDE
COMMUNE  post prim sans sec univ
  band      1    1   19   3    7
  baru      3    5   28   1    2
  gom       0    0   14   1    0
  kasav     0    0    3   0    0
  kinga     0    0    4   0    1
  kingab    0    4   18   0    3
  kinsh     1    2   24   0    3
  kint      0    0    8   0    3
  masi      0    0   16   1    3
  ndjil    54   47  271  10   48
  ngiri    54   54  362  25  106
  sele      0    0   36   0    0
```

Calcul de l'analyse en composantes principales

```
toto=read.table("c:/DATAR/notorietereseau.txt",sep="\t",h=TRUE)


titi1=table(toto$COMMUNE,toto$RESEAU)*100/length(toto$COMMUNE)
titi2=table(toto$SEXE,toto$RESEAU)*100/length(toto$COMMUNE)
titi3=table(toto$ETUDE,toto$RESEAU)*100/length(toto$COMMUNE)
titi4=table(toto$CSP,toto$RESEAU)*100/length(toto$COMMUNE)
titi=rbind(titi1,titi2,titi3,titi4)

titiacp=prcomp(titi)
biplot(titiacp)
abline(v=0)
abline(h=0)
```



Le réseau Vadac se démarque. Il est beaucoup prisé toutes les catégories socio-professionnelles confondues. Ce sont les Hommes d'affaires qui se placent en tête suivis des privés et même des sans emplois.

Celta constitue une classe à part. Ce sont les fonctionnaires qui s'y abonnent plus.

Les réseaux Afric, Smarc et Ccc n'ont pas été beaucoup citésé.

Toutes les variables semblent corrélées sauf celles qui sont aux extrêmes.

Grouper les réseaux en 3 groupes

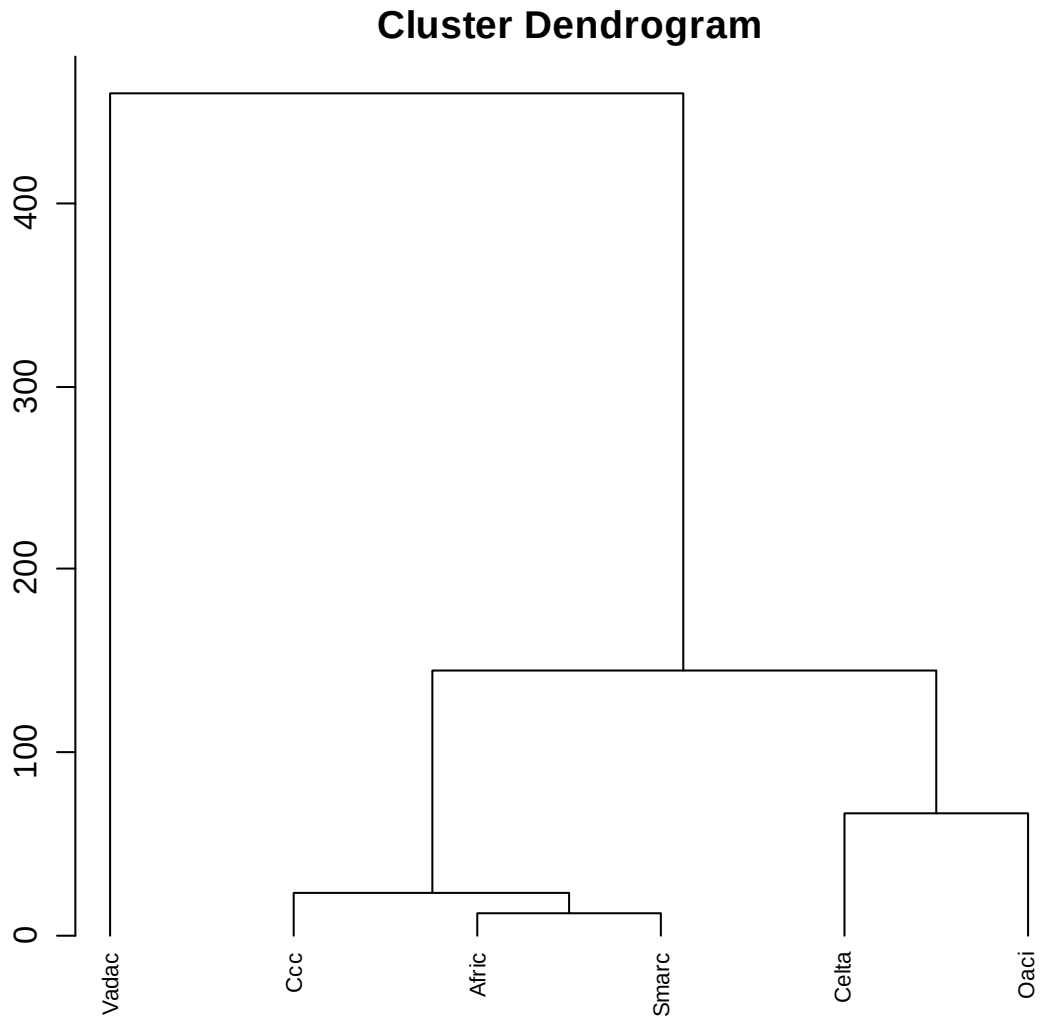
```
> mobile=kmeans(data,3)
> mobile
```

```
K-means clustering with 3 clusters of sizes 3, 1, 2


Clustering vector:
Afric   Ccc Celta  Oaci Smarc Vadac
    1     1     3     3     1     2
```

Le premier groupe est constitué de Afric Ccc et Smarc. Ils sont aux antipodes de toutes les caractères. Ils ne sont pas beaucoup choisis. Le deuxième groupe est Vadac qui est choisi par les Hommes d'affaires, les privés etc. Le troisième groupe est constitué de Celta et Oaci qui sont choisis par les Fonctionnaires (Voir ACP).

Exécuter la classification hiérarchique

```
> hc1 <- hclust(dist(data), method = "ward")
> plot(hc1,hang=-1,cex=0.7)
```



Couper le dendrogramme à la hauteur correspondant à 3 classes

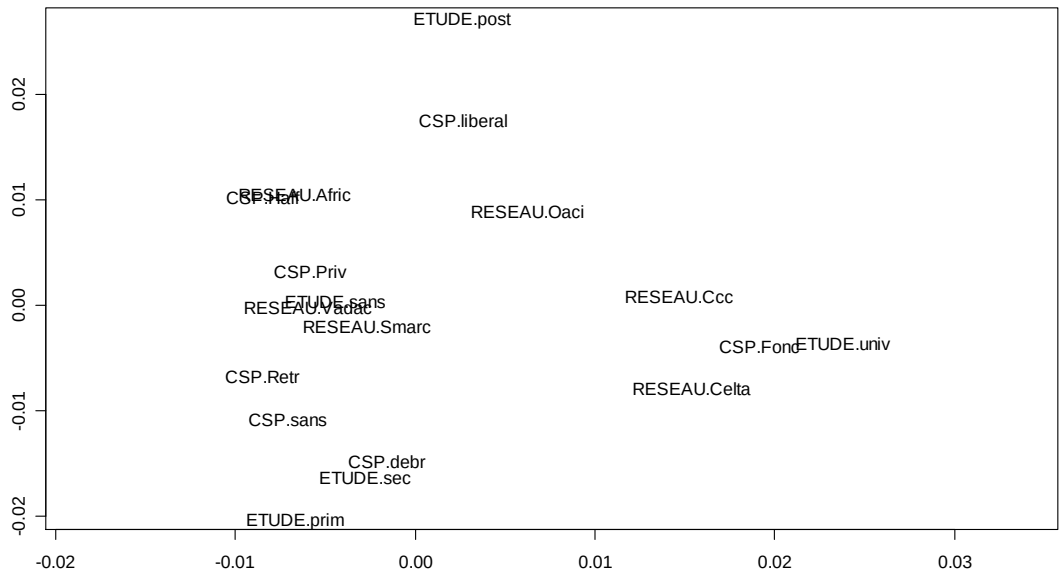
```
> cutree(hc1,3)
Afric   Ccc Celta  Oaci Smarc Vadac
```

```
    1     1     2     2     1     3
>
```

Analyse élémentaire

```
> summary(notoriete)
   RESEAU      ETUDE           CSP
 Afric: 23   post:113   debr   :197
 Ccc  : 52   prim:113   Fonc   :216
 Celta:207   sans:803   Haff   :137
 Oaci :150   sec : 41   liberal:258
 Smarc: 30   univ:176   Priv   :100
 Vadac:784              Retr   :238
                        sans   :100
```

Correspondance entre réseaux et étude et CSP

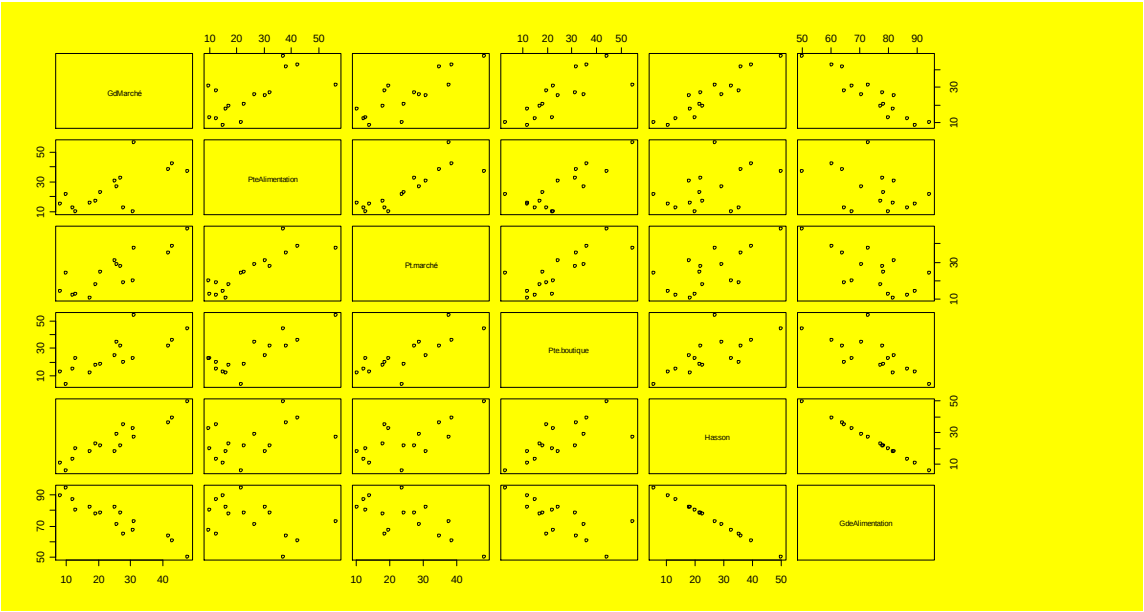


Problème 4. Quels marchés fréquentent-ils ?

Quel marché fréquente le sénateur, le Député, le Fonctionnaire ? les classes d'âge sont considérés dans l'enquête à savoir les personnes de moins de 35 ans (a<35), les personnes entre 35 et 49 ans (a35-49), les personnes entre 50 et 64 ans (a50-64) et les personnes de plus de 65 ans (a65+). Le nombre d'enfants est aussi considéré. Trois niveaux sont considérés : foyer de un enfant (foy_1), foyer de deux enfants (foy_2), foyer de trois enfants (foy_3) et le foyer de quatre enfants et plus (foy4_+).

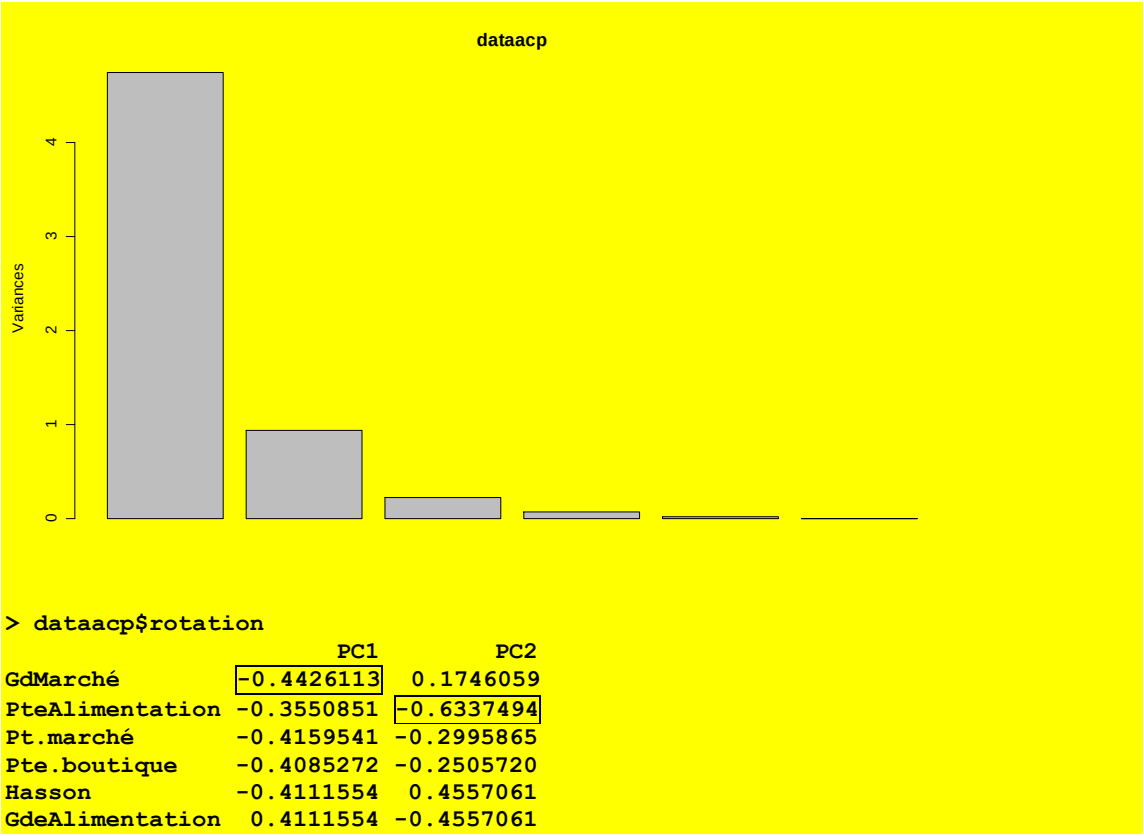
Les habitants de Kingasani ont été ciblés, ceux de Gombe et ceux des quartiers de moins de 2000 habitants et de plus de 5000 habitants.

```
Statistique élémentaires
> Moyenne=mean(data)
> Ecartype=sd(data)
> parametres=data.frame(Moyenne,Ecartype)
> parametres
                 Moyenne Ecartype
GdMarché         25.00000 11.88450
PteAlimentation  25.00000 13.51123
Pt.marché        24.99375 10.98917
Pte.boutique     24.98750 13.20752
Hasson           25.00625 11.54363
GdeAlimentation  74.99375 11.54363
>
Statistique bivariée
> round(cor(data),2)
                GdMarché PteAlimentation Pt.marché
GdMarché            1.00            0.65      0.84
PteAlimentation     0.65            1.00      0.86
Pt.marché           0.84            0.86      1.00
Pte.boutique        0.76            0.81      0.79
Hasson              0.93            0.42      0.68
GdeAlimentation    -0.93           -0.42     -0.68
                Pte.boutique Hasson GdeAlimentation
GdMarché                0.76   0.93           -0.93
PteAlimentation         0.81   0.42           -0.42
Pt.marché               0.79   0.68           -0.68
Pte.boutique            1.00   0.70           -0.70
Hasson                  0.70   1.00           -1.00
GdeAlimentation        -0.70  -1.00            1.00>
> pairs(data)
```

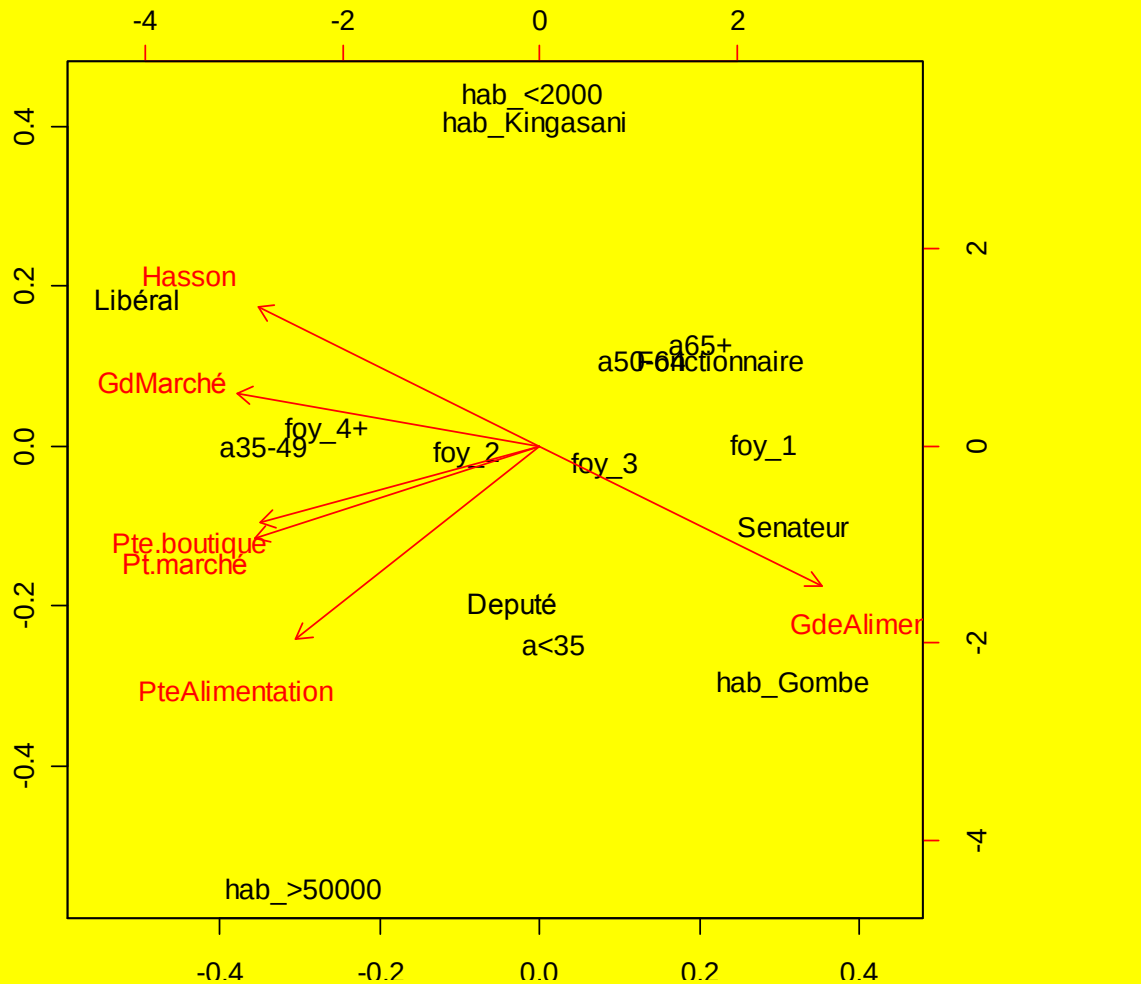


Analyse en composantes principales

```
> dataacp=prcomp(data,scale=T)
>
> Variance= (dataacp$sdev)^2
>
> Inertie=dataacp$sdev^2*100/sum(dataacp$sdev^2)
> cumul=rep(0,length(Inertie))
> cumul[1]=Inertie[1]
> for (i in 2:length(Inertie))
+ {cumul[i]=Inertie[i]+cumul[i-1]}
>
> data.frame(Variance,Inertie,cumul)
      Variance      Inertie     cumul
1 4.738248e+00 7.897080e+01  78.97080
2 9.423701e-01 1.570617e+01  94.67697
3 2.296893e-01 3.828154e+00  98.50512
4 6.800282e-02 1.133380e+00  99.63850
5 2.168975e-02 3.614958e-01 100.00000
6 2.282997e-31 3.804995e-30 100.00000
>
>
> plot(dataacp)
```



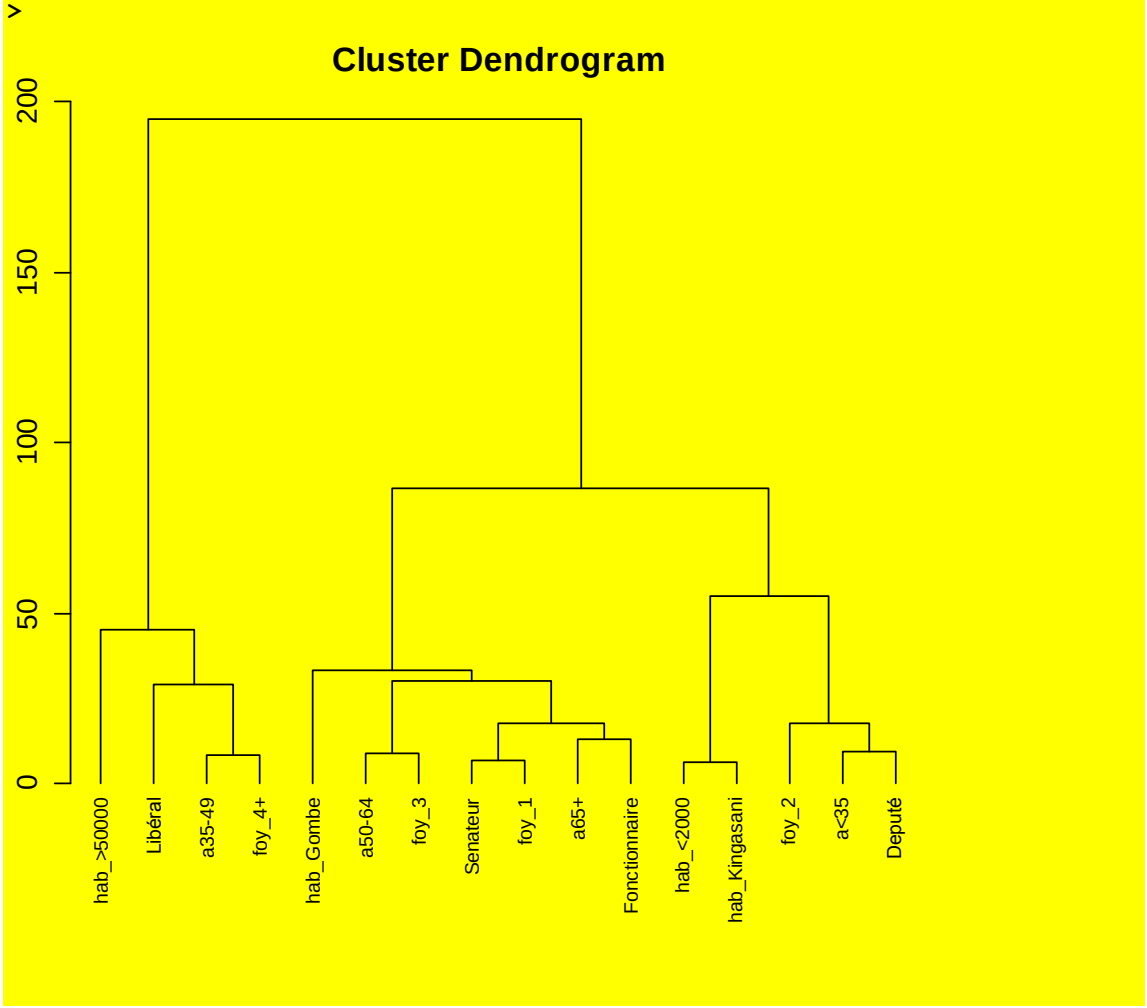

```
> dataacp$rotation
                     PC1        PC2
GdMarché       -0.4426113  0.1746059
PteAlimentation -0.3550851 -0.6337494
Pt.marché      -0.4159541 -0.2995865
Pte.boutique   -0.4085272 -0.2505720
Hasson         -0.4111554  0.4557061
GdeAlimentation  0.4111554 -0.4557061
```



Classification

```
> mobile=kmeans(data,4)
> mobile
K-means clustering with 4 clusters of sizes 6, 4, 2, 4


Clustering vector:
         a<35        a35-49        a50-64          a65+
            4             2             1             1
    Senateur        Deputé       Libéral Fonctionnaire
            1             4             2             1
      foy_1         foy_2         foy_3        foy_4+
            1             4             4             2
   hab_<2000 hab_Kingasani    hab_>50000     hab_Gombe
            3             3             2             1

Within cluster sum of squares by cluster:
[1]  901.5 1321.8   19.8  395.7
> hc1 <- hclust(dist(data), method = "ward")

> plot(hc1,hang=-1,cex=0.7)
> cutree(hc1,4)
         a<35        a35-49        a50-64          a65+
            1             2             3             3
```

```
     Senateur          Deputé        Libéral Fonctionnaire
            3               1              2             3
       foy_1           foy_2          foy_3         foy_4+
            3               1              3             2
  hab_<2000 hab_Kingasani     hab_>50000      hab_Gombe
            4               4              2             3
```



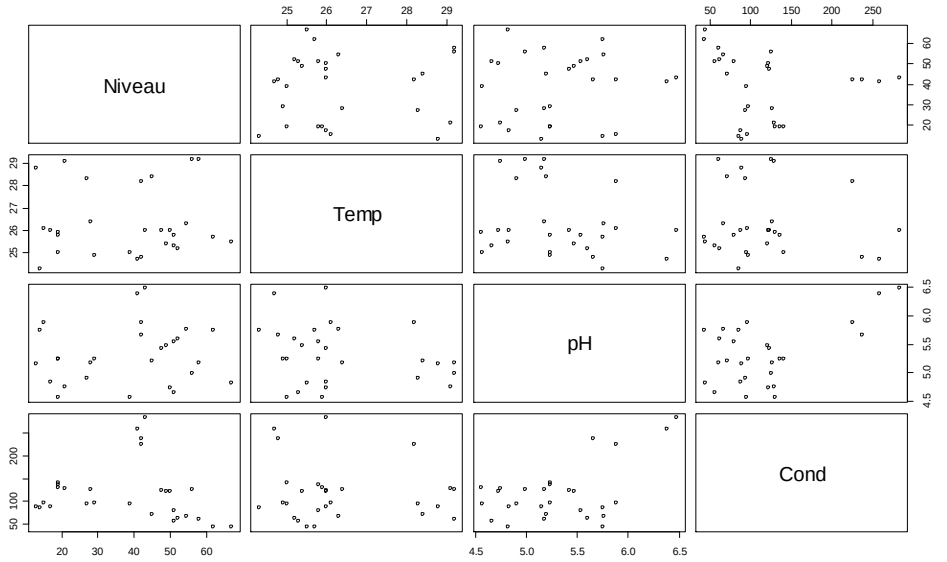
Problème 5. Etude de l'eau de puits par un laboratoire d'éco-toxicologie

Une étude est menée sur 28 puits. On mesure la profondeur d'où est provenue l'échantillon, la température, le pH et la conductivité.

On veut savoir, par rapport à ces caractères considérés simultanément, quels sont les puits qui se ressemblent. S'ils se ressemblent, ils appartiennent à quelle classe ? Les classes sont déterminées par les caractères mesurés. On veut savoir aussi quels sont les caractères qui sont corrélés.

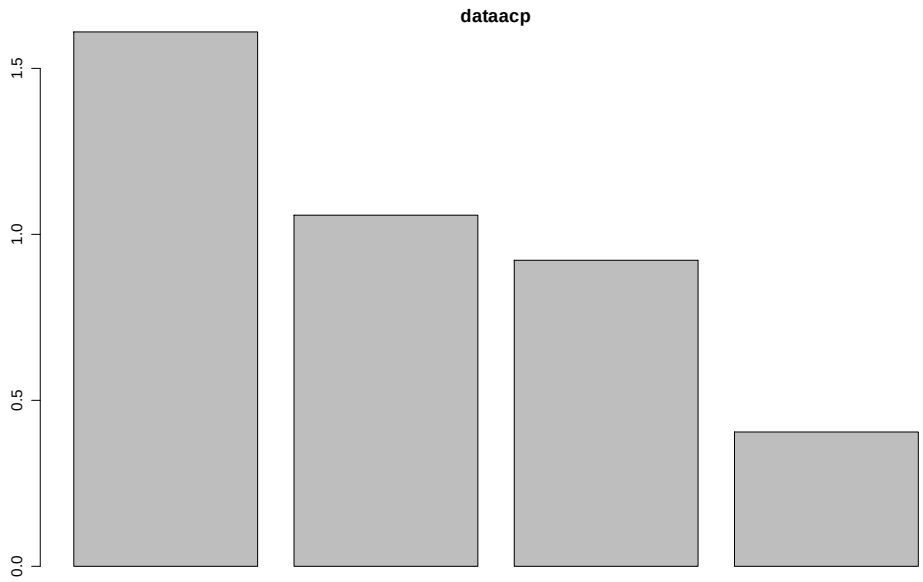
Analyse élémentaire

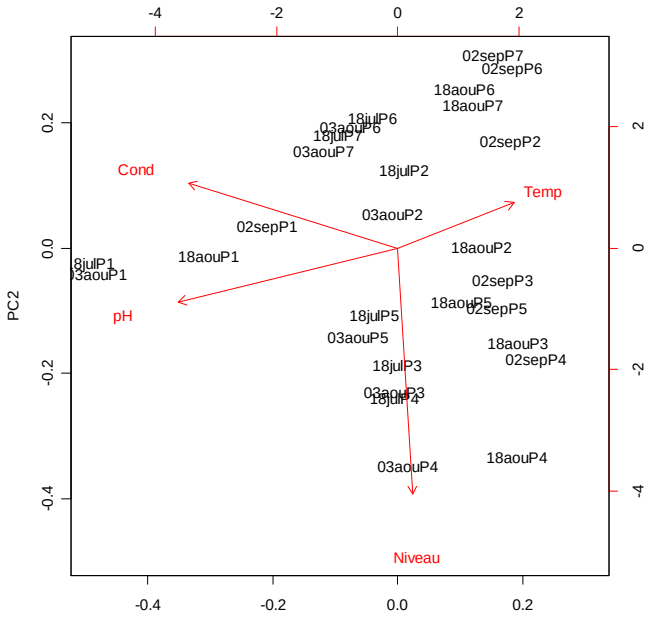
```
> summary(data)
     Niveau           Temp             pH              Cond
 Min.   :13.00   Min.   :24.30   Min.   :4.560   Min.   : 43.0
 1st Qu.:20.50   1st Qu.:25.27   1st Qu.:4.890   1st Qu.: 78.0
 Median :42.00   Median :25.95   Median :5.240   Median : 96.5
 Mean   :38.25   Mean   :26.33   Mean   :5.315   Mean   :117.5
 3rd Qu.:51.00   3rd Qu.:26.85   3rd Qu.:5.683   3rd Qu.:129.3
 Max.   :67.00   Max.   :29.20   Max.   :6.480   Max.   :283.0
>
> Moyenne=mean(data)
> Ecartype=sd(data)
>
>
> parametres=data.frame(Moyenne,Ecartype)
> parametres
         Moyenne   Ecartype
Niveau  38.25000 16.3998080
Temp    26.33214  1.5170765
pH       5.31500  0.5110447
Cond   117.50000 62.7047756
>
> round(cor(data),2)
       Niveau  Temp    pH  Cond
Niveau   1.00  0.01  0.11 -0.15
Temp     0.01  1.00 -0.23 -0.09
pH       0.11 -0.23  1.00  0.52
Cond    -0.15 -0.09  0.52  1.00
>
> pairs(data)
```



Analyse en composantes principales

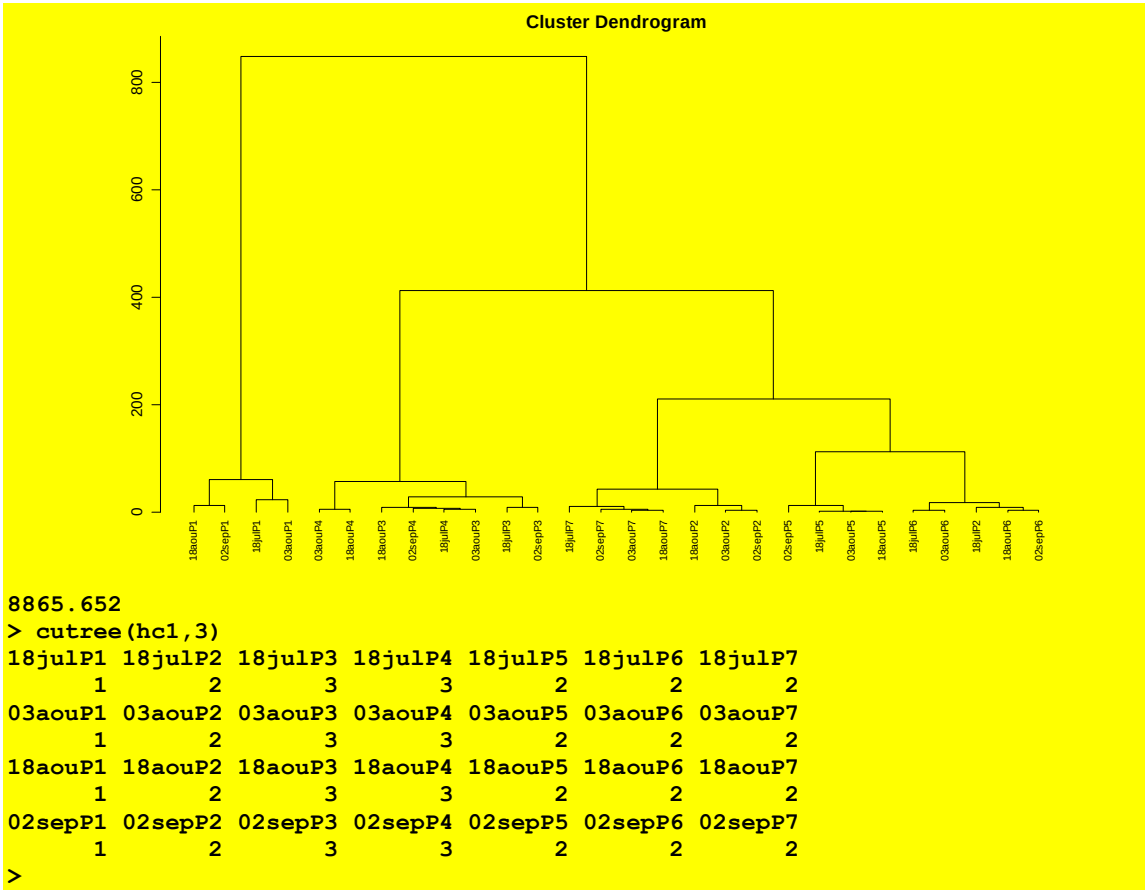
```
> dataacp=prcomp(data,scale=T)
>
> Variance= (dataacp$sdev)^2
>
> Inertie=dataacp$sdev^2*100/sum(dataacp$sdev^2)
> cumul=rep(0,length(Inertie))
> cumul[1]=Inertie[1]
> for (i in 2:length(Inertie))
+ {cumul[i]=Inertie[i]+cumul[i-1]}
>
> data.frame(Variance,Inertie,cumul)
   Variance  Inertie     cumul
1 1.6118167 40.29542  40.29542
2 1.0594660 26.48665  66.78207
3 0.9234969 23.08742  89.86949
4 0.4052204 10.13051 100.00000
```





Classification

```
> mobile=kmeans(data,3)
> mobile
K-means clustering with 3 clusters of sizes 8, 4, 16

Cluster means:
    Niveau    Temp       pH   Cond
1 55.06250 26.4250 5.313750  60.75
2 42.00000 25.9250 6.100000 251.00
3 28.90625 26.3875 5.119375 112.50

Clustering vector:
18julP1 18julP2 18julP3 18julP4 18julP5 18julP6 18julP7
      2       3       1       1       3       3       3
03aouP1 03aouP2 03aouP3 03aouP4 03aouP5 03aouP6 03aouP7
      2       3       1       1       3       3       3
18aouP1 18aouP2 18aouP3 18aouP4 18aouP5 18aouP6 18aouP7
      2       3       1       1       3       3       3
02sepP1 02sepP2 02sepP3 02sepP4 02sepP5 02sepP6 02sepP7
      2       3       1       1       3       3       3

Within cluster sum of squares by cluster:
[1] 1484.098 1970.412
```




```
8865.652
> cutree(hc1,3)
18julP1 18julP2 18julP3 18julP4 18julP5 18julP6 18julP7
      1       2       3       3       2       2       2
03aouP1 03aouP2 03aouP3 03aouP4 03aouP5 03aouP6 03aouP7
      1       2       3       3       2       2       2
18aouP1 18aouP2 18aouP3 18aouP4 18aouP5 18aouP6 18aouP7
      1       2       3       3       2       2       2
02sepP1 02sepP2 02sepP3 02sepP4 02sepP5 02sepP6 02sepP7
      1       2       3       3       2       2       2
>
```

Etude des voitures

Ces données sont tirées de l'ouvrage de Saporta, pages 177 à 181. De très légères modifications ont été introduites : traitement de variables illustratives quantitatives et traitement d'observations illustratives. Les justifications théoriques et les formules sont disponibles dans le même ouvrage, pages 155 à 177.

Traitements réalisés

Réaliser une ACP sur un fichier de données.
Afficher les valeurs propres. Construire le graphiques éboulis des valeurs propres.
Construire le cercle de corrélations.
Projeter les observations dans le premier plan factoriel.
Positionner des variables illustratives quantitatives dans le cercle des corrélations.
Positionner les modalités d'une variable illustrative catégorielle.
Positionner des observations illustratives.

```
>options(digits=3)
>autostoutes=read.table("c:/DATAR/Saporta2006.txt",sep="\t",h=TRUE,row
.names=1)
> autostoutes
```

	CYL	PUISS	LONG	LARG	POIDS	V.MAX	FINITION	PRIX	RPOIDPUIS
AlfasudTI	1350	79	393	161	870	165	2_B	30570	11.01
Audi100	1588	85	468	177	1110	160	3_TB	39990	13.06
Simca1300	1294	68	424	168	1050	152	1_M	29600	15.44
CitroenGsClub	1222	59	412	161	930	151	1_M	28250	15.76
Fiat132	1585	98	439	164	1105	165	2_B	34900	11.28
LanciaBeta	1297	82	429	169	1080	160	3_TB	35480	13.17
Peugeot504	1796	79	449	169	1160	154	2_B	32300	14.68
Renault16TL	1565	55	424	163	1010	140	2_B	32000	18.36
Renault30	2664	128	452	173	1320	180	3_TB	47700	10.31
ToyotaCorolla	1166	55	399	157	815	140	1_M	26540	14.82
Alfetta-1.66	1570	109	428	162	1060	175	3_TB	42395	9.72
Princess-1800	1798	82	445	172	1160	158	2_B	33990	14.15
Datsun-200L	1998	115	469	169	1370	160	3_TB	43980	11.91
Taunus-2000	1993	98	438	170	1080	167	2_B	35010	11.02
Rancho	1442	80	431	166	1129	144	3_TB	39450	14.11
Mazda-9295	1769	83	440	165	1095	165	1_M	27900	13.19
Opel-Rekord	1979	100	459	173	1120	173	2_B	32700	11.20
Lada-1300	1294	68	404	161	955	140	1_M	22100	14.04
Peugeot604	2664	136	472	177	1410	180	<NA>	NA	NA
Peugeot304S	1288	74	414	157	915	160	<NA>	NA	NA

Ce tableau dénommé « autostoutes » contient les informations sur 20 voitures. Nous allons caractériser les 18 premières et allons voir si les deux dernières peuvent être classées dans quelle catégorie. Extrayons les 18 premières et appelons le tableau autos. Mémorisons les deux dernières dans le tableau « ind.illus » pour individus illustratifs.

```
> autos=autostoutes[1:18,]
> ind.illus=autostoutes[19:20,]
> autos
```

	CYL	PUISS	LONG	LARG	POIDS	V.MAX	FINITION	PRIX	RPOIDPUIS
AlfasudTI	1350	79	393	161	870	165	2_B	30570	11.01
Audi100	1588	85	468	177	1110	160	3_TB	39990	13.06
Simca1300	1294	68	424	168	1050	152	1_M	29600	15.44

```
CitroenGsClub 1222    59  412  161   930  151      1_M 28250      15.76
Fiat132       1585    98  439  164  1105  165      2_B 34900      11.28
LanciaBeta    1297    82  429  169  1080  160     3_TB 35480      13.17
Peugeot504    1796    79  449  169  1160  154      2_B 32300      14.68
Renault16TL   1565    55  424  163  1010  140      2_B 32000      18.36
Renault30     2664   128  452  173  1320  180     3_TB 47700      10.31
ToyotaCorolla 1166    55  399  157   815  140      1_M 26540      14.82
Alfetta-1.66  1570   109  428  162  1060  175     3_TB 42395       9.72
Princess-1800 1798    82  445  172  1160  158      2_B 33990      14.15
Datsun-200L   1998   115  469  169  1370  160     3_TB 43980      11.91
Taunus-2000   1993    98  438  170  1080  167      2_B 35010      11.02
Rancho        1442    80  431  166  1129  144     3_TB 39450      14.11
Mazda-9295    1769    83  440  165  1095  165      1_M 27900      13.19
Opel-Rekord   1979   100  459  173  1120  173      2_B 32700      11.20
Lada-1300     1294    68  404  161   955  140      1_M 22100      14.04
```

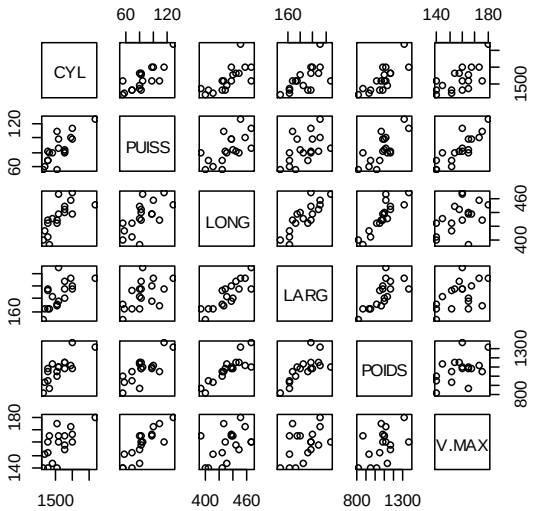
Les variables actives sont : cylindrée (CYL), puissance (PUISS), longueur (LONG), largeur (LARG), poids (POIDS), vitesse maximale (V.MAX),

Les variables illustratives : finition (FINITION), prix (PRIX), rapport poids/puissance (RPOIDPUISS),

```
> ind.illus
            CYL PUISS LONG LARG POIDS V.MAX FINITION PRIX RPOIDPUIS
Peugeot604 2664   136  472  177  1410   180     <NA>   NA        NA
Peugeot304S 1288   74  414  157   915   160     <NA>   NA        NA
```

Variables actives et variables illustratives du tableau des autos actives

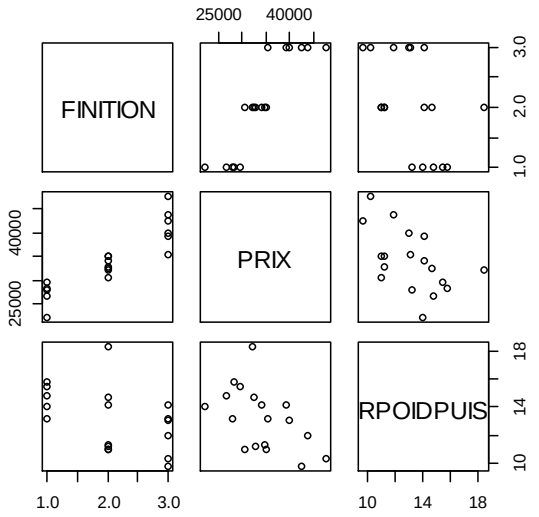
```
autos.actifs <- autos[,1:6]
autos.var.illus <- autos[,7:9]
```



Il y a une corrélation qui se dessine entre la cylindrée et la puissance (graphique de la ligne 2 et la colonne 1, r= 0.8) , entre la cylindrée et la longueur (graphique ligne 3, col 1, r= 0.7).

```
cor(autos.actifs)
          CYL PUISS  LONG  LARG POIDS V.MAX
CYL   1.000 0.797 0.701 0.630 0.789 0.665
PUISS 0.797 1.000 0.641 0.521 0.765 0.844
LONG  0.701 0.641 1.000 0.849 0.868 0.476
LARG  0.630 0.521 0.849 1.000 0.717 0.473
POIDS 0.789 0.765 0.868 0.717 1.000 0.478
V.MAX 0.665 0.844 0.476 0.473 0.478 1.000
```

Faisons la même chose avec le tableau autos.illus. La variable FINITION est une variable qualitative. En général, son introduction dans ce type de graphique n'est pas très indiquée. Néanmoins, on remarquera qu'on peut parfois en tirer des informations utiles : par exemple, ici, selon la finition, les prix sont différents (graphique ligne 2 colonne 1). Il y a une corrélation négative qui se dessine entre le prix et le rapport poids/puissance.



Le centrage et la réduction des données

Nous allons utiliser la commande pricomp de R pour calculer les valeurs propres et les vecteurs propres de la matrice de variance et covariance. Cette commande a pour argument, le tableau de donnée. Une option est cor. En déclarant cor =TRUE, cela veut dire que nous travaillons avec la matrice de corrélation. En déclarant score = TRUE, nous demandons que les coordonnées factorielles soient calculées.

Dans le cas de cet exemple, les variables n'ont pas toutes les mêmes unités. Pour pouvoir avoir des résultats comparables, il faut travailler avec la matrice centrée et réduite.

```
> #centrage et réduction des données --> cor = T
> #calcul des coordonnées factorielles --> scores = T
> acp.autos <- princomp(autos.actifs, cor = T, scores = T)
> #summary
> print(summary(acp.autos))
```

```
Importance of components:
                       Comp.1 Comp.2 Comp.3 Comp.4 Comp.5  Comp.6
Standard deviation      2.103  0.925 0.6108 0.4625 0.3046 0.20806
Proportion of Variance  0.737  0.143 0.0622 0.0357 0.0155 0.00722
Cumulative Proportion   0.737  0.879 0.9417 0.9773 0.9928 1.00000
```

PRINCOMP fournit les écarts-types associés aux axes. Le carré correspond aux variances et les variances sont les valeurs propres associées aux axes.

```
> #obtenir les variances associées aux axes c.-à-d. les valeurs propres
> val.propres <- acp.autos$sdev^2
> print(val.propres)
```

```
Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
4.4209 0.8561 0.3731 0.2139 0.0928 0.0433
```

Nous avons également le pourcentage cumulé. Avec ATTRIBUTES, nous avons la liste des informations que nous pourrons exploiter par la suite.

```
> #quelles les propriétés associées à l'objet ?
> print(attributes(acp.autos))
$names
[1] "sdev"     "loadings" "center"   "scale"    "n.obs"    "scores"   "call"

$class
[1] "princomp"
```

La commande princomp calcule la matrice de corrélation, la diagonalise et calcule les valeurs propres et les vecteurs propres de la matrice de corrélation.

Calcul, intervalles de confiance et Scree Plot

```
> #scree plot (graphique des éboulis des valeurs propres)
>  plot(1:6,val.propres,  type="b",  ylab=  "Valeurs  propres",  xlab=
"Composante",main="% information par composante")
>
text(1:6,val.propres,round(val.propres*100/(sum(val.propres)),2),pos=1
)
```

Les deux premiers axes traduisent 88% de l'information disponible. On se rend compte ici qu'on pouvait s'en tenir uniquement au premier facteur.

Mais c'est moins pratique pour les graphiques ; on suspecte aussi un «effet taille » dans les données. On va donc conserver les deux premiers facteurs.

```
> #intervalle de confiance des val.propres à 95% (cf.Saporta, page
172)
> val.basse <- val.propres * exp(-1.96 * sqrt(2.0/(n-1)))
> val.haute <- val.propres * exp(+1.96 * sqrt(2.0/(n-1)))
> #affichage sous forme de tableau
> tableau <- cbind(val.basse,val.propres,val.haute)
> colnames(tableau) <- c("B.Inf.","Val.","B.Sup")
> print(tableau,digits=3)
```

```
       B.Inf.   Val.  B.Sup
Comp.1 2.2571 4.4209 8.6591
Comp.2 0.4371 0.8561 1.6768
Comp.3 0.1905 0.3731 0.7307
Comp.4 0.1092 0.2139 0.4190
Comp.5 0.0474 0.0928 0.1818
Comp.6 0.0221 0.0433 0.0848
```

Les intervalles de confiance d'Anderson ne sont licites que si le nuage de points est gaussien. On ne l'affiche donc qu'à tire indicatif (cf. formules page 172 de Saporta).

Cercle des corrélations

Variables actives

```
> #**** corrélation variables-facteurs ****
> c1 <- acp.autos$loadings[,1]*acp.autos$sdev[1]
> c2 <- acp.autos$loadings[,2]*acp.autos$sdev[2]
> #affichage
> correlation <- cbind(c1,c2)
> print(correlation,digits=2)
```

```
          c1    c2
CYL   -0.89  0.11
PUISS -0.89  0.38
LONG  -0.89 -0.38
LARG  -0.81 -0.41
POIDS -0.91 -0.22
V.MAX -0.75  0.57
```

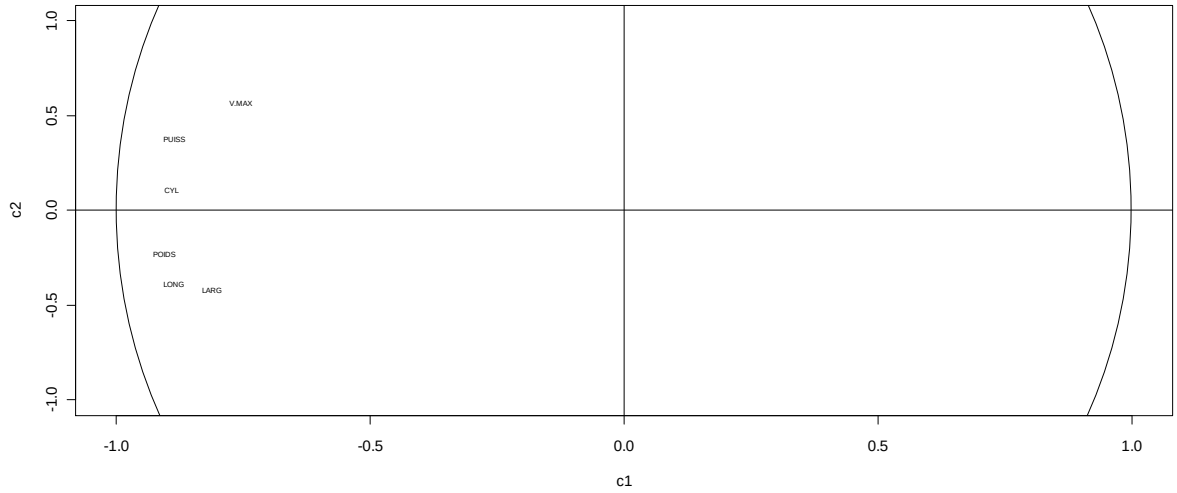
```
> #carrés de la corrélation (cosinus²)
> print(correlation^2,digits=2)
```

```
          c1    c2
CYL   0.80 0.013
PUISS 0.79 0.148
LONG  0.79 0.145
LARG  0.66 0.170
POIDS 0.82 0.050
V.MAX 0.57 0.329
```

```
> #cumul carrés de la corrélation
> print(t(apply(correlation^2,1,cumsum)),digits=2)
```

```
          c1   c2
CYL   0.80 0.81
PUISS 0.79 0.93
LONG  0.79 0.93
LARG  0.66 0.83
POIDS 0.82 0.87
V.MAX 0.57 0.90
```

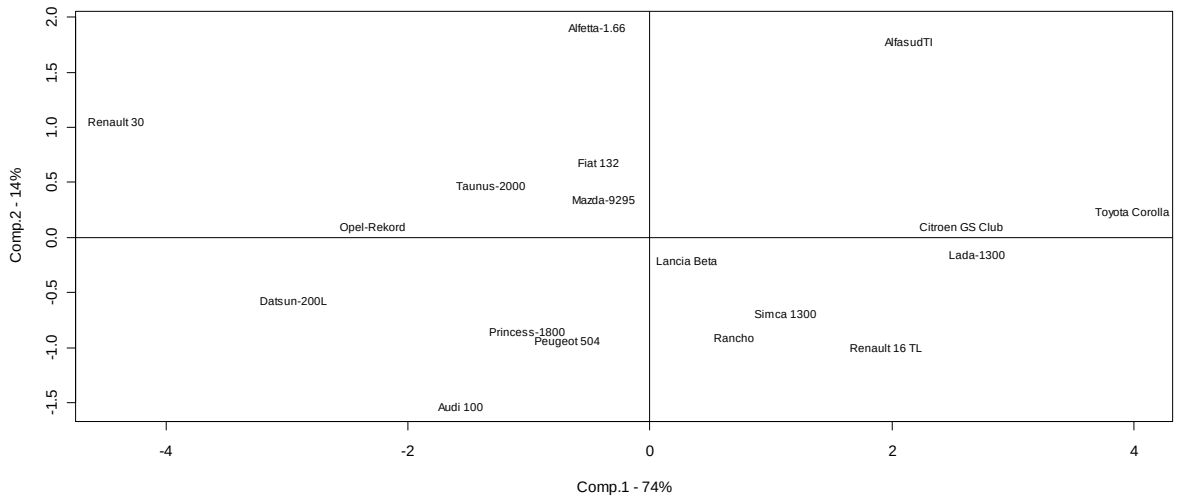
```
> #*** cercle des corrélations - variables actives ***
> plot(c1,c2,xlim=c(-1,+1),ylim=c(-1,+1),type="n")
> abline(h=0,v=0)
> text(c1,c2,labels=colnames(autos.actifs),cex=0.5)
> symbols(0,0,circles=1,inches=F,add=T)
```



Carte des individus sur les 2 premiers axes

Individus actifs

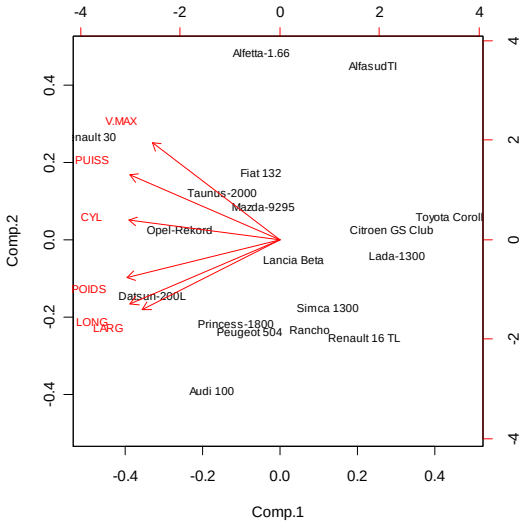
```
> #*** cercle des corrélations - variables actives ***
> plot(c1,c2,xlim=c(-1,+1),ylim=c(-1,+1),type="n")
> abline(h=0,v=0)
> text(c1,c2,labels=colnames(autos.actifs),cex=0.5)
> symbols(0,0,circles=1,inches=F,add=T)
> #l'option "scores" demandé dans princomp est très important ici
>  plot(acp.autos$scores[,1],acp.autos$scores[,2],type="n",xlab="Comp.1
- 74%",ylab="Comp.2 - 14%")
> abline(h=0,v=0)
>
text(acp.autos$scores[,1],acp.autos$scores[,2],labels=rownames(autos.a
ctifs),cex=0.75)
```



Carte des individus et des variables

Individus actifs et variables actives

```
> #*** représentation simultanée : individus x variables
> # cf. Lebart et al., pages 46-48
> biplot(acp.autos,cex=0.75)
```



Attention : N'allons surtout pas chercher une quelconque proximité entre les individus et les variables dans ce graphique. Ce sont les directions qui sont importantes !!! (cf. principe et formules dans Lebart et al., pages 46 à 48)

COSINUS² des individus avec les composantes

Qualité de représentation des individus sur les composantes

```
> #calcul du carré de la distance d'un individu au center de gravité
>             d2              <-              function(x){return(sum(((x-
acp.autos$center)/acp.autos$scale)^2))}
> #appliquer à l'ensemble des observations
> all.d2 <- apply(autos.actifs,1,d2)
> #cosinus^2 pour une composante
> cos2 <- function(x){return(x^2/all.d2)}
> #cosinus^2 pour les composantes retenues (les 2 premières)
> all.cos2 <- apply(acp.autos$scores[,1:2],2,cos2)
> print(all.cos2)
```

	Comp.1	Comp.2
AlfasudTI	0.55621843	0.387670355
Audi 100	0.36533449	0.349406285
Simca 1300	0.58028427	0.210694005
Citroen GS Club	0.97699183	0.001879434
Fiat 132	0.15657878	0.413825852
Lancia Beta	0.08155471	0.033899586
Peugeot 504	0.30920173	0.575488494
Renault 16 TL	0.67353923	0.170535407
Renault 30	0.89243070	0.051919840
Toyota Corolla	0.97521929	0.003425970
Alfetta-1.66	0.04297836	0.820652103
Princess-1800	0.53094720	0.362855147
Datsun-200L	0.77838982	0.028137029
Taunus-2000	0.70481879	0.096496323

```
Rancho          0.24327262 0.410468952
Mazda-9295      0.21733631 0.185337481
Opel-Rekord     0.86189979 0.001789839
Lada-1300       0.92605245 0.002606517
```

La somme pour chaque ligne (individu) vaut 1 si l'on prend l'ensemble des composantes (les 6 composantes).

Cf formules dans Tenenhaus, 2006 ; page 162.

CONTRIBUTION des individus aux composantes

Déterminer les individus qui pèsent le plus dans la définition d'une composante

```
> #contributions à une composante - calcul pour les 2 premières composantes
> all.ctr <- NULL
> for (k in 1:2){all.ctr <- cbind(all.ctr,100.0*(1.0/n)*(acp.autos$scores[,k]^2)/
(acp.autos$sdev[k]^2))}
> print(all.ctr)
                    [,1]        [,2]
AlfasudTI        5.7492537 20.69330670
Audi 100         3.0639513 15.13293304
Simca 1300       1.5746358  2.95251897
Citroen GS Club  8.3243598  0.08269669
Fiat 132         0.2300459  3.13978898
Lancia Beta      0.1163180  0.24968553
Peugeot 504      0.5878172  5.64986748
Renault 16 TL    4.7710990  6.23837245
Renault 30      24.4368837  7.34185620
Toyota Corolla  19.9640247  0.36218506
Alfetta-1.66     0.2407083 23.73566876
Princess-1800    1.3027651  4.59779144
Datsun-200L     10.8701289  2.02916325
Taunus-2000      2.1726682  1.53612980
Rancho           0.6002290  5.23004324
Mazda-9295       0.1869563  0.82332696
Opel-Rekord      6.5887644  0.07065818
Lada-1300        9.2193906  0.13400728
>
```

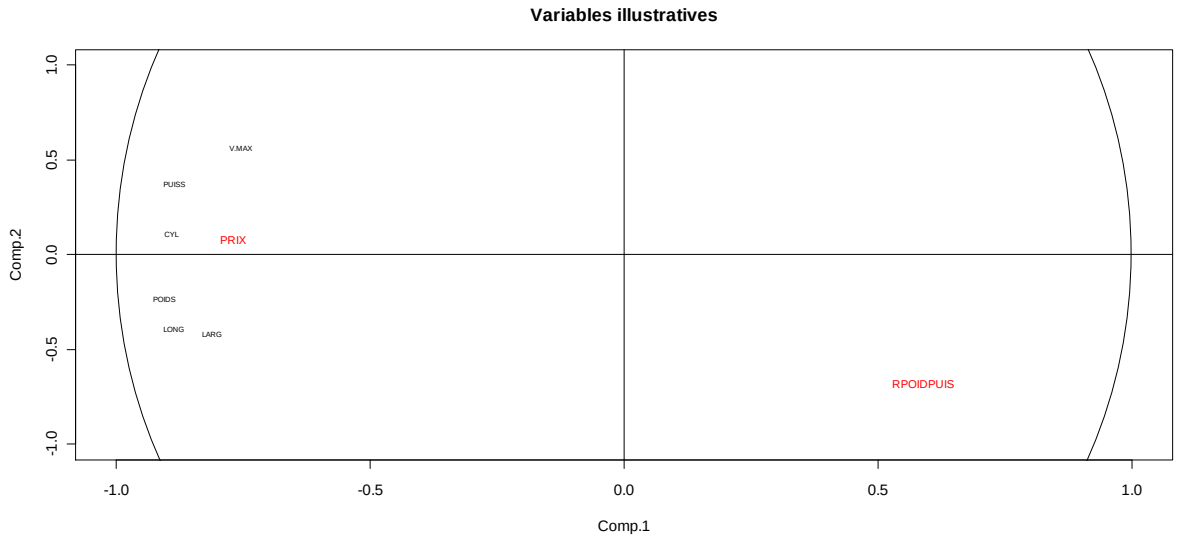
La somme pour chaque colonne (composante) vaut 100.

cf. Saporta, page 175.

Remarque : toutes les observations ont le même poids dans notre exemple.

```
> #*****************************************
> #que faut-il penser de PRIX et R.POIS.PUIS ?
> #*****************************************
> #corrélation de chaque var. illustrative avec le premier axe
> ma_cor_1 <- function(x){return(cor(x,acp.autos$scores[,1]))}
> s1 <- sapply(autos.illus[,2:3],ma_cor_1)
> #corrélation de chaque variable illustrative avec le second axe
> ma_cor_2 <- function(x){return(cor(x,acp.autos$scores[,2]))}
> s2 <- sapply(autos.illus[,2:3],ma_cor_2)
> #position sur le cercle
>       plot(s1,s2,xlim=c(-1,+1),ylim=c(-1,+1),type="n",main="Variables
illustratives",xlab="Comp.1",ylab="Comp.2")
> abline(h=0,v=0)
```

```
> text(s1,s2,labels=colnames(autos.illus[2:3]),cex=1.0)
> symbols(0,0,circles=1,inches=F,add=T)
> #représentation simultanée (avec les variables actives)
>                         plot(c(c1,s1),c(c2,s2),xlim=c(-1,+1),ylim=c(-
1,+1),type="n",main="Variables
illustratives",xlab="Comp.1",ylab="Comp.2")
> text(c1,c2,labels=colnames(autos.actifs),cex=0.5)
> text(s1,s2,labels=colnames(autos.illus[2:3]),cex=0.75,col="red")
> abline(h=0,v=0)
> symbols(0,0,circles=1,inches=F,add=T)
```



La représentation simultanée est autrement plus instructive.

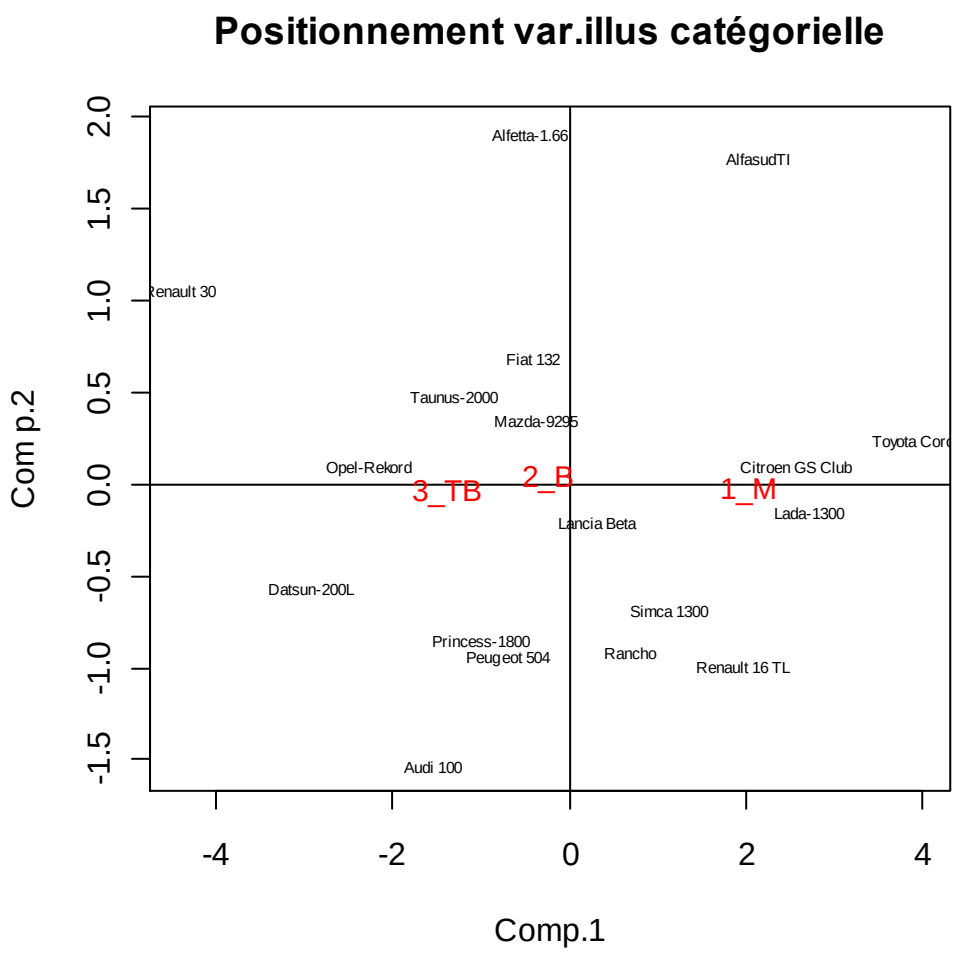
La forte corrélation entre une variable illustrative et un axe est d'autant plus intéressante que la variable n'a pas participé à la construction de l'axe.

Variables qualitatives illustratives

Positionner les groupes associés aux modalités de la variable illustratives

```
>
#*********************************************************************
**********
> #positionner les modalités de la variable illustrative + calcul des
valeurs test
>
#*********************************************************************
**********
> K <- nlevels(autos.illus[,"FINITION"])
> var.illus <- unclass(autos.illus[,"FINITION"])
> m1 <- c()
> m2 <- c()
> for (i in 1:K){m1[i] <- mean(acp.autos$scores[var.illus==i,1])}
> for (i in 1:K){m2[i] <- mean(acp.autos$scores[var.illus==i,2])}
> cond.moyenne <- cbind(m1,m2)
> rownames(cond.moyenne) <- levels(autos.illus[,"FINITION"])
> print(cond.moyenne)
             m1          m2
1_M   2.0003548 -0.02257896
```

```
2_B  -0.2353131  0.04527122
3_TB -1.3924304 -0.03400062
> #graphique
>
plot(c(acp.autos$scores[,1],m1),c(acp.autos$scores[,2],m2),xlab="Comp.
1",ylab="Com            p.2",main="Positionnement            var.illus
catégorielle",type="n")
> abline(h=0,v=0)
>
text(acp.autos$scores[,1],acp.autos$scores[,2],rownames(autos.actifs),
cex=0.5)
> text(m1,m2,rownames(cond.moyenne),cex=0.95,col="red")
>
```



Variables qualitatives illustratives

Moyennes conditionnelles et valeurs test

```
> #*** calcul des valeurs test ***
> #*****************************
> #effectifs par modalité
> nk <- as.vector(table(var.illus))
> #valeur test par composante (les 2 premières)
> vt <- NULL
```

```
> for (j in 1:2){vt <-
cbind(vt,cond.moyenne[,j]/sqrt((val.propres[j]/nk)*((n- nk)/(n-1))))}
> #affichage des valeurs
> tableau <- cbind(cond.moyenne,vt)
> colnames(tableau) <- c("Coord.1","Coord.2","VT.1","VT.2")
> print(tableau)
       Coord.1     Coord.2       VT.1        VT.2
1_M   2.0003548 -0.02257896  2.4327167 -0.06240065
2_B  -0.2353131  0.04527122 -0.3681035  0.16093357
3_TB -1.3924304 -0.03400062 -1.9307662 -0.10713802
```

La valeur test essaie de caractériser la « significativité » de l'écart par rapport à la moyenne globale.

Cf. Saporta, page 177. *On considère qu'il y a un écartement significatif lorsqu'elle est supérieure, en valeur absolue, à 2 voire 3.*

Dans notre exemple, les véhicules se différencient véritablement par la FINITION sur le premier axe factoriel.

Individus illustratifs

Positionner des individus n'ayant pas participé à la construction des axes

Comment situer ces deux nouveaux véhicules ?

Modele	CYL	PUISS	LONG	LARG	POIDS	VMAX
Peugeot 604	2664	136	472	177	1410	180
Peugeot 304 S	1288	74	414	157	915	160

```
#chargement de la seconde feuille de calcul + vérification
ind.illus =
read.table("c:/DATAR/IndIllustratifs.txt",row.names=1,sep="\t",h=T)
summary(ind.illus)
#centrage et réduction en utilisant les moyennes et écarts-type de
l'ACP
ind.scaled <- NULL
for (i in 1:nrow(ind.illus)){ind.scaled <-
rbind(ind.scaled,(ind.illus[i,] - acp.autos$center)/acp.autos$scale)}
print(ind.scaled)
#calcul des coordonnées factorielles (en utilisant les valeurs propres
cf. loadings)
produit.scal <- function(x,k){return(sum(x*acp.autos$loadings[,k]))}
ind.coord <- NULL
for (k in 1:2){ind.coord <-
cbind(ind.coord,apply(ind.scaled,1,produit.scal,k))}
```

```
print(ind.coord)
#*** projection des individus actifs ET illustratifs dans le premier plan factoriel
plot(c(acp.autos$scores[,1],ind.coord[,1]),c(acp.autos$scores[,2],ind.coord[,2]),xlim =c(-6,6),type="n",xlab="Comp.1",ylab="Comp.2")
abline(h=0,v=0)
text(acp.autos$scores[,1],acp.autos$scores[,2],labels=rownames(autos.actifs),cex=0.5)
text(ind.coord[,1],ind.coord[,2],labels=rownames(ind.illus),cex=0.5,col="red")
```

> #chargement de la seconde feuille de calcul + vérification

> ind.illus = read.table("c:/DATAR/IndIllustratifs.txt",row.names=1,sep="\t",h=T)

> summary(ind.illus)

CYL PUISS LONG LARG POIDS

Min. :1288 Min. : 74.0 Min. :414.0 Min. :157 Min. : 915

1st Qu.:1632 1st Qu.: 89.5 1st Qu.:428.5 1st Qu.:162 1st Qu.:1039

Median :1976 Median :105.0 Median :443.0 Median :167 Median :1162

Mean :1976 Mean :105.0 Mean :443.0 Mean :167 Mean :1162

3rd Qu.:2320 3rd Qu.:120.5 3rd Qu.:457.5 3rd Qu.:172 3rd Qu.:1286

Max. :2664 Max. :136.0 Max. :472.0 Max. :177 Max. :1410

VMAX

Min. :160

1st Qu.:165

Median :170

Mean :170

3rd Qu.:175

Max. :180

> #centrage et réduction en utilisant les moyennes et écarts-type de l'ACP

> ind.scaled <- NULL

> for (i in 1:nrow(ind.illus)){ind.scaled <- rbind(ind.scaled,(ind.illus[i,] - acp.autos$center)/acp.autos$scale)}

> print(ind.scaled)

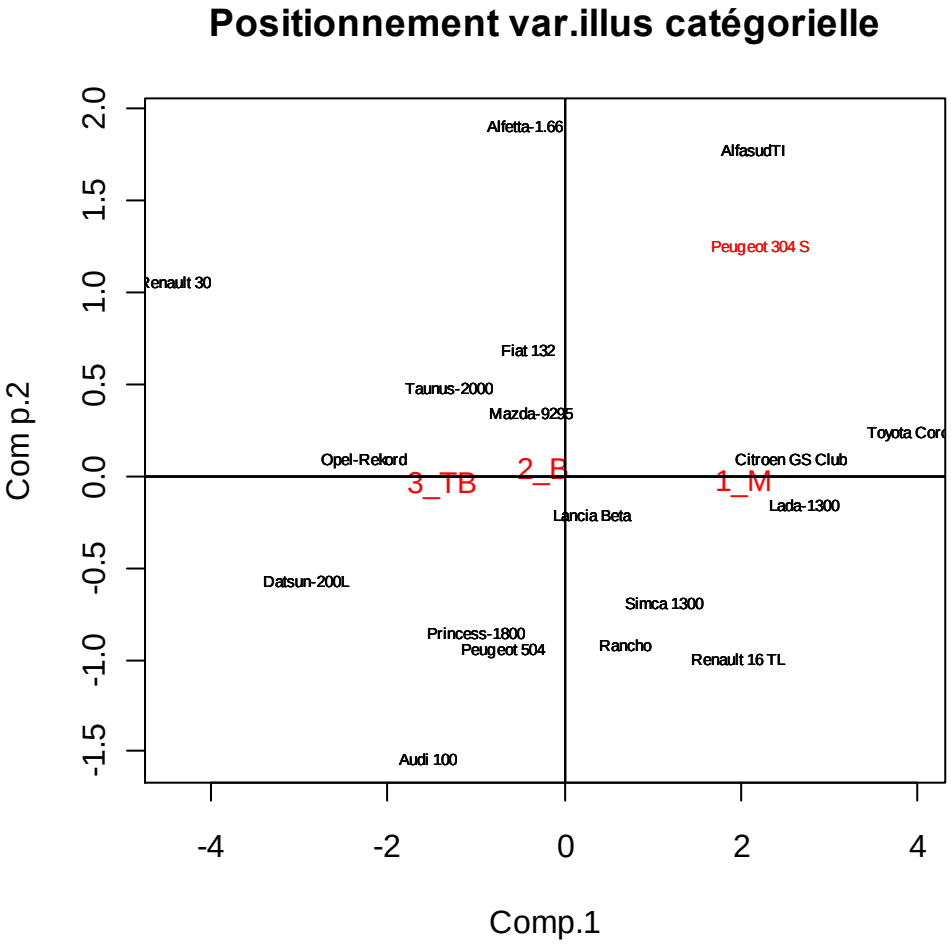
CYL PUISS LONG LARG POIDS VMAX

Peugeot 604 2.8408062 2.5951120 1.7919904 2.001041 2.488122 1.8411267

Peugeot 304 S -0.9457124 -0.5358556 -0.9076315 -1.871942 -1.230913 0.1459717

> #calcul des coordonnées factorielles (en utilisant les valeurs propres cf. loadings)

```
> produit.scal <- function(x,k){return(sum(x*acp.autos$loadings[,k]))}
> ind.coord <- NULL
> for (k in 1:2){ind.coord <- cbind(ind.coord,apply(ind.scaled,1,produit.scal,k))}
> print(ind.coord)
                [,1]     [,2]
Peugeot 604   -5.563292 0.3386093
Peugeot 304 S  2.212241 1.2577790
> #*** projection des individus actifs ET illustratifs dans le premier plan factoriel
plot(c(acp.autos$scores[,1],ind.coord[,1]),c(acp.autos$scores[,2],ind.coord[,2]),xlim
=c(-6,6),type="n",xlab="Comp.1",ylab="Comp.2")
> abline(h=0,v=0)
>
text(acp.autos$scores[,1],acp.autos$scores[,2],labels=rownames(autos.actifs),cex=0.5
)
> text(ind.coord[,1],ind.coord[,2],labels=rownames(ind.illus),cex=0.5,col="red")
>
```



Traitements réalisés

Réaliser une ACP sur un fichier de données.
Afficher les valeurs propres. Construire le graphiques éboulis des valeurs propres.
Construire le cercle de corrélations.
Projeter les observations dans le premier plan factoriel.
Positionner des variables illustratives quantitatives dans le cercle des corrélations.
Positionner les modalités d'une variable illustrative catégorielle.
Positionner des observations illustratives.

Bibliographie

www.marketing-etudiant.fr. Vous y trouverez cours, exposés, mémoires,

Gardat S. 2008, Analyse factorielle des correspondances multiples, Laboratoire de statistique et de probabilité UMR 5583 CNRS-UPS

Benzecri, J.P. 1973. L'analyse des données, Tome 2 : L'analyse des Correspondan:

ces, Dunod, Paris