Introduction au Clustering avec R

Jean-Marie Marion
Institut de Mathématiques Appliquées
Université Catholique de l'Ouest
jean-marie.marion@uco.fr

Le **Clustering (**ou **partitionnement des données)** est une méthode statistique utilisée pour grouper des données en classes.

Les données de chaque classe doivent avoir

des caractéristiques communes

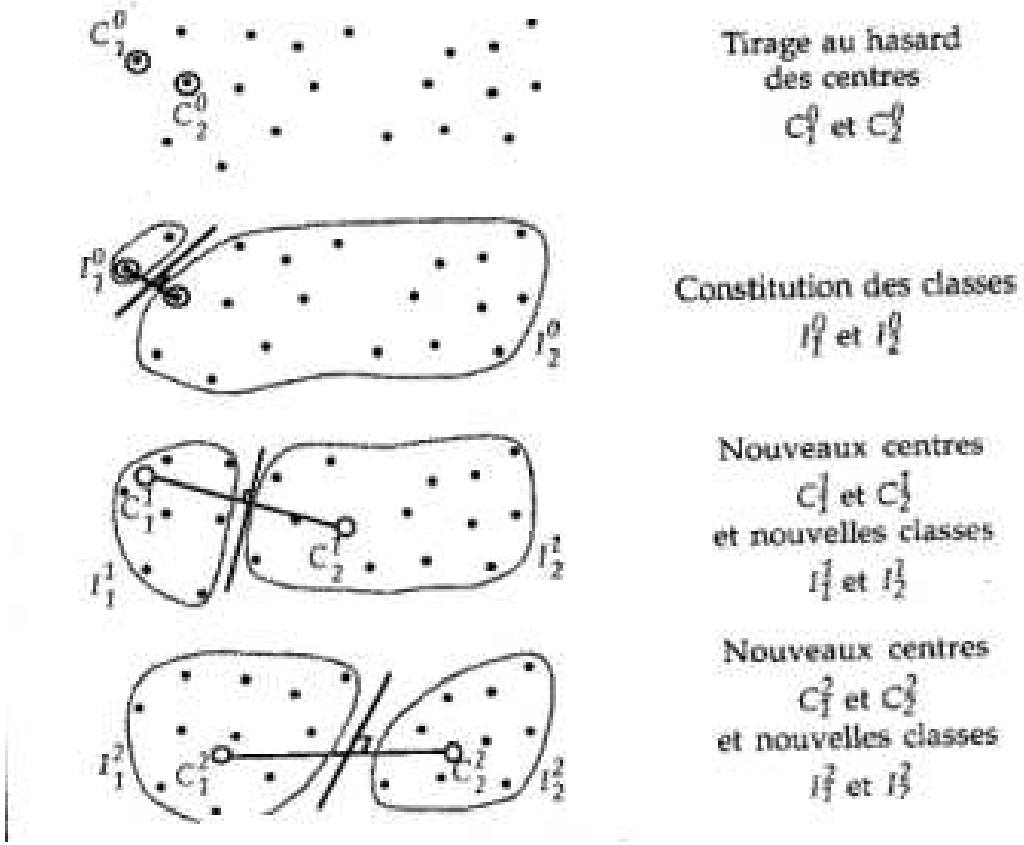
Pour obtenir un bon partitionnement, il faut:

- obtenir des classes les plus homogènes possibles
(inertie intra classe minimale)

- obtenir des classes les plus différenciées possibles
(inertie inter classes maximale)

Méthodes de Clustering

- Méthodes non hiérarchiques ⟶ Partitionnement en k classes



Lebart L- Morineau A- Piron M

- Méthode des centres mobiles

- Choix aléatoire de k points (centres de classes)
- Itérer les 2 étapes suivantes jusqu'à ce que le critère inertie intra classes ne décroisse plus de manière significative:
    - tous les individus sont affectés à la classe dont le centre est le plus proche au sens de la « distance » choisie → k classes d'individus
    - calculer les barycentres des classes créées qui deviennent les k nouveaux centres

- Méthode des k-means

Les barycentres des classes ne sont pas recalculés à la fin des affectations mais après chaque allocation d'un individu à une classe ⟶ algorithme plus rapide

• • •

Avantage de ces méthodes: - classification de données massives

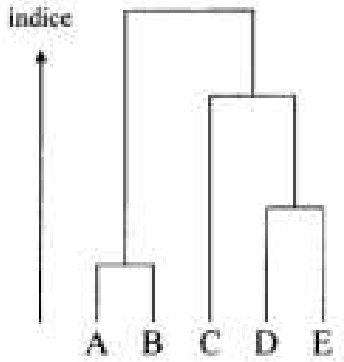
Inconvénients de ces méthodes: - le nombre de classes est imposé au départ
- la répartition en classes dépend du choix initial des centres

- Méthodes hiérarchiques

Fournir un ensemble de partitions de moins en moins fines obtenues par regroupements successifs de parties (partitions emboîtées)

Classification ascendante hiérarchique (CAH)

…



Avantage de ces méthodes: - la lecture du dendogramme permet de déterminer le nombre optimal de classes

Inconvénient de ces méthodes: - non adapté aux données massives (coûteux en temps de calcul)

Pourquoi faire une Classification après une Analyse en Composantes Principales (ACP)?

- **Réduire le « bruit » dans les données afin d'avoir une classification plus stable**

- **Dans le cas d'un grand fichier de données multivariées, réduire la dimension des données à quelques variables continues contenant les informations les plus pertinentes sur ces données.**

Analyse en Composantes Principales (ACP)

L'ACP est une méthode d'Analyse des Données qui permet de décrire un ensemble de données numériques multivariées, d'en réduire la dimensionnalité

Elle transforme les variables originelles, corrélées entre elles, en variables décorrélées les unes des autres appelées « composantes principales »

Le nuage de points construit à partir des premières composantes principales contient la majeure partie de l'information du nuage de départ.

Exemple: **EAUX MINERALES**

```
> setwd("K:/IMA 2018_2019/SCHOOL OF AI_11 juillet 2019")

> donnees<-read.table("eaux minerales_bis.csv",header=TRUE,sep=";",dec=",",row.names=1)
> head(donnees,8)
                  TYPE PG   CA   MG  NA.   K  SUL NO3  HCO3   CL
Evian                M  P 78.0 24.0  5.0 1.0 10.0 3.8 357.0  4.5
Montagne Pyrenees    S  P 48.0 11.0 34.0 1.0 16.0 4.0 183.0 50.0
Cristaline           S  P 71.0  5.5 11.2 3.2  5.0 1.0 250.0 20.0
Fiee des Lois        S  P 89.0 31.0 17.0 2.0 47.0 0.0 360.0 28.0
Volcania             S  P  4.1  1.7  2.7 0.9  1.1 0.8  25.8  0.9
Luchon               M  P 26.5  1.0  0.8 0.2  8.2 1.8  78.1  2.3
Volvic               M  P  9.9  6.1  9.4 5.7  6.9 6.3  65.3  8.4
Alpes/Moulettes      S  P 63.0 10.2  1.4 0.4 51.3 2.0 173.2  1.0
...
    Pyrenees         M  G  48 12.0 31.0 1.0 18.0  4.0 183.0 35.0
    Montcalm         S  P   3  0.6  1.5 0.4  8.7  0.9   5.2  0.6
    Chantereine      S  P 119 28.0  7.0 2.0 52.0  0.0 430.0  7.0
    18 Carats        S  G 118 30.0 18.0 7.0 85.0  0.5 403.0 39.0
    Spring Water     S  G 117 19.0 13.0 2.0 16.0 20.0 405.0 28.0
    Montclar         S  P  41  3.0  2.0 0.0  2.0  3.0 134.0  3.0
```

Individus: 35 **Variables: 10** (8 actives, 2 nominales supplémentaires)

TYPE: M minérale / S source **PG: P plate / G gazeuse**

```
> donnees1<-donnees [,3:10]
> mcor<-cor(donnees1)
> mcor
             CA         MG         NA.          K         SUL         NO3         HCO3          CL
CA    1.00000000  0.5362007  0.3606775  0.4273343  0.5361705  0.16354362  0.90805944  0.09390068
MG    0.53620069  1.0000000  0.7502490  0.7707790  0.8648286 -0.14170715  0.79923224  0.21908387
NA.   0.36067747  0.7502490  1.0000000  0.9318875  0.7944543 -0.10148981  0.66290351  0.40822052
K     0.42733427  0.7707790  0.9318875  1.0000000  0.8418393 -0.14892396  0.72669284  0.12286674
SUL   0.53617051  0.8648286  0.7944543  0.8418393  1.0000000 -0.24986680  0.74990438  0.24460404
NO3   0.16354362 -0.1417072 -0.1014898 -0.1489240 -0.2498668  1.00000000  0.02204385  0.10298603
HCO3  0.90805944  0.7992322  0.6629035  0.7266928  0.7499044  0.02204385  1.00000000  0.11213255
CL    0.09390068  0.2190839  0.4082205  0.1228667  0.2446040  0.10298603  0.11213255  1.00000000
```

```
install.packages("corrplot")
library(corrplot)
corrplot(mcor, type="upper", order="hclust", tl.col="black", tl.srt=45)
```

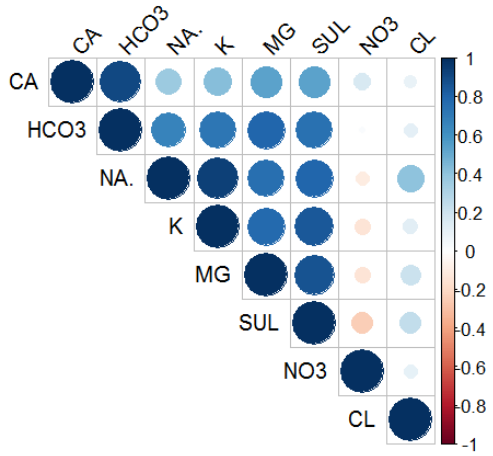


Tableau: individus x variables

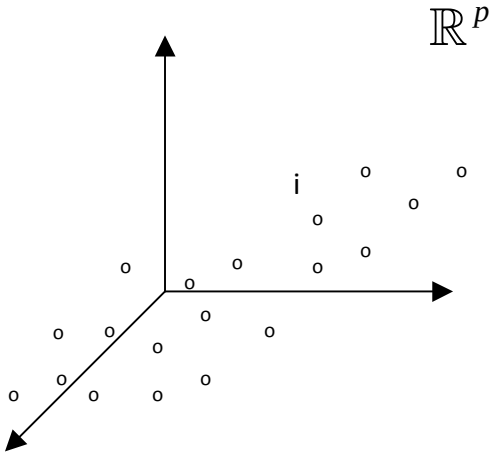
	X_1	.....	X_j	.....	X_p
Individu **1**					
.....			.....		
Individu **i**		.....	x_{ij}	.....	
.....			.....		
Individu **n**					

$X_1, X_2, ..., X_p$ ⟶ *p* variables numériques observées sur n individus (1,…n)

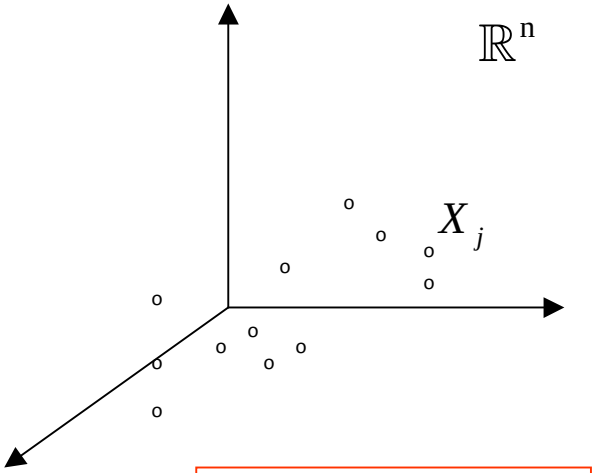
X_{ij} ⟶ valeur de la variable *j* sur l'individu i

(**Dans l'Exemple « Eaux minérales » n = 35, *p* = 8**)

Dans l'Exemple « Eaux minérales» n = 35, *p* = 8

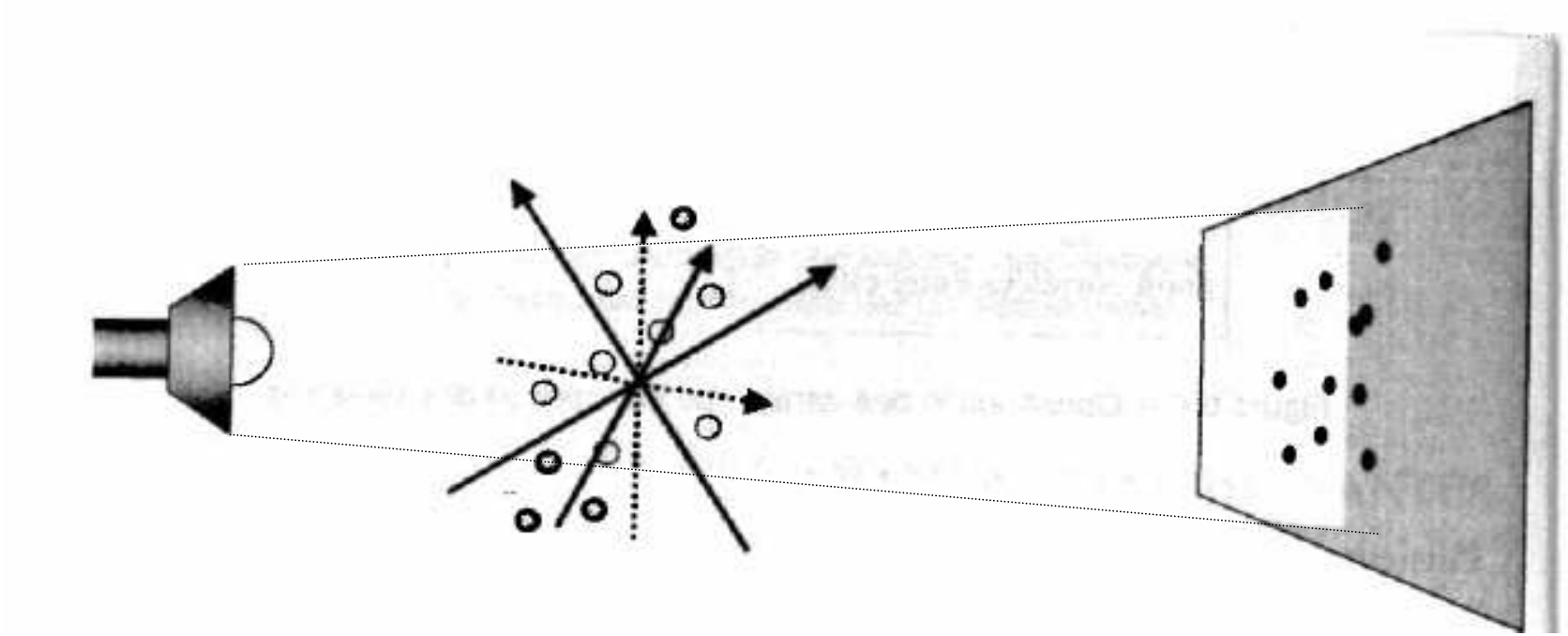


Espace des individus



Espace des variables

On veut réduire l'espace initial en passant à un espace de dimension 2 (plan) par projection tout en gardant l'essentiel de l'information du tableau de départ.



Cas de variables non homogènes

-Variables avec unités différentes

-Variables avec mêmes unités mais échelles différentes

-Variables avec des variances très différentes ….

→ **ACPN**

15

Calcul des moyennes des composants

```
> apply(donnees1,2,mean)
        CA         MG        NA.          K        SUL        NO3       HCO3         CL
63.962857  12.140857  14.678571   3.175714  21.477143   3.985714 240.845714  14.231429
```

Calcul des écart-types des composants

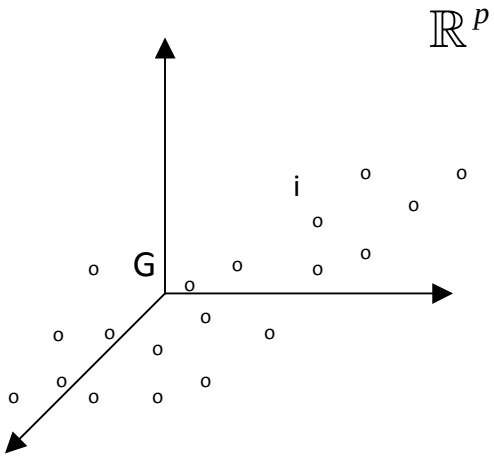
```
> apply(donnees1,2,sd)
        CA         MG        NA.          K        SUL        NO3       HCO3         CL
54.607935  13.966915  26.380372   8.154684  29.884120   6.158120 228.511128  14.719590
```

```
> scale(donnees1)
                                CA          MG          NA.           K          SUL          NO3
Evian                  0.25705317  0.84908820 -0.36688533 -0.26680545 -0.38405490 -0.030157629
Montagne Pyrenees     -0.29231754 -0.08168283  0.73241684 -0.26680545 -0.18327937  0.002319818
Cristaline             0.12886667 -0.47547057 -0.13186211  0.00297813 -0.55136784 -0.484841880
Fiee des Lois          0.45848910  1.35027260  0.08799833 -0.14417655  0.85406085 -0.647229113
Volcania              -1.09623002 -0.74754210 -0.45407137 -0.27906834 -0.68187194 -0.517319327
Luchon                -0.68603322 -0.79766054 -0.52609461 -0.36490858 -0.44428756 -0.354932094
Volvic                -0.99001835 -0.43251191 -0.20009466  0.30955039 -0.48778893  0.375810453
Alpes/Moulettes       -0.01763218 -0.13896105 -0.50335043 -0.34038280  0.99794998 -0.322454648


                              HCO3          CL
Evian                  0.50830910 -0.66112089
Montagne Pyrenees     -0.25314178  2.42999777
Cristaline             0.04006057  0.39189756
Fiee des Lois          0.52143756  0.93539095
Volcania              -0.94107327 -0.90569292
Luchon                -0.71220039 -0.81058157
Volvic                -0.76821517 -0.39616786
Alpes/Moulettes       -0.29602810 -0.89889925
```

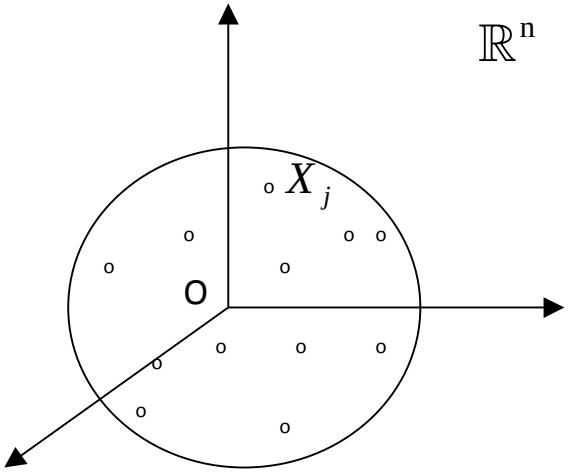
• • •

Chaque variable a comme moyenne 0 et comme variance 1

Dans l'exemple « Eaux minérales » n = 35, *p* = 8



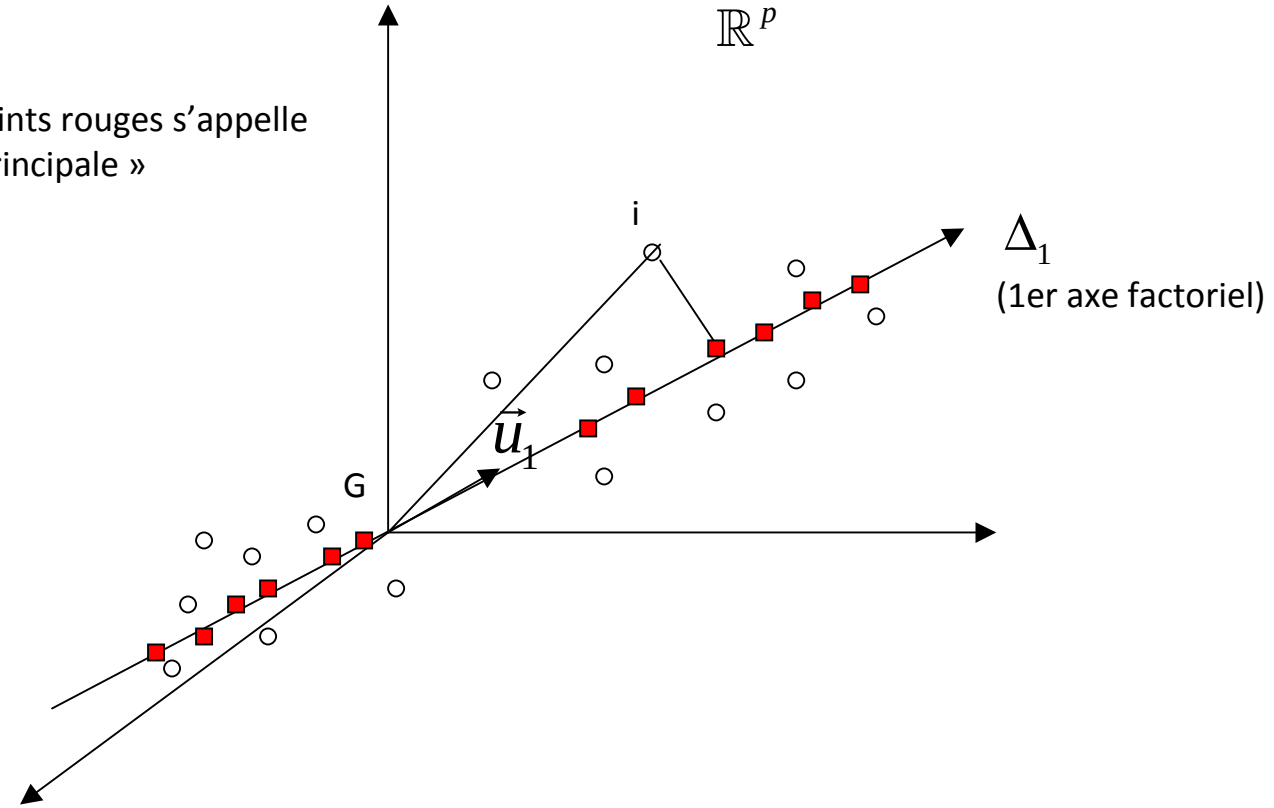
Espace des individus



Espace des variables

Dans l'exemple « Eaux minérales » n = 35, *p* = 8

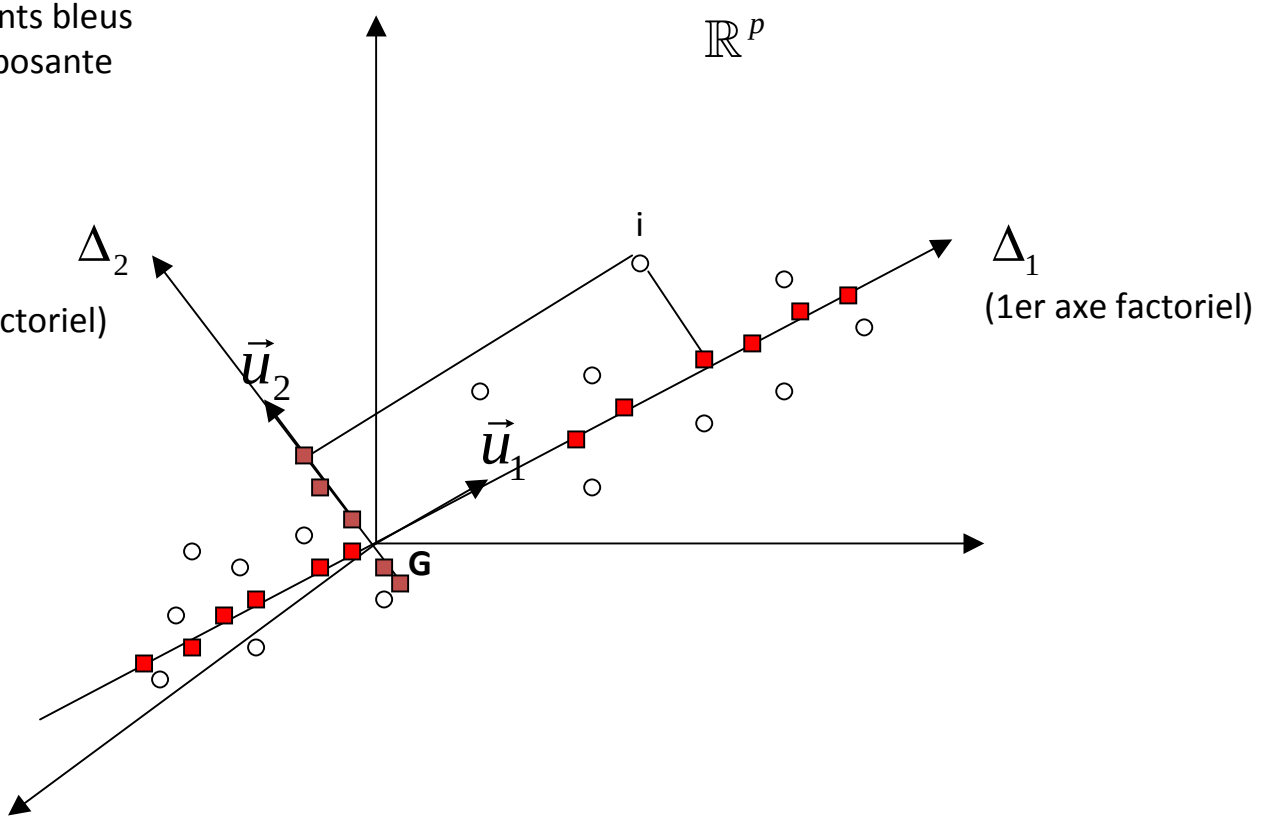
L'ensemble des 35 points rouges s'appelle « 1ère composante principale »



$\vec{u}_1$ se calcule à partir de la matrice des corrélations des 8 variables de même que la variance (notée λ_1) des projections des 35 individus sur Δ_1

Dans l'exemple « Eaux minérales » n = 35, *p* = 8

L'ensemble des 35 points bleus s'appelle « 2ème composante principale »



$\vec{u}_2$ se calcule à partir de la matrice des corrélations des 8 variables de même que la <u>variance</u> (notée λ_2) des projections des 35 individus sur Δ_2

Mise en œuvre avec FactoMineR

```
> library("FactoMineR")
> res<-PCA(donnees,scale.unit=TRUE, graph=TRUE, quali.sup=1:2)
```

```
> res
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 35 individuals, described by 10 variables
*The results are available in the following objects:

   name                 description
1  "$eig"               "eigenvalues"
2  "$var"               "results for the variables"
3  "$var$coord"         "coord. for the variables"
4  "$var$cor"           "correlations variables - dimensions"
5  "$var$cos2"          "cos2 for the variables"
6  "$var$contrib"       "contributions of the variables"
7  "$ind"               "results for the individuals"
8  "$ind$coord"         "coord. for the individuals"
9  "$ind$cos2"          "cos2 for the individuals"
10 "$ind$contrib"       "contributions of the individuals"
11 "$quali.sup"         "results for the supplementary categorical variables"
12 "$quali.sup$coord"   "coord. for the supplementary categories"
13 "$quali.sup$v.test"  "v-test of the supplementary categories"
14 "$call"              "summary statistics"
15 "$call$centre"       "mean of the variables"
16 "$call$ecart.type"   "standard error of the variables"
17 "$call$row.w"        "weights for the individuals"
18 "$call$col.w"        "weights for the variables"
```

Dans l'exemple « Eaux minérales » n = 35, *p* = 8

```
> res$eig
        eigenvalue percentage of variance cumulative percentage of variance
comp 1 4.663954355            58.29942944                           58.29943
comp 2 1.241017692            15.51272115                           73.81215
comp 3 1.082980256            13.53725321                           87.34940
comp 4 0.605172634             7.56465793                           94.91406
comp 5 0.258725516             3.23406895                           98.14813
comp 6 0.129808965             1.62261206                           99.77074
comp 7 0.016928647             0.21160809                           99.98235
comp 8 0.001411934             0.01764918                          100.00000
```

λ_j avec $1 \leq j \leq p$

$$\sum_j \lambda_j = p$$

Nous ne retiendrons que les $\lambda_j \geq 1$

(ici nous récupérons 73,81% de la variance du nuage des individus en projetant sur le plan principal!)

Coordonnées des variables

Corrélations entre variables et composantes principales

Cosinus carrés (qualité de la représentation de la variable *j*)

Contributions des variables à la création des axes

```
> res$var
$coord
          Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
CA    0.6871477  0.55125003 -0.29019511 -0.346038583 -0.13402691
MG    0.9087753 -0.07065107 -0.03232868  0.005577621  0.37671287
NA.   0.8829193 -0.18027465  0.28071358  0.250515998 -0.18247112
K     0.9024314 -0.19171881 -0.01747604  0.333828184 -0.16882231
SUL   0.9260619 -0.18588867 -0.01289597 -0.020907282  0.15073582
NO3  -0.1152593  0.85737080  0.25347911  0.419040764  0.09813435
HCO3  0.9048851  0.29857773 -0.24100659 -0.137296577 -0.06579104
CL    0.2868599  0.06408217  0.89225234 -0.341052305 -0.01931152

$cor
          Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
CA    0.6871477  0.55125003 -0.29019511 -0.346038583 -0.13402691
MG    0.9087753 -0.07065107 -0.03232868  0.005577621  0.37671287
NA.   0.8829193 -0.18027465  0.28071358  0.250515998 -0.18247112
K     0.9024314 -0.19171881 -0.01747604  0.333828184 -0.16882231
SUL   0.9260619 -0.18588867 -0.01289597 -0.020907282  0.15073582
NO3  -0.1152593  0.85737080  0.25347911  0.419040764  0.09813435
HCO3  0.9048851  0.29857773 -0.24100659 -0.137296577 -0.06579104
CL    0.2868599  0.06408217  0.89225234 -0.341052305 -0.01931152

$cos2
          Dim.1       Dim.2       Dim.3       Dim.4       Dim.5
CA   0.47217197 0.303876596 0.0842132041 1.197427e-01 0.0179632126
MG   0.82587252 0.004991574 0.0010451434 3.110985e-05 0.1419125868
NA.  0.77954655 0.032498950 0.0788001163 6.275827e-02 0.0332957097
K    0.81438240 0.036756101 0.0003054121 1.114413e-01 0.0285009732
SUL  0.85759058 0.034554598 0.0001663060 4.371144e-04 0.0227212878
NO3  0.01328470 0.735084687 0.0642516599 1.755952e-01 0.0096303498
HCO3 0.81881702 0.089148662 0.0580841768 1.885035e-02 0.0043284612
CL   0.08228862 0.004106524 0.7961142379 1.163167e-01 0.0003729347

$contrib
          Dim.1      Dim.2       Dim.3        Dim.4     Dim.5
CA   10.1238549 24.4860809  7.77606088 19.786535984  6.942961
MG   17.7075601  0.4022162  0.09650623  0.005140657 54.850634
NA.  16.7142834  2.6187339  7.27622834 10.370307881 12.869125
K    17.4612001  2.9617709  0.02820108 18.414787814 11.015911
SUL  18.3876280  2.7843759  0.01535633  0.072229711  8.782005
NO3   0.2848377 59.2324099  5.93285607 29.015714253  3.722227
HCO3 17.5562828  7.1835126  5.36336433  3.114871521  1.672994
CL    1.7643530  0.3308997 73.51142674 19.220412178  0.144143
```

Interprétation de la « position » des variables

(Cas où les variables sont proches du cercle des corrélations) ⟶ bien représentées en projection

$$d^2\left(X_j, X_{j'}\right) = 2\left[1 - corr(X_j, X_{j'})\right]$$

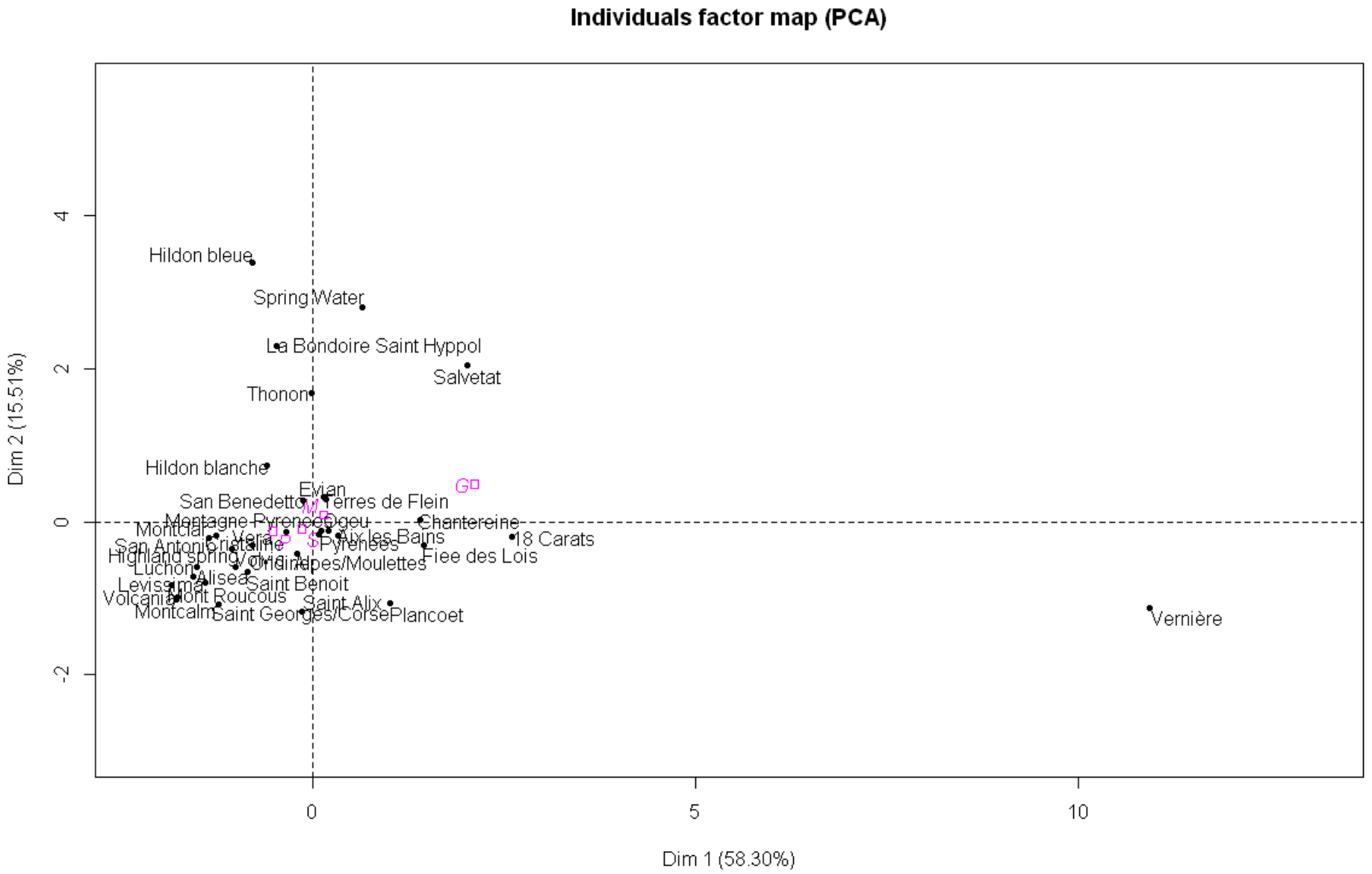
-2 variables fortement corrélées positivement sont proches l'une de l'autre

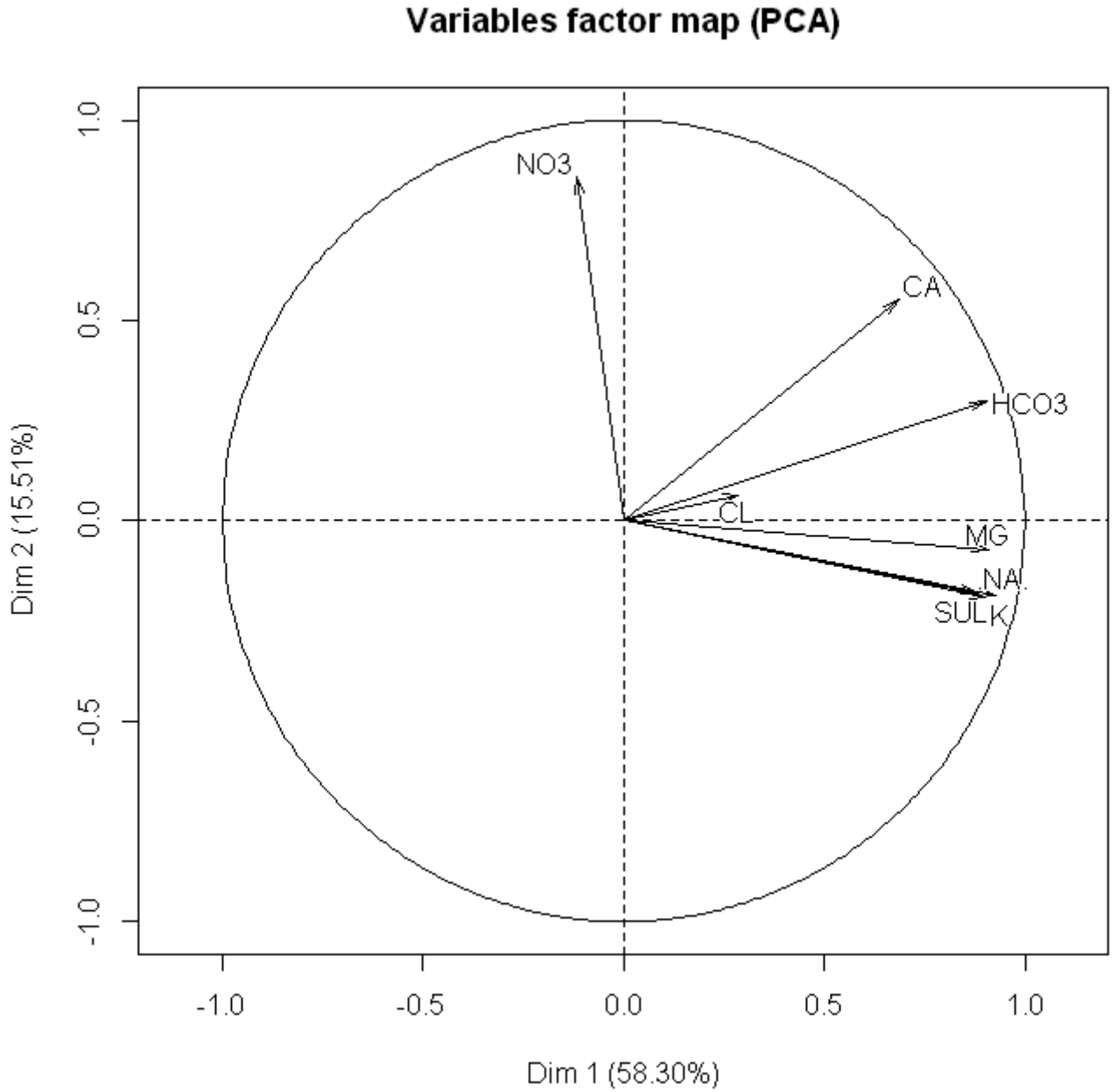
(et réciproquement)

-2 variables fortement corrélées négativement sont les plus éloignées possibles l'une de l'autre

(et réciproquement)

- 2 variables non corrélées sont orthogonales

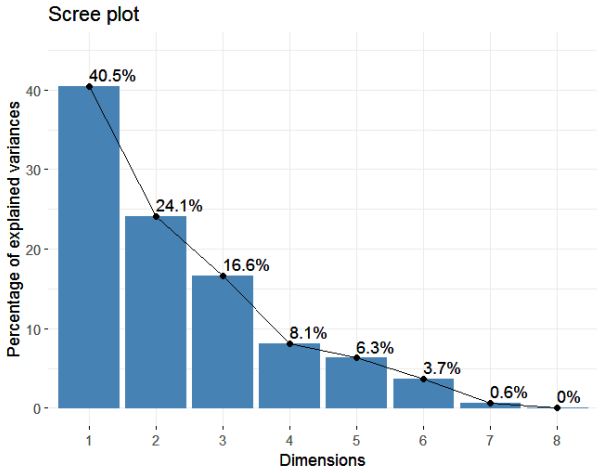

Individuals factor map (PCA)
Hildon bleue
Spring Water
La Bondoire Saint Hyppol
Salvetat
Thonon
Hildon blanche
Evian
San Benedetto
Terres de Flein
Montagne Pyrenees
Ogeu
Chantereine
Montclar
Aix les Bains
18 Carats
Pyrenees
Fiee des Lois
Luchon
Alpes/Moulettes
Alisea
Saint Benoit
Mont Roucous
Saint Alix
Montcalm
Saint Georges/Corse
Plancoet
Vernière
Dim 2 (15.51%)
Dim 1 (58.30%)
4
2
0
-2
0
5
10


Variables factor map (PCA)
NO3
CA
HCO3
CL
MG
NA
SULK
1.0
0.5
0.0
-0.5
-1.0
Dim 2 (15.51%)
Dim 1 (58.30%)

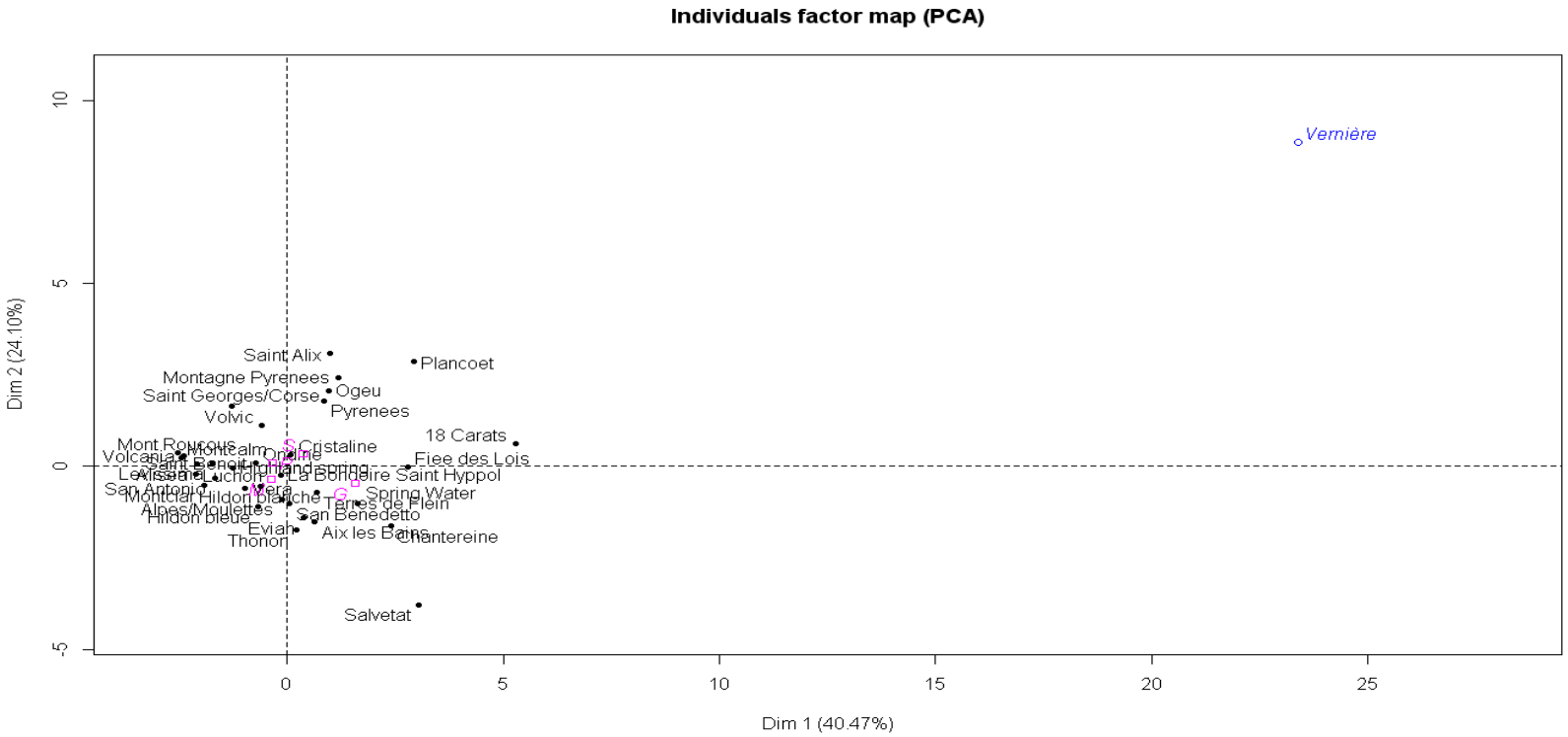
```
> res<-PCA(donnees,scale.unit=TRUE, ind.sup=28, quali.sup=1:2)
```

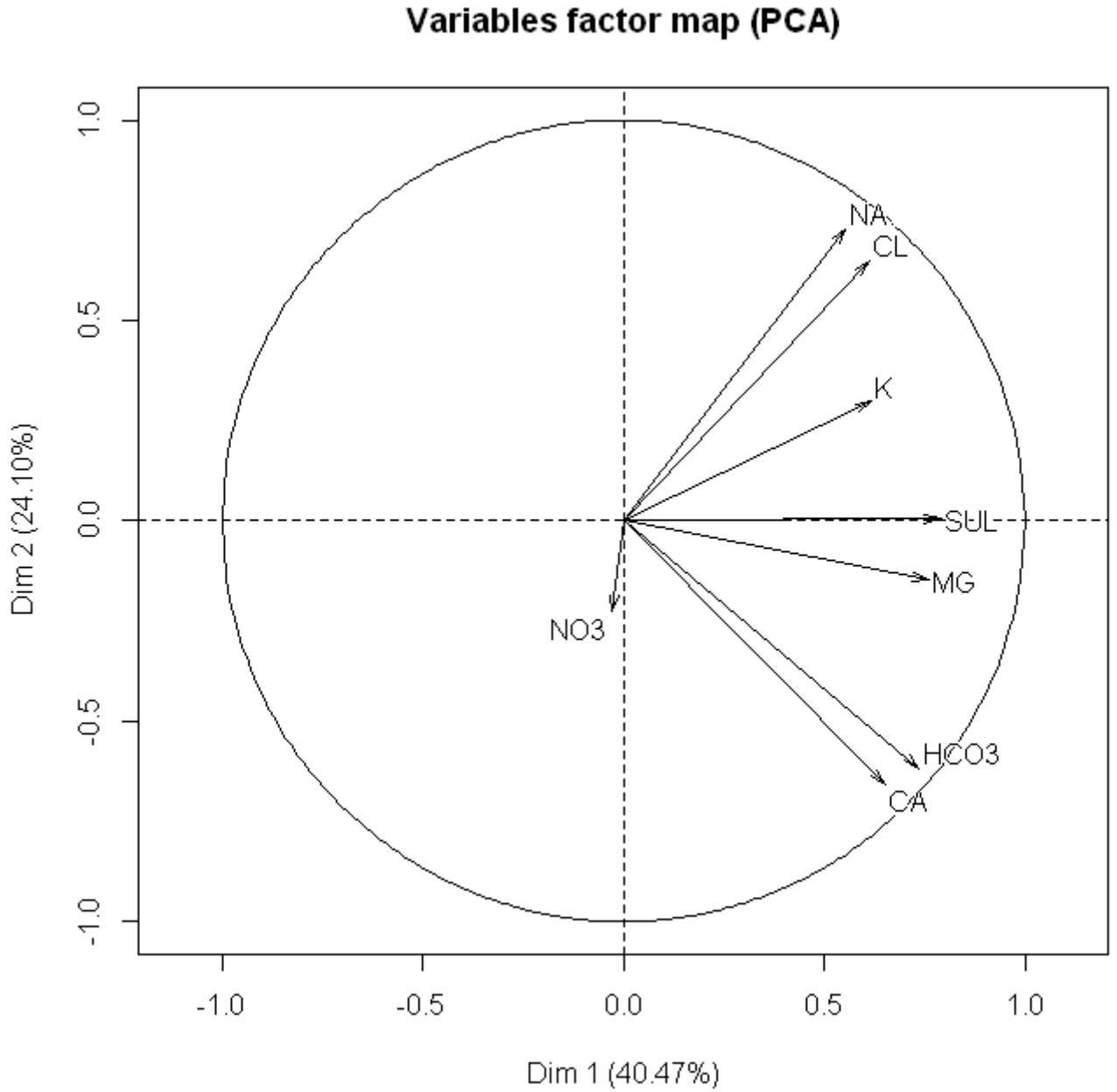
```
>   install.packages("factoextra")
>   library("factoextra")
```

fviz_eig(res,addlabels=TRUE,ylim=c(0,45))

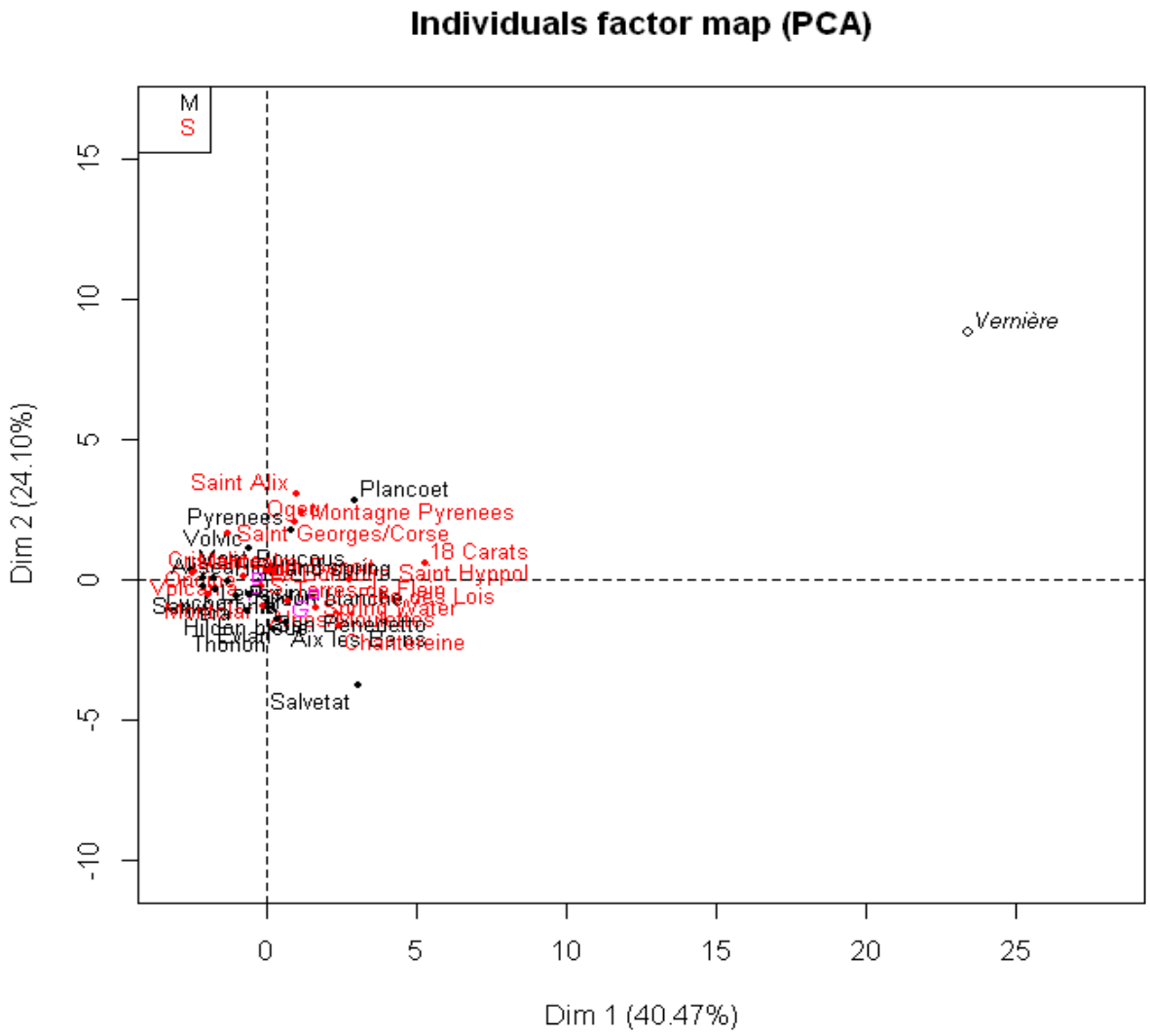


La somme des 4 premières valeurs propres: 89.3%



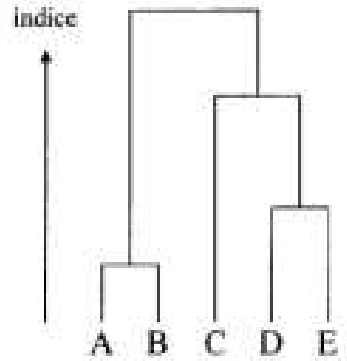

Variables factor map (PCA)
NA
CL
K
SUL
MG
NO3
HCO3
CA
1.0
0.5
0.0
-0.5
-1.0
Dim 2 (24.10%)
Dim 1 (40.47%)

```
> plot(res, cex=0.8,habillage="TYPE")
```



Classification Ascendante Hiérarchique (CAH)

Objectif : construire une hiérarchie (indicée) sur les individus
afin de regrouper en catégories homogènes vis à vis des variables
les individus d'un ensemble ⟶ dendogramme



A partir d'un tableau de données, construire une partition telle que:

- **A l'intérieur de chaque classe les individus se ressemblent**
- **D'une classe à l'autre les individus diffèrent**

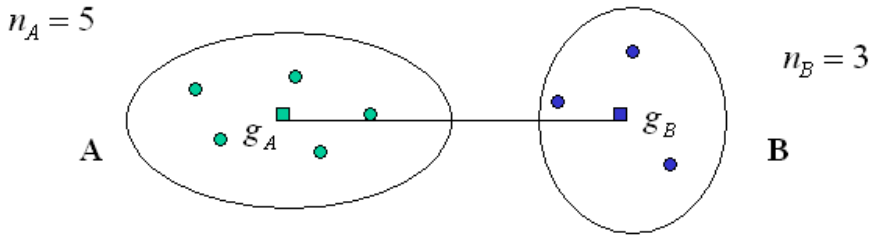
Obtenir les classes les plus homogènes possibles et les plus différenciées possibles

Distance entre individus (permet de voir leur ressemblance)

individus i et l $\qquad d^2(i,l) = \sum_{j=1}^{p} \left(x_{ij} - x_{lj} \right)^2$

Il existe d'autres distances….

Distance entre groupes d'individus



$$d_1(A,B) = \min_{(x,y)\in A\times B} \{d(x,y)\}$$

$$d_2(A,B) = \max_{(x,y)\in A\times B} \{d(x,y)\}$$

$$d_3(A,B) = d(g_A, g_B)$$

Ici nous considèrerons une distance basée sur l'inertie des groupes

La méthode hiérarchique ascendante

Pas 1: Parmi les n individus on sélectionne les deux « plus proches » au sens de la distance retenue. Une première classe est construite avec ces 2 individus. Il reste donc n-2 éléments à classer.

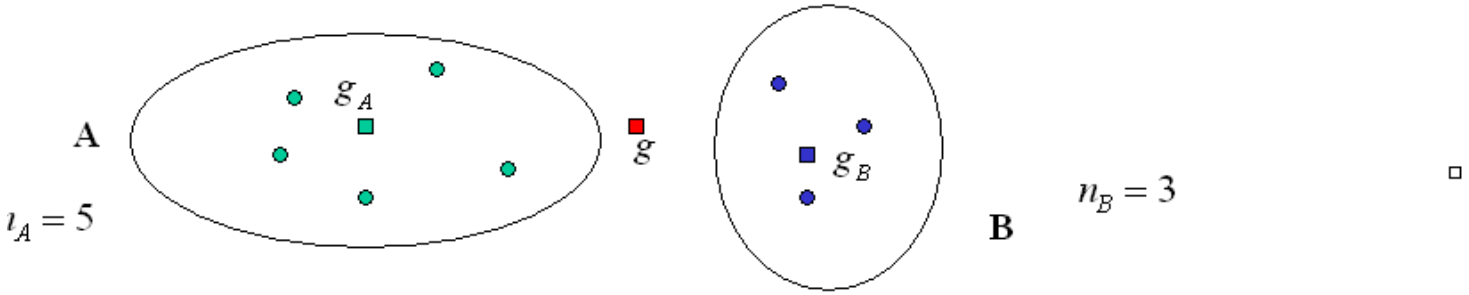
Pas 2: On calcule la « distance » entre la classe construite au Pas1 et les éléments restants. A nouveau les 2 éléments les plus proches sont regroupés dans une nouvelle classe

Pas suivants: Réitérer le Pas2 jusqu'à ce que tous les individus soient regroupés en une seule classe.

« **Stratégie d'agrégation** »

Décomposition de l'inertie (théorème d'Huygens)

Inertie totale = Inertie <u>inter</u>-classes + Inertie <u>intra</u>-classes



Inertie totale (AUB par rapport à g) = Inertie inter (de $\{ g_A, g_B \}$ par rapport à g)

+ Inertie intra (inertie de A par rapport à g_A <u>plus</u> inertie de B par rapport à g_B)

Inertie totale = Inertie inter-classes + Inertie intra-classes

Rechercher une « bonne partition » revient à minimiser la variabilité intra-classes ou maximiser la variabilité inter-classes (**la variabilité totale est fixée par les données**).

Qualité d'une partition = Inertie inter-classes / Inertie totale

Méthode de WARD

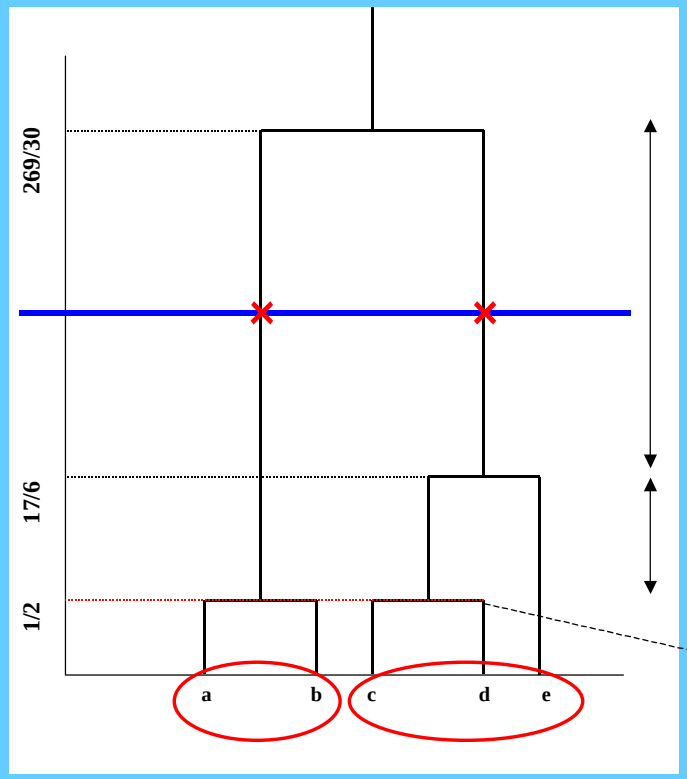
(agrégation suivant l'inertie)

En agrégeant 2 classes l'inertie intra-classe augmente de:

$$D(A,B) = \frac{n_A \times n_B}{n_A + n_B} d^2\left(g_A, g_B\right)$$

→ **Objectif:** choisir les 2 classes à agréger de façon à minimiser l'accroissement d'inertie intra-classe (donc centres de gravité proches et classes d'effectifs faibles).

Dendogramme



Accroissement 3 :
269/30-17/6=184/30=6.13

Accroissement 2 :
17/6-1/2=14/6=2.33

Accroissement 1 : 0,5

Inertie totale du nuage: 12,8

	Accroissement de l'inertie intra classes	Inertie intra classes	
Passage de 5 classes à 4 classes	0,5	0,5	→ 4 classes
Passage de 4 classes à 3 classes	0,5	1	→ 3 classes
Passage de 3 classes à 2 classes	2,83	3,83	→ 2 classes
Passage de 2 classes à 1 classe	8,97	12,8	→ 1 classe

Qualité de la partition 2 classes : 1- (3,83/12,8) = 70%

Qualité de la partition 3 classes : 1- (1/12,8) = 92,2% mais il n'y a que 5 points!!

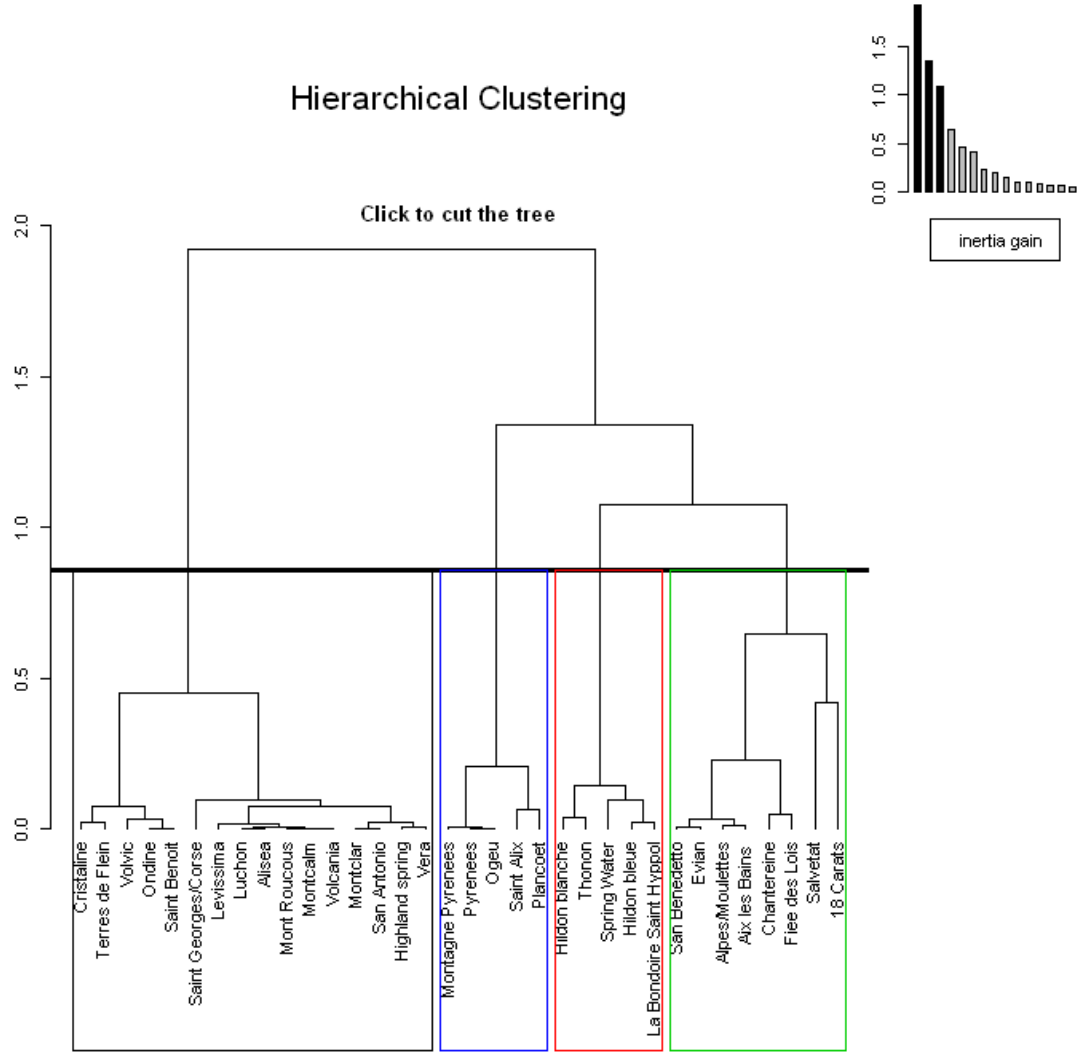
Exemple « Eaux Minérales »

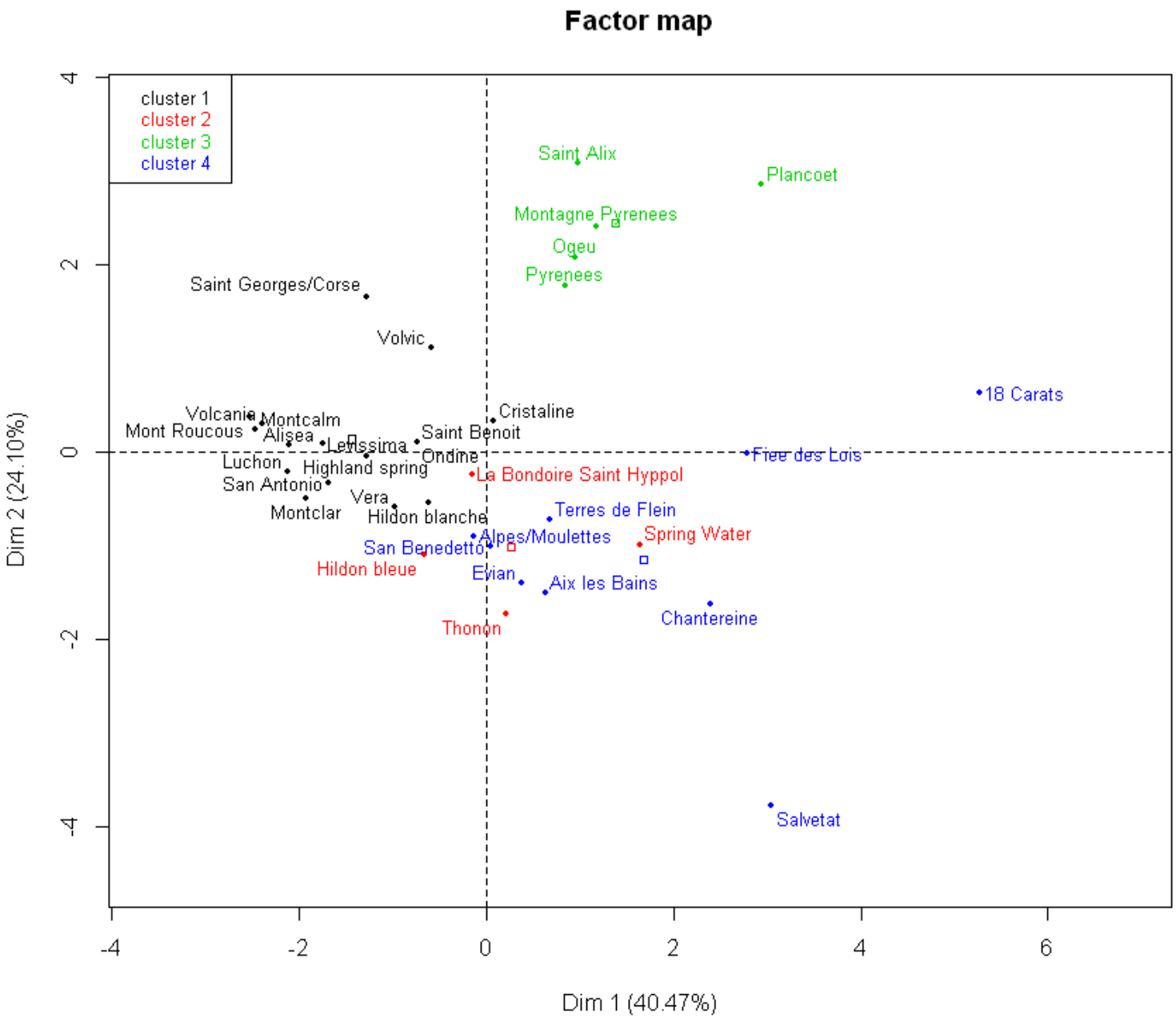
```
> donnees<-read.table("eaux minerales_bis.csv",header=TRUE,sep=";",dec=",",row.names=1)
> head(donnees,8)
                  TYPE PG   CA   MG  NA.   K  SUL NO3  HCO3   CL
Evian                M  P 78.0 24.0  5.0 1.0 10.0 3.8 357.0  4.5
Montagne Pyrenees    S  P 48.0 11.0 34.0 1.0 16.0 4.0 183.0 50.0
Cristaline           S  P 71.0  5.5 11.2 3.2  5.0 1.0 250.0 20.0
Fiee des Lois        S  P 89.0 31.0 17.0 2.0 47.0 0.0 360.0 28.0
Volcania             S  P  4.1  1.7  2.7 0.9  1.1 0.8  25.8  0.9
Luchon               M  P 26.5  1.0  0.8 0.2  8.2 1.8  78.1  2.3
Volvic               M  P  9.9  6.1  9.4 5.7  6.9 6.3  65.3  8.4
Alpes/Moulettes      S  P 63.0 10.2  1.4 0.4 51.3 2.0 173.2  1.0
...
    Pyrenees         M  G  48 12.0 31.0 1.0 18.0  4.0 183.0 35.0
    Montcalm         S  P   3  0.6  1.5 0.4  8.7  0.9   5.2  0.6
    Chantereine      S  P 119 28.0  7.0 2.0 52.0  0.0 430.0  7.0
    18 Carats        S  G 118 30.0 18.0 7.0 85.0  0.5 403.0 39.0
    Spring Water     S  G 117 19.0 13.0 2.0 16.0 20.0 405.0 28.0
    Montclar         S  P  41  3.0  2.0 0.0  2.0  3.0 134.0  3.0
```

Individus: 35 **Variables: 10** (8 actives, 2 nominales supplémentaires)

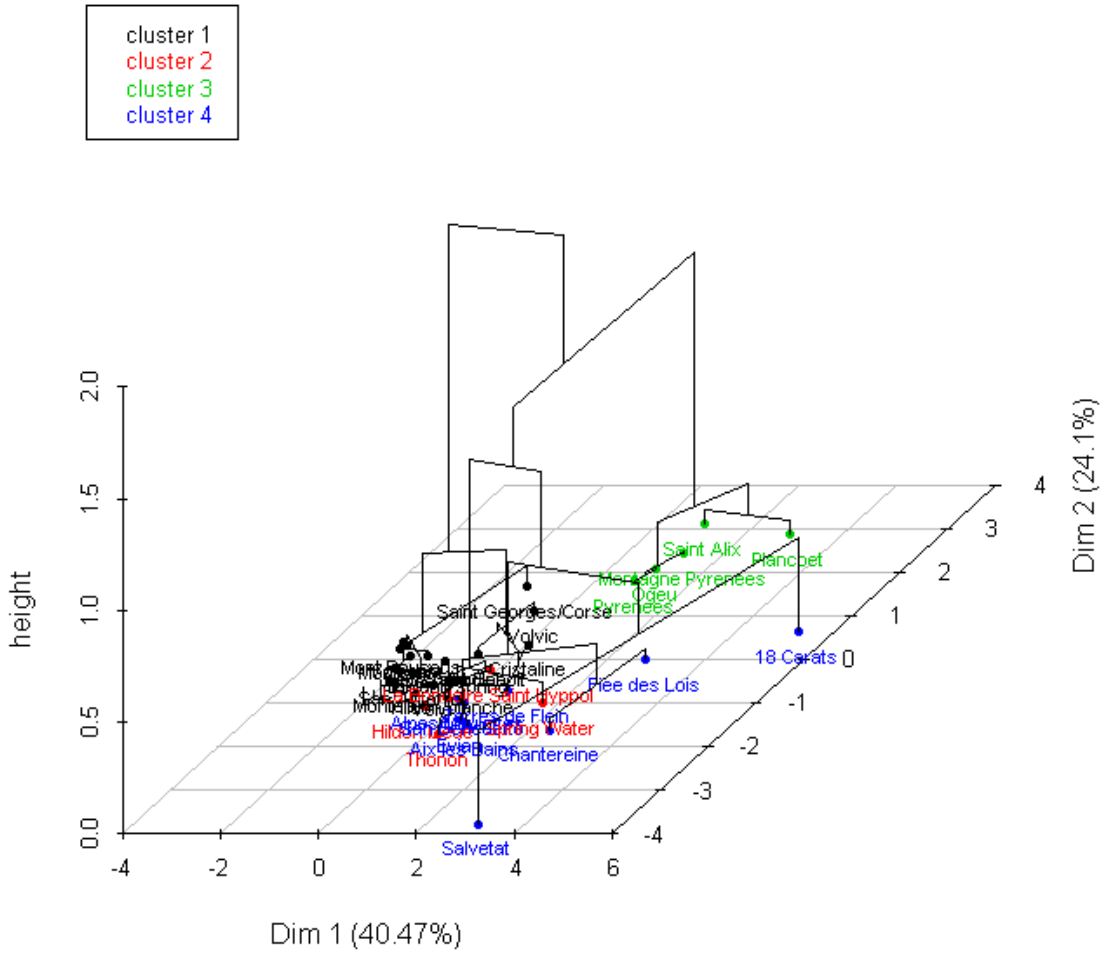
```
> res<-PCA(donnees,scale.unit=TRUE, quali.sup=1:2, ind.sup=28, ncp=4,graph=FALSE)
> reshcpc<-HCPC(res)
```

(5,5,8,16)




Factor map
cluster 1
cluster 2
cluster 3
cluster 4
Saint Alix
Plancoet
Montagne Pyrenees
Ogeu
Pyrenees
Saint Georges/Corse
Volvic
18 Carats
Volcanic
Montcalm
Mont Roucous
Alisea
Cristaline
Saint Benoit
Lessima
Ondine
Fiee des Lois
Luchon
Highland spring
La Bondoire Saint Hyppol
San Antonio
Vera
Montclar
Hildon blanche
Terres de Flein
Alpes/Moulettes
Spring Water
San Benedetto
Hildon bleue
Evian
Aix les Bains
Chantereine
Thonon
Salvetat
Dim 2 (24.10%)
Dim 1 (40.47%)
4
2
0
-2
-4
-4
-2
0
2
4
6

Hierarchical clustering on the factor map



```
> names(reshcpc)
[1] "data.clust" "desc.var"   "desc.axes"  "call"       "desc.ind"
> reshcpc$data.clust
                          TYPE PG    CA    MG    NA.    K  SUL  NO3  HCO3   CL clust
Evian                        M  P  78.0 24.00  5.00 1.00 10.0  3.8 357.0  4.5     4
Montagne Pyrenees            S  P  48.0 11.00 34.00 1.00 16.0  4.0 183.0 50.0     3
Cristaline                   S  P  71.0  5.50 11.20 3.20  5.0  1.0 250.0 20.0     1
Fiee des Lois                S  P  89.0 31.00 17.00 2.00 47.0  0.0 360.0 28.0     4
Volcania                     S  P   4.1  1.70  2.70 0.90  1.1  0.8  25.8  0.9     1
Luchon                       M  P  26.5  1.00  0.80 0.20  8.2  1.8  78.1  2.3     1
Volvic                       M  P   9.9  6.10  9.40 5.70  6.9  6.3  65.3  8.4     1
Alpes/Moulettes              S  P  63.0 10.20  1.40 0.40 51.3  2.0 173.2  1.0     4
Ondine                       S  P  46.1  4.30  6.30 3.50  9.0  0.0 163.5  3.5     1
Thonon                       M  P 108.0 14.00  3.00 1.00 13.0 12.0 350.0  9.0     2
Aix les Bains                M  P  84.0 23.00  2.00 1.00 27.0  0.2 341.0  3.0     4
La Bondoire Saint Hyppol     S  P  86.0  3.00 17.00 1.00  7.0 19.0 256.0 21.0     2
Salvetat                     M  G 253.0 11.00  7.00 3.00 25.0  1.0 820.0  4.0     4
Alisea                       M  P  12.3  2.60  2.50 0.60 10.1  2.5  41.6  0.9     1
San Benedetto                M  P  46.0 28.00  6.80 1.00  5.8  6.6 287.0  2.4     4
Levissima                    M  P  19.8  1.80  1.70 1.80 14.2  1.5  56.5  0.3     1
Vera                         M  P  36.0 13.00  2.00 0.60 18.0  3.6 154.0  2.1     1
San Antonio                  M  P  32.5  6.10  4.90 0.70  1.6  4.3 135.5  1.0     1
Saint Benoit                 S  G  46.1  4.30  6.30 3.50  9.0  0.0 163.5  3.5     1
Plancoet                     M  P  36.0 19.00 36.00 6.00 43.0  0.0 195.0 38.0     3
Saint Alix                   S  P   8.0 10.00 33.00 4.00 20.0  0.5  84.0 37.0     3
Saint Georges/Corse          S  P   5.2  2.43 14.05 1.15  6.0  0.0  30.5 25.0     1
Hildon bleue                 M  P  97.0  1.70  7.70 1.00  4.0 26.4 236.0 16.0     2
Hildon blanche               M  G  97.0  1.70  7.70 1.00  4.0  5.5 236.0 16.0     1
Mont Roucous                 M  P   1.2  0.20  2.80 0.40  3.3  2.3   4.9  3.2     1
Ogeu                         S  P  48.0 11.00 31.00 1.00 16.0  4.0 183.0 44.0     3
Highland spring              M  P  35.0  8.50  6.00 0.60  6.0  1.0 136.0  7.5     1
Terres de Flein              S  P 116.0  4.20  8.00 2.50 24.5  1.0 333.0 15.0     4
Pyrenees                     M  G  48.0 12.00 31.00 1.00 18.0  4.0 183.0 35.0     3
Montcalm                     S  P   3.0  0.60  1.50 0.40  8.7  0.9   5.2  0.6     1
Chantereine                  S  P 119.0 28.00  7.00 2.00 52.0  0.0 430.0  7.0     4
18 Carats                    S  G 118.0 30.00 18.00 7.00 85.0  0.5 403.0 39.0     4
Spring Water                 S  G 117.0 19.00 13.00 2.00 16.0 20.0 405.0 28.0     2
Montclar                     S  P  41.0  3.00  2.00 0.00  2.0  3.0 134.0  3.0     1
```

```
> reshcpc$desc.var
$test.chi2
     p.value df

$quanti.var
          Eta2      P-value
NO3  0.8297696 1.199269e-11
NA.  0.8289453 1.288803e-11
CL   0.6413293 7.629724e-07
MG   0.5836028 6.857841e-06
HCO3 0.5827199 7.074406e-06
CA   0.5142200 6.540412e-05
SUL  0.4751441 2.015778e-04

$quanti
$quanti$`1`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
NA.  -2.845138         5.115625     10.58088       3.761055   10.40370 0.0044392191
CL   -2.940271         6.137500     14.12059       7.338330   14.70498 0.0032792576
SUL  -3.104500         7.068750     17.46176       4.418529   18.13137 0.0019060078
CA   -3.231283        30.418750     60.25588      25.680421   50.01068 0.0012323581
HCO3 -3.638644       105.025000    213.51765      75.265443  161.48846 0.0002740777
MG   -3.756110         3.926875     10.38029       3.242870    9.30534 0.0001725745

$quanti$`2`
      v.test Mean in category Overall mean sd in category Overall sd      p.value
NO3 5.226802            19.35     4.102941       5.105634   6.118943 1.724674e-07

$quanti$`3`
      v.test Mean in category Overall mean sd in category Overall sd     p.value
NA. 5.140123              33.0     10.58088       1.897367   10.40370 2.74559e-07
CL  4.327677              40.8     14.12059       5.491812   14.70498 1.50690e-05

$quanti$`4`
       v.test Mean in category Overall mean sd in category Overall sd      p.value
MG   3.950047         21.04444     10.38029       9.374091    9.30534 7.813588e-05
HCO3 3.753006        389.35556    213.51765     167.303949  161.48846 1.747267e-04
SUL  3.600122         36.40000     17.46176      23.411631   18.13137 3.180684e-04
CA   3.244579        107.33333     60.25588      56.776364   50.01068 1.176243e-03
```

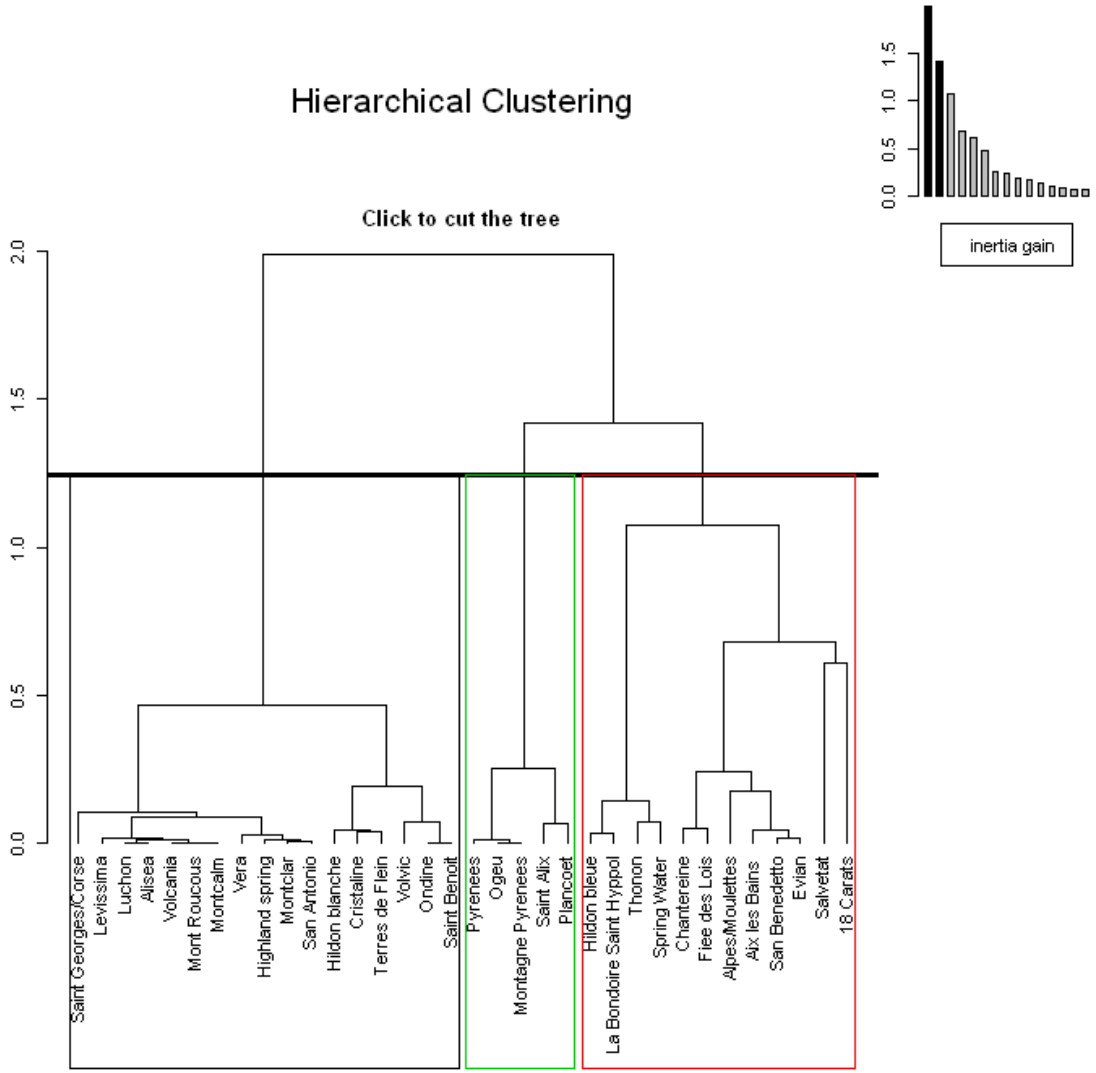
```
> reshcpc$desc.axes
$quanti.var
           Eta2      P-value
Dim.3 0.7243467 1.554611e-08
Dim.2 0.7023642 4.845373e-08
Dim.1 0.6213126 1.698423e-06

$quanti
$quanti$`1`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.1 -4.33694         -1.44083  4.38783e-17      0.7673623   1.799326 1.444799e-05

$quanti$`2`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.3 4.417999         2.430235 8.951683e-17      0.8502652   1.153849 9.961904e-06

$quanti$`3`
        v.test Mean in category Overall mean sd in category Overall sd      p.value
Dim.2 4.194127         2.441705 2.979643e-17      0.4839651   1.388656 2.739245e-05

$quanti$`4`
         v.test Mean in category Overall mean sd in category Overall sd     p.value
Dim.1  3.217552        1.6796828 4.387830e-17      1.7070257   1.799326 0.001292897
Dim.3 -2.167241       -0.7255177 8.951683e-17      0.6255742   1.153849 0.030216508
Dim.2 -2.844872       -1.1461703 2.979643e-17      1.1561547   1.388656 0.004442928
```

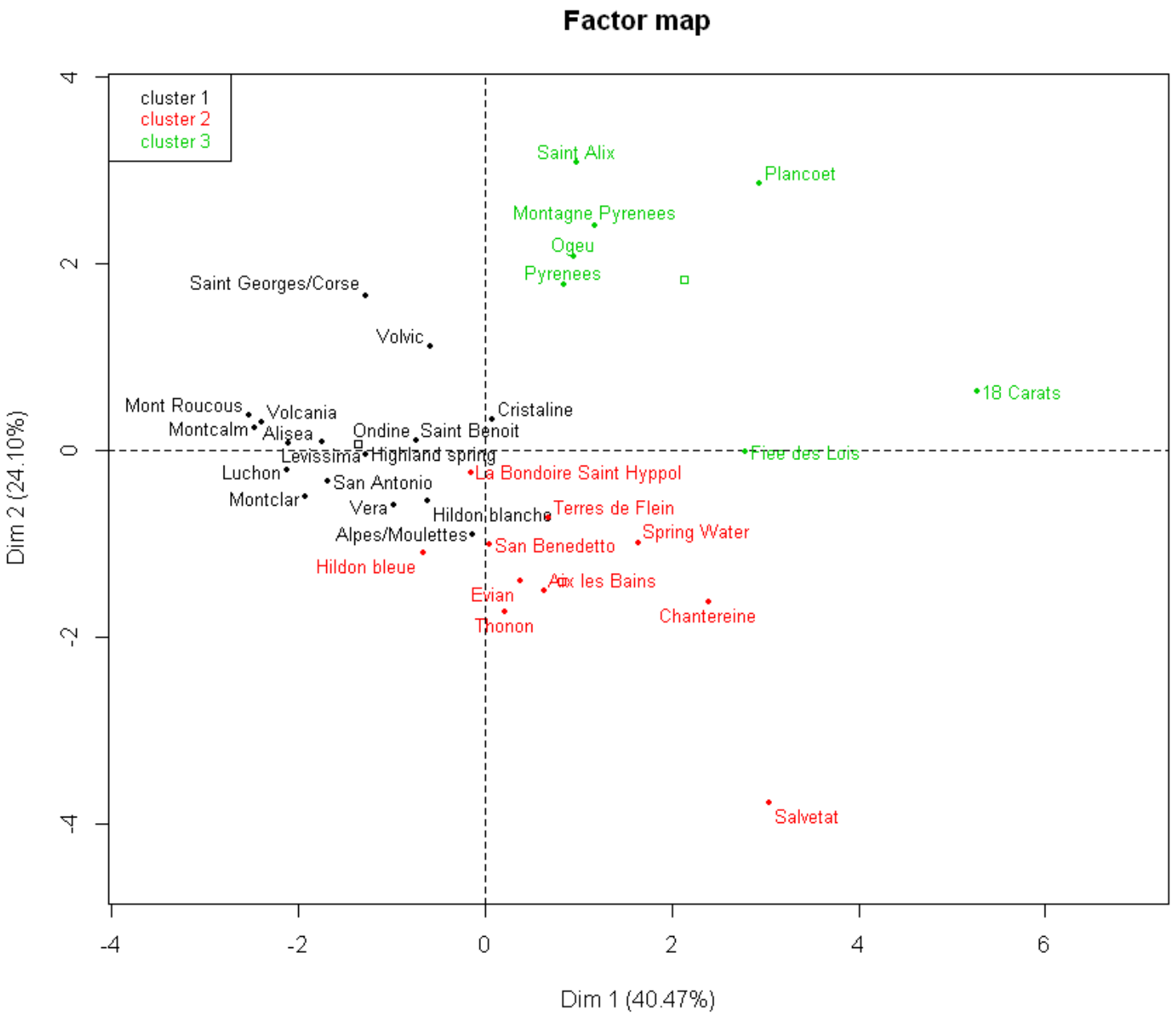
```
> reshcpc$desc.ind
$para
Cluster: 1
      Levissima Highland spring     San Antonio          Alisea          Luchon
      0.6285316       0.7177708       0.7294265       0.7935308       0.8636839
--------------------------------------------------------------------------------
Cluster: 2
La Bondoire Saint Hyppol             Hildon bleue             Spring Water                  Thonon
              0.9422401                1.4063806                1.4108955               1.5674161
--------------------------------------------------------------------------------
Cluster: 3
             Ogeu Montagne Pyrenees          Pyrenees        Saint Alix          Plancoet
        0.9854896         1.0390097         1.0694080         1.1704876         2.2484324
--------------------------------------------------------------------------------
Cluster: 4
    Chantereine  Aix les Bains            Evian Terres de Flein   Fiee des Lois
       1.212961        1.310372        1.521449        1.717499        1.919561

$dist
Cluster: 1
    Montcalm     Volcania Mont Roucous       Alisea       Luchon
    4.370959     4.283323     4.218938     3.981552     3.845677
--------------------------------------------------------------------------------
Cluster: 2
            Hildon bleue La Bondoire Saint Hyppol             Spring Water                  Thonon
                4.069153                 3.399998                 3.263916               2.389983
--------------------------------------------------------------------------------
Cluster: 3
         Plancoet        Saint Alix Montagne Pyrenees             Ogeu          Pyrenees
         4.318517          3.837206          3.824906          3.515190          3.319321
--------------------------------------------------------------------------------
Cluster: 4
     18 Carats       Salvetat    Chantereine  Fiee des Lois Aix les Bains
      4.955503       4.735026       4.477943       3.344010       2.985436
```

```
> reshcpc$call
$t
$t$res
**Results for the Principal Component Analysis (PCA)**
The analysis was performed on 34 individuals, described by 10 variables
*The results are available in the following objects:

   name                 description
1  "$eig"               "eigenvalues"
2  "$var"               "results for the variables"
3  "$var$coord"         "coord. for the variables"
4  "$var$cor"           "correlations variables - dimensions"
5  "$var$cos2"          "cos2 for the variables"
6  "$var$contrib"       "contributions of the variables"
7  "$ind"               "results for the individuals"
8  "$ind$coord"         "coord. for the individuals"
9  "$ind$cos2"          "cos2 for the individuals"
10 "$ind$contrib"       "contributions of the individuals"
11 "$ind.sup"           "results for the supplementary individuals"
12 "$ind.sup$coord"     "coord. for the supplementary individuals"
13 "$ind.sup$cos2"      "cos2 for the supplementary individuals"
14 "$quali.sup"         "results for the supplementary categorical variables"
15 "$quali.sup$coord"   "coord. for the supplementary categories"
16 "$quali.sup$v.test"  "v-test of the supplementary categories"
17 "$call"              "summary statistics"
18 "$call$centre"       "mean of the variables"
19 "$call$ecart.type"   "standard error of the variables"
20 "$call$row.w"        "weights for the individuals"
21 "$call$col.w"        "weights for the variables"
```

```
$t$nb.clust
[1] 4

$t$within
 [1] 7.1481088964 5.2289601480 3.8865564159 2.8104473936 2.1645572927 1.7109912064 1.2904909427 1.0583492012
 [9] 0.8514998444 0.7066454781 0.6103622015 0.5152081569 0.4374954043 0.3614510194 0.2965756469 0.2464427793
[17] 0.2093563646 0.1752606125 0.1425396158 0.1188533590 0.0963890515 0.0741947537 0.0557947399 0.0410189295
[25] 0.0312878669 0.0225733956 0.0150787060 0.0088680215 0.0063703495 0.0038982610 0.0023819148 0.0009855806
[33] 0.0000000000

$t$inert.gain
 [1] 1.9191487484 1.3424037321 1.0761090223 0.6458901009 0.4535660863 0.4205002637 0.2321417415 0.2068493567
 [9] 0.1448543663 0.0962832766 0.0951540446 0.0777127526 0.0760443849 0.0648753725 0.0501328676 0.0370864147
[17] 0.0340957521 0.0327209967 0.0236862568 0.0224643075 0.0221942978 0.0184000137 0.0147758105 0.0097310626
[25] 0.0087144713 0.0074946897 0.0062106844 0.0024976721 0.0024720885 0.0015163462 0.0013963341 0.0009855806
[33] 0.0000000000

$t$quot
[1] 0.7432752 0.7231202 0.7701825 0.7904578 0.7542359 0.8201136 0.8045547 0.8298833
```

```
> restot<-PCA(donnees,scale.unit=TRUE, quali.sup=1:2, ind.sup=28, ncp=8,graph=FALSE)
> reshcpc<-HCPC(restot)
```




Factor map
cluster 1
cluster 2
cluster 3
Saint Alix
Plancoet
Montagne Pyrenees
Ogeu
Pyrenees
Saint Georges/Corse
Volvic
18 Carats
Mont Roucous
Volcania
Montcalm
Alisea
Ondine
Saint Benoit
Cristaline
Levissima
Highland spring
Fee des Lois
Luchon
San Antonio
La Bondoire Saint Hyppol
Montclar
Vera
Hildon blanche
Terres de Flein
Alpes/Moulettes
Spring Water
San Benedetto
Hildon bleue
Evian
Aix les Bains
Chantereine
Thonon
Salvetat
Dim 2 (24.10%)
Dim 1 (40.47%)
4
2
0
-2
-4
-4
-2
0
2
4
6

Méthode des k-means

```
> donnees1<-donnees [,3:10]
> donnees2<-scale(donnees1)
> eaux.kmeans<-kmeans(donnees2,centers=4,iter.max=100,nstart=1)
> eaux.kmeans
K-means clustering with 4 clusters of sizes 5, 1, 8, 21

Cluster means:
            CA         MG        NA.           K        SUL        NO3        HCO3          CL
1  1.249582929 -0.1718960 -0.1947877 -0.19322811 -0.2836671  1.8990027  0.75512421  0.09297619
2  2.308037171  4.2857812  5.2812534  5.61938185  4.5684081 -0.6472291  4.06612271  0.25602421
3  0.005258263  0.4910993  0.4244227 -0.02154765  0.5236178 -0.3833499  0.05154797  1.39396350
4 -0.409429425 -0.3502427 -0.3667951 -0.21337429 -0.3494769 -0.2752850 -0.39305369 -0.56536253

Clustering vector:
                   Evian        Montagne Pyrenees                Cristaline             Fiee des Lois
                       4                        3                        4                        3
               Volcania                   Luchon                   Volvic           Alpes/Moulettes
                       4                        4                        4                        4
                Ondine                   Thonon           Aix les Bains  La Bondoire Saint Hyppol
                       4                        1                        4                        1
              Salvetat                   Alisea            San Benedetto                 Levissima
                       1                        4                        4                        4
                  Vera              San Antonio             Saint Benoit                 Plancoet
                       4                        4                        4                        3
             Saint Alix      Saint Georges/Corse             Hildon bleue           Hildon blanche
                       3                        4                        1                        4
           Mont Roucous                    Ogeu          Highland spring                 Vernière
                       4                        3                        4                        2
         Terres de Flein                Pyrenees                 Montcalm              Chantereine
                       4                        3                        4                        3
              18 Carats             Spring Water                 Montclar
                       3                        1                        4

Within cluster sum of squares by cluster:
[1] 23.76843  0.00000 21.52913 29.69065
 (between_SS / total_SS =  72.4 %)
```

```
> eaux.kmeans<-kmeans(donnees2,centers=3,iter.max=100,nstart=1)
> eaux.kmeans
K-means clustering with 3 clusters of sizes 19, 15, 1

Cluster means:
          CA         MG         NA.          K         SUL        NO3        HCO3         CL
1  0.4353577  0.2126217  0.03293346 -0.1202962  0.05324829  0.2886341  0.3030905  0.5238377
2 -0.7053222 -0.5550396 -0.39379928 -0.2222503 -0.37200837 -0.3224546 -0.6549895 -0.6805961
3  2.3080372  4.2857812  5.28125343  5.6193819  4.56840812 -0.6472291  4.0661227  0.2560242

Clustering vector:
                  Evian        Montagne Pyrenees               Cristaline            Fiee des Lois
                      1                        1                        1                        1
              Volcania                   Luchon                   Volvic          Alpes/Moulettes
                      2                        2                        2                        2
               Ondine                   Thonon          Aix les Bains  La Bondoire Saint Hyppol
                      2                        1                        1                        1
             Salvetat                   Alisea           San Benedetto                Levissima
                      1                        2                        1                        2
                 Vera              San Antonio            Saint Benoit                 Plancoet
                      2                        2                        2                        1
           Saint Alix      Saint Georges/Corse            Hildon bleue          Hildon blanche
                      1                        2                        1                        1
        Mont Roucous                    Ogeu          Highland spring                Vernière
                      2                        1                        2                        3
      Terres de Flein                Pyrenees                 Montcalm             Chantereine
                      1                        1                        2                        1
            18 Carats             Spring Water                 Montclar
                      1                        1                        2

Within cluster sum of squares by cluster:
[1] 94.83012 10.59693  0.00000
 (between_SS / total_SS =  61.2 %)
```

```
> donnees2<-scale(donnees1)
> donnees3<-donnees2[-28,]
> eaux.kmeans<-kmeans(donnees3,centers=4,iter.max=100,nstart=1)
> eaux.kmeans
K-means clustering with 4 clusters of sizes 5, 4, 18, 7

Cluster means:
          CA          MG        NA.           K         SUL        NO3        HCO3          CL
1 -0.4827661  0.03287361  0.6945099 -0.07059921  0.03757371 -0.2412610 -0.3292869  1.80498037
2  0.6965497 -0.19444932 -0.1707168 -0.23614823 -0.38405490  2.4949638  0.3102881  0.28999255
3 -0.5652734 -0.46739593 -0.3667800 -0.21468817 -0.40227342 -0.2584019 -0.5335560 -0.58337114
4  1.0706450  0.67725355 -0.2098422 -0.06534369  0.57440922 -0.4964410  0.8490240  0.00854061

Clustering vector:
                  Evian        Montagne Pyrenees               Cristaline            Fiee des Lois
                      4                        1                        3                        4
               Volcania                   Luchon                   Volvic          Alpes/Moulettes
                      3                        3                        3                        3
                Ondine                   Thonon           Aix les Bains  La Bondoire Saint Hyppol
                      3                        2                        4                        2
              Salvetat                   Alisea          San Benedetto                Levissima
                      4                        3                        3                        3
                  Vera              San Antonio             Saint Benoit                 Plancoet
                      3                        3                        3                        1
            Saint Alix      Saint Georges/Corse             Hildon bleue           Hildon blanche
                      1                        3                        2                        3
         Mont Roucous                     Ogeu          Highland spring          Terres de Flein
                      3                        1                        3                        4
             Pyrenees                 Montcalm              Chantereine                18 Carats
                      1                        3                        4                        4
         Spring Water                 Montclar
                      2                        3

Within cluster sum of squares by cluster:
[1]  2.921424  5.554072 19.604202 24.534450
 (between_SS / total_SS =  64.3 %)
```

```
> donnees2<-scale(donnees1)
> donnees3<-donnees2[-28,]
> eaux.kmeans<-kmeans(donnees3,centers=3,iter.max=100,nstart=1)
> eaux.kmeans
K-means clustering with 3 clusters of sizes 7, 11, 16

Cluster means:
          CA         MG        NA.             K        SUL        NO3        HCO3         CL
1 -0.1379705  0.3990451  0.5266361 -0.004029235  0.4525098 -0.3456528 -0.05934066  1.6632839
2  0.8280658  0.1565425 -0.2662596 -0.205491003 -0.1458621  0.7625873  0.55764030 -0.1892698
3 -0.6531854 -0.5500665 -0.3774282 -0.208173511 -0.3832183 -0.3326038 -0.61154884 -0.6135652

Clustering vector:
                  Evian        Montagne Pyrenees               Cristaline            Fiee des Lois
                      2                        1                        3                        1
              Volcania                   Luchon                   Volvic          Alpes/Moulettes
                      3                        3                        3                        3
               Ondine                   Thonon           Aix les Bains  La Bondoire Saint Hyppol
                      3                        2                        2                        2
             Salvetat                   Alisea            San Benedetto                Levissima
                      2                        3                        2                        3
                 Vera              San Antonio             Saint Benoit                 Plancoet
                      3                        3                        3                        1
           Saint Alix      Saint Georges/Corse             Hildon bleue           Hildon blanche
                      1                        3                        2                        2
         Mont Roucous                    Ogeu          Highland spring          Terres de Flein
                      3                        1                        3                        2
            Pyrenees                  Montcalm              Chantereine                 18 Carats
                      1                        3                        2                        1
         Spring Water                  Montclar
                      2                        3

Within cluster sum of squares by cluster:
[1] 14.19033 48.68554 12.95326
 (between_SS / total_SS =  48.6 %)
```