Solène CATELLA Hugo LE BAHER

Mise en oeuvre d'une ACP sur des Eaux

Sommaire

Introduction

Gazeuse, dé-ionisée ou encore riche en magnésium, l'eau, élément naturel et central dans notre vie quotidienne, peut revêtir plusieurs formes. Sa composition minérale, sa provenance, sa nature mais aussi son potentiel hydrogène sont autant de paramètres qui entrent en jeu dans sa caractérisation. Analyser une eau revient donc à s'intéresser à l'ensemble des attributs et propriétés susceptibles de la définir.

Le jeu de données dont nous disposons recense près de 100 eaux (95 exactement), déclinées sous 12 variables à la fois qualitatives (nom, nature, pays) et quantitatives (composition ionique, sodique, etc.). L'intérêt d'une ACP est, entre autres, de caractériser des observations dites "multivariées", c'est-à-dire multidimensionnelles, et d'en dégager des relations de ressemblance/dissemblance. En d'autres termes, l'ACP permet d'avoir une vue globale et synthétique des données : c'est la raison pour laquelle son usage est ici parfaitement approprié.

Avant de procéder à l'ACP en tant que telle, nous explorerons le jeu de données au travers d'analyses univariées et bivariées. Une classification de type non-supervisée sera menée en dernier lieu, via la méthode des K-means.

Le présent document a été réalisé à partir de RMarkdown.

1. Analyse univariée et bivariée

1.1. Etude du jeu de données

La commande `head` nous donne un premier aperçu du jeu de données étudié. Il s'agit de données relatives aux caractéristiques et propriétés minérales de diverses eaux.

```
head(waters)
```

```
##               Nom Nature  Ca   Mg Na   K Cl NO3 SO4   HCO3  PH   Pays
## 1       Abatilles   plat  16  8.0 75 3.0 95   0   8  112.0 8.2 France
## 2   Aix-Les-Bains   plat  72 38.0 14 2.0  6   1  81  329.0 7.4 France
## 3            Alet   plat  63 23.0 13 1.3 11   2  14  300.0 7.4 France
## 4         Alpille   plat  41  3.0  2 0.0  3   3   2  134.0  NA France
## 5 Amelie le Reine   <NA> 390 27.5 45 2.8 19   2  36 1376.6  NA France
## 6       Aquarelle   plat  70  2.1  2  NA NA   4  NA  210.0  NA France
```

```
dim(waters)
```

```
## [1] 95 12
```

Le jeu de données comprend 95 observations réparties sur 12 variables, dont 3 catégorielles (Nom, Nature, Pays) et 9 numériques (Ca, Mg, Na, K, Cl, NO3, SO4, HCO3, PH).

Dans ce jeu de données, on distingue 84 observations actives (eaux françaises) pour 11 observations supplémentaires (eaux marocaines).

On note que certaines données sont incomplètes. En effet, toutes les variables ne sont pas renseignées pour chacune des différentes eaux. Par exemple, les concentrations en K, Cl et SO4 de l'eau Aquarelle, ainsi que son PH, ne sont pas connus (cf. tableau). Il convient dans un premier temps d'étudier la fiabilité des différentes variables, de façon à détecter les valeurs manquantes, incohérentes, extrêmes ou aberrantes : c'est l'objet des résumés unidimensionnels.

1.2. Analyse exploratoire des données

1.2.1. Recherche des données manquantes

On s'intéresse en premier lieu aux données manquantes, représentées via le symbole NA.

```
colSums(is.na(waters))
```

```
##    Nom Nature     Ca     Mg     Na      K     Cl    NO3    SO4   HCO3
##      0      3      0      0      0      3      4     19      2      2
##     PH   Pays
##     19      0
```

Parmi les variables étudiées, on remarque que le NO3 et le PH sont celles qui enregistrent le plus grand nombre de valeurs manquantes (19 chacune), suivies des variables Cl, K, Nature, SO4 et HCO3. Ces données manquantes peuvent alors être :

- Supprimées, de façon à ne conserver que les observations complètes ;
- Imputées par une valeur plausible ou aléatoire (méthodes d'imputation simples et multiples).

L'étude statistique unidimensionnelle peut nous aider à mieux appréhender la distribution des différentes variables, et préférer l'une ou l'autre de ces méthodes. Dans la mesure où l'Analyse en Composantes Principales (ACP) est une méthode appliquée à des variables quantitatives continues, nous nous intéresserons particulièrement à ces dernières, après avoir brièvement analysé les variables qualitatives.

1.2.2. Analyse unidimensionnelle : variables qualitatives

Le jeu de données comprend 3 variables qualitatives : le nom de l'eau (Nom), sa nature (Nature) et son origine (Pays).

Il peut ici être intéressant de s'intéresser aux propriétés minérales des eaux plates et gazeuses.

```
summary(still.waters)
```

```
##        Ca               Mg               Na                K
##  Min.   :  2.40   Min.   :  0.50   Min.   :   0.80   Min.   :  0.000
##  1st Qu.: 12.02   1st Qu.:  3.00   1st Qu.:   3.00   1st Qu.:  0.750
##  Median : 63.00   Median : 10.00   Median :   8.10   Median :  1.700
##  Mean   : 92.47   Mean   : 19.95   Mean   :  46.23   Mean   :  5.811
##  3rd Qu.: 96.00   3rd Qu.: 24.00   3rd Qu.:  26.10   3rd Qu.:  3.000
##  Max.   :596.00   Max.   :110.00   Max.   :1110.00   Max.   :120.000
##                                                      NA's   :2
##        Cl              NO3              SO4                HCO3
##  Min.   :  0.600   Min.   : 0.000   Min.   :   0.200   Min.   :   2.4
##  1st Qu.:  3.825   1st Qu.: 1.000   1st Qu.:   5.075   1st Qu.:  66.0
##  Median : 10.350   Median : 2.000   Median :  13.000   Median : 210.0
##  Mean   : 31.055   Mean   : 2.753   Mean   : 115.122   Mean   : 303.1
##  3rd Qu.: 19.550   3rd Qu.: 3.450   3rd Qu.:  41.325   3rd Qu.: 342.5
##  Max.   :285.000   Max.   :19.000   Max.   :1530.000   Max.   :3800.0
##  NA's   :3         NA's   :10       NA's   :1          NA's   :2
##        PH
##  Min.   :5.800
##  1st Qu.:6.900
##  Median :7.200
##  Mean   :7.125
##  3rd Qu.:7.500
##  Max.   :8.200
##  NA's   :12
```

```
summary(gaz.waters)
```

```
##        Ca              Mg               Na               K
##  Min.   :  6.7   Min.   :  1.60   Min.   :   3.0   Min.   :  0.500
```

```
##  1st Qu.: 92.0   1st Qu.: 13.75   1st Qu.:  93.5   1st Qu.:  6.475
##  Median :147.3   Median : 55.90   Median : 252.0   Median : 40.400
##  Mean   :148.6   Mean   : 62.81   Mean   : 418.2   Mean   : 45.269
##  3rd Qu.:204.0   3rd Qu.: 84.25   3rd Qu.: 473.0   3rd Qu.: 51.425
##  Max.   :420.0   Max.   :243.00   Max.   :1945.0   Max.   :192.200
##                                                    NA's   :1
##        Cl              NO3              SO4              HCO3
##  Min.   :  3.00   Min.   : 0.000   Min.   :  5.00   Min.   :  30.5
##  1st Qu.: 18.75   1st Qu.: 0.750   1st Qu.: 25.00   1st Qu.: 614.0
##  Median : 39.00   Median : 1.000   Median : 47.35   Median :1390.0
##  Mean   :132.10   Mean   : 2.574   Mean   : 99.48   Mean   :1609.7
##  3rd Qu.:230.00   3rd Qu.: 2.000   3rd Qu.:156.00   3rd Qu.:1918.5
##  Max.   :649.00   Max.   :18.300   Max.   :549.20   Max.   :6722.2
##  NA's   :1        NA's   :8        NA's   :1
##        PH
##  Min.   :5.200
##  1st Qu.:6.000
##  Median :6.300
##  Mean   :6.343
##  3rd Qu.:6.800
##  Max.   :7.700
##  NA's   :6
```

Ces résumés nous indiquent que les eaux gazeuses sont en moyenne plus riches que les eaux plates tous minéraux confondus, à l'exception des nitrates (NO3) et sulfates (SO4). Les composants minéraux qui semblent être les plus discriminants dans la distinction des eaux plates et gazeuses sont le HCO3, le Na et le Ca (comparaison des médianes). On note également que les eaux gazeuses enregistrent un PH légèrement inférieur aux eaux plates.

En termes de répartition des échantillons, on remarque que les eaux plates sont en quantité supérieure, représentant à peu près 2/3 du jeu de données. C'est le diagramme en mosaïque (mosaic plot) qui nous permet d'aboutir à de telles conclusions, en faisant la représentation des effectifs relatifs à un tableau de contingence, ici défini par les variables Pays et Nature.

```
table(waters$Nature, waters$Pays)
```

```
##
##        France Maroc
##   gaz      27     0
##   plat     54    11
```

```
mosaicplot(~waters$Pays+waters$Nature, data=waters, color=TRUE)
```

Comme on peut le voir, les eaux gazeuses sont exclusivement françaises : il n'y a pas d'eaux marocaines gazeuses. On peut aussi relever que les eaux marocaines sont très peu représentées vis-à-vis des eaux françaises, raison pour laquelle elles sont sans doute traitées dans cette analyse comme des observations supplémentaires.

1.2.3. Analyse unidimensionnelle : variables quantitatives

Nous nous intéressons désormais aux variables quantitatives.

La commande `summary` fournit un résumé détaillé des variables étudiées, ici numériques (minimum, premier quartile, médiane, moyenne, troisième quartile et maximum).

```
summary(waters[sapply(waters, is.numeric)])
```

```
##        Ca              Mg              Na               K
```

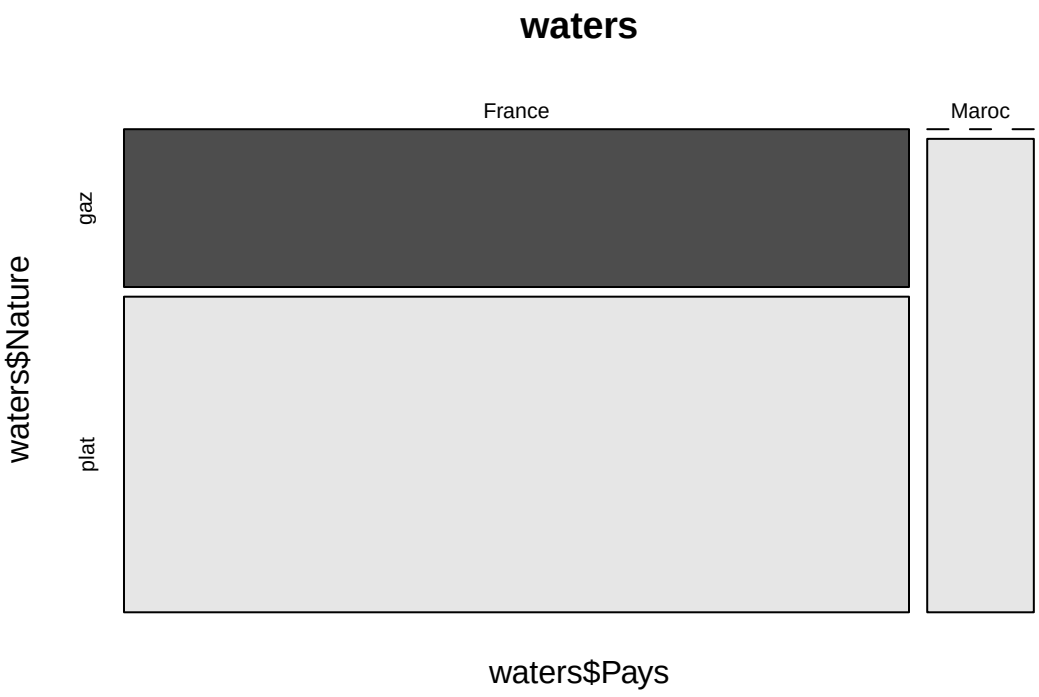


Figure 1: Répartition des eaux par Pays et Nature

```
##  Min.   :  2.4   Min.   :  0.50   Min.   :   0.8   Min.   :  0.00
##  1st Qu.: 20.9   1st Qu.:  4.35   1st Qu.:   3.7   1st Qu.:  1.00
##  Median : 72.0   Median : 14.00   Median :  13.0   Median :  2.00
##  Mean   :111.1   Mean   : 31.92   Mean   : 151.4   Mean   : 16.82
##  3rd Qu.:146.2   3rd Qu.: 45.00   3rd Qu.: 111.5   3rd Qu.: 12.50
##  Max.   :596.0   Max.   :243.00   Max.   :1945.0   Max.   :192.20
##                                                    NA's   :3
##        Cl              NO3             SO4             HCO3
##  Min.   :  0.60   Min.   : 0.000   Min.   :   0.2   Min.   :   2.4
##  1st Qu.:  4.40   1st Qu.: 1.000   1st Qu.:   6.0   1st Qu.: 121.0
##  Median : 15.00   Median : 2.000   Median :  18.0   Median : 306.0
##  Mean   : 59.54   Mean   : 2.689   Mean   : 108.7   Mean   : 691.4
##  3rd Qu.: 39.10   3rd Qu.: 3.000   3rd Qu.:  60.0   3rd Qu.: 820.0
##  Max.   :649.00   Max.   :19.000   Max.   :1530.0   Max.   :6722.2
##  NA's   :4        NA's   :19       NA's   :2        NA's   :2
##        PH
##  Min.   :5.200
##  1st Qu.:6.500
##  Median :7.000
##  Mean   :6.905
##  3rd Qu.:7.400
##  Max.   :8.200
##  NA's   :19
```

En moyenne, les eaux sont principalement concentrées en HCO3 (691.4 mg/L), Na (151.4 mg/L), Ca (111.1 mg/L) et SO4 (108.7 mg/L). Cependant, dans la mesure où un grand nombre de variables enregistrent des valeurs extrêmes telles HCO3 (maximum à 6722.2 mg/L) ou Na (maximum à 1945 mg/L), la médiane apparaît ici comme une meilleure mesure centrale, car plus robuste que la moyenne. Exception faites des variables HCO3 et Ca, on remarque que les concentrations oscillent alors entre 2 mg/L et 18 mg/L, tous composants minéraux confondus.

Un examen plus fin, variable par variable, nous permettra de mieux apprécier leurs distributions respectives. On étudiera tout particulièrement les variables PH et NO3, enregistrant ici le plus grand nombre de données manquantes (19).

Etude de la variable PH

Le tracé de la boîte à moustache de la variable PH laisse apparaître une distribution symétrique. On note que les valeurs de sa moyenne (6.905) et sa médiane (7.000) sont très proches. Sa distribution, homogène et faiblement dispersée, semble s'approcher de celle d'une loi Normale (voire figure ci-contre).

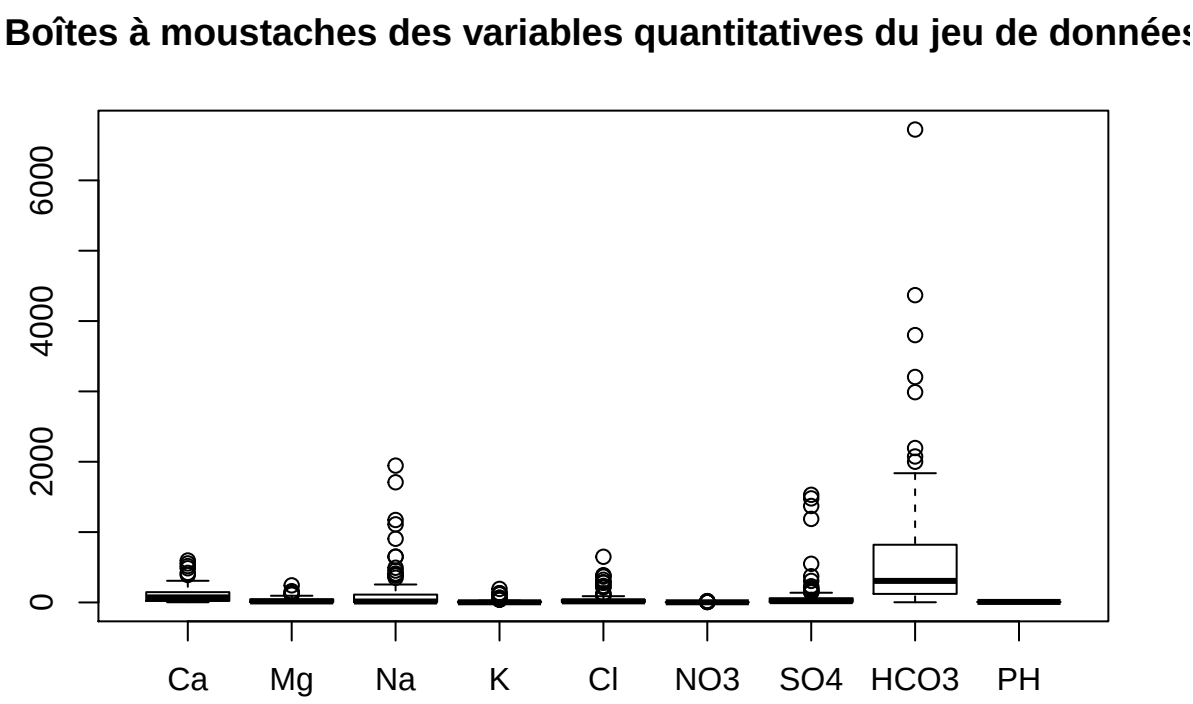
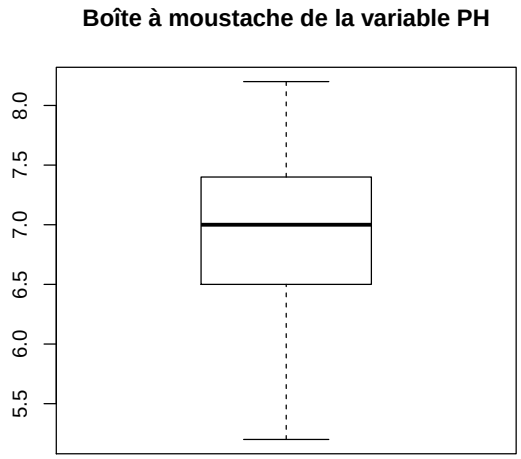


Figure 2: Figure 2



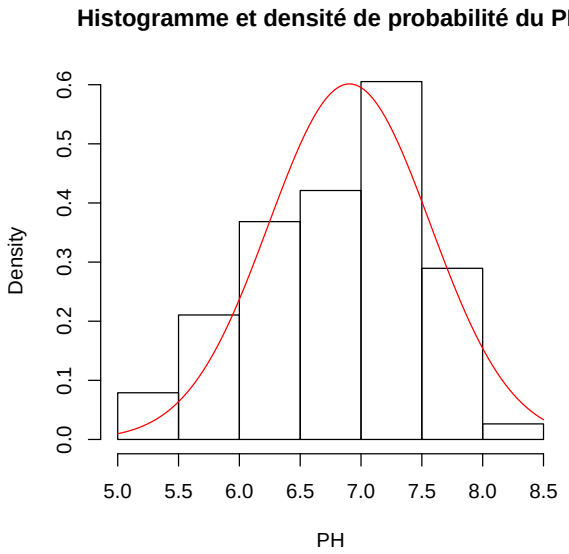
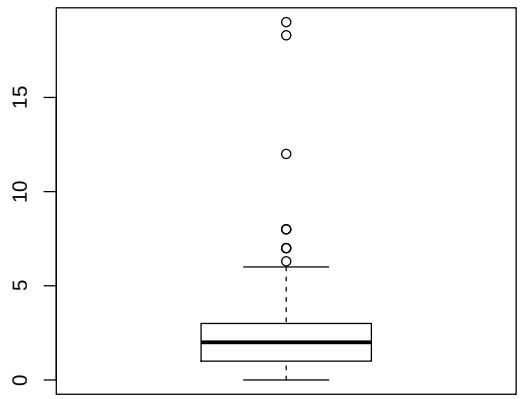


Figure 3

L'histogramme du PH, approché par une courbe de densité en forme de "cloche", vient confirmer cette hypothèse (voir figure ci-contre).

Etude de la variable NO3

La distribution de la variable NO3 est ici quelque peu différente : l'histogramme indique une distribution non-symétrique et non-homogène.



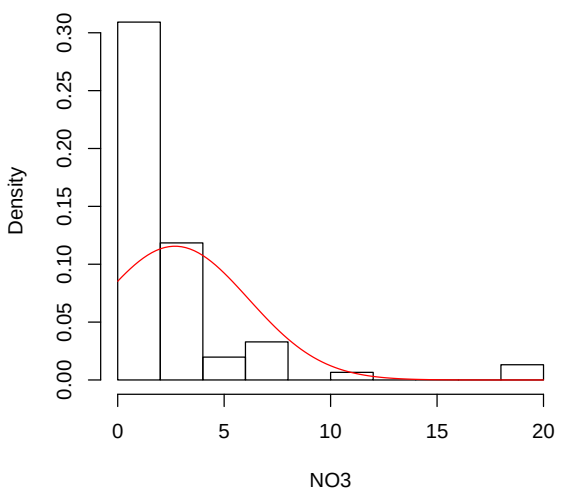


Figure 4

Le tracé de la boîte à moustache de cette variable révèle la présence de nombreuses valeurs extrêmes (symbolisées par un "°") : ces valeurs s'écartent fortement de la valeur moyenne obtenue.

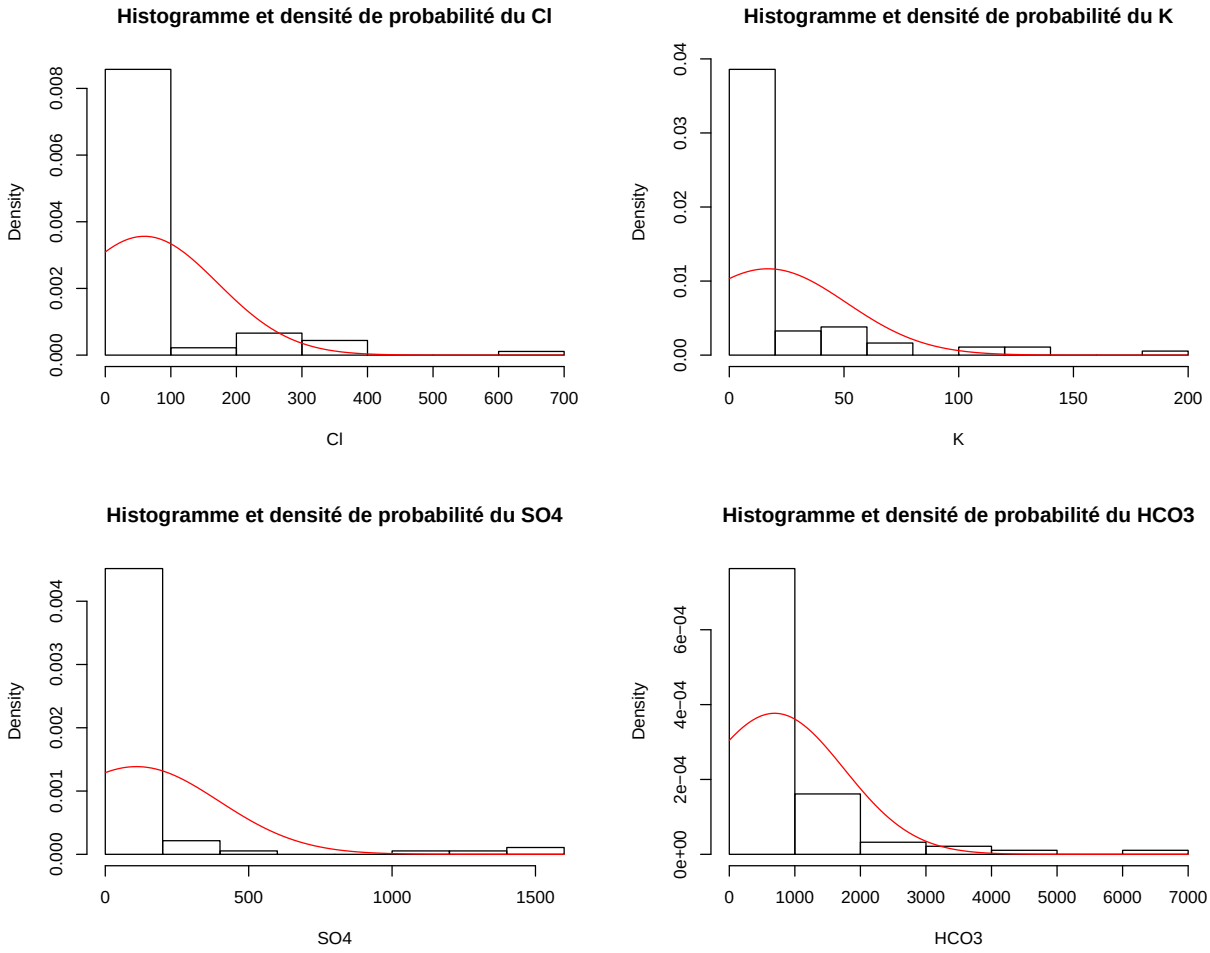


Figure 3: Figure 5

Etude des variables Cl, K, SO4 et HCO3 :

Les variables Cl, K, SO4 et HCO3 témoignent d'une distribution à nouveau non symétrique, peu homogène, enregistrant un grand nombre de valeurs extrêmes. Le tracé de leur densité de probabilité s'éloigne ici d'une loi normale.

Avant de poursuivre notre analyse par une étude bidimensionnelle, il peut donc être pertinent de s'intéresser aux données suspectes (outliers) susceptibles d'être à l'origine de ces distorsions, afin de les corriger et/ou supprimer puisqu'elles influeront nécessairement sur les résultats obtenus dans le cadre de notre ACP.

1.2.4. Recherche des valeurs extrêmes/outliers

Le repérage des données suspectes peut être mené via plusieurs méthodes. Le test de Grubbs est le test le plus courant, mais n'étant applicable que si la distribution de l'échantillon suit une loi Normale ou est voisine d'une loi Normale, il ne sera ici pas retenu.

Graphiquement, le tracé des boxplots révèle la présence d'outliers pour les variables SO4, K, Na, Cl et HCO3, et dans une moindre mesure NO3, Ca et Mg. Dans le cadre de variables continues, il s'agit des observations dont la valeur excède 1.5 fois l'écart interquartile. La commande `boxplot.stats(data)$out` permet de dégager ces valeurs, en considérant chaque variable prise une à une (par exemple ici, SO4) :

Nuage de points croisant deux à deux les variables du jeu de données

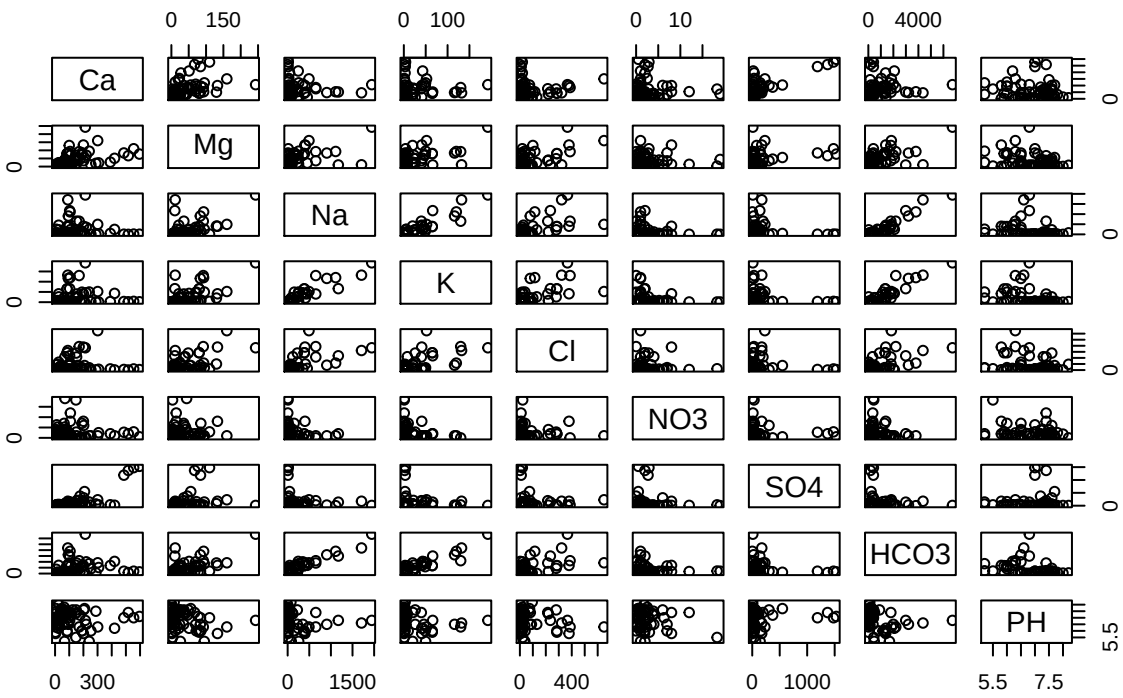


Figure 4: Figure 6

```
boxplot.stats(waters$SO4)$out
```

```
##  [1]  205.0  195.0 1187.0 1371.0 1479.0  150.0  143.0  230.0  372.0  173.0
## [11]  549.2 1530.0  158.0  174.0  306.0
```

La commande which permet d'identifier les indices des outliers :

```
which(waters$SO4 %in% boxplot.stats(waters$SO4)$out)
```

```
##  [1] 13 15 19 20 44 46 61 63 65 68 70 72 77 79 80
```

Les outliers sont donc très présents dans ce jeu de données. Il pourrait ainsi être judicieux de générer un deuxième jeu de données sans ces extrema, ou en tout cas essayer de s'en affranchir autant que faire se peut, de façon à ne pas biaiser la suite de notre étude. Passer ces données en observations supplémentaires dans le cadre de notre ACP est une façon de neutraliser ces valeurs.

1.2.5. Analyse bidimensionnelle

Intéressons-nous désormais à l'analyse bidimensionnelle. L'objectif ici est de pouvoir mettre en évidence, pour chaque paire de variables considérée, l'existence de variations simultanées, aussi appelées liaisons. La matrice des scatterplots permet de visualiser les liaisons existantes entre variables quantitatives prises deux à deux.

On observe ainsi graphiquement l'existence de corrélations entre les variables Na, K et HCO3. Cette hypothèse peut être vérifiée via l'étude numérique des corrélations inter-variables (`commande cor`).

```
round(cor(waters[sapply(waters, is.numeric)], use="complete.obs"),3)
```

```
##          Ca     Mg     Na      K     Cl    NO3    SO4   HCO3     PH
## Ca    1.000  0.662  0.085  0.140  0.118  0.017  0.845  0.286 -0.100
## Mg    0.662  1.000  0.443  0.579  0.627 -0.029  0.463  0.588 -0.369
## Na    0.085  0.443  1.000  0.867  0.629 -0.130 -0.080  0.910 -0.310
## K     0.140  0.579  0.867  1.000  0.616 -0.178 -0.090  0.883 -0.433
## Cl    0.118  0.627  0.629  0.616  1.000 -0.022 -0.043  0.534 -0.235
## NO3   0.017 -0.029 -0.130 -0.178 -0.022  1.000 -0.059 -0.076 -0.176
## SO4   0.845  0.463 -0.080 -0.090 -0.043 -0.059  1.000 -0.042  0.088
## HCO3  0.286  0.588  0.910  0.883  0.534 -0.076 -0.042  1.000 -0.408
## PH   -0.100 -0.369 -0.310 -0.433 -0.235 -0.176  0.088 -0.408  1.000
```

Les variables les plus corrélées positivement sont les variables HCO3 et NA (0.910), HCO3 et K (0.883) et Na et K (0.867) : elles varient simultanément de façon semblable.

On remarque que le PH est corrélé négativement à toutes les autres variables, exception faite du composant S04. Le comportement de cette variable est donc, de façon générale, opposé à celui des autres variables.

Il faut cependant rester vigilant quant aux résultats obtenus. En effet, la présence de valeurs extrêmes peut venir perturber notre analyse, ici portée sur l'étude des corrélations. En effet, le coefficient de corrélation de Pearson (méthode par défaut) est très sensible aux valeurs de données extrêmes. Il pourra donc être intéressant de mettre en oeuvre une analyse sans prendre en compte ces données. C'est justement l'objet de notre troisième ACP, où nous passerons les différents extrema recensés en observations supplémentaires. L'idée sera alors de comparer, d'une ACP à l'autre, les groupes de variables formés et la nature des axes en découlant.

2. Analyse multivariée

Trois jeux de données seront ici considérés :

- un premier jeu de données diminué des données manquantes (approche simpliste) ;
- un deuxième jeu de données où l'ensemble des données manquantes seront substituées par des valeurs approximatives, et ce afin de conserver le maximum d'information possible ;
- un troisième jeu de données brute où, comme nous l'avons écrit plus haut, les outliers seront passés en observations supplémentaires.

Il s'agira donc de travailler sur les données brutes d'abord (1), puis sur un jeu de données transformé ensuite (2). Dans la mesure où les résultats d'une ACP peuvent être biaisés par la présence de données suspectes et que celles-ci sont ici nombreuses, une troisième ACP, où seront passées en individus supplémentaires l'ensemble des données suspectes, sera réalisée. À noter qu'il aurait également été envisageable de neutraliser ces données à l'aide d'une ACP sur les rangs, plus robuste qu'une ACP classique.

2.1. ACP sur le premier jeu de données

Nous considérons ici le jeu de données brutes, diminué des données manquantes. La commande `na.omit` permet d'éliminer d'un jeu de données toutes les lignes incomplètes (contenant des valeurs "NA").

```
complete.waters <- na.omit(waters)
dim(complete.waters)
```

```
## [1] 62 12
```

On note que 33 observations ont ainsi été retirés du jeu de données initial.

De plus, dans la mesure où l'on souhaite ici que les eaux marocaines soient traitées comme des observations supplémentaires et non pas actives, il faut veiller à les retirer au préalable. Une fois ce traitement réalisé, le jeu de données à traiter ne comporte alors plus que 55 lignes.

Avant de débuter l'ACP à proprement parler, il convient de faire un choix de métrique. Dans la mesure où la variable PH ne s'exprime pas dans la même unité que les autres variables (concentrations en mg/L), on préfèrera ici mettre en oeuvre une ACP normée (ACPN). Éliminer la variable PH du jeu de données conduirait en effet à une perte d'information non négligeable. De plus, les variances des variables étant très éloignées les unes des autres, une ACP non normée n'apparait pas judicieuse.

2.1.1. Sélection des axes et plans retenus

L'ACPN est effectuée par un appel à la procédure `dudi.pca`.

La première étape consiste à sélectionner les axes à retenir. Cette sélection s'appuie sur l'examen des valeurs propres, où chaque valeur propre correspond à la part d'inertie projetée sur un axe donnée.

Dans le cadre d'une ACP normée, on peut utiliser le critère de Kaiser selon lequel ne seront retenus que les facteurs ayant une valeur propre supérieure à 1. On remarque ici que seules les trois premières valeurs propres sont supérieures à 1. L'histogramme des valeurs propres traduit ce fait.

```
round(cumsum(100*auto.acp$eig/sum(auto.acp$eig)), 3)
```

```
## [1]  45.646  69.043  82.023  89.180  95.360  97.601  99.066  99.990 100.000
```

Cumulées, elles représentent alors plus de 82.02% de la variabilité totale. On ne conservera donc ici que les trois premiers axes.

2.1.2. Projection des variables et observations dans un plan donné

Analyse des variables

```
round(auto.acp$co,3)
```

```
##       Comp1  Comp2  Comp3  Comp4  Comp5  Comp6  Comp7  Comp8  Comp9
## Ca   -0.378 -0.886 -0.032 -0.067 -0.131 -0.065  0.167  0.132  0.010
## Mg   -0.792 -0.460 -0.055  0.084  0.261 -0.237 -0.073 -0.149  0.005
## Na   -0.864  0.304  0.142  0.083 -0.296  0.187  0.024 -0.100  0.016
## K    -0.907  0.262  0.103 -0.070 -0.094 -0.077 -0.230  0.158  0.001
## Cl   -0.765  0.114  0.013  0.467  0.391  0.140  0.083  0.072 -0.005
## NO3   0.105 -0.018 -0.903  0.335 -0.243 -0.024 -0.047  0.005  0.000
## SO4  -0.123 -0.944  0.103  0.012 -0.108  0.226 -0.138 -0.039 -0.011
## HCO3 -0.917  0.190  0.036 -0.079 -0.294 -0.103  0.129 -0.039 -0.019
## PH    0.562 -0.092  0.554  0.534 -0.254 -0.138 -0.022  0.005  0.000
```

On remarque que :

- La première composante traduit l'existence de fortes corrélations entre HCO3 (-0.917), K (-0.907), Na (-0.864), et dans une moindre mesure Mg (-0.792) et Cl (-0.765).
- La deuxième composante traduit l'existence de fortes corrélations entre SO4 (-0.944) et Ca (-0.886.
- Enfin, la troisième composante traduit l'existence de corrélations négatives entre NO3 (-0.903) et PH (0.554)

NB : On note que la variabilité des six dernières composantes principales (non retenues ici) est beaucoup plus faible que celle des composantes retenues. Autrement dit, la distance qu'elles traduisent entre les observations est négligeable.

Le tracé du cercle des corrélations peut nous aider à apprécier ces corrélations, ainsi que la bonne représentation des différentes variables (proximité vis-à-vis du bord du cercle). Dans la mesure où l'on a décidé ici de retenir 3 axes, il faut alors tracer trois graphiques distincts pour chaque nuage : le nuage projeté sur le plan (axe1, axe2), celui projeté sur le plan (axe1, axe3) et enfin celui sur le plan (axe2, axe3).

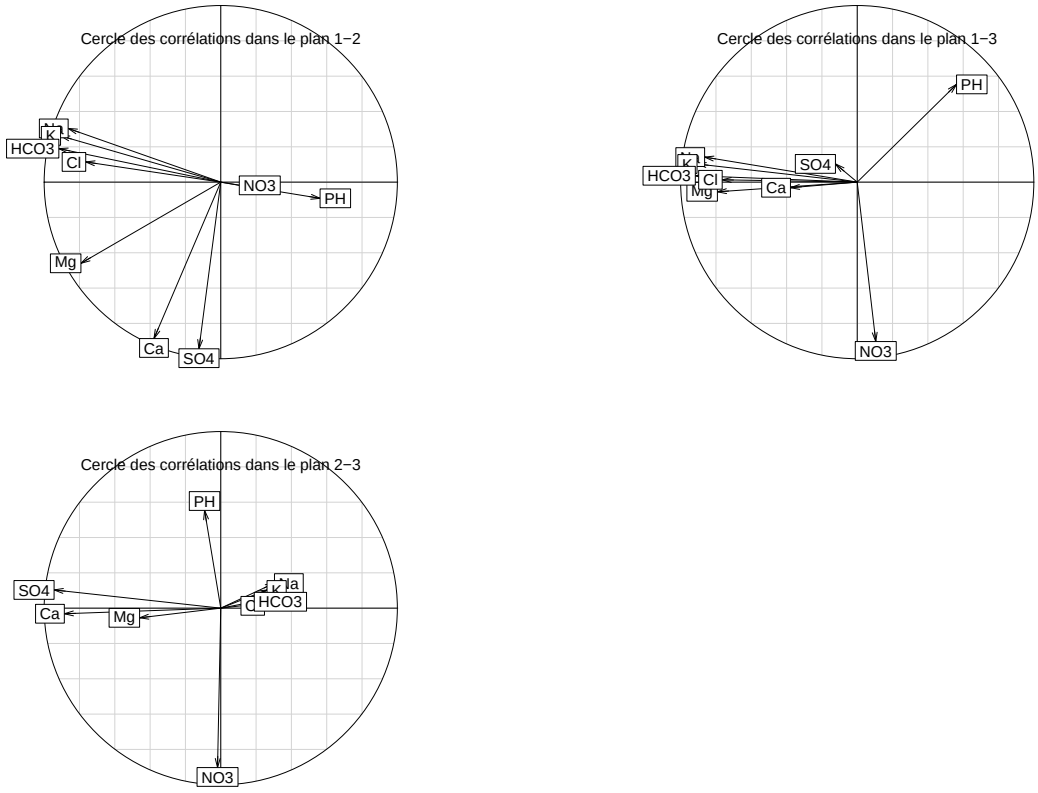


Figure 7

Les plans 1-2 et 1-3 permettent permettent de distinguer 2 groupes de variables distincts, tels que décrits plus haut : Na, K, HCO3, Cl et Mg d'une part, Ca et SO4 d'autre part. Les variables NO3 et PH sont assez mal représentées dans ce plan (éloignées du bord du cercle de corrélation), à la différence des plans 1-3 et 2-3 où leur représentation est bonne. L'existence de corrélations négatives entre ces deux variables est facilement visualisable dans ces configurations. La conservation des trois premiers axes est donc ici nécessaire si l'on ne souhaite pas aboutir à de mauvaises interprétations.

Interprétation des axes :

- Axe 1 : cet axe est lié aux concentrations en Na, K, HCO3, Cl et Mg. Il caractérise la teneur sodique et minérale des eaux ;
- Axe 2 : cet axe est lié aux concentrations en Ca et SO4. Il caractérise les eaux riches/pauvres en sulfate et calcium ;
- Axe 3 : cet axe est lié aux concentrations en NO3. Il caractérise le pouvoir nitrique des eaux (acides ou non).

L'analyse des répartition des observations par nature (eaux plates ou gazeuses) dans chacun des plans permettra de mettre en confrontation les variables quantitatives ici présentées avec la variable qualitative Nature (cf. "Analyse croisée des variables et observations").

Analyse des observations

Qualité de représentation des observations :

La commande `obs.inertie$row.re` permet d'apprécier la qualité de représentation des différents individus sur les différents axes. La qualité de représentation d'un individu dans un plan donné s'évalue en sommant les qualités de représentation de ce même individu sur les axes du plan donné. Sont affichées ici les 30 premières qualités.

NB : Il faut se reporter au jeu de données initial pour établir la correspondance entre un échantillon (numéro) et l'eau associée.

```
head(round(obs.qlt[,1:3],3),30)
```

```
##     Axis1  Axis2  Axis3
## 1   0.344  0.078  0.367
## 2   0.582  0.001  0.187
## 3   0.803  0.054  0.052
## 10 -0.857  0.047  0.020
## 12 -0.184 -0.016 -0.582
## 14 -0.634  0.006  0.035
## 15 -0.552  0.048  0.134
## 16 -0.422 -0.033  0.000
## 19 -0.009 -0.982  0.000
## 20  0.000 -0.954  0.012
## 21  0.831  0.057  0.052
## 22  0.902  0.024  0.019
## 23  0.742  0.053  0.064
## 24  0.466  0.000  0.423
## 28  0.636  0.062  0.198
## 29  0.727  0.076  0.168
## 30  0.802  0.044 -0.004
## 31  0.562  0.132  0.185
## 32  0.307  0.220  0.205
## 33 -0.011  0.059 -0.208
## 34  0.535  0.153 -0.186
## 36  0.747  0.138  0.087
## 37  0.849  0.030  0.074
## 38  0.850  0.031  0.074
## 40  0.369  0.006 -0.331
## 41  0.793  0.040 -0.061
## 44 -0.026 -0.962  0.000
## 48  0.692  0.101  0.190
## 49  0.711  0.060  0.124
## 50  0.649  0.160  0.008
```

Plan 1-2 : les individus les mieux représentés dans le plan 1-2 sont les eaux Contrex (99,06%), Hepar (98,82%), Courmayeur (95,46%), Talians (95,32%), Christalline Aurelie (92,61%), Zilia (92,14%) et Arvie (90,36%). Les individus les moins bien représentés dans le plan 1-2 sont les eaux Perrier (1,97%), Salvetat (2,55%) et Christalline St Sophie (6,99%)

Plan 1-3 : les individus les mieux représentés dans le plan 1-3 sont les eaux Perrier (94,54%), Montagne Ecrins (92,38%), Montagne Alpes (92,37%) et Christalline Aurelie (92,11%). Les individus les moins bien représentés dans le plan 1-3 sont les eaux Contrex (0,88%), Courmayeur (1,23%), Talians (2,45%), Hepar (2,60%) et Salvetat (5,92%)

Plan 2-3 : les individus les mieux représentés dans le plan 2-3 sont les eaux Contrex (98,18%), Courmayeur (96,65%), Talians (96,51%), Hépar (96,22%) et Perrier (92,85%). Les individus les moins bien représentés dans le plan 2-3 sont les eaux Rozana (1,31%), Chateldon (3,36%), Christalline Aurelie (4,24%), Christalline

St JB (4,86%) et Arvie (6,66%)

Contribution des observations aux axes :

Nous nous intéressons ici aux observations dont la contribution est supérieure à la contribution moyenne par axe. Voici un extrait des contributions des observations aux axes 1, 2 et 3.

```
head(round(obs.ctr[,1:3],3),40)
```

```
##    Axis1 Axis2 Axis3
## 1  0.009 0.004 0.033
## 2  0.004 0.000 0.005
## 3  0.007 0.001 0.001
## 10 0.133 0.014 0.011
## 12 0.005 0.001 0.059
## 14 0.021 0.000 0.004
## 15 0.024 0.004 0.020
## 16 0.019 0.003 0.000
## 19 0.001 0.142 0.000
## 20 0.000 0.172 0.004
## 21 0.007 0.001 0.002
## 22 0.008 0.000 0.001
## 23 0.012 0.002 0.003
## 24 0.005 0.000 0.015
## 28 0.007 0.001 0.008
## 29 0.008 0.002 0.007
## 30 0.009 0.001 0.000
## 31 0.006 0.003 0.007
## 32 0.001 0.002 0.003
## 33 0.000 0.003 0.020
## 34 0.009 0.005 0.011
## 36 0.012 0.004 0.005
## 37 0.011 0.001 0.003
## 38 0.011 0.001 0.003
## 40 0.008 0.000 0.025
## 41 0.006 0.001 0.002
## 44 0.003 0.222 0.000
## 48 0.014 0.004 0.014
## 49 0.016 0.003 0.010
## 50 0.010 0.005 0.000
## 51 0.010 0.005 0.002
## 52 0.006 0.006 0.015
## 55 0.011 0.002 0.010
## 57 0.114 0.027 0.011
## 58 0.002 0.000 0.437
## 60 0.003 0.003 0.000
## 61 0.032 0.001 0.008
## 62 0.041 0.007 0.001
## 63 0.133 0.005 0.000
## 64 0.000 0.000 0.003
```

Les eaux Arvie (échantillon 10), Rozana (échantillon 63), Parot (échantillon 57) et Vichy Celestins (échantillon 78) enregistrent les valeurs (absolues) de la première composante les plus élevées. Ces eaux sont naturellement riches en Na, K, HCO3, Cl et Mg, plutôt acides (PH $<$ 7) et gazeuses.

Un examen identique de la seconde composante révèle que les eaux Talians (échantillon 72), Hepar (échantillon 44), Courmayeur (échantillon 20) et Contrex (échantillon 19) ont des valeurs très supérieures à la moyenne

sur cette composante. Ces eaux sont naturellement riches en Ca, Mg et SO4, plutôt basiques ou alcalines (PH $>$ 7) et plates.

Enfin, les eaux Perrier (échantillon 58), Thonon (échantillon 74), Ste Marguerite (échantillon 68) et Badoit (échantillon 12) enregistrent les valeurs les plus élevées de la troisième composante. Ces eaux ont la particularité d'être fortement concentrées en NO3.

L'ACP a donc permis de mettre en évidence des groupements de propriétés minérales propres à certains types d'eaux, difficilement perceptible au premier coup d'oeil.

En résumé, la contribution des individus aux différents axes (par ordre décroissant de contribution) est la suivante :

Axe 1 : Arvie, Rozana, Parot, Vichy Celestins

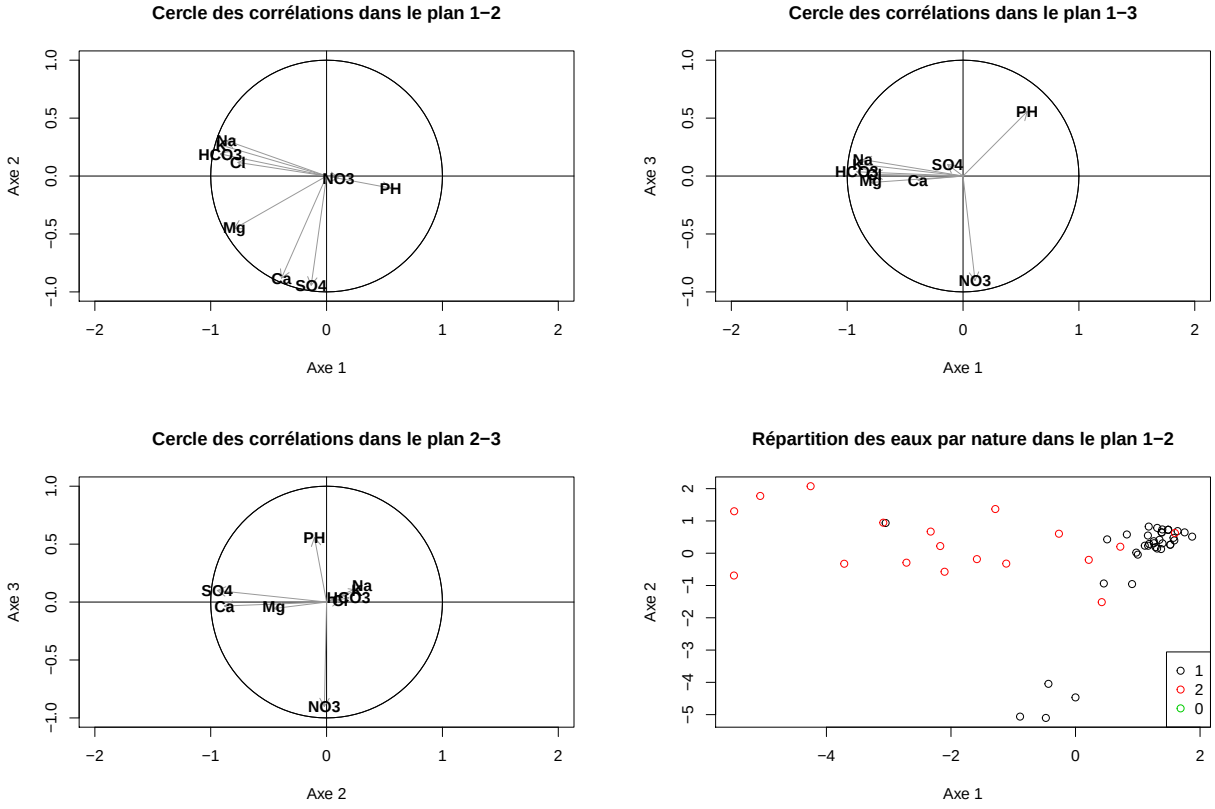
Axe 2 : Talians, Hepar, Courmayeur, Contrex

Axe 3 : Perrier, Thonon, Ste Marguerite, Badoit

Analyse croisée des variables et observations

Le tracé des graphiques de répartion des eaux par Nature (plates, gazeuses) dans les différents plans considérés est présenté ci-contre. Il permet notamment de mettre en relation les variables quantitatives précédemment étudiées avec la variable qualitative Nature, en considérant les individus bien représentés seulement.

En rouge, sont désignées les observations de nature gazeuse; en noir, les observations de nature plate. Les cercles de corrélations ici dessinés sont ceux issus de notre fonction ACP.



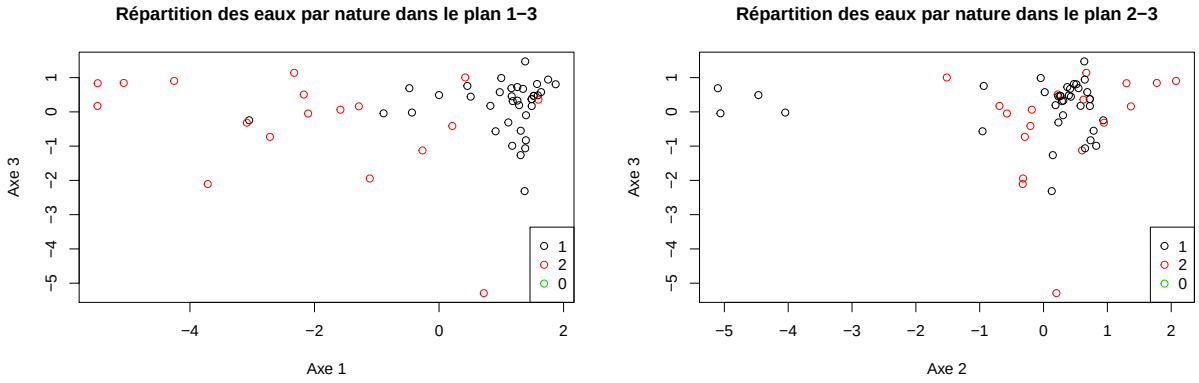


Figure 8

On remarque que c'est dans le plan 1-2 que la séparation entre eaux plates et gazeuses est la meilleure. Or, nous avons précédemment déduit que l'axe 1 était associé aux concentrations en Na, K, HCO3, Cl et Mg d'une part, et que les composants minéraux qui semblaient être les plus discriminants dans la distinction des eaux plates et gazeuses étaient le HCO3 et le Na d'autre part. L'analyse graphique est donc concordante avec les résultats de notre analyse.

2.1.3. Analyse des observations supplémentaires

Les observations supplémentaires sont des eaux provenant du Maroc, exclusivement plates. Comparativement aux eaux françaises, les eaux marocaines sont plus riches en Na.

```
##                Nom Nature     Ca   Mg     Na    K    Cl  NO3  SO4  HCO3  PH
## 85        Sidi Ali   plat  12.02  8.7  25.5  2.8  14.2  0.1 41.7 103.7 6.5
## 86    Sidi Harazem   plat  70.00 40.0 120.0  8.0 220.0  4.0 20.0 335.0 7.3
## 87       Ain Saiss   plat  63.50 35.5   8.0  1.0  19.8  7.0  3.8 372.0 6.5
## 88       Oulm\xe8s   plat 148.80 48.6 224.0 26.0 280.0  2.8 14.3 890.9 7.6
## 89     Ain Soltane   plat  70.00 44.0   4.0  1.0    NA 19.0  3.7 402.0  NA
## 90       Ain Atlas   plat  17.80 13.6  50.0  8.0  14.2  5.2 12.9 250.1  NA
## 91      Ain Ifrane   plat  67.70 40.6   3.0  1.0  10.7  5.2  5.1 402.6  NA
## 92 Ain Chefchaoun   plat  80.40 16.2  14.5  1.0  17.5  0.3 20.9 309.8  NA
## 93           Bahia   plat   8.00  7.3  46.0  1.0   7.8  0.1 15.7  42.7 6.5
## 94            Ciel   plat  25.70 23.8 224.0 26.0 130.0  1.3  6.2  27.5 6.5
## 95          Mazine   plat  11.20  9.7  52.0  1.0  88.8  3.4 20.6  42.7 6.5
##     Pays
## 85 Maroc
## 86 Maroc
## 87 Maroc
## 88 Maroc
## 89 Maroc
## 90 Maroc
## 91 Maroc
## 92 Maroc
## 93 Maroc
## 94 Maroc
## 95 Maroc
```

```
##              Nom     Nature          Ca               Mg
##  Ain Atlas     :1   gaz : 0   Min.   :  8.00   Min.   : 7.30
##  Ain Chefchaoun:1   plat:11   1st Qu.: 14.91   1st Qu.:11.65
```

```
##  Ain Ifrane    :1             Median : 63.50   Median :23.80
##  Ain Saiss     :1             Mean   : 52.28   Mean   :26.18
##  Ain Soltane   :1             3rd Qu.: 70.00   3rd Qu.:40.30
##  Bahia         :1             Max.   :148.80   Max.   :48.60
##  (Other)       :5
##        Na               K                Cl              NO3
##  Min.   :  3.00   Min.   : 1.000   Min.   :  7.80   Min.   : 0.1
##  1st Qu.: 11.25   1st Qu.: 1.000   1st Qu.: 14.20   1st Qu.: 0.8
##  Median : 46.00   Median : 1.000   Median : 18.65   Median : 3.4
##  Mean   : 70.09   Mean   : 6.982   Mean   : 80.30   Mean   : 4.4
##  3rd Qu.: 86.00   3rd Qu.: 8.000   3rd Qu.:119.70   3rd Qu.: 5.2
##  Max.   :224.00   Max.   :26.000   Max.   :280.00   Max.   :19.0
##                                    NA's   :1
##       SO4              HCO3             PH            Pays
##  Min.   : 3.70   Min.   : 27.5   Min.   :6.500   France: 0
##  1st Qu.: 5.65   1st Qu.: 73.2   1st Qu.:6.500   Maroc :11
##  Median :14.30   Median :309.8   Median :6.500
##  Mean   :14.99   Mean   :289.0   Mean   :6.771
##  3rd Qu.:20.30   3rd Qu.:387.0   3rd Qu.:6.900
##  Max.   :41.70   Max.   :890.9   Max.   :7.600
##                                  NA's   :4
```

Parmi ces 11 observations supplémentaires, 4 enregistrent des valeurs manquantes, notamment au niveau de leur PH. Dans la mesure où nous avons précédemment vu que la distribution du PH s'apparente à celle d'une loi normale, on peut raisonnablement substituer ces valeurs manquantes par la valeur médiane de cette variable, qui vaut ici 7.0. Cela évite ainsi d'éliminer un tiers des observations supplémentaires.

Etant donné que nous n'avons pas réalisé d'estimation pour les données manquantes enregistrées au niveau de la variable Cl (cf. deuxième jeu de données), nous avons ici pris le parti de supprimer l'observation Ain Soltane (échantillon 89). Ces individus supplémentaires (ici en rouge) ont été représentés dans chacun des plans 1-2, 1-3 et 2-3.

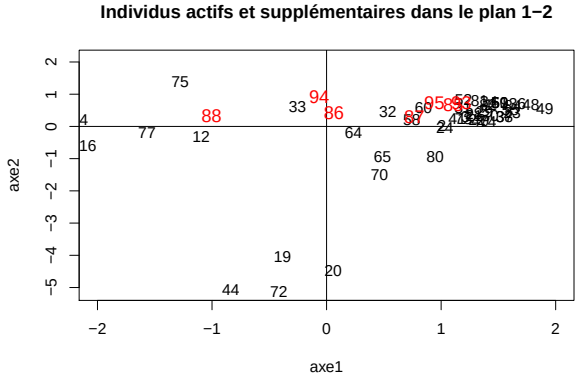
Les cercles de corrélations ont été ici ajoutés une nouvelle fois de façon à plus facilement superposer les résultats d'analyse.

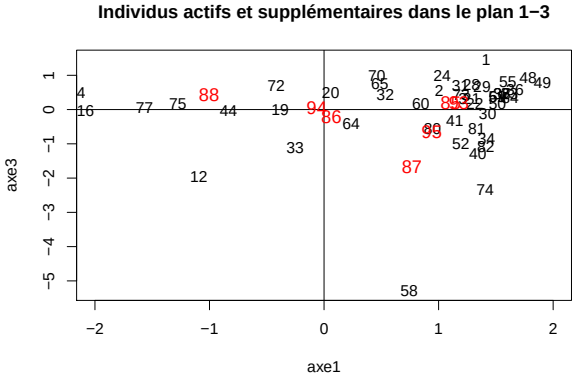
La qualité de représentation des individus supplémentaires s'obtient à partir des commandes suivantes :

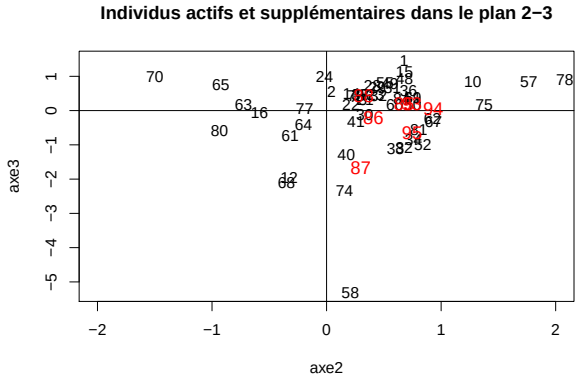
```
ligsup <- suprow(auto.acp,new.supp.waters[3:11])
acpcos2sup <- ligsup$lisup^2/apply(ligsup$lisup^2, 1, sum)
round(acpcos2sup[,1:3]*1000)
```

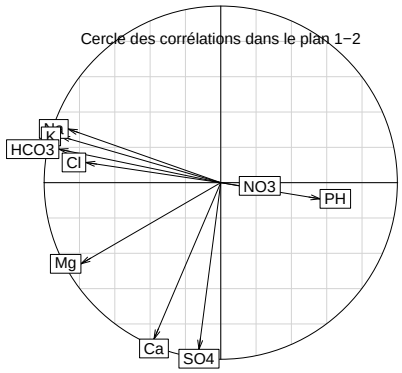
```
##    Axis1 Axis2 Axis3
## 85   383   144    14
## 86     1    59    15
## 87   163    25   763
## 88   210    22    39
## 93   396   158    13
## 94     2   352     1
## 95   319   201   152
```

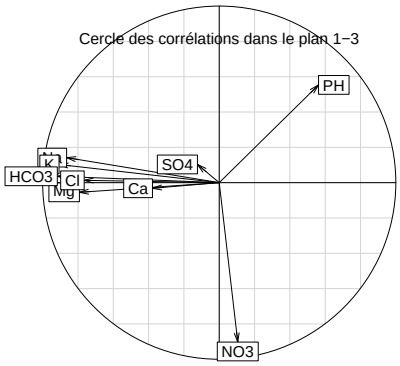
Seules les observations bien représentées peuvent être correctement analysées. L'observation #86 est, à titre d'exemple, très mal représentée, quel que soit le plan. Même chose pour l'observation #94 dans le plan 1-3. Nous nous focaliserons donc sur les autres observations.


Individus actifs et supplémentaires dans le plan 1−2
axe1
axe2


Individus actifs et supplémentaires dans le plan 1−3
axe1
axe3


Individus actifs et supplémentaires dans le plan 2−3
axe2
axe3


Cercle des corrélations dans le plan 1−2
K
HCO3
Cl
NO3
PH
Mg
Ca
SO4


Cercle des corrélations dans le plan 1−3
PH
SO4
HCO3
Cl
Mg
Ca
NO3

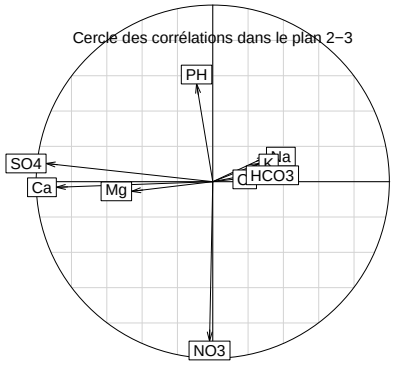

Cercle des corrélations dans le plan 2−3
PH
SO4
Ca
Mg
HCO3
NO3

Figure 9

Interprétation :

On remarque que l'observation #88 (Oulmès) est une eau caractérisée par une forte concentration en HC03, Cl, K, Na et Mg. Cette eau se distingue des autres observations supplémentaires, isolée à la fois dans les plans 1-2 et 1-3. Si l'on jette un coup d'oeil au tableau d'analyse numérique inséré plus haut, on note en effet qu'elle enregistre des concentrations beaucoup plus importantes que les autres eaux pour les variables ici considérées.

Dans le plan 1-3, on distingue un groupe défini par les observations #95, #90, #91 et #87 : ces eaux sont comparativement plus concentrées en NO3 que les autres.

Dans le plan 2-3, c'est l'eau Ciel (#94) qui se distingue, enregistrant de fortes concentrations en Na, K et Cl, ainsi que l'eau Ain Saiss (#87), qui enregistre la deuxième plus forte concentration en NO3.

2.2. ACP sur le deuxième jeu de données

Nous considérons ici un deuxième jeu de données où les valeurs manquantes ne sont plus supprimées mais substituées/imputées par des valeurs approximatives.

Pour ce faire, on peut utiliser l'algorithme des k plus proches voisins, déjà implémenté au sein du package DMwR. Pour chaque valeur manquante, l'algorithme identifie les k observations les plus proches en terme de distance euclidienne et en calcule la moyenne pondérée.

La fonction `knnImputation` permet de générer un nouveau jeu de données suivant cette méthode. En voici un aperçu :

```
## Loading required package: lattice
```

```
## Loading required package: grid
```

```
##               Nom Nature  Ca   Mg Na        K        Cl NO3      SO4
## 1       Abatilles   plat  16  8.0 75 3.000000 95.000000   0  8.00000
## 2   Aix-Les-Bains   plat  72 38.0 14 2.000000  6.000000   1 81.00000
## 3            Alet   plat  63 23.0 13 1.300000 11.000000   2 14.00000
## 4         Alpille   plat  41  3.0  2 0.000000  3.000000   3  2.00000
## 5 Amelie le Reine    gaz 390 27.5 45 2.800000 19.000000   2 36.00000
## 6       Aquarelle   plat  70  2.1  2 1.010236  7.193424   4 20.67155
##     HCO3       PH   Pays
## 1  112.0 8.200000 France
## 2  329.0 7.400000 France
## 3  300.0 7.400000 France
## 4  134.0 7.223073 France
## 5 1376.6 6.567166 France
## 6  210.0 7.116851 France
```

On remarque par exemple que l'eau Aquarelle, originellement incomplète, a ici été imputée au niveau des variables K, Cl, SO4 et PH par les valeurs 1.01, 7.19, 20.67 et 7.12.

Nous pouvons désormais appliquer notre ACP sur ce nouveau jeu de données, en suivant la même méthode.

2.2.1. Sélection des axes et plans retenus

```
round(new.auto.acp$eig,3)
```

```
## [1] 4.225 1.903 1.149 0.712 0.482 0.268 0.190 0.068 0.003
```

```
round(cumsum(100*new.auto.acp$eig/sum(new.auto.acp$eig)), 3)
```

```
## [1]  46.945  68.087  80.857  88.771  94.122  97.095  99.207  99.963 100.000
```

L'étude des valeurs propres selon le critère de Kaiser nous amène à conserver les trois premiers axes. L'inertie cumulée est sensiblement la même que celle obtenue dans le cadre de notre première ACP.

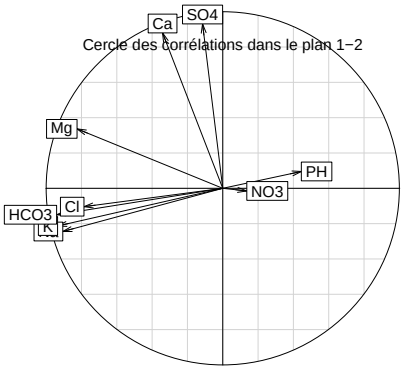
2.2.2. Projection des variables et observations dans un plan donné

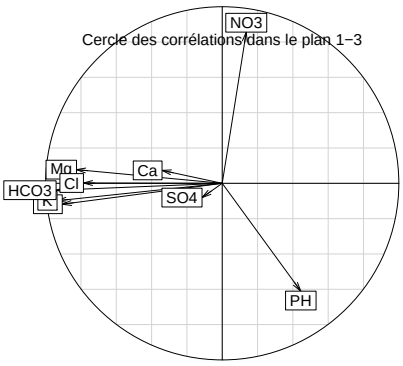
Analyse des variables

```
round(new.auto.acp$co,3)
```

```
##       Comp1  Comp2  Comp3
## Ca   -0.341  0.880  0.073
## Mg   -0.825  0.337  0.077
## Na   -0.906 -0.245 -0.118
## K    -0.935 -0.213 -0.099
## Cl   -0.786 -0.104  0.002
## NO3   0.136 -0.015  0.856
## SO4  -0.114  0.931 -0.081
## HCO3 -0.941 -0.150 -0.039
## PH    0.445  0.095 -0.611
```

On remarque que les corrélations entre variables sont ici identiques à celles précédemment établies. Le tracé du cercle des corrélations vient confirmer la grande proximité des résultats obtenus dans ces deux ACP, en notant toutefois que la corrélation entre Mg et le premier groupe de variables distingué (Na, K, HCO3 et Cl) est plus marquée.





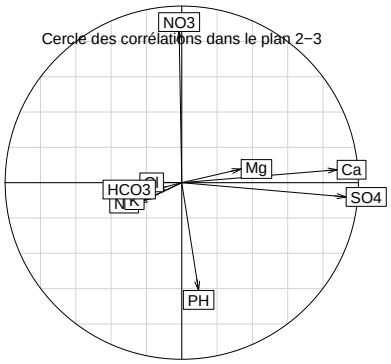


Figure 10

L'interprétation des axes est donc par conséquent la même.

Analyse des observations

C'est l'analyse des observations qui va ici le plus nous importer. En effet, cette deuxième ACP va nous permettre de situer des observations qui, auparavant, avaient été éliminées du jeu de données. Là encore, il convient d'étudier la bonne qualité de représentation de ces observations.

Pour en faciliter le repérage, la liste de ces observations est rappelée ci-contre :

```
sort(unique(which(is.na(active.waters), arr.ind = TRUE)[,1]))
```

```
##  [1]  4  5  6  7  8  9 11 13 17 18 25 26 27 35 39 42 43 45 46 47 53 54 56
## [24] 59 66 71 76 79 83
```

Les qualités sont résumées dans la table ci-contre :

```
round(obs.qlt[4:13,1:3],3)
```

```
##     Axis1  Axis2  Axis3
## 4   0.840 -0.125 -0.006
## 5  -0.009  0.246  0.020
## 6   0.809 -0.073  0.041
## 7   0.589 -0.204 -0.001
## 8  -0.515  0.000  0.081
## 9   0.142 -0.539  0.065
## 10 -0.821 -0.023 -0.013
## 11 -0.331  0.034  0.058
## 12 -0.099  0.017  0.751
## 13 -0.105  0.377  0.250
```

```
##     Axis1  Axis2  Axis3
## 30  0.674 -0.031 -0.002
## 31  0.392 -0.107 -0.368
## 32  0.250 -0.106 -0.378
## 33 -0.005 -0.038  0.242
## 34  0.493 -0.190  0.210
## 35  0.712 -0.204 -0.001
## 36  0.618 -0.143 -0.219
## 37  0.679 -0.019 -0.201
## 38  0.679 -0.020 -0.201
```

```
## 39  0.804 -0.062  0.061
## 40  0.266 -0.004  0.277
## 41  0.724 -0.042  0.025
## 42  0.726 -0.086 -0.013
## 43 -0.831 -0.064 -0.012
## 44 -0.021  0.958  0.000
## 45 -0.919 -0.010 -0.008
## 46  0.479  0.009 -0.490
## 47  0.611 -0.221  0.002
## 48  0.515 -0.089 -0.369
## 49  0.524 -0.048 -0.261
## 50  0.574 -0.184 -0.042
```

```
##    Axis1  Axis2 Axis3
## 76 0.852 -0.092 0.039
```

Plan 1-2 : Parmi les individus imputés, les mieux représentés dans le plan 1-2 sont les eaux #4 (96,47%), #76 (94,44%), #45 (92,95%), #54 (92,07%), #35 (91,67%), #27 (91,30%).

Plan 1-3 : Parmi les individus imputés, les mieux représentés dans le plan 1-3 sont les eaux #46 (96,82%), #45 (92,69%), #76 (89,06%) et #39 (86,46%). À l'inverse, l'individu imputé #5 est assez mal représenté dans ce plan (2,85%).

Plan 2-3 : Parmi les individus imputés, l'individu le mieux représenté dans le plan 2-3 est l'eau #13 (62,76%). Les individus imputés les moins bien représentés dans ce plan sont les eaux #45 (1,76%) et #43 (7,58%)

De façon générale, les individus qui étaient bien représentés dans la première ACP le sont aussi ici; même chose du côté des individus mal représentés.

Contribution des observations aux axes :

De la même façon, nous nous intéresserons ici aux observations dont la contribution est supérieure à la contribution moyenne par axe, en portant notre attention sur les observations imputées avant tout.

```
round(obs.ctr[40:80,1:3],3)
```

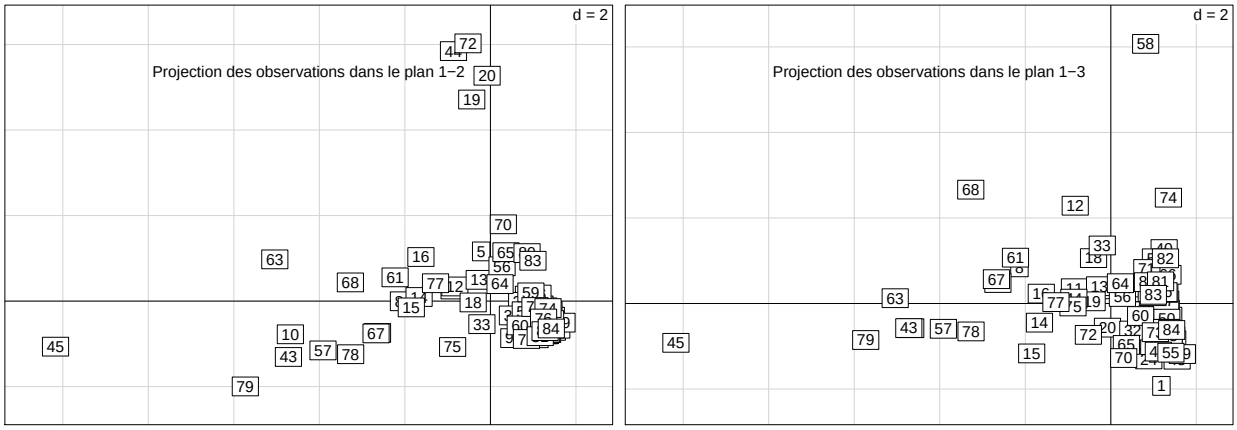
```
##    Axis1 Axis2 Axis3
## 40 0.004 0.000 0.017
## 41 0.003 0.000 0.000
## 42 0.005 0.001 0.000
## 43 0.063 0.011 0.003
## 44 0.002 0.214 0.000
## 45 0.291 0.007 0.009
## 46 0.003 0.000 0.013
## 47 0.005 0.004 0.000
## 48 0.007 0.003 0.018
## 49 0.008 0.002 0.014
## 50 0.005 0.003 0.001
## 51 0.005 0.003 0.004
## 52 0.003 0.005 0.012
## 53 0.003 0.002 0.000
## 54 0.004 0.003 0.001
## 55 0.006 0.001 0.014
## 56 0.000 0.004 0.000
## 57 0.043 0.008 0.004
## 58 0.002 0.000 0.382
## 59 0.003 0.000 0.001
## 60 0.001 0.002 0.001
```

```
## 61 0.014 0.002 0.012
## 62 0.020 0.003 0.002
## 63 0.072 0.006 0.000
## 64 0.000 0.001 0.002
## 65 0.000 0.008 0.010
## 66 0.005 0.004 0.005
## 67 0.020 0.004 0.003
## 68 0.030 0.001 0.073
## 69 0.005 0.003 0.004
## 70 0.000 0.020 0.017
## 71 0.002 0.005 0.007
## 72 0.001 0.227 0.006
## 73 0.003 0.000 0.005
## 74 0.005 0.000 0.063
## 75 0.002 0.007 0.000
## 76 0.004 0.001 0.001
## 77 0.005 0.001 0.000
## 78 0.030 0.010 0.004
## 79 0.093 0.025 0.008
## 80 0.002 0.008 0.003
```

Les eaux Hydroxydase (échantillon 45) et Vichy St Yorre (échantillon 79) enregistrent les valeurs (absolues) de la première composante les plus élevées, en plus des eaux précédemment identifiées lors de la première ACP. Ces eaux sont naturellement riches en Na, K, Cl, HCO3 et éventuellement Mg, plutôt acides (PH < 7) et gazeuses.

Un examen identique de la seconde composante ne révèle aucune indication supplémentaire. Les eaux imputées ne contribuent pas à cet axe. Même remarque pour le troisième axe.

Analyse croisée des variables et observations



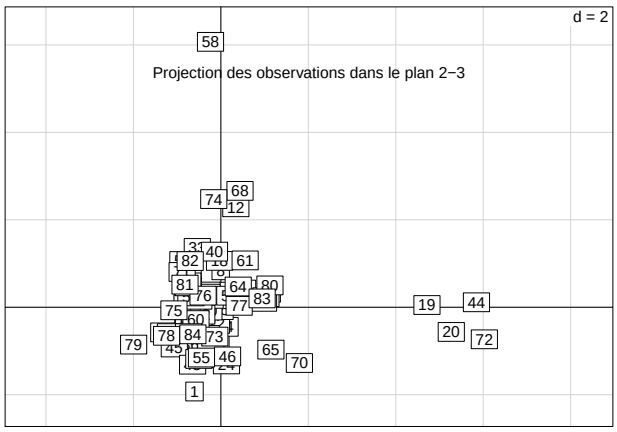


Figure 11

L'idée ici est d'analyser les individus qui ont fait l'objet d'une imputation.

Comme on peut le voir dans le plan 1-2, l'eau Hydroxydase (échantillon 45) se distingue des autres échantillons de par ses fortes concentrations en Cl, HCO3, K, Mg et Na. C'est cette eau qui enregistre la plus haute composition en HCO3 (6722.2 mg/L) : c'est un outlier pour cette variable. L'eau Vichy St Yorre (échantillon 79) enregistre également de fortes concentrations pour ces variables, même si, dans une moindre mesure. Ces deux échantillons se distinguent facilement sur le premier graphique.

L'interprétation liée aux projections des individus imputés sur les autres plans ne pourra pas être aussi concluante, dans la mesure où leur qualité de représentation n'est pas (ou moins) significative, exception faite de l'échantillon 45.

Cette étude s'est donc avérée enrichissante, dans la mesure où elle a permis de repositionner quelques individus préalablement non étudiés, bien que la qualité de représentation de la majorité d'entre eux, trop médiocre, a limité la portée de notre analyse.

2.3. ACP sur le troisième jeu de données

Nous considérons ici un troisième jeu de données, diminué des outliers. Ces individus, séparés des autres de par leurs valeurs "hors norme", ont été repérés un peu plus tôt au cours de notre analyse.

En voici un aperçu :

```
##       Ca   Mg    Na    K   Cl NO3   SO4   HCO3  PH
## 2   72.0 38.0  14.0  2.0  6.0 1.0  81.0  329.0 7.4
## 3   63.0 23.0  13.0  1.3 11.0 2.0  14.0  300.0 7.4
## 21  93.0  8.1   8.8  2.6 18.0 2.0   5.2  306.0 7.4
## 22 106.0  3.8   3.5  1.8  3.8 2.0  58.9  272.0 7.2
## 23  64.5  3.5  12.0  0.5 20.0 2.5   6.0  195.0 7.8
## 24 124.0 25.0  11.0  3.5 16.0 0.0  60.0  420.0 7.6
## 28  44.0 24.0  23.0  2.0  5.0 1.0   3.0  287.0 7.6
## 29  71.0  5.5  11.2  3.2 20.0 1.0   5.0  250.0 7.5
## 30  82.0  7.4   7.3  1.9 14.0 3.9  18.0  263.0 7.5
## 31  40.0 11.0  47.0  3.0 70.0 1.0   8.0  177.0 7.5
## 32  63.0 26.0  99.0 21.0 33.0 2.0  60.0  493.0 7.4
## 33  67.0 26.0  84.0 20.0 32.0 2.0  61.0  473.0 5.2
## 34   6.4  1.2   3.0  0.5  3.0 4.0   5.0   20.0 6.5
## 36   4.1  1.7   2.7  0.9  0.9 0.8   1.1   25.8 7.3
## 37  63.0 10.2   1.4  0.4  1.0 2.0  51.3  173.2 7.6
## 38  63.0 10.0   1.4  0.4  1.0 2.0  51.0  173.0 7.6
```

```
## 41  78.0 24.0   5.0  1.0  4.5 3.8  10.0  357.0 7.2
## 47   4.0  1.0   3.0  0.8  0.8 2.1   0.4   23.4 6.7
## 48   6.5  2.0   4.4  1.7  1.0 0.5   0.2   44.0 7.7
## 49  26.5  1.0   0.8  0.2  2.3 1.8   8.2   78.1 8.0
## 50   3.0  0.6   1.5  0.4  0.6 1.0   8.7    5.2 6.8
## 51   3.6  1.8   3.6  0.6  0.9 0.5   1.2   25.8 6.9
## 52   2.4  0.5   3.1  0.4  3.0 3.0   2.0    6.3 5.9
## 55  46.1  4.3   6.3  3.5  3.5 1.0   9.0  163.5 7.7
## 60  24.0 16.0  32.0  4.9 38.0 0.0  50.0  121.0 6.4
## 61 241.0 95.0 255.0 49.7 38.0 1.0 143.0 1685.4 5.2
## 64 253.0 11.0   7.0  3.0  4.0 1.0  25.0  820.0 6.0
## 69   3.6  1.8   3.6  0.6  0.9 0.5   1.2   24.4 6.9
## 73 116.0  4.4   9.0  2.4 15.5 1.0  25.5  331.0 7.2
## 77 190.0 72.0 154.0 49.0 18.0 0.0 158.0 1170.0 6.0
## 81   2.7  1.0   2.4  0.5  1.2 2.4   1.0   13.0 6.3
## 84  11.0  5.1  15.0  1.3 15.0 2.2   5.0   67.7 7.5
```

Traiter ces données non plus comme des observations actives mais comme des observations supplémentaires permettra alors de s'affranchir de leur influence sur les résultats d'ACP obtenus. On espère ainsi que les axes ne soient plus tirés vers ces éléments qui étaient susceptibles d'introduire des distorsions.

2.3.1. Sélection des axes et plans retenus

```
round(auto.acp$eig,3)
```

```
## [1] 4.007 2.256 1.267 0.712 0.349 0.163 0.147 0.098 0.001
```

```
round(cumsum(100*auto.acp$eig/sum(auto.acp$eig)), 3)
```

```
## [1]  44.521  69.587  83.668  91.581  95.454  97.262  98.893  99.988 100.000
```

Comme pour les deux analyses précédentes, conserver trois axes semble être la meilleure option puisque seules les trois premières valeurs propres sont supérieures à 1 (critère de Kaiser).

On obtient une inertie cumulée de 83,67%, ce qui est satisfaisant.

2.3.2. Projection des variables et observations dans un plan donné

Analyse des variables

```
round(auto.acp$co,3)
```

```
##       Comp1  Comp2  Comp3
## Ca    0.387 -0.840  0.119
## Mg   -0.363 -0.783  0.353
## Na   -0.897  0.007 -0.261
## K    -0.927 -0.089 -0.115
## Cl   -0.715 -0.244  0.203
## NO3   0.295  0.529  0.700
## SO4   0.545 -0.765 -0.056
## HCO3 -0.913 -0.053 -0.127
## PH    0.576  0.004 -0.705
```

On remarque que :

- La première composante traduit l'existence de fortes corrélations entre K (-0.927), HCO3 (-0.913), Na (-0.897) et Cl (-0.715). Mg ne semble ici plus être corrélée (-0.363), comme c'était le cas dans les deux précédentes analyses.
- La deuxième composante traduit l'existence de fortes corrélations entre Ca (-0.840), Mg (-0.783) et SO4 (-0.765).
- Enfin, la troisième composante traduit l'existence de corrélations négatives entre NO3 (0.700) et PH (-0.705), qui apparaissent ici parfaitement symétriques.

Seule les corrélations mettant en jeu la variable Mg sont donc ici différentes. Pour le reste, les corrélations sur chacune des différentes composantes restent inchangées.

Il semblerait également que cette configuration soit propice à une meilleure représentation de la variable PH, désormais plus proche du bord du cercle des corrélations dans les plans 1-3 et 2-3, comme on peut le voir ci-contre :

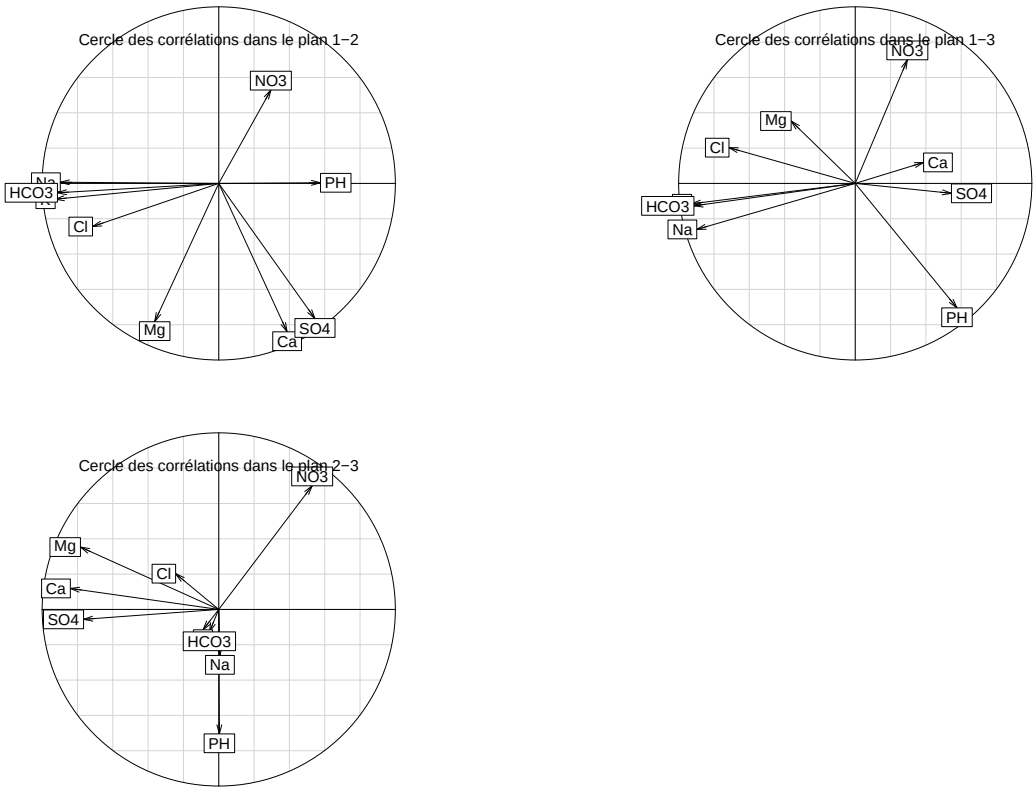


Figure 12

Interprétation des axes :

- Axe 1 : cet axe est lié aux concentrations en K, HCO3, Na et Cl. Il caractérise la teneur sodique des eaux (riches en sodium ou non) ;
- Axe 2 : cet axe est lié aux concentrations en Ca, Mg et SO4. Il caractérise la teneur minérale des eaux (fortement minéralisées ou non) ;
- Axe 3 : cet axe est lié aux concentrations en NO3. Il caractérise le pouvoir nitrique des eaux (acides ou non).

Analyse des observations

Qualité de représentation des observations :

La qualité de représentation des observations dans les différents plans est donnée par la commande `obs.inertie$row.re`.

```
round(obs.inertie$row.re,3)
```

```
##      Axis1   Axis2   Axis3
## 16  -6.872  -3.424  -0.004
## 19  50.312 -46.825   0.356
## 20  57.348 -37.842  -1.334
## 44  34.743 -60.582   1.321
## 72  41.458 -53.389  -0.374
## 63 -40.750 -26.893   7.187
## 10 -85.690  -3.347  -0.335
## 14 -33.866  -0.150  -7.523
## 15 -27.377   2.374 -49.426
## 57 -76.081   0.005  -5.865
## 62 -73.871   0.296   6.277
## 67 -71.914   0.274   7.976
## 68 -31.513  -4.749  53.049
## 75 -17.006  37.462  -6.427
## 78 -54.930   4.468 -14.999
## 1   21.014  20.546 -37.207
## 12   0.003   5.469  60.453
## 40  44.811  50.391  -0.109
## 58   8.427  34.720  48.013
## 74  36.883  49.419   2.546
## 80  75.682  12.356  -2.895
## 82  30.716  62.682  -0.031
## 65  55.113   0.480 -18.067
## 70  57.707  -1.634 -22.918
```

Plan 1-2 : les individus les mieux représentés dans le plan 1-2 sont les eaux #19 (97,14%), #44 (95,32%), #40 (95,20%), #20 (95,19%) et #82 (93,4%)

Plan 1-3 : les individus les mieux représentés dans le plan 1-3 sont les eaux #10 (86,03%), #68 (84,56%), #57 (81,95%) et #70 (80,63%).

Plan 2-3 : les individus les mieux représentés dans le plan 2-3 sont les eaux #58 (82,73%), et, dans une moindre mesure, les eaux #12 (65,92%), #82 (62,71%) et #44 (61,90%).

Contribution des observations aux axes :

De façon analogue, nous nous intéresserons ici aux observations dont la contribution est supérieure à la contribution moyenne par axe.

```
round(obs.ctr[,1:3],3)
```

```
##    Axis1 Axis2 Axis3
## 16 0.003 0.003 0.000
## 19 0.044 0.072 0.001
## 20 0.068 0.080 0.005
## 44 0.048 0.148 0.006
## 72 0.062 0.141 0.002
## 63 0.080 0.094 0.045
## 10 0.134 0.009 0.002
## 14 0.010 0.000 0.007
## 15 0.012 0.002 0.067
## 57 0.130 0.000 0.032
```

```
## 62 0.036 0.000 0.010
## 67 0.037 0.000 0.013
## 68 0.030 0.008 0.159
## 75 0.008 0.030 0.009
## 78 0.091 0.013 0.078
## 1  0.023 0.040 0.128
## 12 0.000 0.003 0.067
## 40 0.030 0.060 0.000
## 58 0.017 0.123 0.303
## 74 0.035 0.083 0.008
## 80 0.035 0.010 0.004
## 82 0.022 0.078 0.000
## 65 0.019 0.000 0.020
## 70 0.028 0.001 0.035
```

Les eaux Parot (échantillon 57), Vichy Celestins (échantillon 78), Rozana (échantillon 63), Courmayeur (échantillon 20) et Talians (échantillon 72) enregistrent les valeurs (absolues) de la première composante les plus élevées. Ces eaux sont naturellement riches en Na, K, HCO3, et Cl, plutôt acides (PH < 7) et gazeuses.

Un examen identique de la deuxième composante révèle que les eaux Talians (échantillon 72), Perrier (échantillon 58), Rozana (échantillon 63), Courmayeur (échantillon 20), Volvic (échantillon 82) ont des valeurs très supérieures à la moyenne sur cette composante. Ces eaux sont naturellement riches en Ca, Mg et SO4, plutôt basiques ou alcalines (PH > 7) et plates.

Enfin, les eaux Perrier (échantillon 58), Ste Marguerite (échantillon 68) et Abatilles (échantillon 1) enregistrent les valeurs les plus élevées de la troisième composante. Ces eaux ont la particularité d'être ou bien acides et très fortement concentrées en NO3, ou bien basique et très faiblement concentrées en NO3 (les variables PH et NO3 sont anti-corrélées).

Les analyses ici faites sont donc sensiblement les mêmes que celles que nous avions obtenues lors de nos précédentes ACP, à la différence près que l'anti-corrélation existante entre les variables NO3 et PH apparaît ici comme plus marquée. Autre point sur laquelle cette ACP diffère : la corrélation entre la variable Mg et la deuxième composante, plus marquée qu'auparavant, bien que cette différenciation ne soit pas unanime pour tous les échantillons.

Cette ACP aura donc surtout permis de mettre en évidence l'anti-corrélation très marqué entre le PH d'une eau et sa concentration en NO3.

2.3.3. Analyse des observations supplémentaires

L'idée ici est d'analyser les observations de type "outlier" ayant été soustraites des individus actifs.

En étudiant ceux qui contribuaient le plus aux axes dans le contexte de la première analyse, on observe que leurs qualités de représentation étaient très concentrées sur un seul axe. Par exemple les individus 22, 37, 38 et 84 contribuaient tous à plus de 78% au premier axe. On peut ainsi comparer avec les nouvelles valeurs, issues de l'analyse actuelle :

```
##    Axis1 Axis2 Axis3
## 22  46.1  40.1  13.8
## 37  44.2  34.9  20.9
## 38  44.1  34.9  20.9
## 84  36.2  46.4  17.3
```

On peut voir que désormais, les qualités dont bien plus réparties sur les trois axes. La forte influence de certains individus sur un axe en particulier, due à leurs valeurs extrêmes est le biais que nous souhaitions réduire. On peut donc confirmer qu'il a été absorbé.

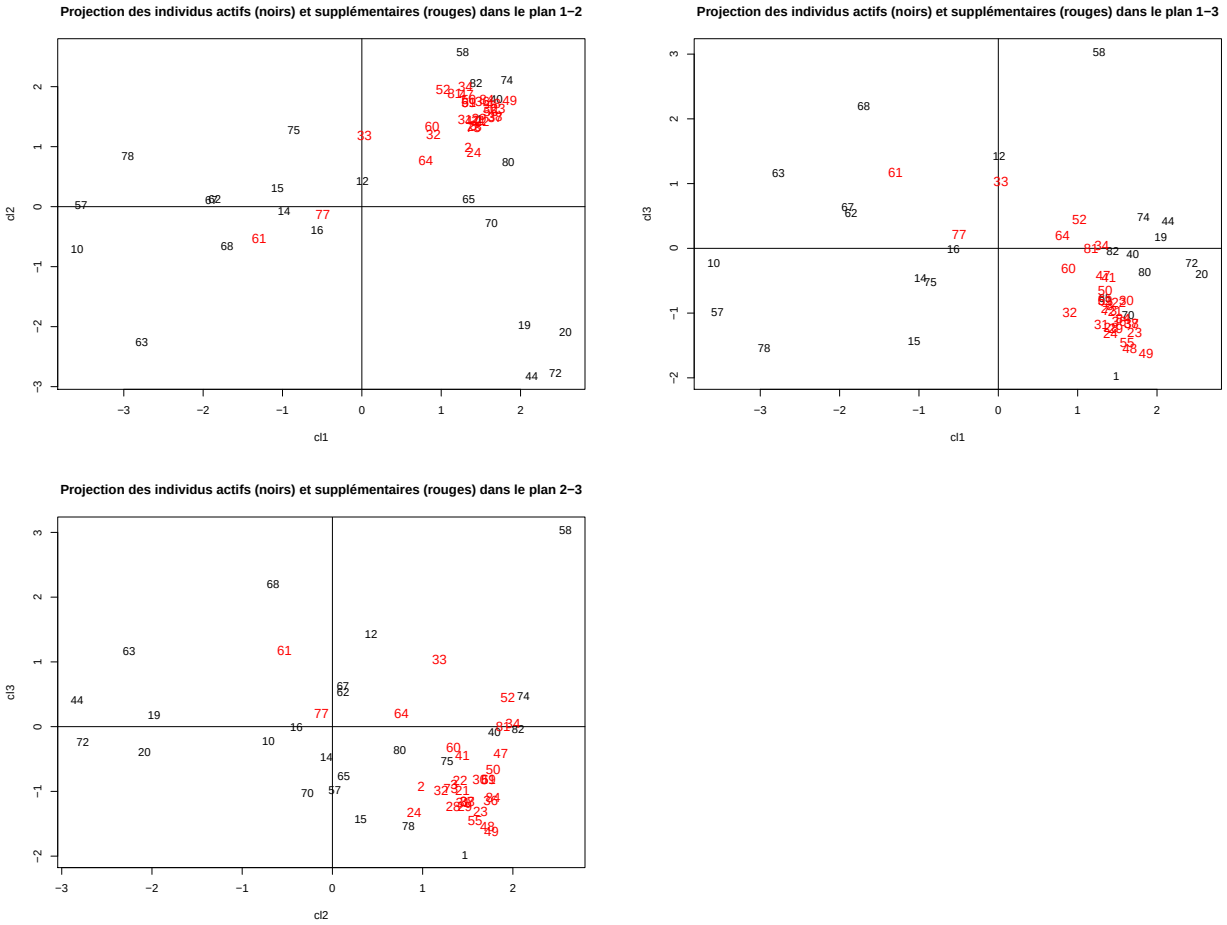


Figure 13

En représentant toutes les observations, on peut voir que les individus supplémentaires ont des positions plus équilibrées ce qui confirme cette observation.

2.4. Classification non supervisée de type k-means

Pour approfondir notre analyse, on peut choisir de classer nos individus par catégories, par classes. Ainsi, on pourrait voir apparaître différents types d'eau dans nos données.

Plusieurs méthodes s'offrent à nous. Cependant, nous n'avons pas vraiment de types d'eaux prédéfinis, ni de données d'entraînement sur lesquelles s'appuyer : le choix d'une méthode de partitionnement non supervisée semble donc raisonnable.

La fonction `kmeans(data, k)` déjà implémentée a été conçue à dessein. À partir des données d'entrée et d'un nombre de classes k, l'algorithme regroupe les individus selon leurs distances relatives, jusqu'à trouver les distances minimales.

Même si l'algorithme est déjà implémenté, il nous faut quand même trouver un k optimal. En effet, ce choix est primordial : si on prend un k trop petit, le risque est de se retrouver avec des classes peu représentatives, assez arbitraires; si le k choisi est trop grand, on peut obtenir des classes qui ne s'appliquent qu'à notre jeu de données, ou alors avoir beaucoup de classes mais avec très peu d'individus.

Un nombre classique de classes est $\sqrt{2/n}$, notre n étant le nombre d'individus (c'est-à-dire 63 après avoir filtré les données); un k optimal s'approcherait de 6. On va confirmer cette hypothèse avec la méthode suivante :

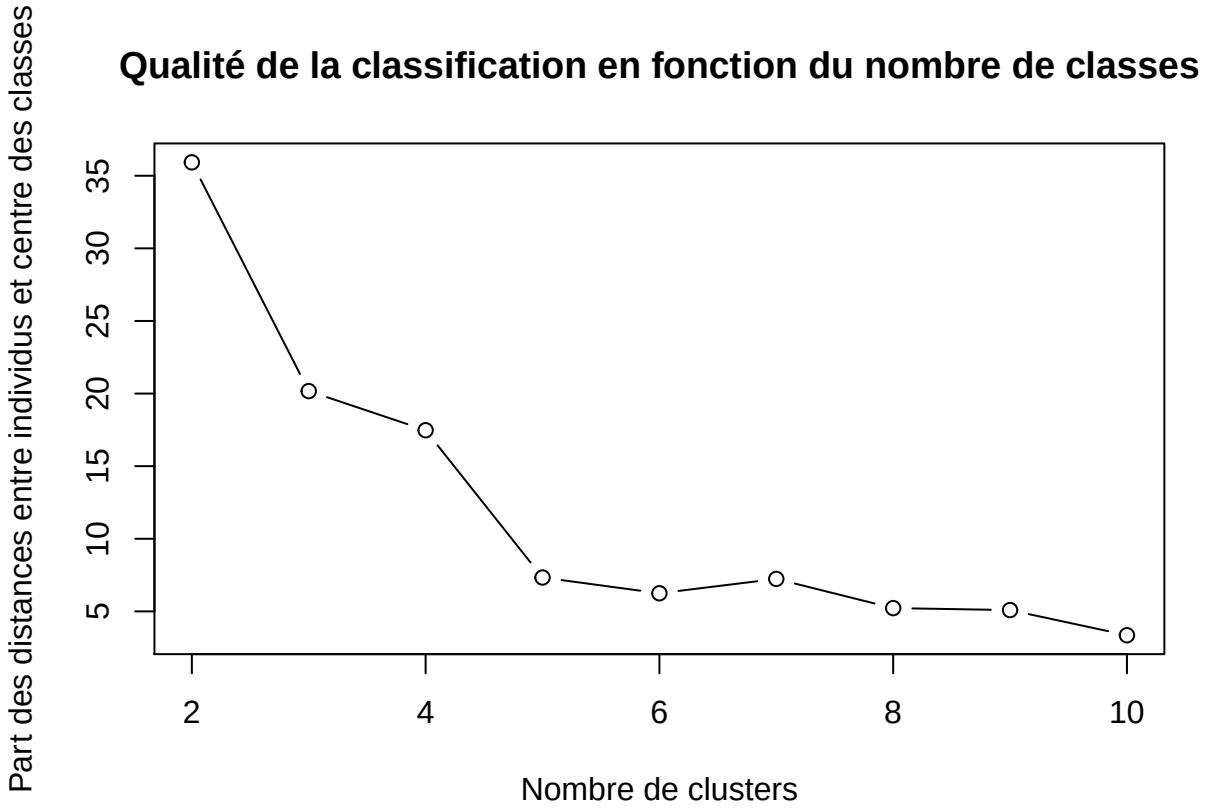


Figure 5: Figure 14

Quand on applique l'algorithme, il nous renvoie la somme des carrées des distances entre les individus et le centre des classes. Une part plus petite par rapport à la somme de toutes les distances étant meilleure, on peut dessiner le graphique de cette part, en fonction de k. En observant les décrochages de la courbe, on peut alors décider d'un k satisfaisant.

Il faut aussi préciser que la fonction kmeans incorpore de l'aléatoire. On préfère donc fixer la graine de l'aléatoire afin d'avoir toujours les mêmes résultats.

Le graphique montre qu'à k=5, la qualité de représentation ne progresse plus beaucoup; 5 semble donc être un choix raisonnable. Une fois executé, l'algorithme nous revoie un tableau des individus classés.

En appliquant les résultats de ce tableau sur une représentation issue de l'ACP, on peut observer que la classification groupe bien des individus similaires, c'est à dire proches les uns des autres même quand on réduit les dimensions. Si deux classes semblent se mélanger (bleue et rouge) sur le plan 1-2, on peut confirmer leur distinction avec une autre dimension (plan 2-3).

Ainsi, les résultats de notre analyse en composantes principales et de notre classification sont cohérents l'un par rapport à l'autre.

Conclusion

Les différentes ACP mises en oeuvres ont tout d'abord permis de dégager des groupes de variables traduisant des similarités de comportements, c'est-à-dire évoluant dans le même sens. Trois grands ensembles ont ainsi pu être mis en avant, tel que celui défini par les variables PH et NO3, semblant ici indiquer qu'une eau à

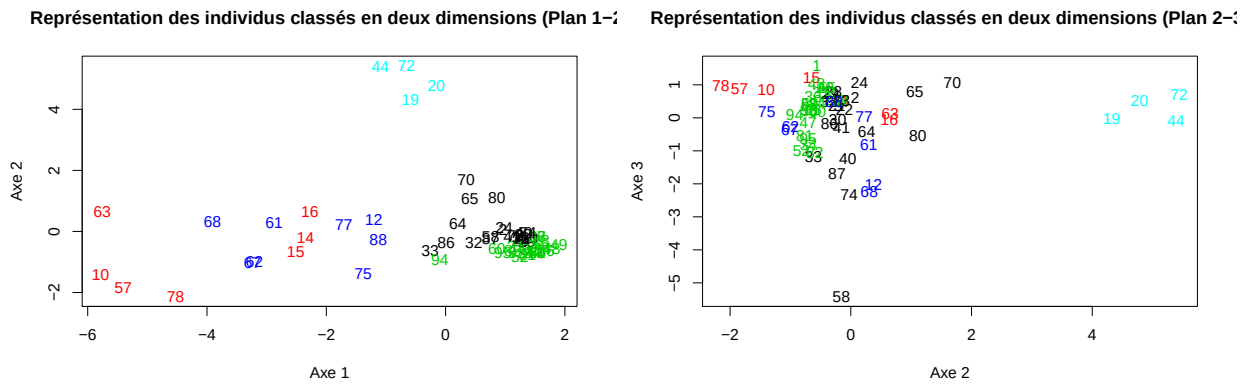


Figure 6: Figure 15

faible PH (acide) est généralement fortement concentrée en NO3 et inversement.Ces analyses ont ensuite facilité le regroupement d'observations partageant des caractéristiques communes. La première ACP, réalisée sur les données brutes, a par exemple révélé que les eaux Talians, Hépar, Courmayeur et Contrex étaient des eaux riches en calcium et en magnésium.

Cette méthode d'analyse exploratoire des données a donc entre autres permis de dresser une typologie des eaux, les différenciant sur base de leur composition minérale. Ce "typage" trouve notamment écho dans le domaine de la santé, où la consommation en eau doit être adaptée tant à l'âge qu'aux situations. Mener une ACP dans ce contexte a donc tout son intérêt. Mais la qualité du jeu de données utilisé en entrée peut conditionner et éventuellement dégrader la qualité des résultats obtenus en sortie. Neutraliser le nombre important de données manquantes et extrêmes est donc un travail préliminaire essentiel visant en partie à répondre à cet aléa, bien qu'il ne soit jamais possible de s'en affranchir complètement.

Code source

ACP "automatique"

ACP "manuelle"

Classification non supervisée (K-Means)