UNIVERSITÉ DE NANTES

Faculté des sciences et techniques

Master Ingénierie Statistique

Etude de l'eaux en France et au Maroc en 2018

Auteurs :

Victor Vasse

Aymeric Lechevranton

Unité d'enseignement :

Classification non-supervisée

Référent :

L. Bellanger

19 décembre 2018

Table des matières

1 Introduction

Malgrès le fait que l'analyse graphique puisse induire une intuition à propos des groupes présents dans un jeu de données, cette dernière n'est pas suffisante dès que l'intuition relève de la subjectivité. De plus, la taille du jeu de données peut représenter une limite quant à l'interprétation des résultats sans méthode statistique. En effet, des outils tel que K-means ou la classification hiérarchique ont été développé au cours du XXè siècle. La miniaturisation des ordinateurs associée à des puissances de calcul toujours plus élevées permettent la mise en oeuvre de ces algorithmes pouvant être couteux, aussi bien en terme d'espace mémoire que de capacité de calculs.

L'objectif de ce rapport est de mettre en oeuvre les différentes procédures de classification sur le jeu de données Eaux2018. Les variables de ce dernier représentent principalement des caractéristiques chimiques des eaux étudiées pour le Maroc et la France, quelles soient gazeuzes ou plates. Les minéraux sont effectivement commun à la composition chimique de beaucoup d'eau. En effet, l'eau contient des ions dissous comme notamment le calcium (Ca++), le magnésium (Mg++), le sodium (Na+), le potassium (K+), les carbonates (CO3−), les bicarbonates (HCO3-), les sulfates (SO4−), les chlorures (Cl-) et les nitrates (NO3-). Ils proviennent pour l'essentiel du lessivage des sols par les eaux de pluie. En outre, la teneur d'une eau dépend directement de la nature des roches du bassin versant, c'est-à-dire des roches que l'eau rencontre quand elle converge vers l'exutoire (cours d'eau, lac, mer, etc).[1]

L'objectif principal de la classification est de construire des groupes regroupant des objets (variables ou individus) de sorte que deux objets d'un même groupe se ressemblent le plus possible mais au contraire, deux objets de groupes distincts diffèrent le plus possible. Dans cet objectif, nous allons dans un premier temps présenter le jeu de données brutes et celui utilisé pour la méthode de classification. Dans un second temps, nous tenterons d'effectuer une analyse exhaustive des méthodes de classification applicables sur notre de jeu de données. Dans un troisième temps, nous étudierons la validité externe afin de décider d'une méthode de classification qui se révèlera être la plus cohérente sur Eaux2018. Dans un dernier temps, nous introduirons les méthodes de classification supervisées en mettant en oeuvre tout d'abord les k plus proches voisins (knn), ensuite l'analyse factorielle et enfin les arbres de décision. Nous calculerons également les taux d'erreur a ociés à ces méthodes de classifications supervisées.

2 Présentatisson du jeu de données

2.1 Jeu de données brutes

Dans cette partie nous allons présenter les élements principaux de notre jeu de données sans modification.

Table 1 – Données brutes

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max	N/A
Ca	95	111.130	125.498	2	20.9	146.2	596	0
Mg	95	31.918	41.054	0	4.3	45	243	0
Na	95	151.356	334.302	1	3.7	111.5	1,945	0
K	92	16.824	34.217	0	1	12.5	192.2	3
Cl	91	59.54	111.916	0.6	4.4	39.1	649	4
NO3	76	2.689	3.452	0	1	3	19	19
SO4	93	108.691	287.776	0.2	6	60	1530	2
HCO3	93	691.446	1058.45	2.4	121	820	6722.2	2
PH	76	6.905	0.663	5.2	6.5	7.4	8.2	19

Nous remarquons dans un premier temps que certaines variables possèdent des valeurs manquantes (N/A). De plus, les écart types associés aux différentes variables présentent des différences d'échelle. Ainsi, la réduction pourrait être utile. De plus nous pouvons observer un résumé pour les variables qualitatives. A noter que R ne considère pas les N/A comme étant une classe.

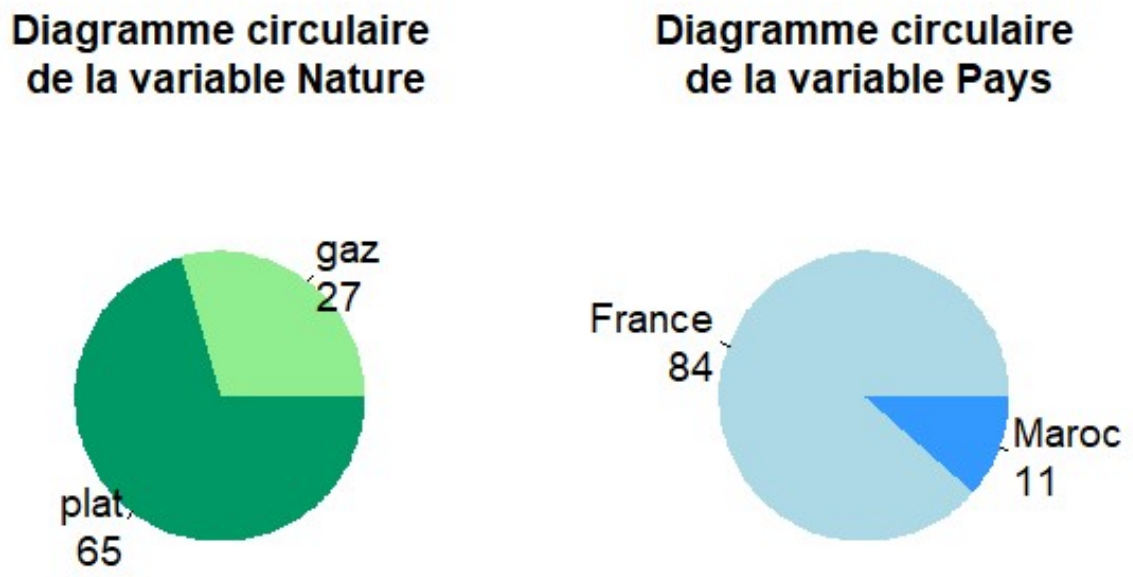


Figure 1 – Diagrammes circulaires des variables qualitatives

2.2 Jeu de données utilisé pour les méthodes de classification

Dans cette partie, nous présenterons de manière bref notre jeu de données utilisées pour les méthodes de classification. La stratégie étant de supprimer toutes les lignes pour lesquelles il y a des données manquantes puis de procéder à un centrage et une réduction.

Nous remarquons dans un premier temps que les variables ne possèdent plus de valeurs

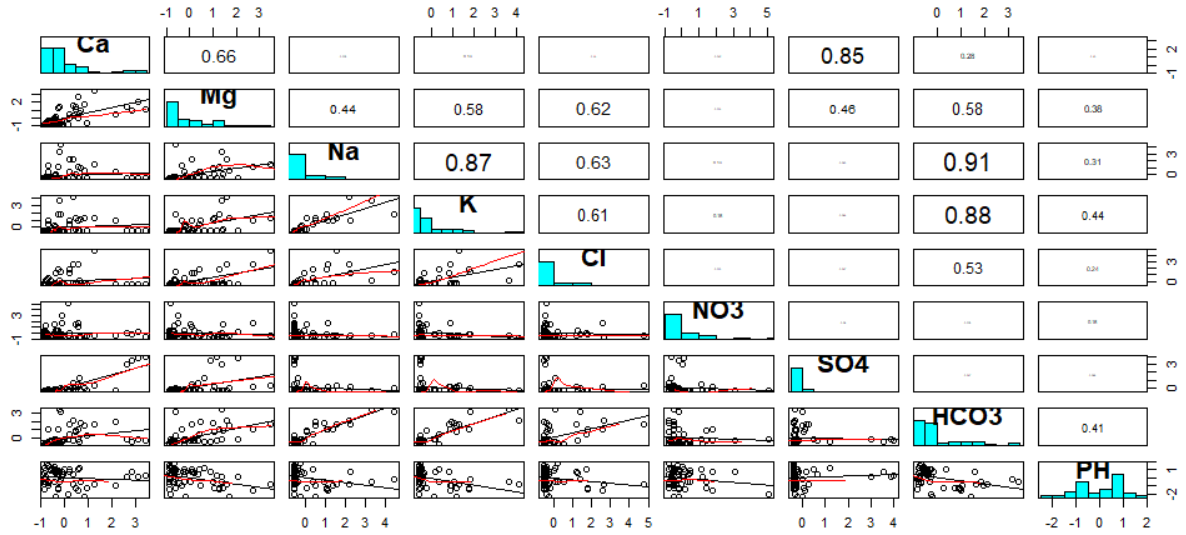


FIGURE 2 – Graphique analyse bivariée des données

TABLE 2 – Données centrées réduites, utilisées pour les méthodes de classification

Statistic	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max	NA
Ca	62	0	1	−0.858	−0.693	0.315	3.41	0
Mg	62	0	1	−0.892	−0.744	0.466	3.449	0
Na	62	0	1	−0.546	−0.53	0.058	4.445	0
K	62	0	1	−0.577	−0.548	0.169	4.139	0
Cl	62	0	1	−0.535	−0.506	−0.215	4.796	0
NO3	62	0	1	−0.790	−0.469	0.101	5.086	0
SO4	62	0	1	−0.413	−0.390	−0.240	4.007	0
HCO3	62	0	1	−0.792	−0.648	0.182	3.460	0
PH	62	0	1	−2.406	−0.676	0.808	1.832	0

manquantes (N/A). De plus nous pouvons observer un résumé pour les variables qualitatives. A noter que R ne considère pas les N/A comme étant une classe.

En appliquant notre stratégie quant à la modification de notre jeu de données, nous obtenons des dimensions restreintes : 62x9. Nous perdons donc 33 lignes vis-à-vis de la table de départ.

3 Existence de Cluster

Cette partie aura pour prinpical objectif de conclure de l'existence de groupes au sein de jeu de données Eaux2018. Dans un premier temps nous allons effectuer le test de Hopkins qui tend à déterminer de manière statistique l'existence d'une partition et dans un second temps tenter de visualiser le jeu de données afin d'obtenir davantage d'informations pouvant guider notre décision.

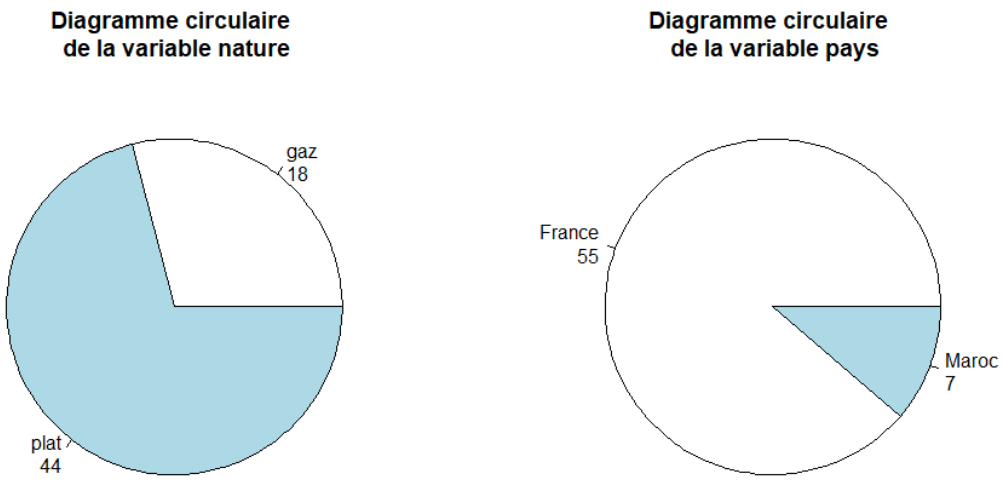


FIGURE 3 – Diagrammes circulaires des variables qualitatives

3.1 Statistique de Hopkins et algorithme VAT

Sous l'hypothèse nulle, l'échantillon est distribué selon une loi uniforme, impliquant la non-existence de groupe d'après la statistique de Hopkins. Ainsi, par construction du test, on rejète H_0 si la statistique de Hopkins est proche de 0 (en particulier nettement inférieure à 0.5). Dans ce cas, nous pouvons conclure de l'existence de classes dans notre jeu de données. La valeur exacte du test est 0.1655068. Nous rejetons donc l'hypothèse nulle selon laquelle l'échantillon serait uniformément distribué. D'après la statistique d'Hopkins, nous concluons de l'existence de classes.

De plus, l'algorithme VAT nous permet de distinguer trois possibles classes au sein du jeu de données, ce qui vient renforcer l'aspect "clusterisable" de celui-ci.

3.2 Analyse graphique

Lors de l'analyse du jeu de données, nous avons remarqué qu'il y avait deux types de nature : gaz et plat. De plus, les eaux proviennent du Maroc et de la France. Or, si en France les deux types d'eaux sont présentes, au Maroc, l'échantillon ne fournit pas d'information quant aux eaux gazeuzes. Ainsi, de manière heuristique nous pouvons construire trois groupes, à savoir : Eaux plates Françaises (FP), eaux gazeuzes françaises (FG) et eaux plates marocaines (MP). Il s'agira ainsi de représenter les individus dans le plan factoriel 1/2 grâce à une ACP sur les données "na.omit" contenant 62 observations pour 9 variables. Grâce à la ggplot2, nous avons contruit un graphique contenant l'essentiel des informations.

Grâce à ce graphique, nous constatons que l'apparition de 3 groupes distincts n'est pas

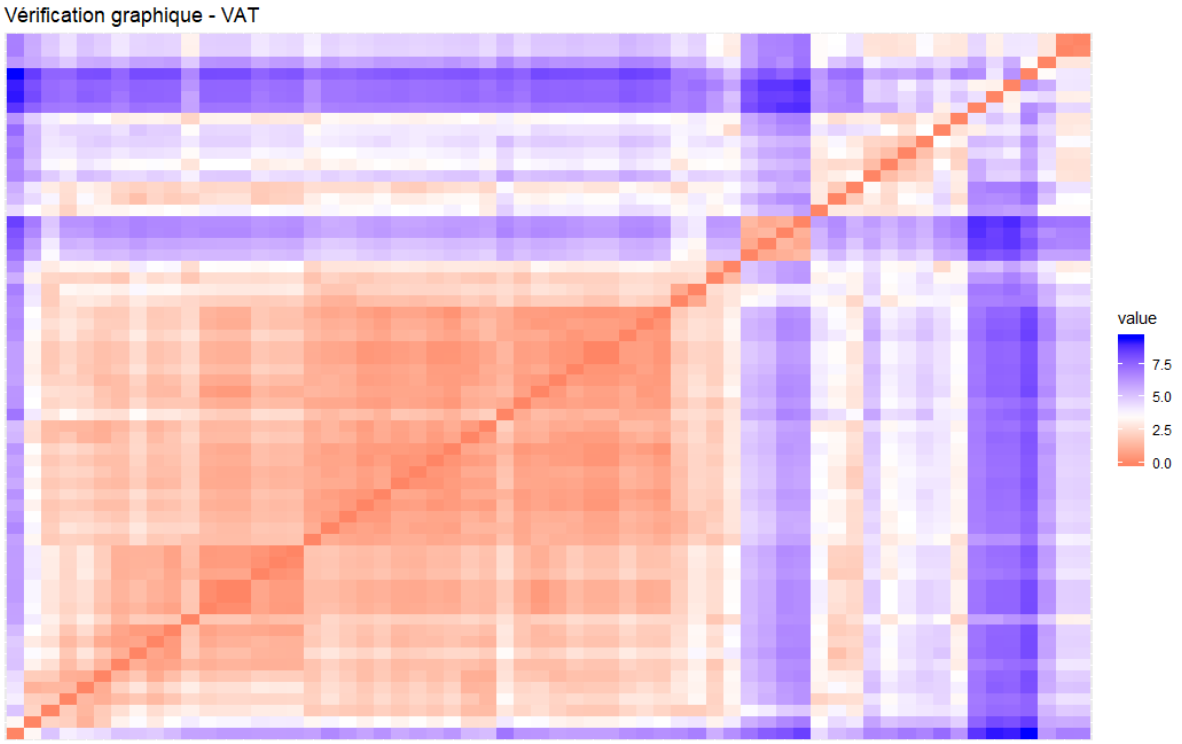


FIGURE 4 – Algorithme VAT

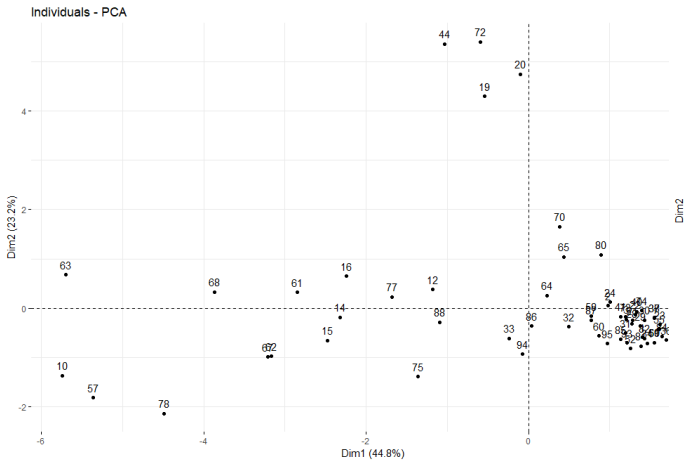


FIGURE 5 – Représentation individus plan factoriel 1/2

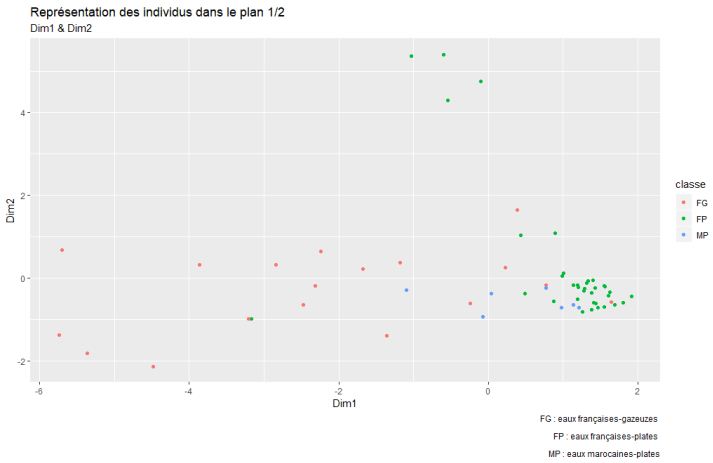


FIGURE 6 – Représentation individus plan factoriel 1/2

évidente. En effet, on remarque une séparation plus nette entre les eaux plates, quelles soient Marocaines ou Françaises, et les gazeuzes. De plus, nous notons l'apparition d'individus particulier, dans le sens où ils apparaissent clairement dans un autre groupe.

4 Méthodes de classification non supervisée

4.1 Mesures de sélection du nombre optimal de K

Le graphique des silouhettes, développée en 1987 par Rousseeuw, est un moyen graphique de sélectionner le nombre de classes présent dans notre jeu de donné. La silouette pour l'observation i est définie comme suit : S_i = $\frac{b_i - a_i}{max(a_i, b_i)}$, avec,

- a_i est la distance moyenne entre i et toutes les observations de sa classe C_k,
- b_i = $\min_{C_j \neq C_k}(d(i, C_j))$

L'objectif est alors de trouver le k pour lequel la moyenne des silhouettes est maximale.

La méthode du coude est un autre moyen graphique de sélectionner le nombre de classes présent dans notre jeu de données. Dans un premier temps, cela consiste à calculer la variabilité intra classe totale puis dans un second temps de la représenter sur un graphique en fonction du nombre de groupes (k). Enfin, dans un dernier temps, choisir le k pour lequel la pente de la courbe diminue.

La méthode de la statistique de GAP, présente en Figure 6, a été défini par R. Tibshirani, G. Walther et T. Hastie en 2001 et s'applique sur toutes les méthodes de classification. Intuitivement, si nous définissons I_w comme dans le cours, nous pouvons définir pour chaque nombre de k, sa valeur moyenne. Par positivité de la variance (stricte dans notre cas), nous pouvons définir le logarithme de I_w. Dans un second il s'agira de définir l'espérance : $E_n[log(I_w)]$ de notre échantillon de taille n. Nous considérons ici la distribution empirique. Alors, la statistique de GAP est simplement définie comme suit [4] :

$$Gap_n(k) = E_n[log(I_w)] - log(I_w)$$

Empiriquement, $E_n[log(I_w)]$ est estimé à partir d'une moyenne des $log(I_w)$ calculée sur B sous-échantillons simulé à partir de la technique de MonteCarlo. Posons sd(k) l'écart type associé aux B $log(I_w)$. De plus, s_k = $sd_k\sqrt{1 + 1/B}$. Il s'agira, pour finir, de choisir le nombre de cluster via : $\widehat{k}$ est égal au plus petit k tel que :

$$\text{GAP(k)} \geq GAP(k+1) - \text{s}_{k+1}$$

Pour finir, nous pouvons noter que nous attribuons une importance équivalente à toutes les méthodes. En ce sens, l'objectif sera de déterminer le nombre de groupe optimal via une règle de majorité, autrement dit, si deux indicateurs fournissent un nombre optimal de groupes : n, alors nous choisiront $\widehat{k} = n$.

4.2 Classification par partition

K-means

La méthode "Kmeans" a été introduite par Hartigan et Wong en 1979. Son principe est relativement simple. Il s'agit dans un premier temps de choisir les k premiers centres initiaux par tirage pseudo aléatoire, dans un second temps d'affecter chaque objet au centre le plus proche (en ayant défini une distance au préalable) puis dans un dernier temps, recalculer les centres initiaux de chaque classes et itérer le processus jusqu'à stabilisation de l'algorithme. Nous allons donc appliquer cet algorithme à notre jeu de données Eaux2018.

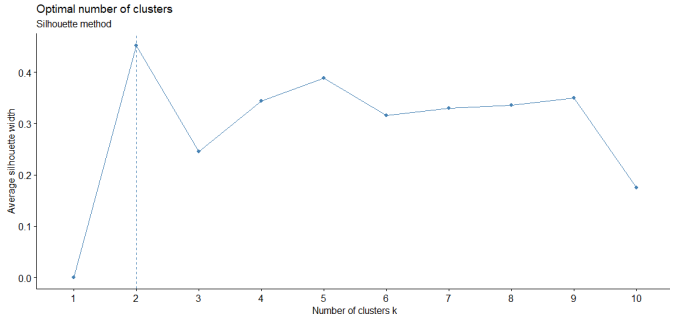


FIGURE 7 – Sélection du nombre de groupes, méthode silouhette

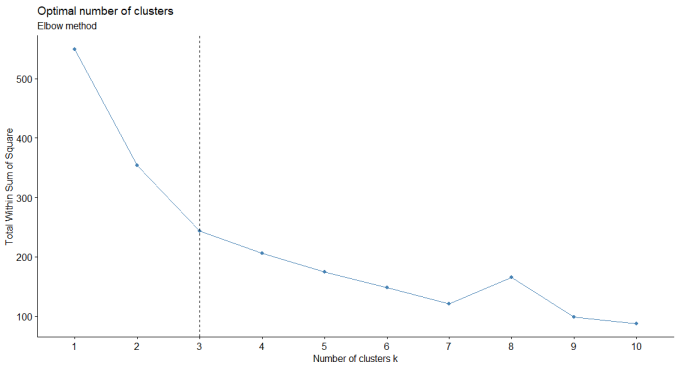


FIGURE 8 – Sélection du nombre de groupes, méthode du coude

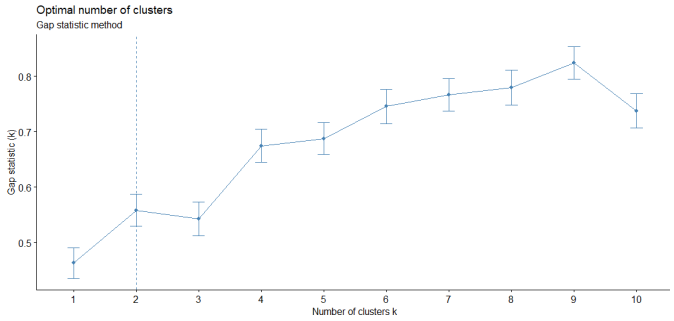


FIGURE 9 – Sélection du nombre de groupes, critère GAP

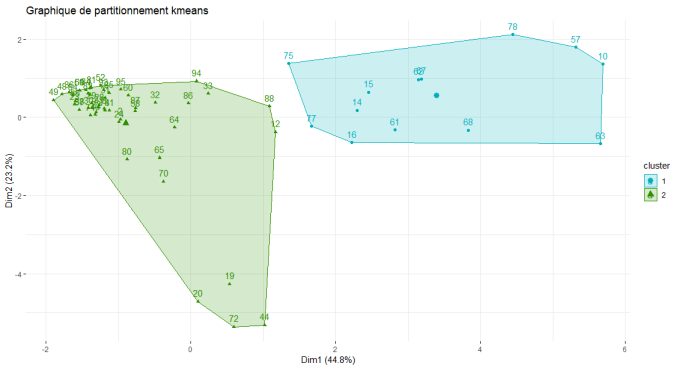


FIGURE 10 – Représentation des groupes via Kmeans

Ainsi, nous constatons qu'avec Kmeans, le nombre optimal de groupes selon la mesure de la silouhette est 2, avec une valeur moyenne de silouhette d'environ 0.45.

Sur la Figure 5, nous constatons l'évolution de la variabilité intra-classe totale. L'objectif étant d'observer l'emplacement des fortes réductions sur la courbe, nous constatons que le nombre optimal de groupes est également 2.

Sur la figure 6, nous observons la statistique de GAP le critère étant défini ci-dessus, le k optimal est 2. Néanmoins, nous constatons que certaines observations (partie basse de la figure 8) sont "extrême", d'où l'idée d'appliquer les K-médoïdes.

K-médoïdes : PAM

Dans le premier projet effectué en analyse de données, nous avons effectué une analyse univariée et bivariée permettant de mettre en avant des individus qui sont extrêmes. En ce sens, la méthode des K-médoïdes peut-être cohérente. De plus, étant donné la volumétrie de notre jeu de données, la complexité algorithmique n'est pas un problème.

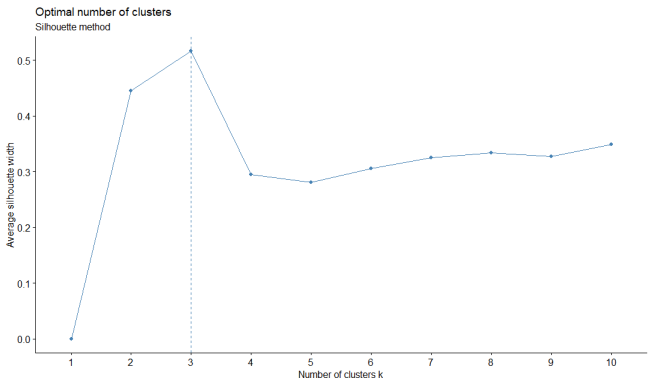


FIGURE 11 – Sélection du nombre de groupes, méthode silouhette

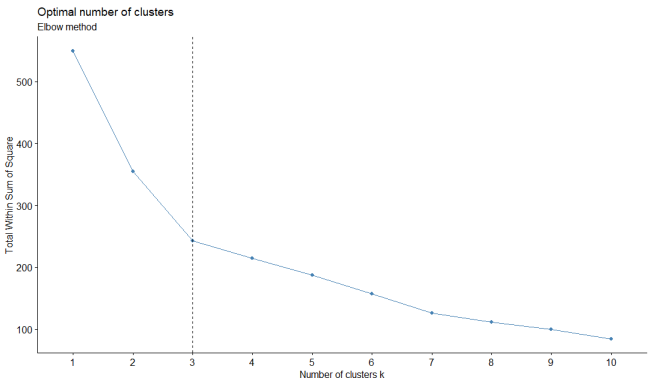


FIGURE 12 – Sélection du nombre de groupes, méthode du coude

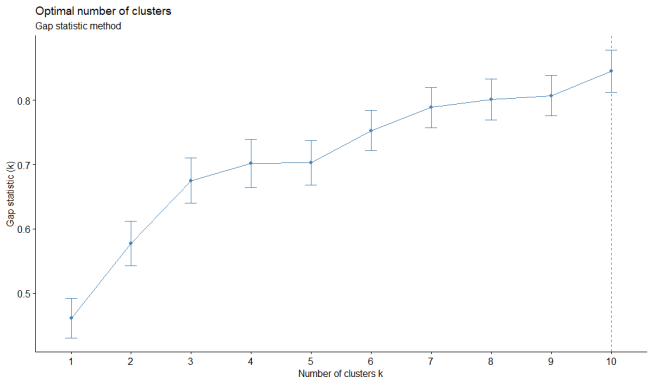


FIGURE 13 – Sélection du nombre de groupes, critère GAP

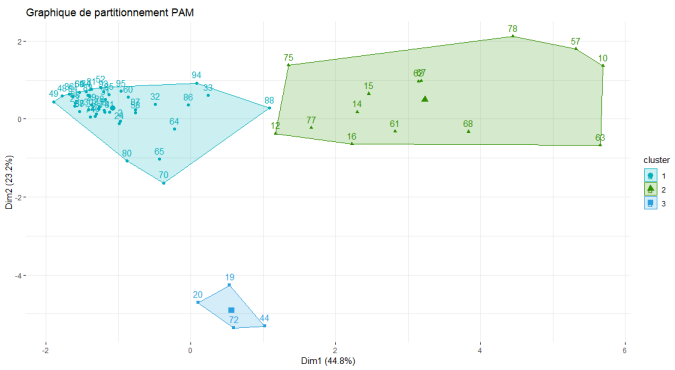


FIGURE 14 – Représentation des groupes via PAM

La premier élément remarquable est que la méthode de la statistique de GAP renvoie un nombre optimal de cluster de 10. En invoquant le critère de l'interprétabilité, nous pouvons l'ignorer. En effet, lors de la représentation du jeu de données dans un plan factoriel de dimension 2, nous constatons que 10 groupes n'est pas un choix juficieux. De plus, le second élément est que la méthode de silouhette fournit les mêmes résultats que la méthode Elbow. Nous pouvons ainsi représenter les résultats de la classification PAM sur notre plan factoriel. L'avantage de l'algorithme PAM est vérifié. En effet, ce dernier est moins sensible aux individus extrêmes dans le sens ou chaque observation qui possède des caractéristiques "hors-normes" est classée dans un groupe distinct, ne déplaçant pas le centre de gravité des autres classes de manière exagérée.

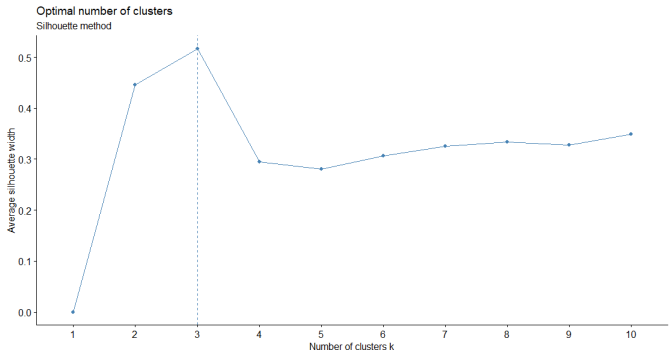


FIGURE 15 – Sélection du nombre de groupes, méthode silouhette

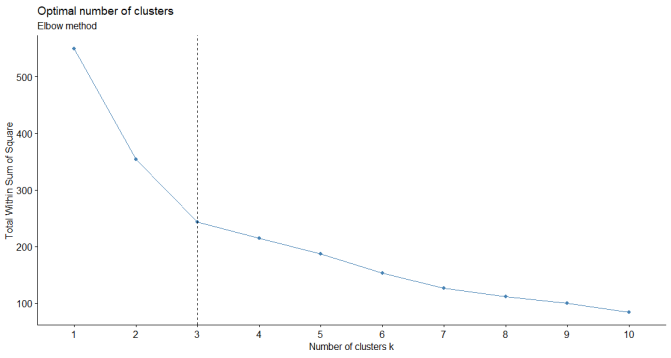


FIGURE 16 – Sélection du nombre de groupes, méthode du coude

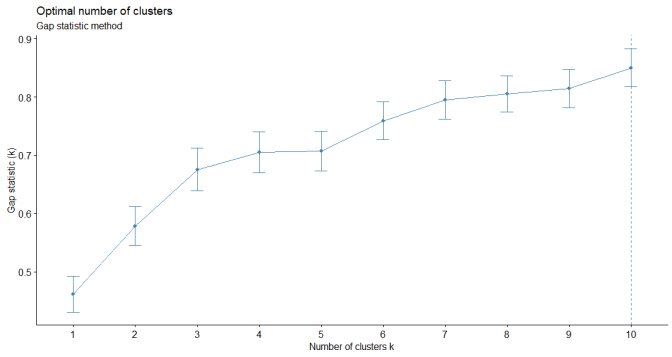


FIGURE 17 – Sélection du nombre de groupes, critère GAP

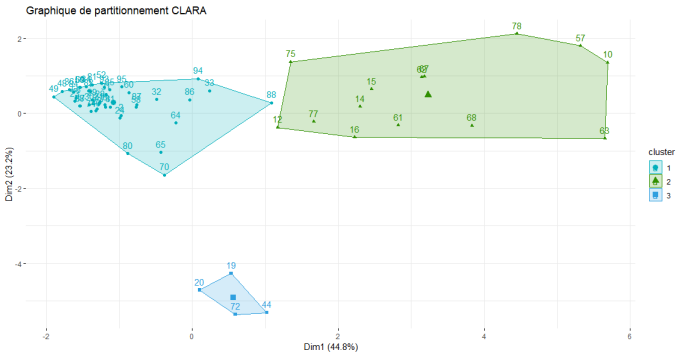


FIGURE 18 – Représentation des groupes via CLARA

K-médoïdes : CLARA

Comme dans le cas précédent, nous retrouvons les mêmes résultats. En effet, CLARA n'est qu'une succession d'exécution de PAM. Ainsi, pour un échantillon de petite taille, il est probable que la solution soit semblable.

Classification floue

L'idée de cette méthode de classification non-supervisée est de répondre à la question suivante : exite-t-il des observations pour lesquelles la classification est litigieuse ? Ainsi, on peut définir un degré d'appartenance à une classe afin de créer des classes "empiétantes".

Par la règle d'arbitrage définie plus tôt, nous choisissons k=2. En effet, la méthode des silouhettes et du coude nous donnent un k optimal de 2. Nous pouvons donc représenter ces deux groupes dans le plan factoriel 1/2. De plus, afin d'avoir une idée de l'appartenance des observations à un groupe, nous pouvons calculer le coefficient de partition de Dunn via : fanny.res$coeff (ou : mean(apply(fanny.res$membership$\hat{2}$, 1, sum))). Ce dernier est égal à 0.54823017. Via le calcul, on s'aperçoit que lorsque ce coefficient tend vers 1 cela signifie que toutes les observations sont dans une unique classe et inversement. Ainsi, au voisinage de 0.5 (1/k), on peut conclure que l'appartenance n'est pas nette malgrè une visualisation graphique relativement claire. Pertinence de cette méthode.

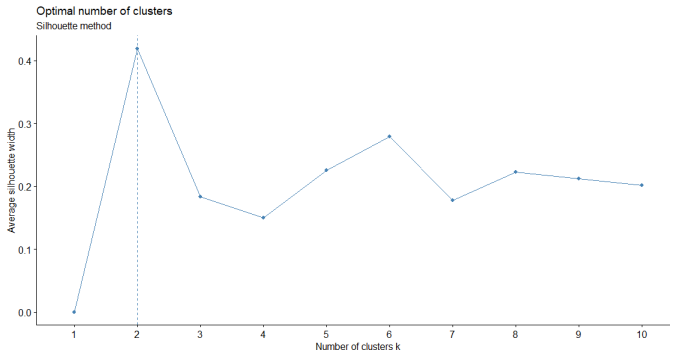


FIGURE 19 – Sélection du nombre de groupes, méthode silouhette

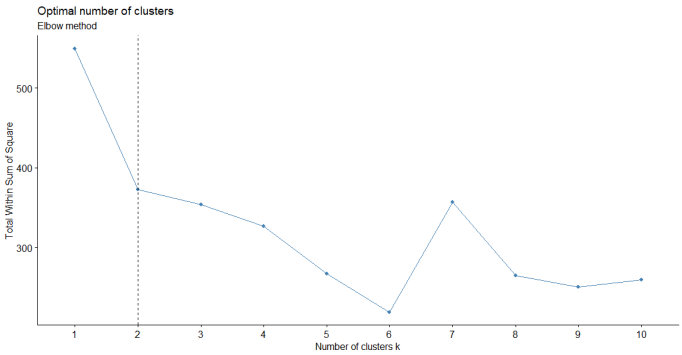


FIGURE 20 – Sélection du nombre de groupes, méthode du coude

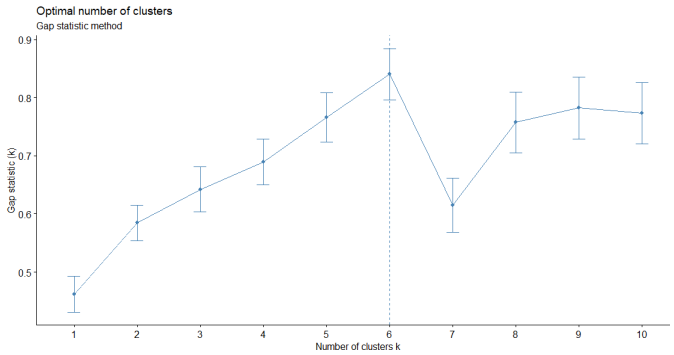


FIGURE 21 – Sélection du nombre de groupes, critère GAP

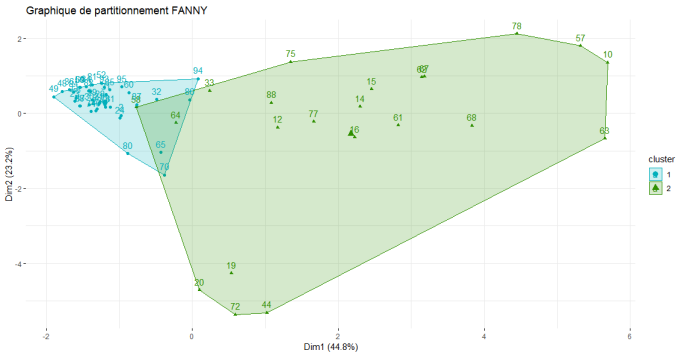


FIGURE 22 – Représentation des groupes via FANNY

4.3 Classification hiérarchique

Les méthodes de classification hiérarchiques aboutissent à la construction d'un dendrogramme. Ce graphique permet de représenter l'ensemble des individus de manière indicé. Le but sera alors de trouver la valeur de cet indice afin de couper l'arbre pour faire tomber les feuilles. Ainsi, dans un premier temps, l'algorithme CAH (resp. CDH) considère chaque observation comme étant une classe distinctes, soit n classes. Dans un second temps, il s'agit de trouver les deux individus qui sont le plus proches (resp. éloignés), dans un dernier temps itérer le processus sur les n-1 individus.

Classification ascendante hiérarchique

En classification ascendante hiérarchique (CAH), il existe différentes méthodes de regroupement des observations :

- Single linkage : distance minimale entre deux groupes.
- Complete linkage : distance maximale entre deux groupes.
- Average linkage : valeur moyenne des distances entre deux groupes.
- Mcquitty : la moyenne des moyennes contenant à la base les moyennes des distances entre deux groupes.

— Ward : perte d'inertie minimale/distance au carré entre les barycentres de deux groupes et pondérés par leur effectif respectif.

— Median : valeur médiane des distances entre deux groupes.

— Centroïde : distance entre les barycentres de deux groupes.

Afin de réaliser une CAH sur notre jeu de données (62x9), nous avons sélectionné une de ces méthodes par rapport aux résultats que nous avons obtenus par les indices de la validation interne (corrélation cophénétique, indice de Dunn, Silhouette et connectivité) et ceci pour des coupes de la CAH de 2 à 9 clusters. Le tableau 23 présente ces résultats sous la forme d'un tableau à gains : pour chaque indice calculé, un rang a été effectué parmi les méthodes avec 6 points attribué à la meilleure méthode et 1 point pour la méthode qui obtenait le moins bon résultat pour l'indice en question.

	average	single	complete	ward.D	ward.D2	mcquitty
k=2	17	20	18	5	5	14
3	13	17	16	10	9	13
4	16	16	18	3	8	17
5	17	16	17	4	7	17
6	15	18	16	4	7	18
7	18	21	9	5	7	18
8	19	21	9	5	7	18
9	20	22	11	5	6	14

FIGURE 23 – Score obtenu pour chaque méthode et pour un nombre de groupes donné

Par rapport à ce tableau, nous avons choisie de sélectionner la méthode d'agrégation "complete" qui nous permet d'obtenir de bons score tout en ayant un nombre de cluster petit, ce qui nous permettra de pouvoir interpréter les résultats plus facilement.

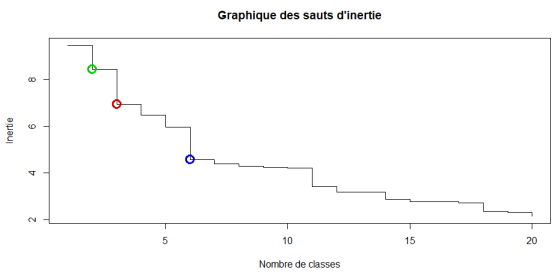


FIGURE 24 – Sélection du nombre de groupes, méthode sauts d'inertie

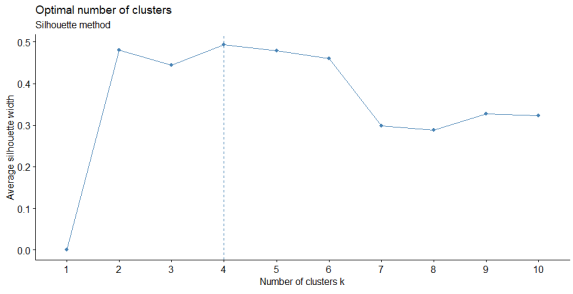


FIGURE 25 – Sélection du nombre de groupes, méthode des silouhettes

En considérant le même critère de sélection, nous pouvons sélectionner k=4. En effet, la méthode de sauts d'inertie, des silouhettes et du coude fournissent des conclusions similaires. Nous pouvons donc représenter le dendogramme ainsi que le nuage de point avec 4 groupes.

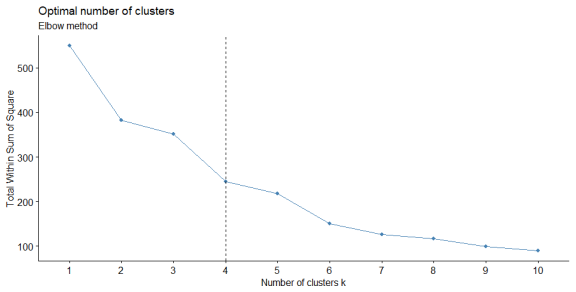


FIGURE 26 – Sélection du nombre de groupes, méthode du coude

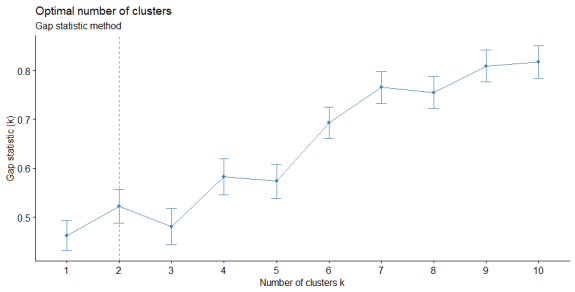


FIGURE 27 – Sélection du nombre de groupes, critère de GAP

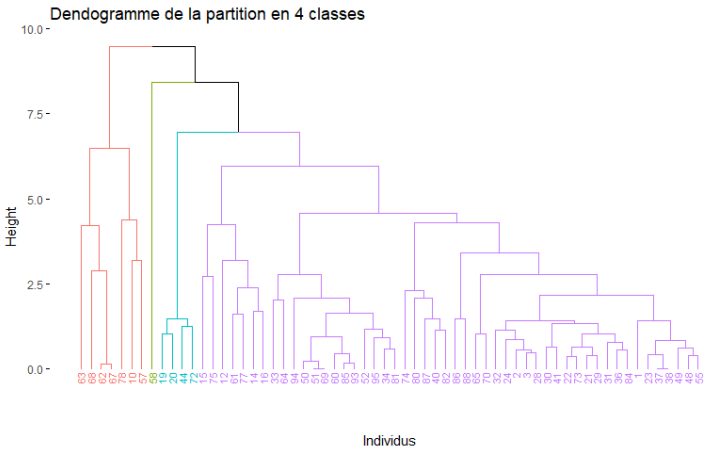


FIGURE 28 – Représentation du jeu de données via le dendogramme

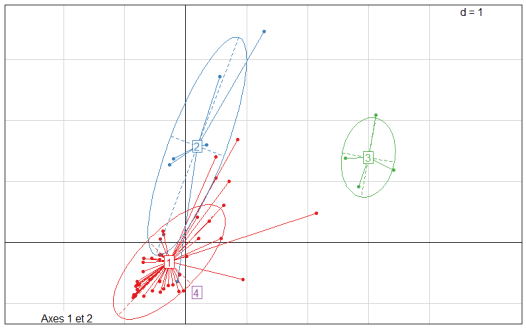


FIGURE 29 – Représentation du jeu de données via le plan factoriel 1/2

Nous constatons que l'algorithme a détecté un groupe d'un seul individus, posant ainsi des questions quant à la robustesse de cette méthode. De plus, nous pouvons appliquer des méthodes de classification mixtes, autrement dit, en appliquant une classification type k-means sur le jeu de données cela permet de réduire le temps de calcul de la CAH pour une grosse volumétrie et donc potentiellement beaucoup de groupes. Dans notre cas, cela n'est pas utile car la volumétrie du jeu de données est faible : 6 ko. Ainsi, l'exécution de CAH s'effectue en moins d'une seconde.

Classification descendante hiérarchique

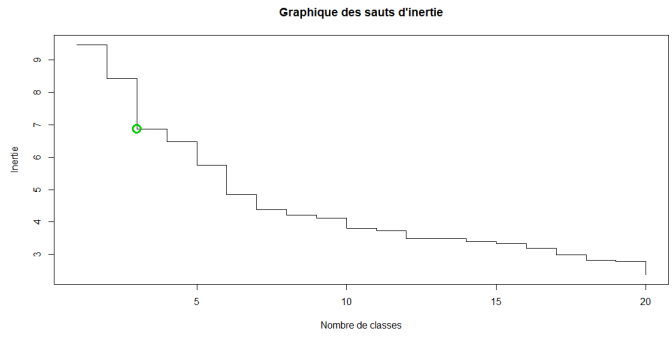


FIGURE 30 – Sélection du nombre de groupes, méthode sauts d'inertie

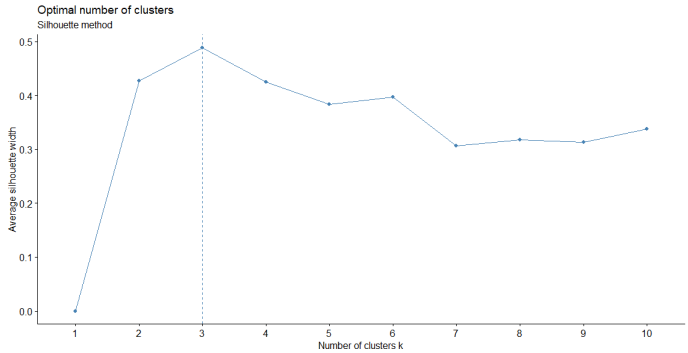


FIGURE 31 – Sélection du nombre de groupes, méthode des silouhettes

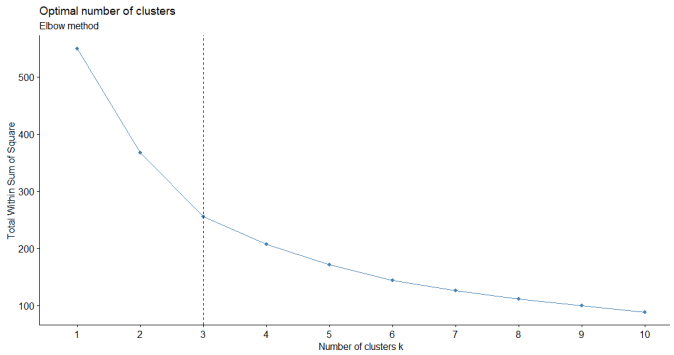


FIGURE 32 – Sélection du nombre de groupes, méthode du coude

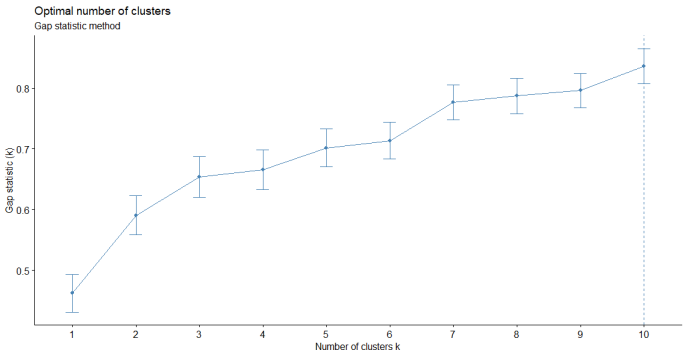


FIGURE 33 – Sélection du nombre de groupes, critère de GAP

Nous constatons un nombre de groupe optimal de 3. En effet, lié à la règle de sélection, 3 mesures parmis 4 fournissent ces conclusions. Nous pouvons donc représenter ces groupes sur le plan factoriel 1/2 :

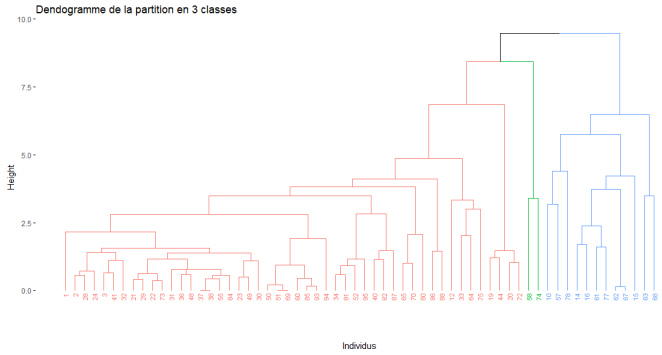


FIGURE 34 – Représentation du jeu de données via le dendogramme

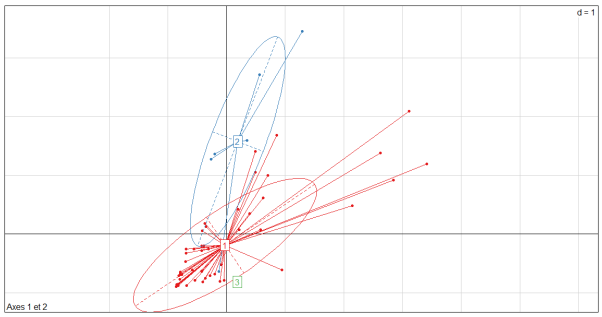


FIGURE 35 – Représentation du jeu de données via le plan factoriel 1/2

4.4 Mesure de validité interne et externe

Les validités internes et externes permettront de choisir l'algorithme le plus efficace afin de créer des groupes pour Eaux2018. Pour ce faire, nous allons utiliser les outils suivant :

Validité interne

- Indice de Dunn : Relativement simple à calculer, cet indice calcule le rapport entre la distance minimale qui sépare deux observations classées dans 2 groupes différents et la distance maximale qui sépare deux individus classés dans le même groupe. En quelque sorte, cet indice est une mesure de compacité des groupes. Ainsi, nous pouvons noter que, par construction, les algorithmes hiérarchiques fourniront un indice élevé dès lors que nous utilisons la méthode 'complete' sous r, favorable à la compacité des classes.
- Indice des silouhettes : Cet indice mesure la similarité des observations qui sont regroupées dans un groupe. En effet, plus cet indice sera proche de 1, plus les individus seront classés

correctement. De plus, au voisinage de 0, l'observation est dite "entre deux classes" et, au voisinage de 1, elle est dite "mal classéeé.

— Indice de connectivité : Par construction, cet indice est croissant avec le nombre de groupes. En effet, lorsque deux observations sont classées dans le même groupe, la distance qui les sépare est réduite à 0. Ainsi, pour les algorithmes ou on a choisi k=2, l'indice de connectivité sera mécaniquement réduit.

— Corrélation cophénétique : Cet indice donne un niveau de qualité d'une classification hiérarchique et se base sur la moyenne des distances euclidiennes entre deux points, la distance dendogrammatique de ces deux points et la moyenne de cette dernière distance :

$$c = \frac{\sum_{i<j}(x(i,j) - \bar{x})(t(i,j) - \bar{t})}{\sqrt{[\sum_{i<j}(x(i,j) - \bar{x})^2][\sum_{i<j}(t(i,j) - \bar{t})^2]}}.$$

Plus la valeur obtenue est proche de 1 (>0.75), plus la classification est "bonne".

Validité externe

— Rand ajusté : "Ajusté" dans le sens où il permet de comparer des méthodes entre différents groupes. Cet indice mesure le pourcentage de concordance entre 2 partitions.

— Kappa de Cohen : Cet indice évalue l'accord/la concordance entre partition.

— Meila's variation information [5]

Validité interne

Dans un premier temps, nous remarquons ci-dessus que l'algorithme de classification floue : FANNY admet relativement beaucoup de silouhettes négatives. Ces dernières seront donc potentiellement mal classées, ou auront tendance à appartenir à 2 groupes (caractéristique de l'algorithme FANNY). Dans un second temps, les graphiques représentant la présence de 3 groupes fournissent des conclusions encourageantes. En effet, on constate que lors de l'apparition d'un troisième groupe (en gris) les silouhettes associées sont pour PAM et CLARA toutes au dessus de la moyenne, signifiant ainsi qu'elles sont relativement bien classées. Dans un dernier temps, les algorithmes PAM et CLARA fournissent une moyenne de silouhettes égale à 0.52, soit le maximum.
Afin de compléter notre analyse, nous pouvons fournir un ensemble d'indicateurs (définient ci-dessus).

Comment anticipé, l'algorithme AGNES fournit un indice de Dunn élevé. De même, l'algorithme K-means, ou l'on a choisi k=2, nous retourne un indice de connectivité faible. Etant donné que ceci relève de la structure des méthodes et des indices, nous devons davantage faire confiance aux autres indicateurs. K-means et PAM sont donc les deux indicateurs que l'on retient d'après le critère de validité interne.

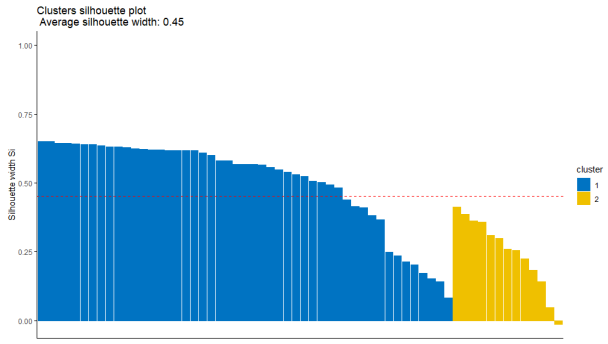


FIGURE 36 – Graphique des silouettes pour l'algorithme K-MEANS

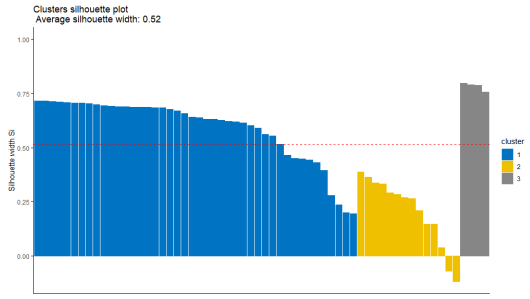


FIGURE 37 – Graphique des silouettes pour l'algorithme PAM

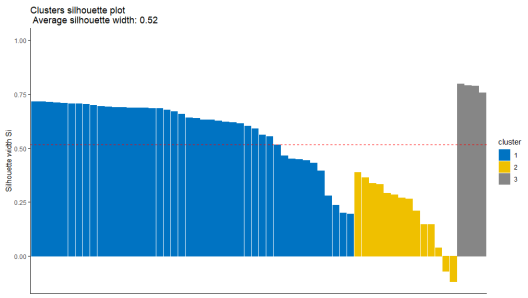


FIGURE 38 – Graphique des silouettes pour l'algorithme CLARA

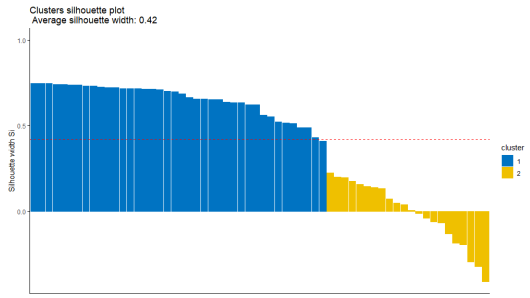


FIGURE 39 – Graphique des silouettes pour l'algorithme FANNY

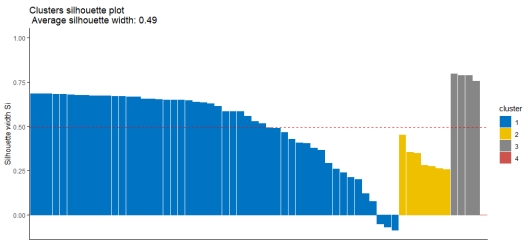


FIGURE 40 – Graphique des silouettes pour l'algorithme AGNES

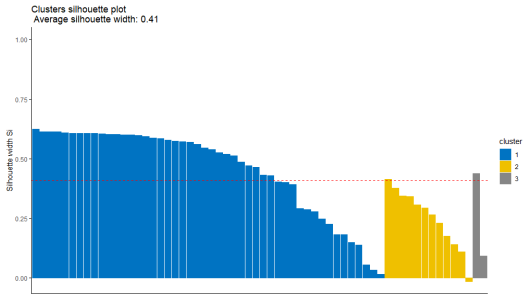


FIGURE 41 – Graphique des silouettes pour l'algorithme DIANA

	kmeans	pam	clara	fanny	agnes	diana
Incide de Dunn	0.278	0.320	0.320	0.154	0.418	0.187
Indice de connectivité	7.969	11.706	11.706	17.093	18.233	16.156
Silhouette	0.452	0.516	0.516	0.419	0.493	0.410

FIGURE 42 – Indices de validité interne

Validité externe par rapport à la variable nature

D'après les la Figure 43, nous remarquons clairement que l'algorithme kmeans et l'algorithme fanny ont plus de bon résultats aux indices de validité externe que les autres algorithmes : autrement dit, vis-à-vis de la variable nature, les algorithmes kmeans et fanny classent relativement

	kmeans	pam	clara	fanny	agnes	diana
Rand ajusté	0.570	0.492	0.492	0.538	0.200	0.523
Meila's VI	0.612	0.773	0.773	0.720	1.003	0.763
Kappa de Cohen	0.702	0.042	0.042	0.706	0.110	0.240

FIGURE 43 – Indices de validité externe

bien les eaux dans "gaz" et "plat". C'est d'ailleurs kmeans qui obtient le meilleur indice de rand ajusté (0.570) et de variation d'information de Meila (0.612). En revanche, c'est fanny qui obtient le meilleur indice de Kappa de Cohen (0.706). Il est également notable que l'algorithme agnes obtient les moins bon résultats pour les indices composants la validation externe.

kmeans	pam	clara	fanny	agnes	diana
4	3	3	3	2	3

FIGURE 44 – Score moyen par classement après validation interne et externe

Pour savoir quel algorithme obtient les meilleures indices en validité externe et interne, nous avons utilisé la même méthode de classement que précédement [1]. Pour chaque indice, l'algorithme obtenant le meilleur score se voit attribué 6 points et ainsi de suite jusqu'à 1 point pour l'algorithme obtenant le moins bon score à ce même indice. Ainsi, la Figure 44 représente le score moyen après avoir calculé ces scores pour chacun des indices de validité interne et externe. Il en résulte que kmeans obtient le meilleur score. Notre choix final se porte donc sur l'algorithme kmeans par la constance de ses bons résultats. Dans la partie qui suit, nous allons analyser les deux clusters formés par kmeans pour tenter d'interpréter leurs dynamiques internes.

4.5 Analyse et interprétation des résultats

Les graphiques en étoile des moyennes et médianes des variables des 2 clusters formés par kmeans nous montrent bien que les deux clusters n'ont pas du tout les mêmes profils. En effet, nous pouvons voir que les variables différenciant les deux clusters sont HCO3, Na, K, Cl Mg et PH. Au contraire, les variables NO3, SO4 et Ca ne sont pas différentes d'un cluster à l'autre. Cette remarque entre en résonnance avec l'analyse en composante principale qui avait notamment aboutie au fait que NO3 ne participait ni à l'axe 1, ni à l'axe 2. De plus, SO4 et Ca participaient grandement à l'axe 2 [2]. Nous pouvons en conclure que les deux clusters sont pricipalement différenciés par rapport à l'axe

1. Classement des méthodes d'agrégation en classification ascendante hiérarchique.
2. Biplot du projet Analyse de données visible en annexe.

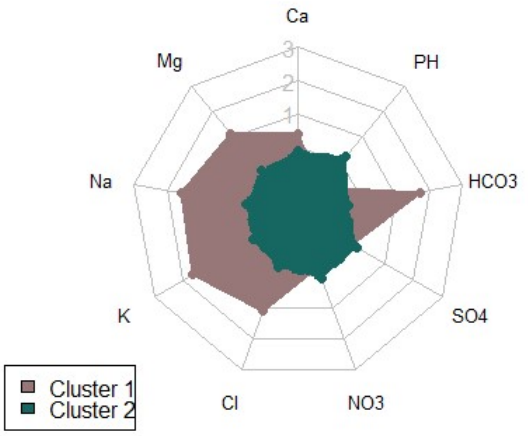


FIGURE 45 – Graphique en étoile des moyennes des variables de chacun des 2 clusters

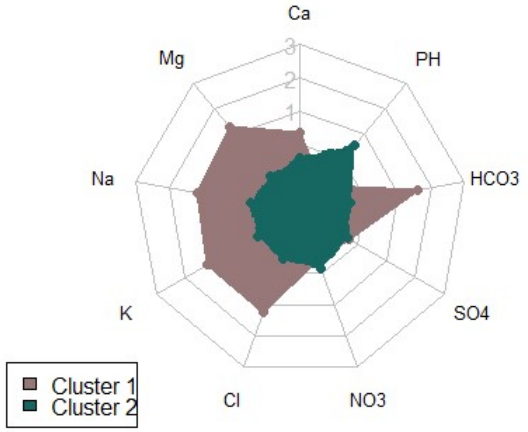


FIGURE 46 – Graphique en étoile des médianes des variables de chacun des 2 clusters

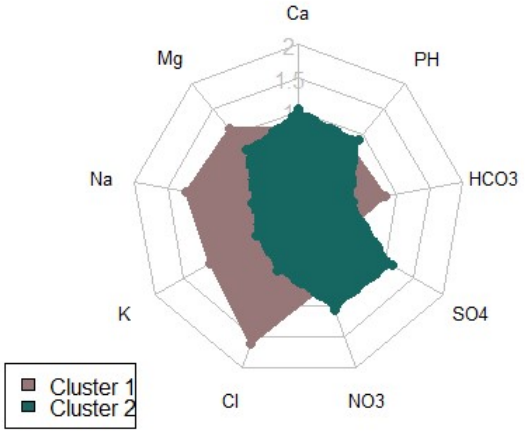


FIGURE 47 – Graphique en étoile des écart-types des variables de chacun des 2 clusters

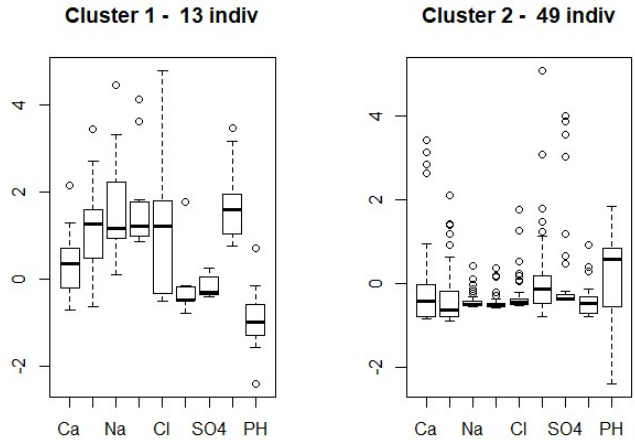


FIGURE 48 – Diagrammes en boîte des variables de chacun des deux clusters

1 et non à l'axe 2. Nous pouvons voir également, grâce au graphique en étoile des écart-types, que les écart-types des variables varient également entre les deux clusters. Le cluster 1 est plus éparpillé pour les variables composants l'axe 1, tandise que le cluster 2 est plus éparpillé pour les variables composants l'axe 2. Ces interprétations sont corroborantes avec le graphique 10 représentant les deux clusters de kmeans dans le plan des composantes principales 1 et 2.

Enfin, grâce aux diagrammes à boîte, nous remarquons que le nombre d'observations extrêmes sont bien plus nombreuses dans le cluster 2 que dans le cluster 1. Graphiquement, cela est nottament due aux quatres individus qui sont éloignés du nuage de points et plus spécifiquement du centre de gravité du cluster 2.

A travers cette matrice de confusion, nous pouvons voir que les eaux gazeuses françaises [3] sont davantages dans le cluster 1 (2/3 dans le cluster 1), tandise que les eaux plates qu'elles soient

3. Dans notre jeu de données, il n'y a pas d'eau gazeuse marocaine.

Table 3 – Matrice de confusion entre les deux clusters de kmeans et les trois groupes hypothétiques

	gaz fr	plat fr	plat mar
cluster 1	12	1	0
cluster 2	6	36	7

françaises ou marocaines, sont très majoritairement dans le cluster 2. En conclusion, les deux clusters obtenus par l'algorithme kmeans tendent à représenter les deux groupes que sont les eaux plates et les eaux gazeuses.

5 Méthodes de classification supervisée : prévoir la nature de l'eau

Les méthodes de classification supervisée permettent, grâce à un ensemble de données dit "d'apprentissage" d'attribuer une appartenance à un groupe pour une nouvelle observation. Naturellement, cela suppose que les groupes soient prédéfinies par l'utilisateur. Dans notre cas, nous allons considérer eaux plates versus eaux gazeuzes. Aussi, le but sera de construire un modèle qui est généralisable. En ce sens, des méthodes telles que la validation-croisée ou le bootstrap permettent de s'arrurer que le modèle n'apprend pas trop du jeu de données. En effet, la qualité d'un modèle ne peut être déterminer par sa performance sur l'ensemble d'apprentissage. Le but étant de maximiser la performance, l'algorithme apprendra, dans le cas extrême, parfaitement la structure du jeu de données et aura un taux d'erreur proche de zéro. Cependant, lors de l'implémentation d'un nouvel individus n'appartenant pas au groupe d'apprentissage cela posera des problèmes de précision. L'enjeu sera donc de trouver un algorithme qui arrive le mieux à prévoir la nature de l'eau, grâce aux méthodes : K plus proche voisins, analyse discriminante et Arbres de décision binaire. **Une valeur seuil de validation** vis-à-vis du terme d'erreur sera la valeur pour laquelle l'algorithme ne prédira que des eaux plates. En effet, nous sommes en présence d'un déséquilibre du cardinal des classe : #(plat) = 44 et #(gaz) = 18. Donc, si on construit un "stupid bot" qui ne prédit que des eaux plat, l'erreur sera d'environ 30%. Aussi, nous devrons être attentif à l'écart entre le taux d'erreur fournit par les différentes méthodes appliquées et ceux fournit par ce 'stupid bot'.

5.1 Méthodes de calcul des taux d'erreur

Méthode échantillon/test

Cette méthode consiste à faire entrainer notre modèle sur un échantillon de notre jeu de données (ici 80%) et de tester sa précision sur un échantillon dit de test (20%) grâce à la matrice de confusion.

Validation croisée stratifiée

Cette méthode consiste à répartir les données aléatoirement [4] en q blocs [5]. A partir de ces blocs, on applique sur $q-1$ blocs l'algorithme de classification en question pour modéliser le classement des individus. Ensuite, nous appliquons le modèle créé au dernier bloc qui n'a pas été utilisé pour l'entraînement, ce bloc est appelé "bloc test". Nous calculons l'erreur des prévisions par rapport aux valeurs réelles des observations et répétons ce processus q fois pour que chaque bloc soit une fois bloc test.

Validation croisée - Leave or out (LOO)

Sur le même principe que la méthode de validation croisée expliquée précèdement, la méthode LOO n'utilise pas de bloc mais utilise un individu test et le reste des observations pour l'entraînement. Suite à l'obtention de la prévision pour la valeur test, nous calculons l'erreur entre la prévision et la réelle valeur de l'observation (0 ou 1). Nous répétons ce processus n fois (n : nombre d'observations) et obtenons ainsi une erreur moyenne.

Technique du Bootstrap

La technique de bootstrap est une technique de rééchantillonage. En effet, on applique l'algorithme sur un sous-ensemble d'entraînement de notre jeu de données créé à partir d'un tirage aléatoire avec remise. L'algorithme donne alors une prévision pour tout le jeu de données, puis nous calculons l'erreur comise. Cette technique permet de simuler la potentielle arrivée de nouvelles données dans notre modèle.

5.2 Méthode des k plus proche voisins

Cette méthode consiste à créer des groupes d'observations similaires et à partir de ces groupes déterminer la valeur de la variable manquante par rapport aux variables présentes pour une certaine observation. La méthode kNN, ou aussi kPPV [6], est une méthode non-paramétrique. L'avantage de cette méthode est donc qu'elle n'a pas d'hypothèse sur les distributions, elle est simple à mettre en œuvre et elle donne une probabilité d'erreur faible. Cependant, elle a des inconvéniants comme son temps de calcul important pour rechercher des k plus proche voisins et la place mémoire prise par le stockage de l'ensemble des prototypes.
Dans un contexte de classification d'une nouvelle observation x, l'idée fondatrice simple est de faire voter les plus proches voisins de cette observation. La classe de x est déterminée en fonction de la

4. La répartition aléatoire est un processus qui nous permet d'obtenir des blocs qui tendent à être représentatifs de l'ensemble des données et de sa diversité : stratification.
5. Terme anglais : q-folds
6. kNN : k nearest neighbors ; kPPV : k plus proches voisins

classe majoritaire parmi les k-plus proches voisins de l'observation x. La méthode kNN est donc une méthode à base de voisinage, non-paramétrique.
La décision est en faveur de la classe majoritairement représentée par les k-voisins. Soit k le nombre d'observations issues du groupe des plus proches voisins appartenant à la classe r, sachant qu'il existe c classes. On a alors :

$$\sum_{r=1}^{c} k_r = k$$

Ainsi, une nouvelle observation est prédite dans la classe l avec :

$$l = \max_r k_r$$

Présentation des résultats Knn

Dans un premier temps, nous pouvons présenter les résultats grâce à la méthode de séparation train/test. En effet, nous avons effectué une boucle afin de constater le terme d'erreur calculé à partir de la matrice de confusion. Nous pouvons représenter cette matrice pour chaque nombre de voisins considéré pour $k \in [1, 9]$.

TABLE 4 – *Matrice de confusion pour Knn - Train/Test*

	k=1		k=2		k=3		k=4		k=5		k=6		k=7	
	plat	gaz	plat	gaz	plat	gaz	plat	gaz	plat	gaz	plat	gaz	plat	gaz
plat	6	2	7	1	7	1	7	1	7	1	7	1	7	1
gaz	1	3	1	3	0	4	0	4	0	4	1	3	0	4
Taux erreur	0.25		0.1667		0.0833		0.0833		0.0833		0.1667		0.0833	

	k=8		k=9	
	plat	gaz	plat	gaz
plat	7	1	7	1
gaz	1	3	1	3
Taux erreur	0.1667		0.1667	

Nous observons donc une erreur de 0.083 par méthode de train/test pour $k_1 = 3, 4, 5, 7$.

Dans un second temps, afin de limiter les biais d'échantillonage du calcul de l'erreur, nous pouvons appliquer une validation croisée. En suivant la méthode méthode que précédemment pour le calcul de l'erreur, nous pouvons constater une erreur moyenne de classement pour $k_2 = 3, 5, 6, 7$ égale à 0.094, correspondant au minimum. On remarque que $k_1 \bigcap k_2$ = 3, 5, 7.

Dans un troisième temps, nous pouvons effectuer une validation via bootstrap. L'erreur minimale est égale à 0.0587.

Nous observons une différence significative avec la validation croisées. En effet, dans la méthode des validations croisées, les sous-échantillons de données utilisées ne se chevauchent pas : deux observations similaires ne peuvent pas se retrouver dans le même bloc, ce qui est un élément fondamental de son succès. Cependant, dans le cas de la technique du bootstrap il y a un chevauchement avec les données initiales. En effet, à chaque échantillonage, nous avons environ 2/3 du jeu de données initial qui se retrouve dans celui créé par le bootstrap, ce qui entraine une sous-estimation du terme d'erreur lié à l'overfitting !

Dans un dernier temps, nous pouvons faire une vérification via une validation croisée (leave or out). En suivant la méthode méthode que précédemment pour le calcul de l'erreur, nous pouvons constater une erreur moyenne de classement pour $k_3 = 3, 5, 6, 7$ égale à 0.097, correspondant au minimum. On remarque que $k_1 \bigcap k_2 \bigcap k_2 \bigcap k_2$ = 3, 5, 7.

On remarque une convergence des critères de validation.

5.3 Analyse discriminante linéaire

Cette méthode est une méthode prédictive au même titre que les plus proche voisins. Cependant, elle répond à certaines hypothèses vis-à-vis des distributions (hypothèses de multi-normalité), vis-à-vis de la variance des termes d'erreurs (homoscédasticité) et non colinéarité des variables. Dans la pratique, cette méthode possède une fonction de classement qui s'exprime comme une combinaison linéaire des variables explicatives[6].

Afin de tester la multi-normalité, nous allons effectuer un test de shapiro-Wilk[7]. Sous H_0, les variables sont distribuées comme une loi multi normale. Nous remarquons que la p-value est égale à 6.899e-14, nous rejetons ainsi H_0, on ne peut donc pas affirmer que l'échantillon suit une loi multi normale.

De plus, nous pouvons tester l'homoscédasticité. La question est la suivante, la variance pour les eaux plates et gazeuzes sont-elles similaires pour les variables prédictives ? On peut construire le test de MANOVA, afin de constater si les variances sont semblables pour chaque type de Nature. Sous H_0, les variances des groupes 'plat' et gaz' sont similaires. La p-value associée à ce test est très proche de 1. Ainsi, nous ne pouvons rejeter le fait que les variances soient égales.

Présentation des résultats : Analyse discriminante linéaire

Dans un premier temps, nous pouvons présenter les résultats grâce à la méthode de séparation train/test. Nous pouvons ainsi calculer la matrice de confusion :

TABLE 5 – *Matrice de confusion pour Analyse discriminante linéaire - Train/Test*

	gaz	plat
gaz	4	0
plat	1	7

Le terme d'erreur est donc de 8.33%.

Dans un second temps, afin de limiter les biais d'échantillonage du calcul de l'erreur, nous pouvons appliquer une validation croisée. En suivant la méthode méthode que précédemment pour le calcul de l'erreur, nous pouvons constater une erreur moyenne de classement égale à 10.94 %.

Dans un troisième temps, nous pouvons effectuer une validation via bootstrap. L'erreur moyenne est égale à 7.98%.

Dans un dernier temps, nous pouvons faire une vérification via une validation croisée (leave or out). En suivant la méthode méthode que précédemment pour le calcul de l'erreur, nous pouvons constater une erreur moyenne de classement égale à 8.06%.

5.4 Arbres de décision binaires : CART

L'objectif de ces méthodes est de construire un arbre de décision contenant un ensemble de divisions. Ces divisions sont elles-mêmes définies par une valeur seuil pour les variables quantitatives et un partage en deux groupes des modalités pour les variables qualitatives. Afin de définir ces règles, nous devons définir un critère de division que l'on va conserver durant la durée de cette analyse : L'indice de GINI : $D_{(k)}$. Cette fonction est non-négative, nulle si le noeud est homogène, c'est-à-dire que tous les individus appartiennent à la même modalité et est maximale lorsque la probabilité d'appartenir à une classe est semblable. En effet, nous pouvons voir cette dynamique via la formule du coefficient de GINI :

$$D_{(k)} = 1 - \sum_{l=1}^{m} p_{lk}^2$$

Avec p_{lk} qui correspond à la probabilité qu'un élément du kième noeud appartienne à la lième catégorie.

Nous pouvons décrire le fonctionnement de la méthode de CART. La fonction rpart (R) construit un arbre contenant l'ensemble des individus. Après cette étape, l'arbre est découpé par le plus petit arbre avec la plus petite 'miss-classification "loss" :

- Les données sont divisées en n (default = 10) blocs construits de manière aléatoire : K1 to K10.
- Il utilise alors 10-fold-cross-validation et adapte chaque sous-arbre $T_1 ... T_m$ sur chaque bloc d'entrainement.

— La "miss-classification loss" correspondante : R_m (valeur de risque) est alors calculée en comparant les classes prédites pour les blocs de validation et celles des réelles classes. Et cette valeur de risque pour tout les sous-arbre est propagée pour tous les blocs.

— Le paramètre de complexité donnant le plus petit risque total par rapport à tout le jeu de données est alors sélectionné.

— Le jeu de données entier est alors "fité" en utilisant ce paramètre de complexité et l'arbre associé est alors le meilleur.

Notons que pour notre méthode, $R_m = D_{(k)}$, étant l'indice de GINI.

Présentation des résultats CART

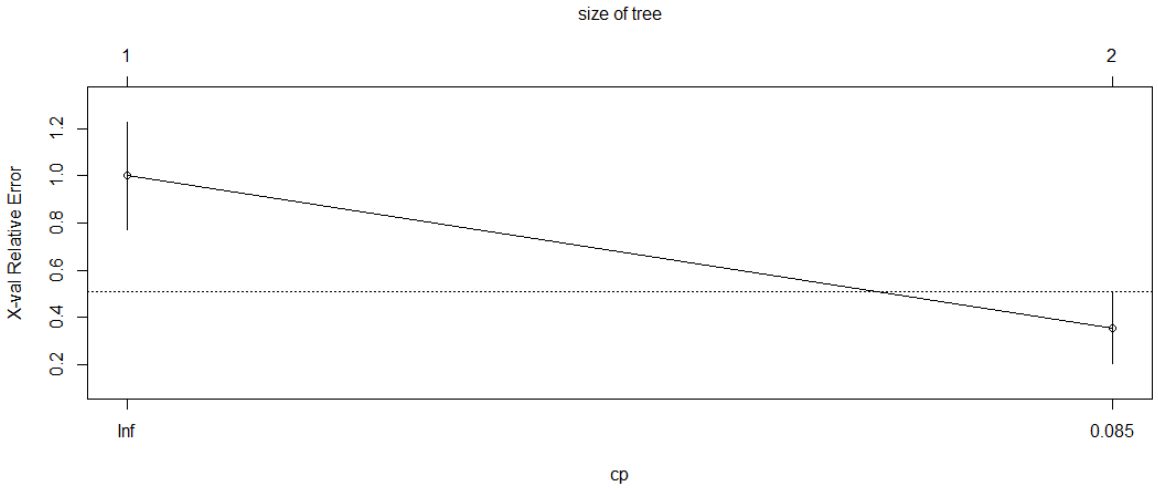


FIGURE 49 – Erreur et écart type des sous arbres élagués

Dans ce graphique, chaque point représente un arbre, avec l'estimation de l'écart-type de l'erreur de validation croisée. On choisit l'arbre qui minimise l'erreur de validation croisée. Le meilleur sous-arbre élagué de l'arbre maximal est composé de 2 feuilles et 1 variable : HCO3.

Nous pouvons ainsi représenter l'arbre optimal.

Nous constatons ainsi que l'arbre est simpliste dans le sens ou il n'admet qu'une règle de division : HCO3 >= -0.23 (sur les données centrées-réduites). Autrement dit, lorsque l'eau aura un taux de HCO3 est supérieure à -0.23, l'eau est prédite comme étant gazeuze. Nous pouvons ainsi calculer la matrice de confusion afin de calculer le taux d'erreur sur l'échantillon de test qui est de taille : 12x10.

TABLE 6 – *Matrice de confusion pour CART - Train/Test*

	gaz	plat
gaz	4	0
plat	1	7
Taux erreur	0.0833	

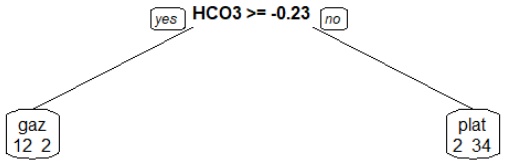


FIGURE 50 – Représentation de l'arbre optimal

On remarque grâce à l'arbre de décision que HCO3 est l'unique variable de décision. En effet, en comparant les valeurs de la variables HCO3 de l'échantillon test avec celui d'entrainement, l'arbre est capable, pour 11 observations sur 12 de prédire correctement la nature de l'eau. Ainsi, le choix quant à la préparation des données se révèle être cohérent dans le sens ou supprimer la variable HCO3 était une option sachant le nombre de valeurs manquantes qu'elle possédait. ()

5.5 Résultats classification supervisée : Taux d'erreur

La table ci-dessous permet de résumer les informations contenu dans l'ensemble de l'analyse des méthodes supervisées.

	kNN	Analyse.discriminante	Arbre.binaire
Méthode échantillon/test	0.0833	0.0833	0.0833
Validation croisée stratifiée	0.0938	0.1094	0.1094
Validation croisée par LOO	0.0968	0.0806	0.1129
Technique du bootstrap	0.0587	0.0798	0.1018

FIGURE 51 – Tableau des taux d'erreur

Dans un premier temps, on remarque que la technique du bootstrap fournit des erreurs inférieures à celles associées aux méthodes d'échantillonages (Validation croisée stratifiée et LOO). De plus, avec la méthode échantillon/test, les trois classification supervisées fournissent des résultats similaires ce qui indique que, sous risque de biais d'échantillonnage, les performances sont semblables. Néanmoins, en limitant ce biais on constate que Knn est relativement performant. Cependant, par

soucis d'interprétation, étant donné la structure de l'arbre de décision, il pourrait être intéressant de présenter ces résultats à un public novice.

6 Conclusion

Dans ce projet, nous avons dans un premier temps justifier l'existence de groupes grâce au test de Hopkins et à l'algorithme VAT. Dans un second, nous avons justifié la sélection d'un nombre de groupes à considérer dans notre jeu de données à l'aide de critères statistiques et d'interprétabilité. Dans un troisième temps, nous avons évalué la validité interne et externe des différents algorithmes de classification non supervisée, nous permettant de choisir le meilleur d'entre eux. Puis, dans un dernier temps, l'objectif fut de prédire la nature de l'eau à l'aide de méthodes de classification supervisées. Malgré l'existence de cluster relativement évidente, leurs déterminations n'étaient pas aussi facile. En effet, l'apparition d'observations "plat" dans "gaz", via la représentation du nuage de points dans le plan factoriel 1/2, et vice-versa a été une difficulté. Aussi, l'idéal aurait été de contacter un professionnel ayant des connaissances poussées dans la composition des eaux afin qu'ils puissent nous transmettre quelques recommandations. Dans ce projet, nous avons pris la responsabilité de ne pas modifier la variable "Nature", malgré la constatation de certaines observations aberrantes. Conscients que cela puisse altérer la qualité de notre modèle, nous ne souhaitions pas intégrer du biais dans notre modèle sous prétexte que la répartition des observations ne soit pas "idéale". Effectivement, nous considérons que la responsabilité de la modification des valeurs d'un jeu de données relève du métier et non du modélisateur, ou du moins tout changement doit se faire en étroite collaboration avec l'expert métier.

Par ailleurs, les résultats liés à la classification non-supervisées montre que K-means est le meilleur algorithme afin de trouver le nombre de groupes. En effet, les différents indicateurs de validité interne et externe étaient meilleurs pour cet algorithme.

Les méthodes de classification supervisées fournissent des résultats pouvant être améliorés à l'aide d'un contact métier qui préparerait les données d'une meilleure façon. Nous pouvons également ajouter que les outils de validation utilisées favorisent la robustesse vis-à-vis du surapprentissage des données d'entrainement. Enfin, les voies d'amélioration sont les suivantes :

- Nécessité de contacter un professionnel afin de nous transmettre des recommandations quant à la préparation des données ou d'émettre des hypothèses à l'aide d'une interprétation opérationnelle.
- Tester des méthodes de classification supervisées telles que les forêts aléatoires.

7 Annexe

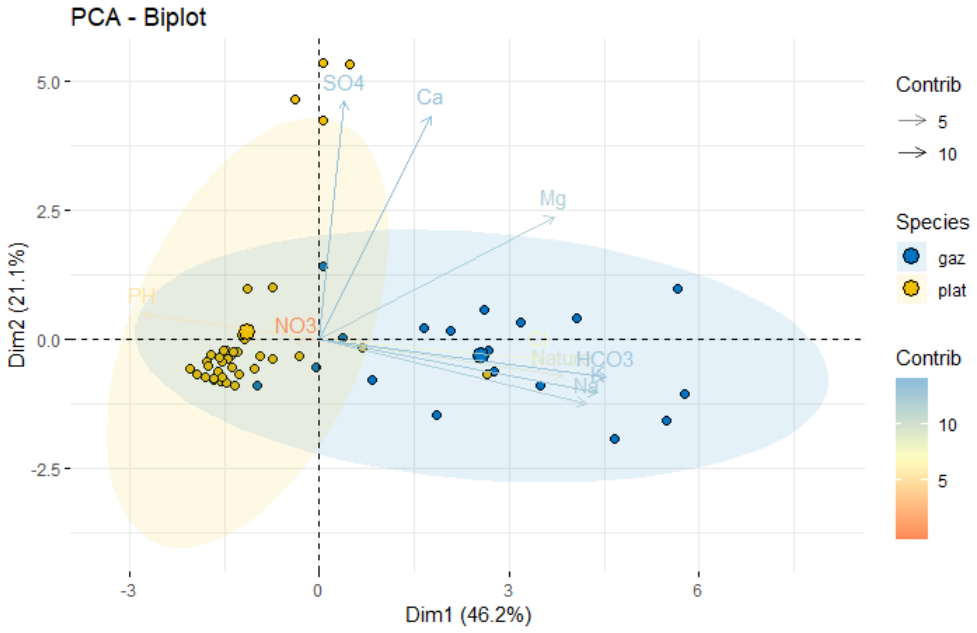
Tableau des données

Table 7 – Jeu de données utilisé

Num indiv	Ca	Mg	Na	K	Cl	NO3	SO4	HCO3	PH
1	16.00	8.0	75.0	3.0	95.0	0.0	8.0	112.0	8.2
2	72.00	38.0	14.0	2.0	6.0	1.0	81.0	329.0	7.4
3	63.00	23.0	13.0	1.3	11.0	2.0	14.0	300.0	7.4
10	170.00	92.0	650.0	130.0	387.0	0.0	31.0	2195.0	6.3
12	190.00	85.0	150.0	10.0	40.0	7.0	40.0	1300.0	6.0
14	220.00	70.0	350.0	46.0	21.0	0.0	6.0	2000.0	6.5
15	152.00	36.0	651.0	40.0	215.0	1.0	195.0	1799.0	7.4
16	420.00	51.0	245.0	43.0	3.0	1.0	18.0	2075.0	6.2
19	486.00	84.0	9.1	3.2	10.0	2.7	1187.0	403.0	7.0
20	517.00	67.0	1.0	2.0	1.0	2.0	1371.0	168.0	7.4
21	93.00	8.1	8.8	2.6	18.0	2.0	5.2	306.0	7.4
22	106.00	3.8	3.5	1.8	3.8	2.0	58.9	272.0	7.2
23	64.50	3.5	12.0	0.5	20.0	2.5	6.0	195.0	7.8
24	124.00	25.0	11.0	3.5	16.0	0.0	60.0	420.0	7.6
28	44.00	24.0	23.0	2.0	5.0	1.0	3.0	287.0	7.6
29	71.00	5.5	11.2	3.2	20.0	1.0	5.0	250.0	7.5
30	82.00	7.4	7.3	1.9	14.0	3.9	18.0	263.0	7.5
31	40.00	11.0	47.0	3.0	70.0	1.0	8.0	177.0	7.5
32	63.00	26.0	99.0	21.0	33.0	2.0	60.0	493.0	7.4
33	67.00	26.0	84.0	20.0	32.0	2.0	61.0	473.0	5.2
34	6.40	1.2	3.0	0.5	3.0	4.0	5.0	20.0	6.5
36	4.10	1.7	2.7	0.9	0.9	0.8	1.1	25.8	7.3
37	63.00	10.2	1.4	0.4	1.0	2.0	51.3	173.2	7.6
38	63.00	10.0	1.4	0.4	1.0	2.0	51.0	173.0	7.6
40	108.00	14.0	3.0	1.0	9.0	8.0	13.0	350.0	7.4
41	78.00	24.0	5.0	1.0	4.5	3.8	10.0	357.0	7.2
44	555.00	110.0	14.0	4.0	18.8	2.9	1479.0	403.0	7.0
48	6.50	2.0	4.4	1.7	1.0	0.5	0.2	44.0	7.7
49	26.50	1.0	0.8	0.2	2.3	1.8	8.2	78.1	8.0
50	3.00	0.6	1.5	0.4	0.6	1.0	8.7	5.2	6.8
51	3.60	1.8	3.6	0.6	0.9	0.5	1.2	25.8	6.9
52	2.40	0.5	3.1	0.4	3.0	3.0	2.0	6.3	5.9
55	46.10	4.3	6.3	3.5	3.5	1.0	9.0	163.5	7.7
57	94.00	83.5	905.0	116.0	80.0	1.0	25.0	3204.8	6.5
58	147.30	3.4	9.0	1.0	21.5	18.3	33.0	430.0	5.5
60	24.00	16.0	32.0	4.9	38.0	0.0	50.0	121.0	6.4

Num indiv	Ca	Mg	Na	K	Cl	NO3	SO4	HCO3	PH
61	241.00	95.0	255.0	49.7	38.0	1.0	143.0	1685.4	5.2
62	85.00	80.0	385.0	65.0	285.0	1.9	25.0	1350.0	6.0
63	301.00	160.0	493.0	52.0	649.0	1.0	230.0	1837.0	6.3
64	253.00	11.0	7.0	3.0	4.0	1.0	25.0	820.0	6.0
65	176.00	46.0	28.0	5.0	37.0	0.0	372.0	312.0	7.2
67	85.00	80.0	385.0	65.0	285.0	1.9	25.0	1350.0	5.9
68	200.00	133.0	400.0	41.0	379.0	8.0	173.0	1390.0	5.8
69	3.60	1.8	3.6	0.6	0.9	0.5	1.2	24.4	6.9
70	208.00	55.9	43.6	2.7	74.3	0.5	549.2	219.6	7.7
72	596.00	77.0	7.0	2.0	8.0	0.5	1530.0	290.0	7.1
73	116.00	4.4	9.0	2.4	15.5	1.0	25.5	331.0	7.2
74	108.00	14.0	3.0	1.0	9.0	12.0	13.0	350.0	7.4
75	25.20	21.3	453.0	40.8	27.2	1.0	38.9	1403.0	6.2
77	190.00	72.0	154.0	49.0	18.0	0.0	158.0	1170.0	6.0
78	103.00	10.0	1172.0	66.0	235.0	2.0	138.0	2989.0	6.8
80	202.00	36.0	3.8	2.0	7.2	6.0	306.0	402.0	7.6
81	2.70	1.0	2.4	0.5	1.2	2.4	1.0	13.0	6.3
82	11.50	8.0	11.6	5.7	13.4	6.3	8.1	71.0	7.0
84	11.00	5.1	15.0	1.3	15.0	2.2	5.0	67.7	7.5
85	12.02	8.7	25.5	2.8	14.2	0.1	41.7	103.7	6.5
86	70.00	40.0	120.0	8.0	220.0	4.0	20.0	335.0	7.3
87	63.50	35.5	8.0	1.0	19.8	7.0	3.8	372.0	6.5
88	148.80	48.6	224.0	26.0	280.0	2.8	14.3	890.9	7.6
93	8.00	7.3	46.0	1.0	7.8	0.1	15.7	42.7	6.5
94	25.70	23.8	224.0	26.0	130.0	1.3	6.2	27.5	6.5
95	11.20	9.7	52.0	1.0	88.8	3.4	20.6	42.7	6.5

Biplot obtenu par ACP avec 2 composantes retenues



Bibliographie

[1] CNRS, *Eau potable : composition chimique.*
https ://www.cnrs.fr/cw/dossiers/doseau/decouv/potable/compoChim.html ;
Site consulté le 28/10/2018.

[2] Lise Bellanger, Richard Tomassone, *Exploration de données et méthodes statistiques.*
Editions ellipse, 2014.

[3] Régis Bourbonnais, *Econométrie.*
Editions dunod, 2016.

[4] R. Tibshirani, G. Walther et T. Hastie, *Estimating the number of clusters in a dataset via the gap statistic*, pp 411-423.
Royal Statistical Society, 2001.

[5] M. Meilã, *Comparing Clusterings – an information based distance.*
Elsevier Science, 2006.

[6] A.B. Dufour, D. Chessel J.R. Lobry, *Analyse discriminante linéaire.*
Université de Lyon, 2015.

[7] WikiStat, *Analyse de variance multivariée — MANOVA.*
https ://www.math.univ-toulouse.fr/ besse/Wikistat/pdf/st-m-modmixt5-manova.pdf ;
Site consulté le 08/12/2018.