Méthodes de classification non-supervisées et supervisées - Eaux2018

Lechevranton & Vasse

December 19, 2018

Outline

1. Introduction : jeu de données
2. Existence de cluster
3. Classification non-supervisée
4. Classification supervisée : Prévoir la nature de l'eau

Introduction : jeu de données

	Nom	Nature	Ca	Mg	Na	K	Cl	NO3	SO4	HCO3	PH	Pays
1	Abatilles	plat	16.00	8.00	75.00	3.00	95.00	0.00	8.00	112.00	8.20	France
2	Aix-Les-Bains	plat	72.00	38.00	14.00	2.00	6.00	1.00	81.00	329.00	7.40	France
3	Alet	plat	63.00	23.00	13.00	1.30	11.00	2.00	14.00	300.00	7.40	France
4	Alpille	plat	41.00	3.00	2.00	0.00	3.00	3.00	2.00	134.00		France
5	Amelie le Reine		390.00	27.50	45.00	2.80	19.00	2.00	36.00	1376.60		France
6	Aquarelle	plat	70.00	2.10	2.00			4.00		210.00		France

Figure 1: Premières lignes des données brutes

Introduction : jeu de données

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

Figure 2: Summary des données brutes

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

Figure 3: Summary des données CR

Existence de clusters : Intuition graphique

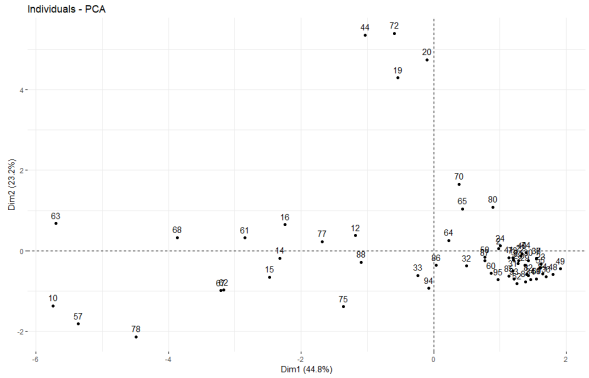

Figure 4: représentation individus plan factoriel 1/2

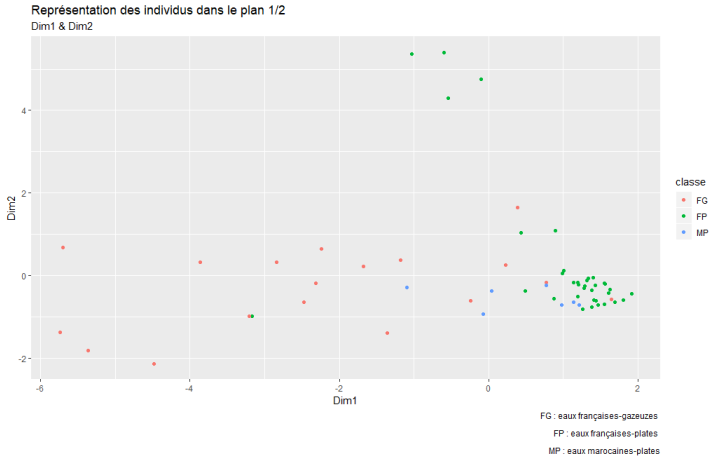

Figure 5: ggplot représentation des individus

Existence de clusters : Algorithme VAT

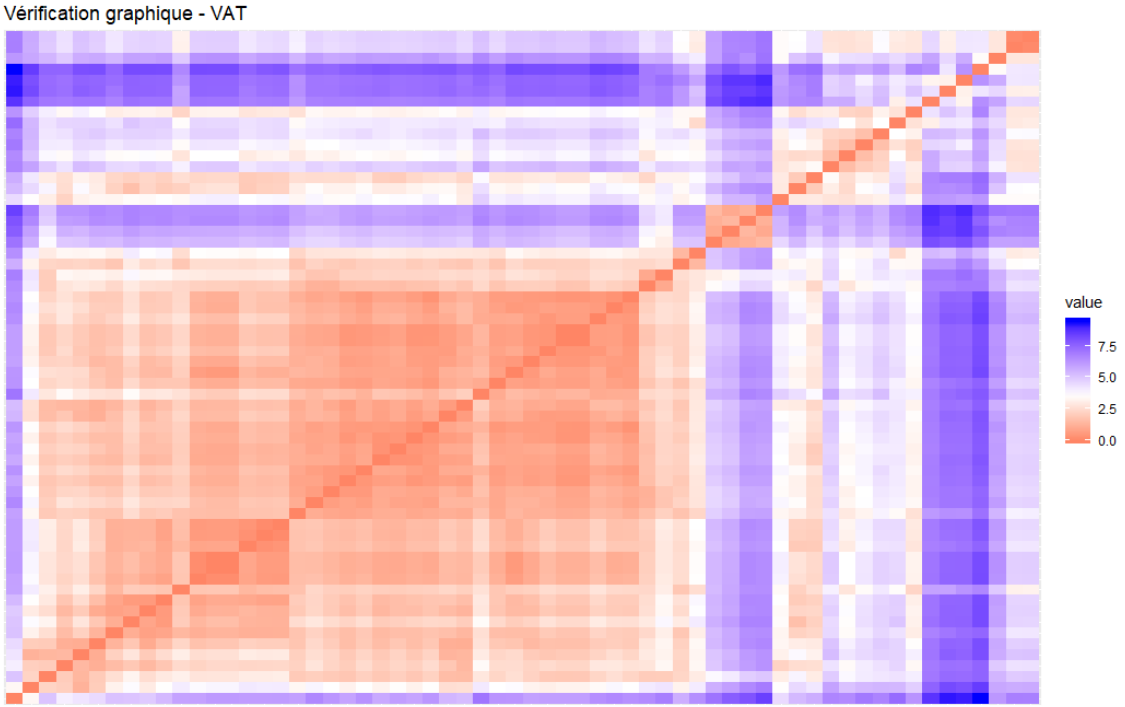

Figure 6: Représentation graphique algorithme VAT

Existence de clusters : Statistique de Hopkins (H)

H_0 : l'échantillon est distribué selon une loi uniforme
H_1 : Non H_0

Dans notre jeu de données, H $= 0.1655068 \leq 0.5 \Longrightarrow$ On rejette H_0.

Classification non-supervisée : Mesure de sélection de K

Graphique des Silhouettes

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,
- $b_i = \min_{C_j \neq C_k}(d(i, C_j))$

Méthode du coude

Représente graphiquement la variabilité intra-classe totale en fonction du nombre de K, puis cibler un k tel que la pente diminue fortement.

Classification non-supervisée : Mesure de sélection de K

Statistique GAP

L'objectif est de sélectionner k étant égal au plus petit k tel que :

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

- $Gap_n(k) = E_n[log(I_w)] - log(I_w)$
- $s_k = sd_k\sqrt{1 + 1/B}$

Classification non-supervisée : K-means

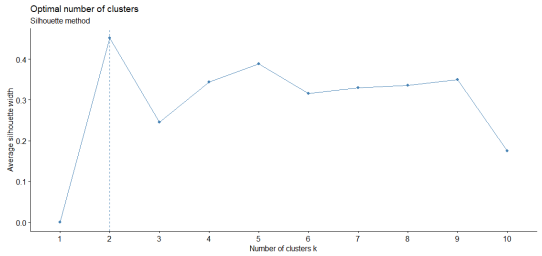

Figure 7: Sélection du nombre de groupes, méthode silouhette

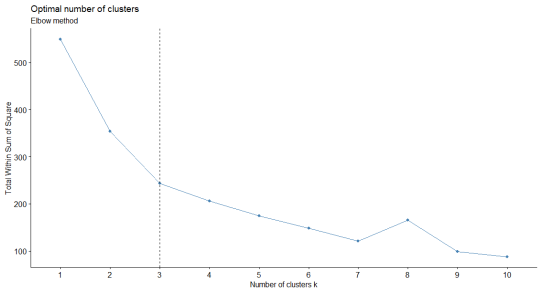

Figure 8: Sélection du nombre de groupes, méthode du coude

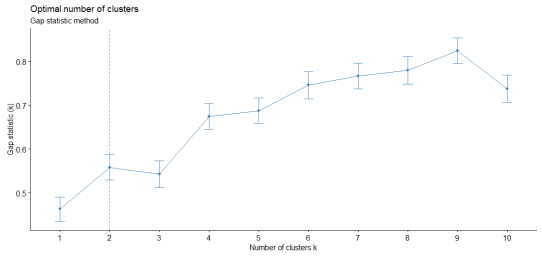

Figure 9: Sélection du nombre de groupes, critère GAP

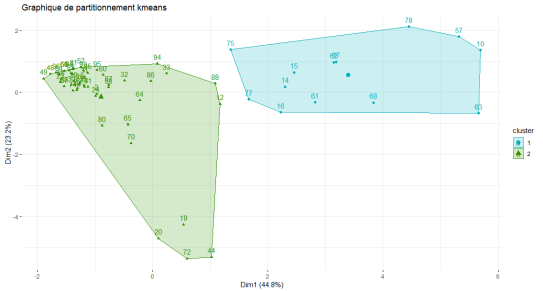

Figure 10: Représentation des groupes via Kmeans

Classification non-supervisée : PAM

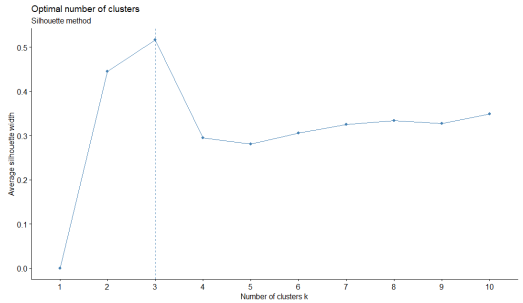

Figure 11: Sélection du nombre de groupes, méthode silouhette

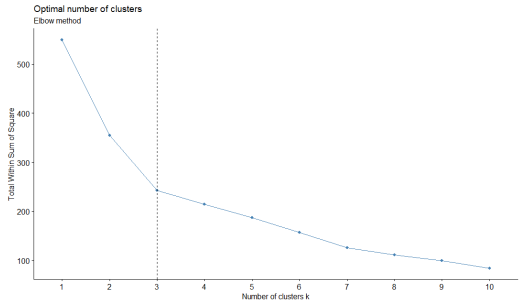

Figure 12: Sélection du nombre de groupes, méthode du coude

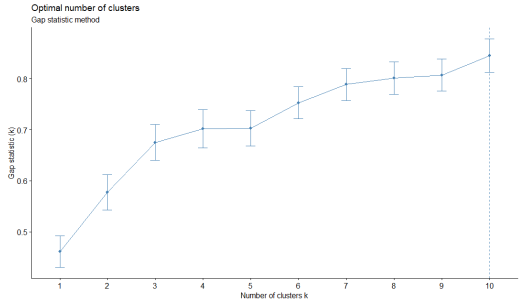

Figure 13: Sélection du nombre de groupes, critère GAP

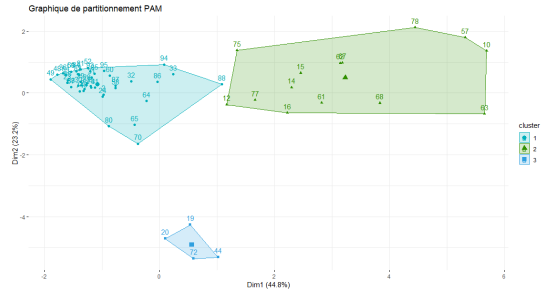

Figure 14: Représentation des groupes via PAM

Classification non-supervisée : FANNY

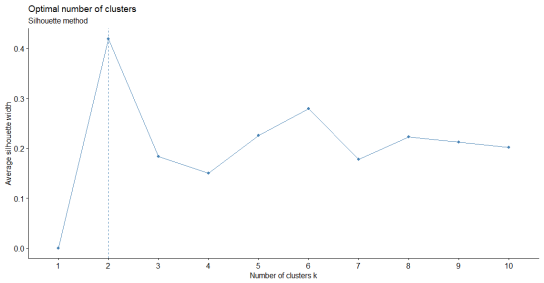

Figure 15: Sélection du nombre de groupes, méthode silouhette

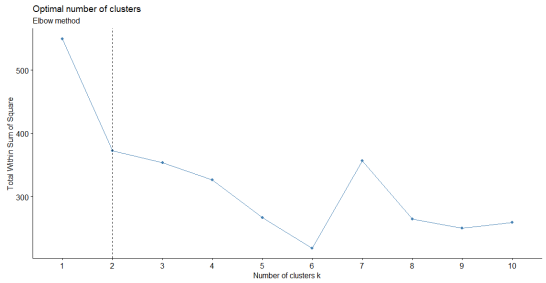

Figure 16: Sélection du nombre de groupes, méthode du coude

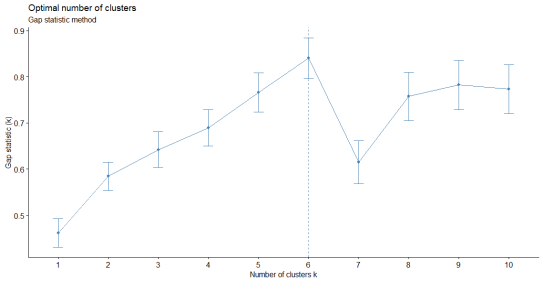

Figure 17: Sélection du nombre de groupes, critère GAP

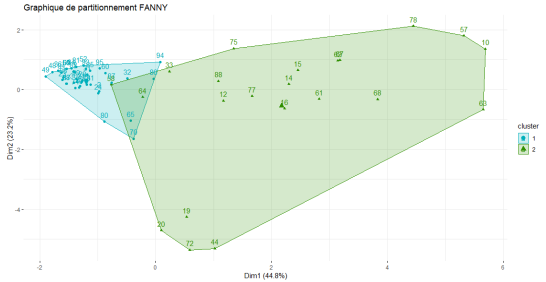

Figure 18: Représentation des groupes via FANNY

Classification non-supervisée : AGNES

	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

Classification non-supervisée : AGNES

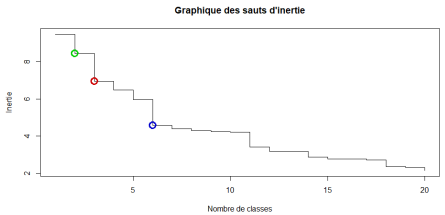

Figure 19: Sélection du nombre de groupes, méthode sauts d'inertie

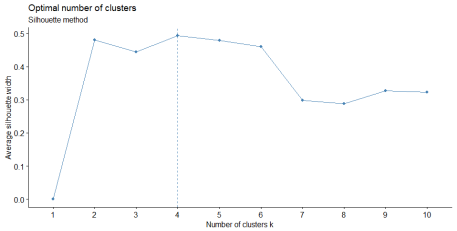

Figure 20: Sélection du nombre de groupes, méthode des silouhettes

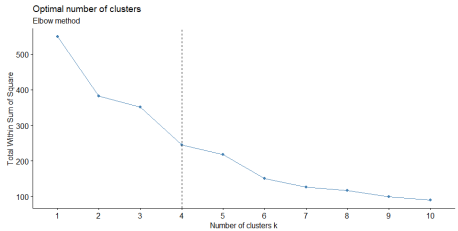

Figure 21: Sélection du nombre de groupes, méthode du coude

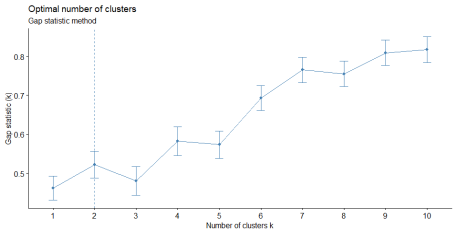

Figure 22: Sélection du nombre de groupes, critère de GAP

Classification non-supervisée : AGNES

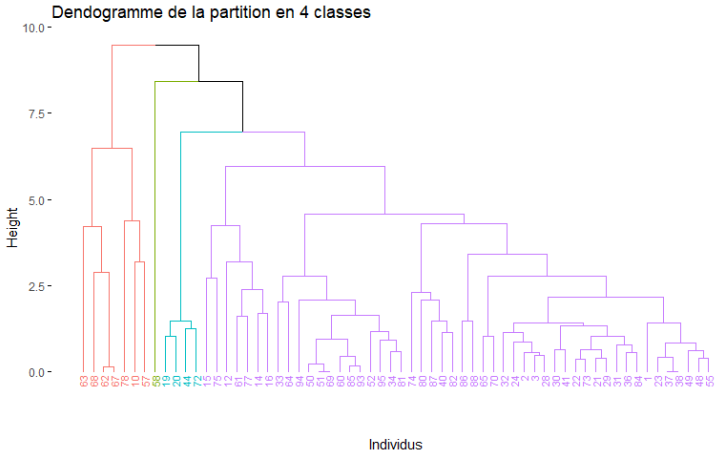

Figure 23: Dendogramme AGNES

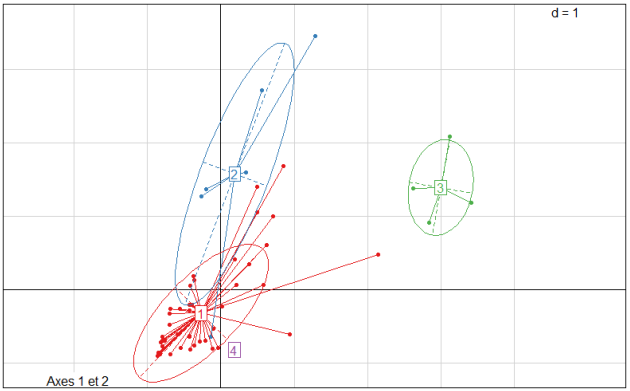

Figure 24: Représentation des individus selon AGNES

Classification non-supervisée : DIANA

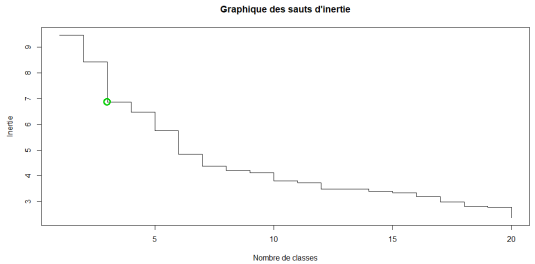

Figure 25: Sélection du nombre de groupes, méthode sauts d'inertie

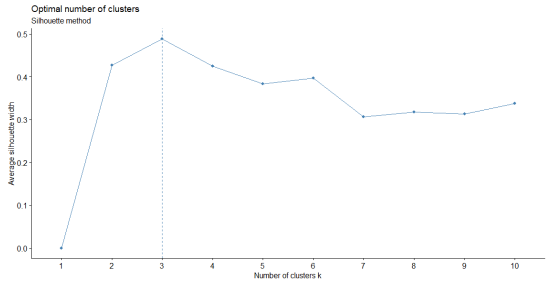

Figure 26: Sélection du nombre de groupes, méthode des silouhettes

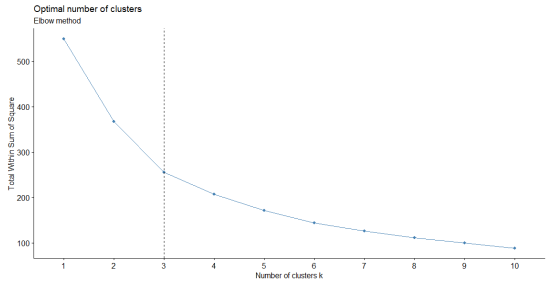

Figure 27: Sélection du nombre de groupes, méthode du coude

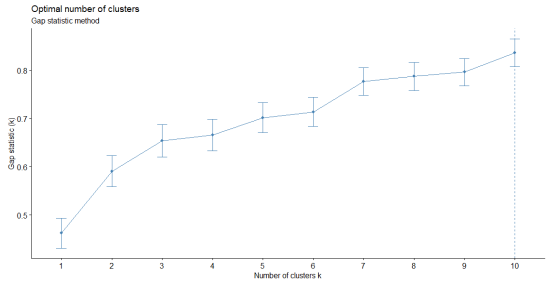

Figure 28: Sélection du nombre de groupes, critère de GAP

Classification non-supervisée : DIANA

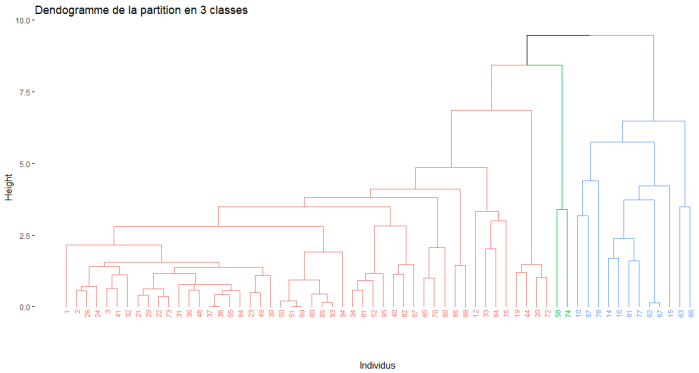

Figure 29: Dendogramme DIANA

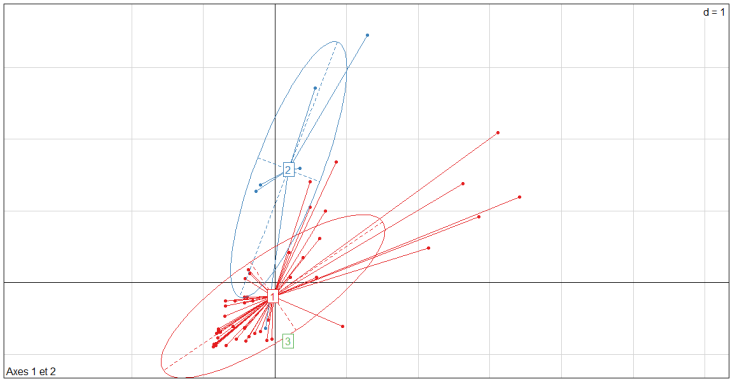

Figure 30: Représentation des individus selon DIANA

Classification non-supervisée : Critères de validité internes

	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

Classification non-supervisée : Critères de validité externes

	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

Classification non-supervisée : Choix de l'algorithme

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

Classification non-supervisée : Choix de l'algorithme

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

Choix : kmeans pour k=2

Classification non-supervisée : Analyse et interprétation des résultats

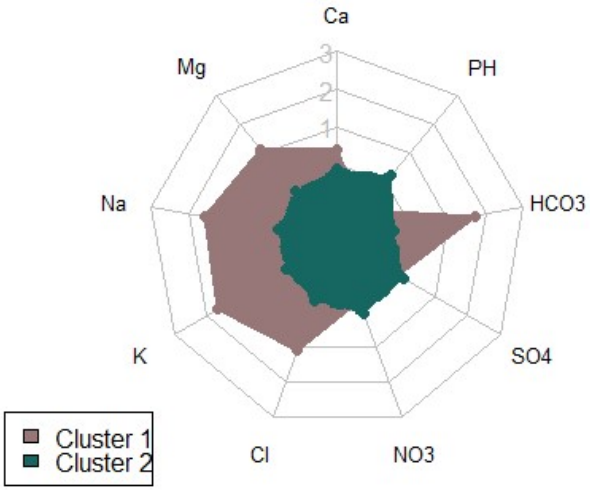

Figure 31: Graphique en étoile des moyennes

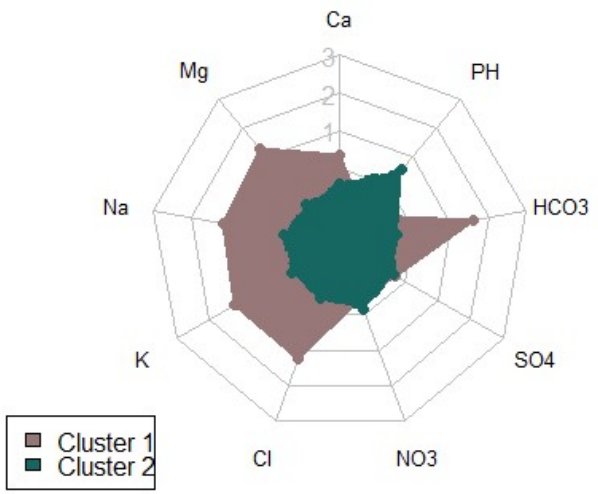

Figure 32: Graphique en étoile des médianes

Classification non-supervisée : Analyse et interprétation des résultats

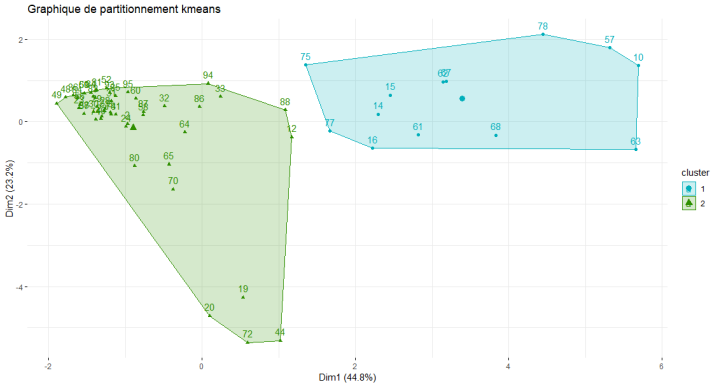

Figure 33: Graphique des deux clusters par kmeans

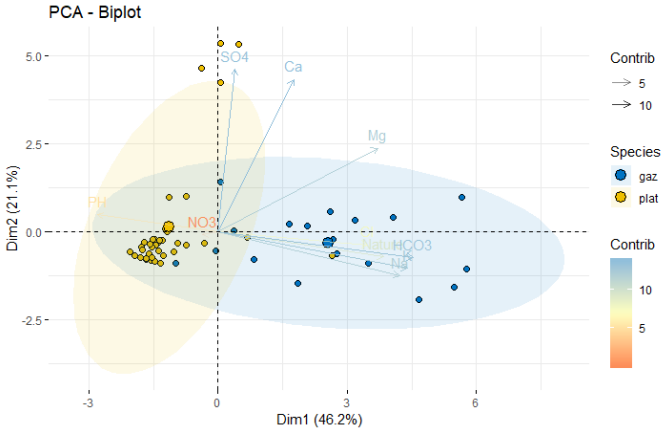

Figure 34: Biplot de l'ACP

Classification non-supervisée : Analyse et interprétation des résultats

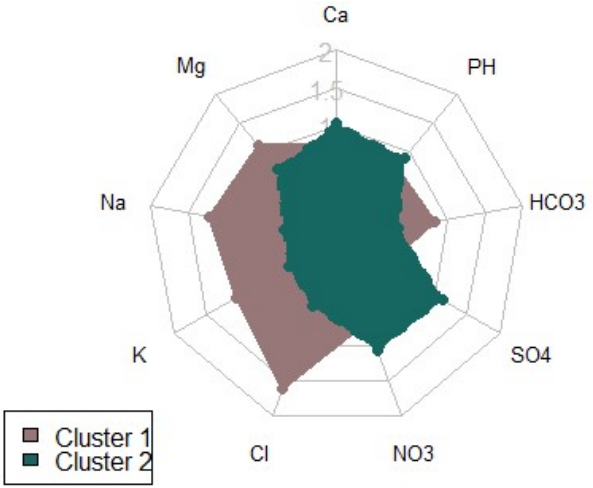

Figure 35: Graphique en étoile des écart-types

Classification non-supervisée : Analyse et interprétation des résultats

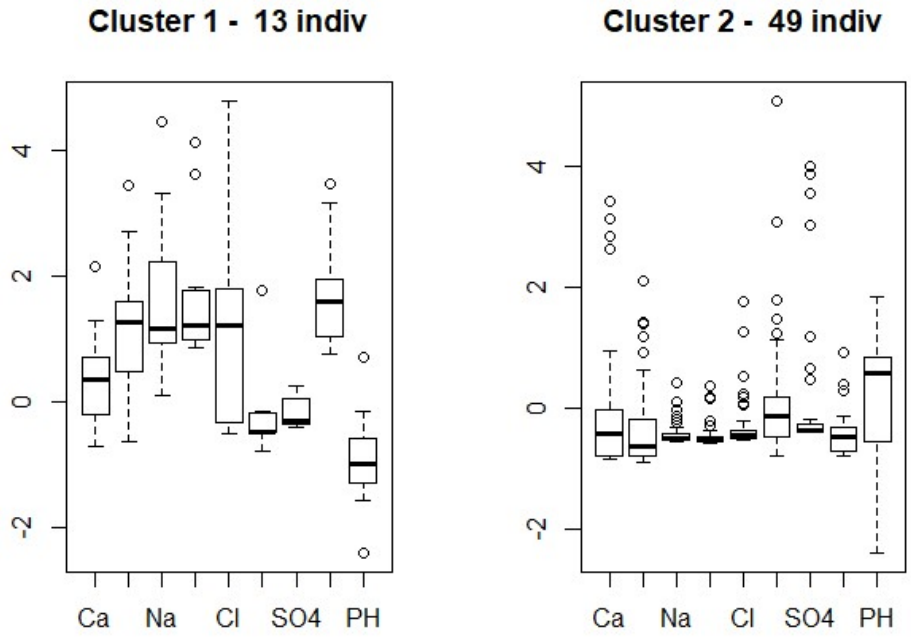

Figure 36: Diagramme en boîte des variables des deux clusters

Classification non-supervisée : Analyse et interprétation des résultats

Table 1: Matrice de confusion entre les deux clusters de kmeans et les trois groupes hypothétiques

	gaz français	plat français	plat marocain
cluster 1	12	1	0
cluster 2	6	36	7

Calcul du taux d'erreur

Validation croisée stratifiée

- On divise le jeu de données en q blocs "harmonieux"
- On entraine le modèle sur q-1 blocs
- On applique le modèle sur le q_ième bloc

Validation croisée LOO

- Validation croisée
- nombre de blocs = nombre d'observations

Bootstrap

- Echantillonnages : Tirages avec remises
- Calcul des erreurs en faisant la moyenne des erreurs à chaque tirage

Table 2: *Matrice de confusion, Knn : Méthode test/train*

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

	k=6		k=7		k=8		k=9	
	plat	gaz	plat	gaz	plat	gaz	plat	gaz
plat	7	1	7	1	7	1	7	1
gaz	1	3	0	4	1	3	1	3
Erreur	0.1667		0.0833		0.1667		0.1667	

Arbres binaires : CART

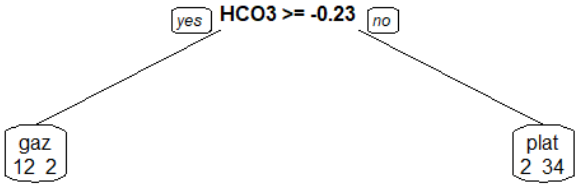

Figure 37: Arbre optimal

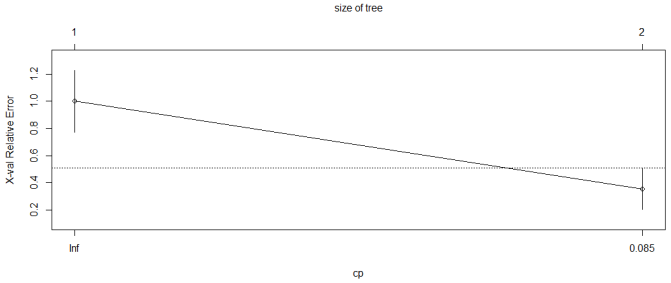

Figure 38: Erreur et écart-type des sous-arbres élagués

Classification supervisée : Taux d'erreur

	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 39: Tableau des taux d'erreur