Analyses factorielles :
classifications et partitionnements

Marie-Camille CAUMON
Ingénieur de recherche
GeoRessources - UMR 7359
Entrée 3B - bureau A508
+33 3 72 74 55 37
marie-camille.caumon@univ-lorraine.fr
http://georessources.univ-lorraine.fr/

Plan du cours

1. Introduction
2. Classifications : principes
3. Classification ascendante hiérarchique
    1. Principe
    2. Dissimilarité
    3. Règles d’agrégation
    4. Inertie
    5. Construction du dendrogramme
4. Méthode de partitionnement k-means

Principes :
- Les individus d'une même classe sont les plus similaires possibles
= variance inter-classe faible

- Les classes sont les plus dissemblables possibles
= variance inter-classe forte

approche non supervisée = aucun individu n'est attribué à une classe au départ. La répartition évolue avec le nombre de classe

approche supervisée = on doit connaître pour chaque individu la classe dans laquelle il peut rentrer.

Classification ascendante hiérarchique = méthode pas à pas

1er pas : on a autant de groupe que d'individus

Dernier pas : un seul groupe

-> Chaque classe d'une partition appartient à une classe de partition suivante

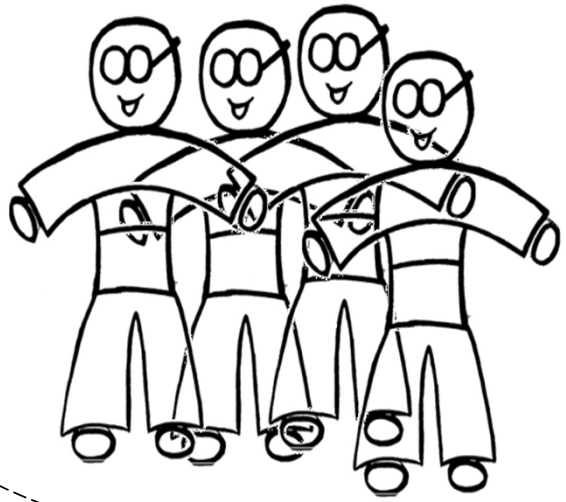
La partition en 4 classes et obtenue en regroupant 2 classes de la partition k+1

Résultat : dandrogramme

1. Introduction

population

variance intra-classe faible

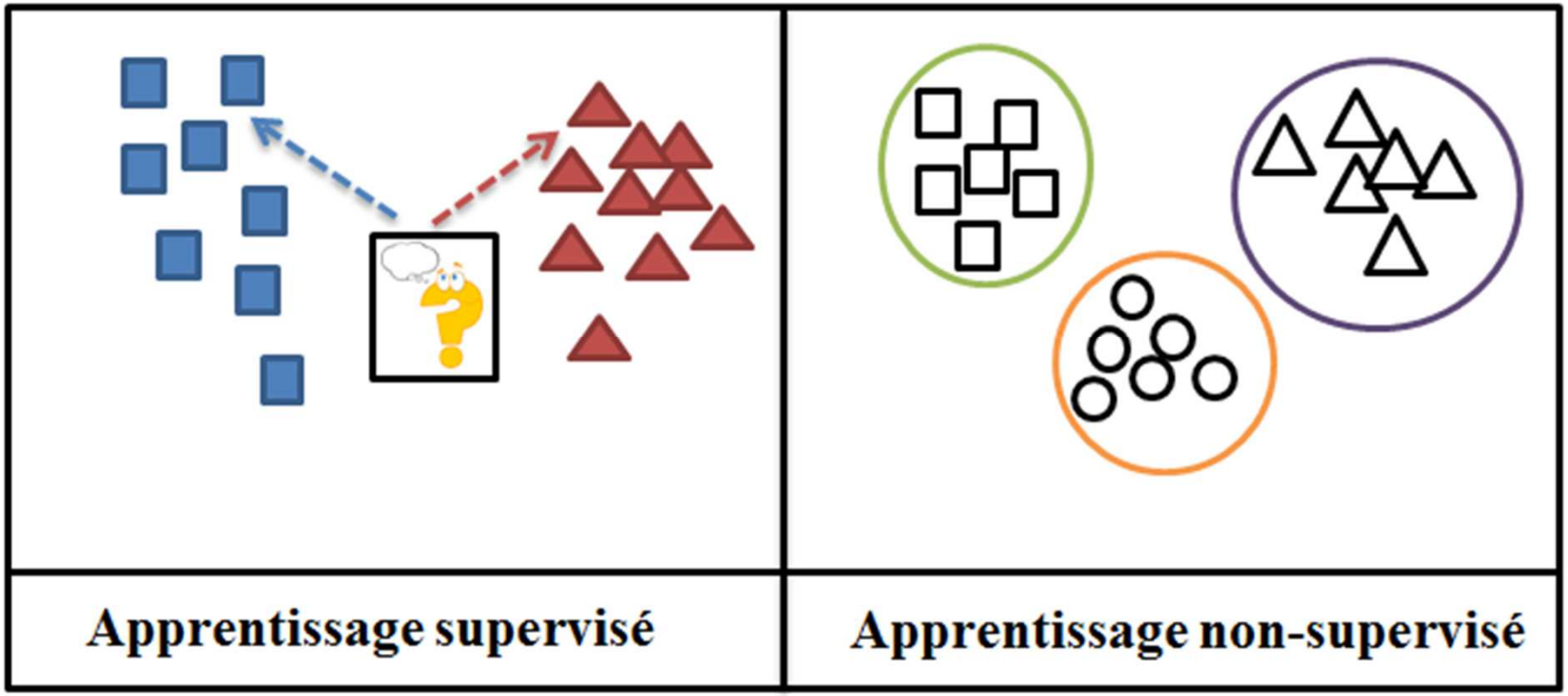


variance intra-classe faible

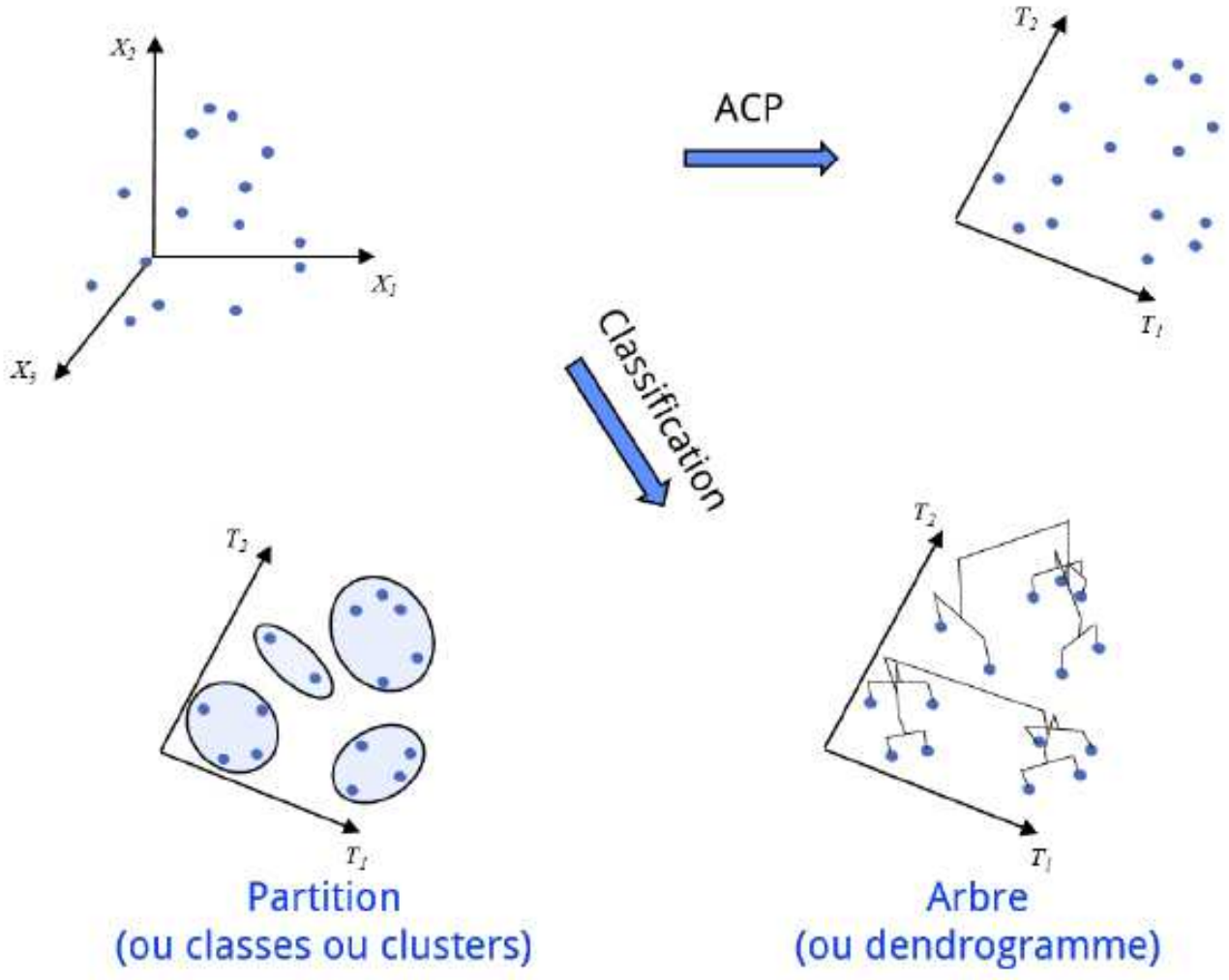
Classe 1

Classe 2

1. Introduction



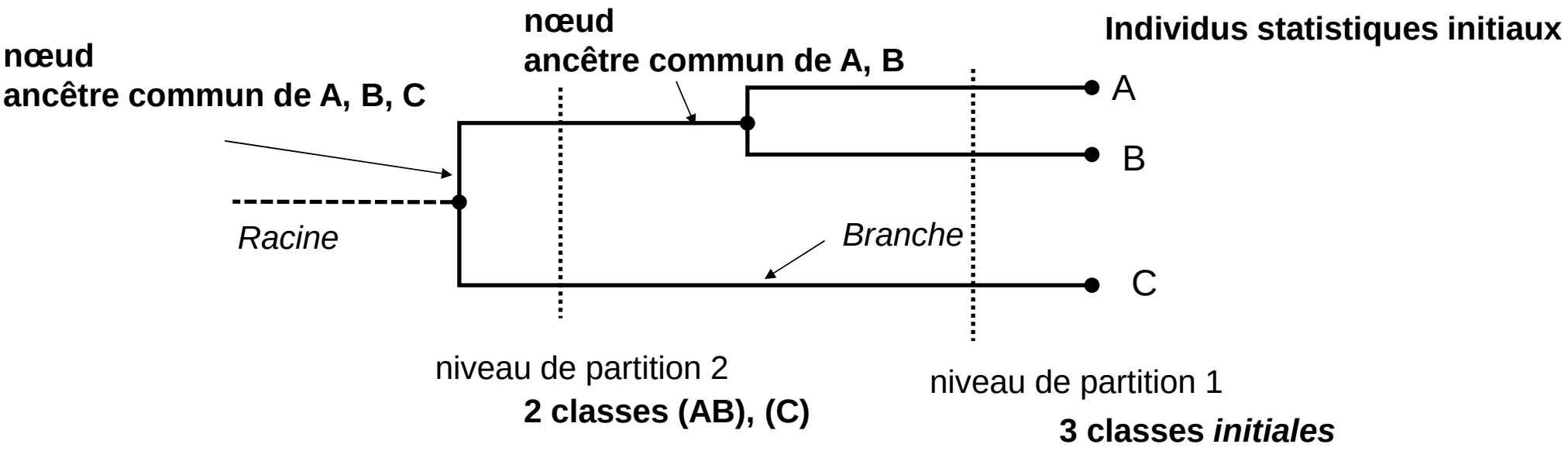
1. Introduction



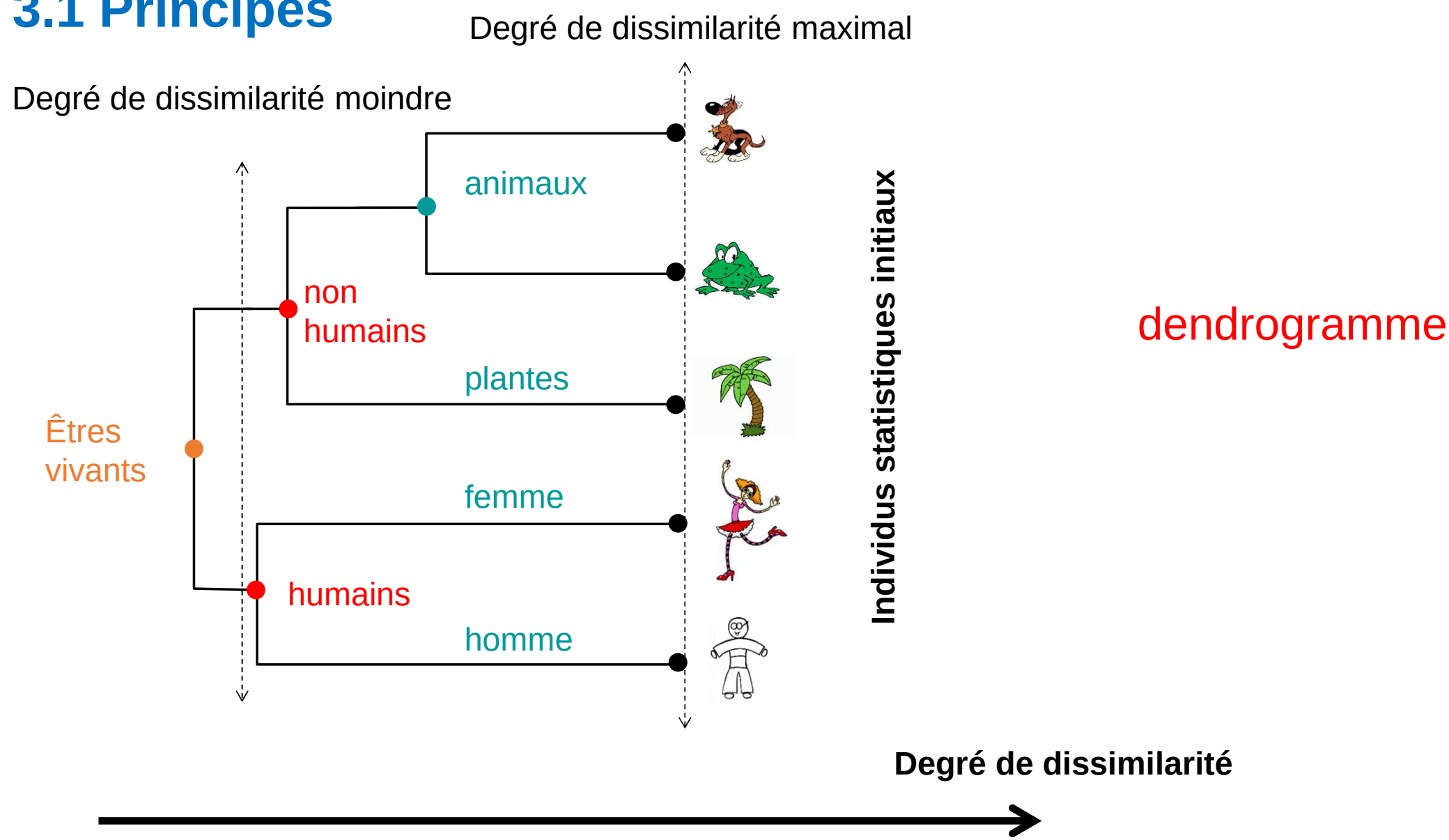
Plan du cours

1. Introduction
2. Classifications : principes
3. Classification ascendante hiérarchique
    1. Principe
    2. Dissimilarité
    3. Règles d’agrégation
    4. Inertie
    5. Construction du dendrogramme
4. Méthode de partitionnement k-means

3.1 Principes

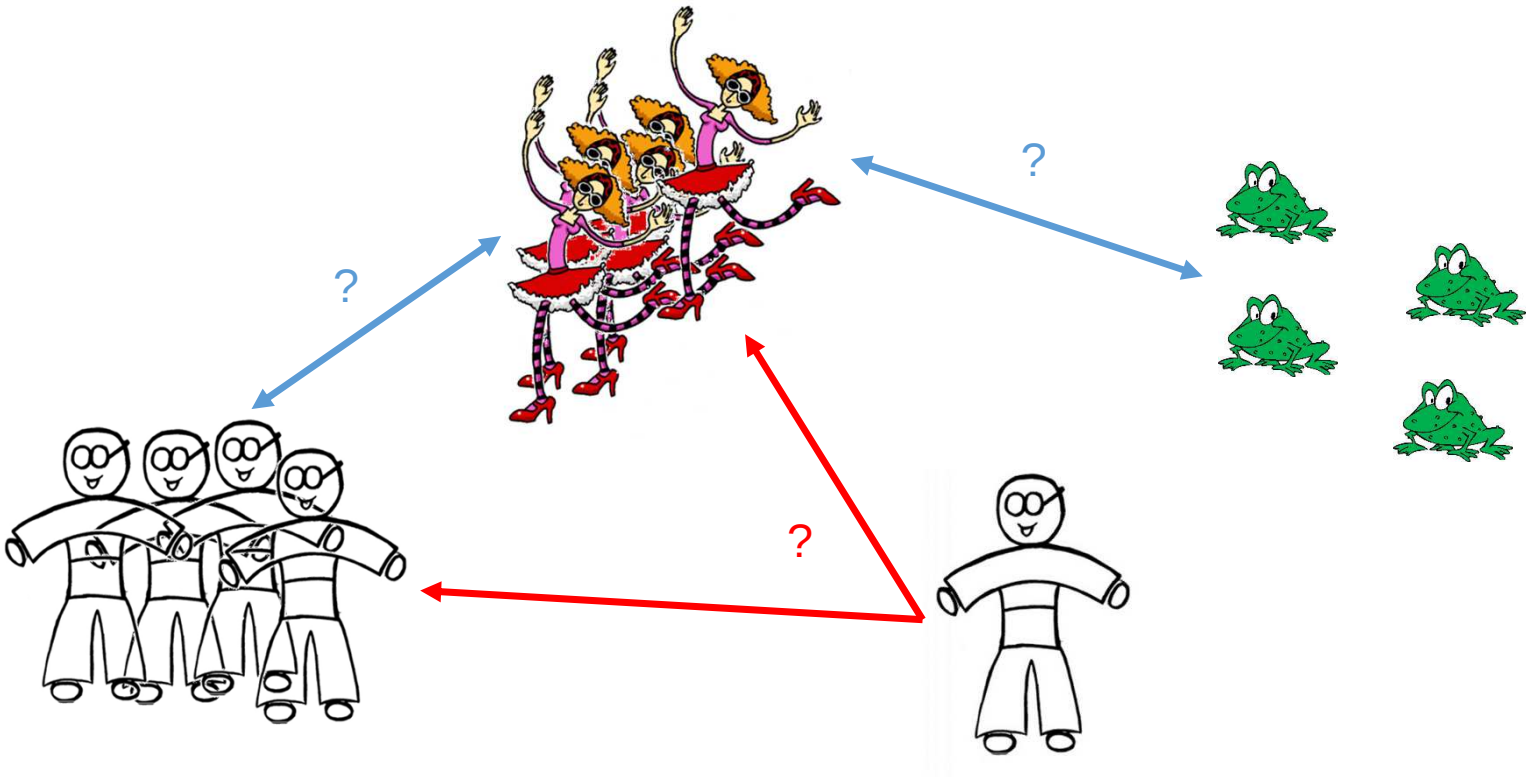


3.1 Principes



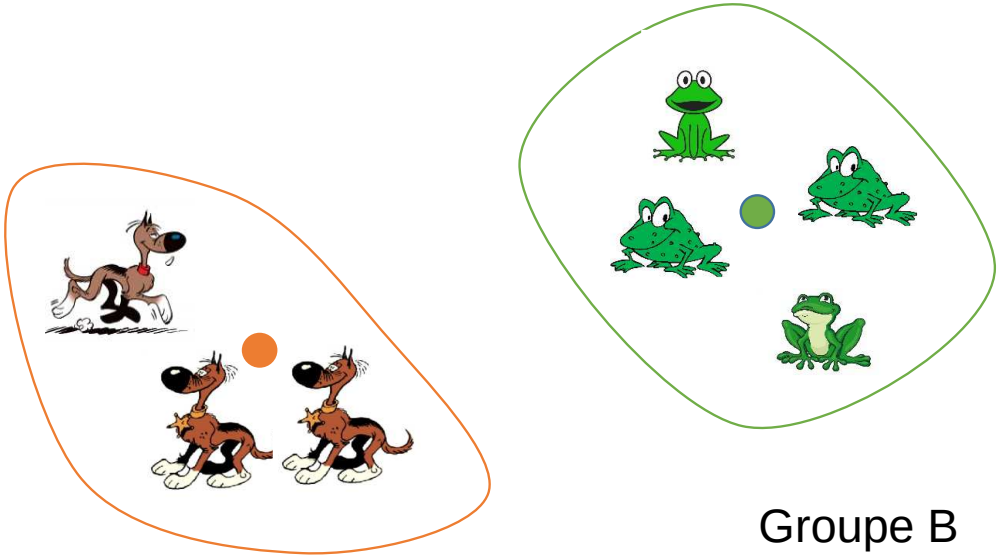
3.1 Principes

Critère 1 : critère de dissimilarité
Critère 2 : critère d'agrégation



Dissimilarité : mesure de distance entre les individus. -> distance euclidienne

3.3 Règles d’agrégation



Groupe A
g_A ; N_A

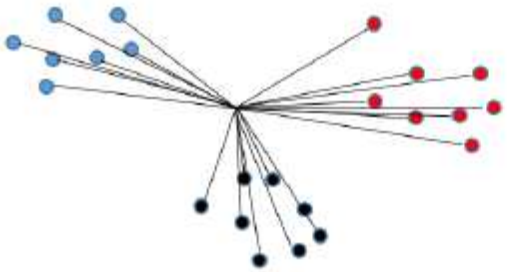
Groupe B
g_B ; N_B

- Distance des liens moyen
- Distance de gravité entre les points rouges et verts

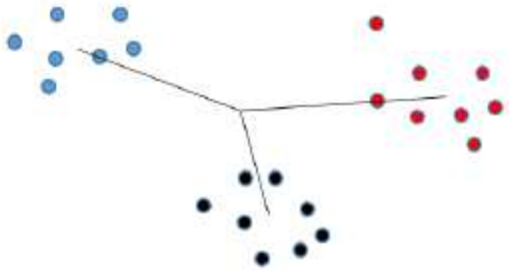
Critère de Ward minimiser l'intertie intraclasse lors de l'agrégation. -> l'inertie augmente d'une valeur(Na x NB) /(Na + NB) d²GAGB
= > Distance de Ward

3.4 Inertie

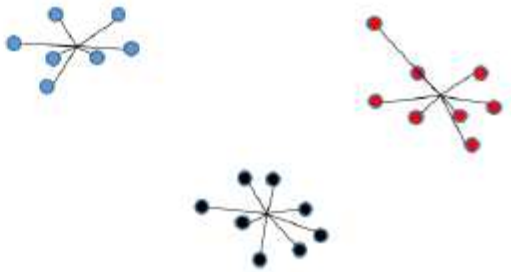
$$I_T = I_B + I_W$$



(a) Inertie totale I_T

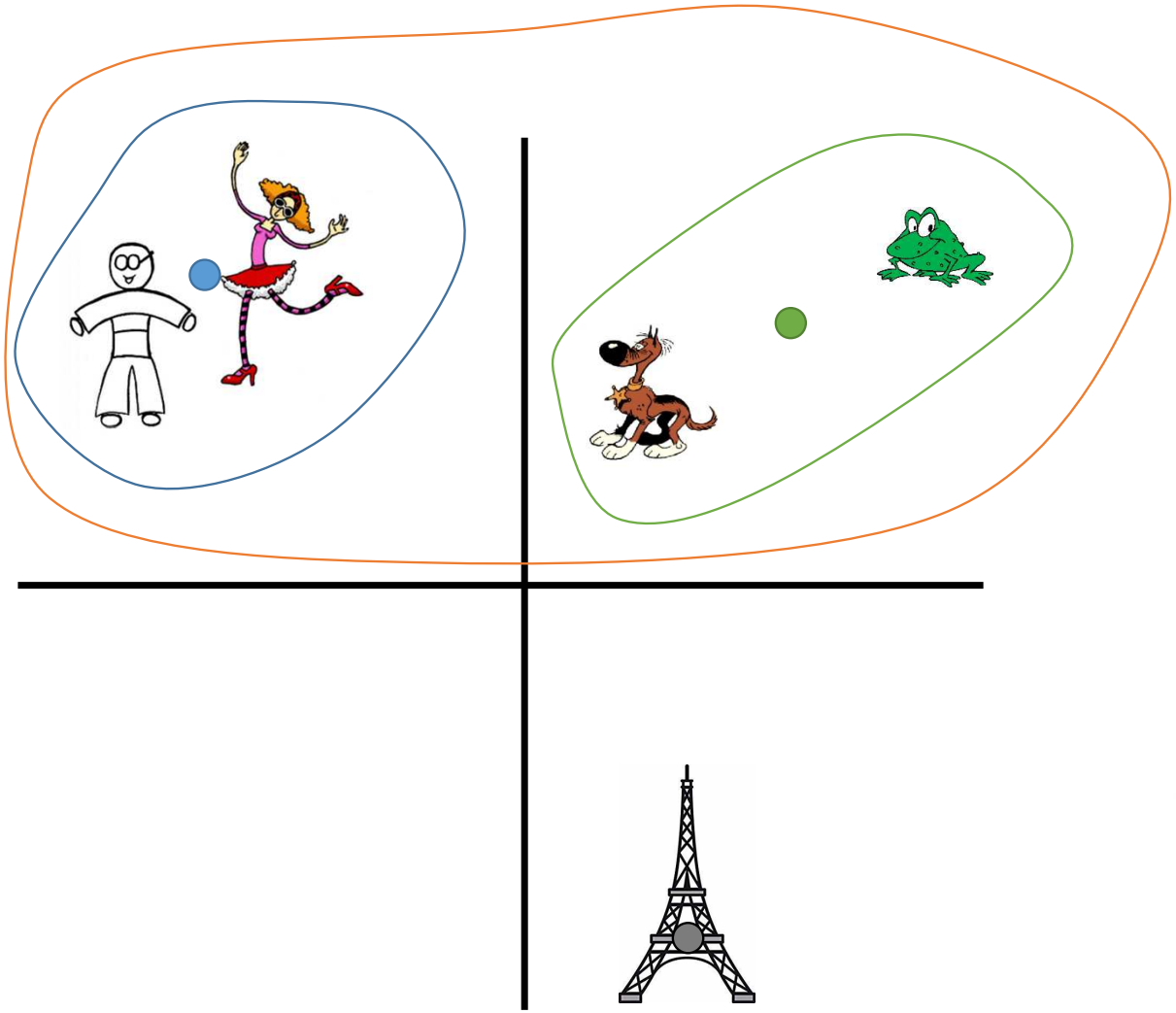


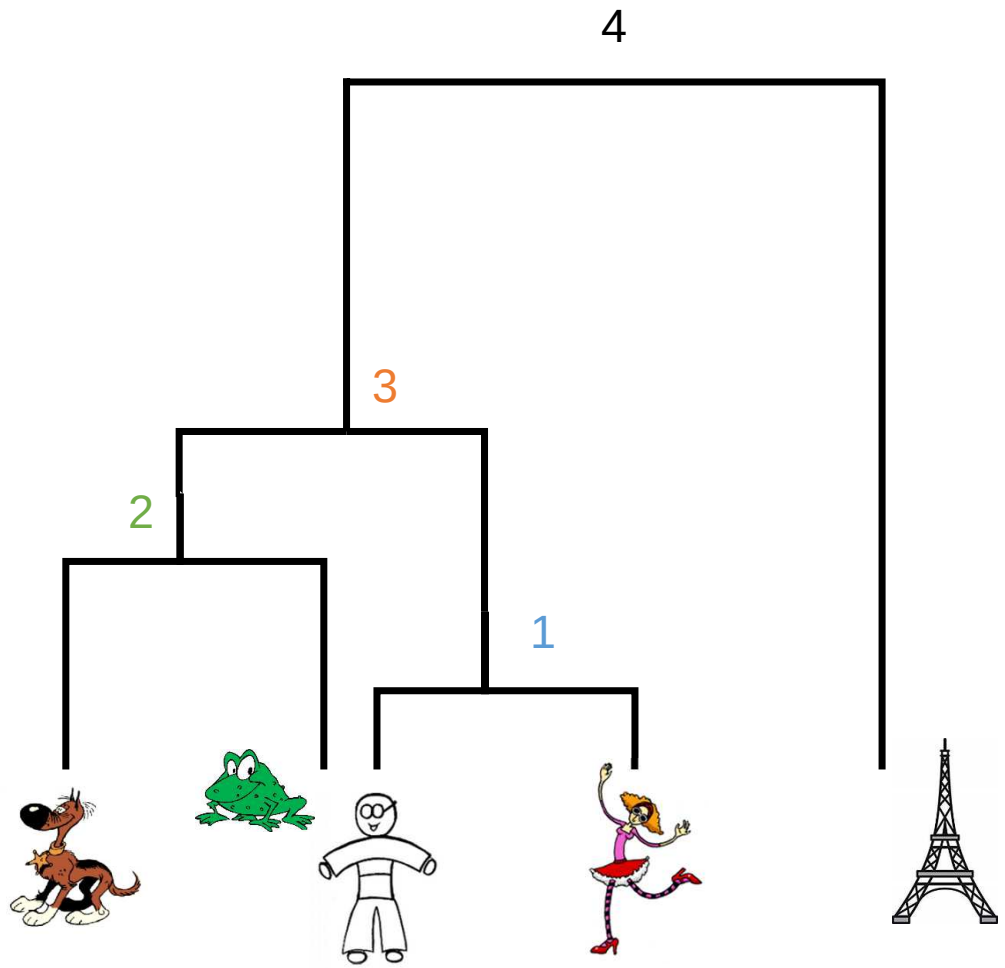
(b) Inertie inter-classes I_B



(c) Inertie intra-classes I_W

3.5 Construction du dendrogramme





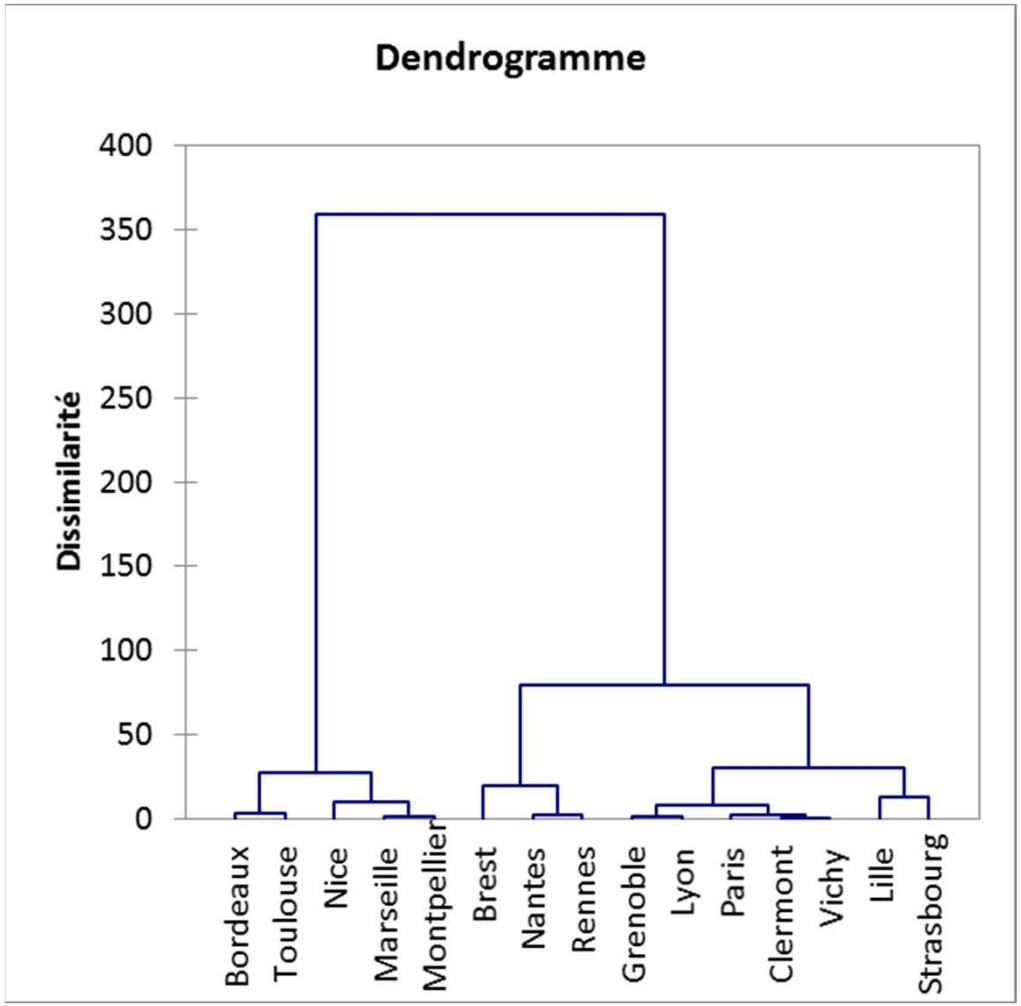
Exemple

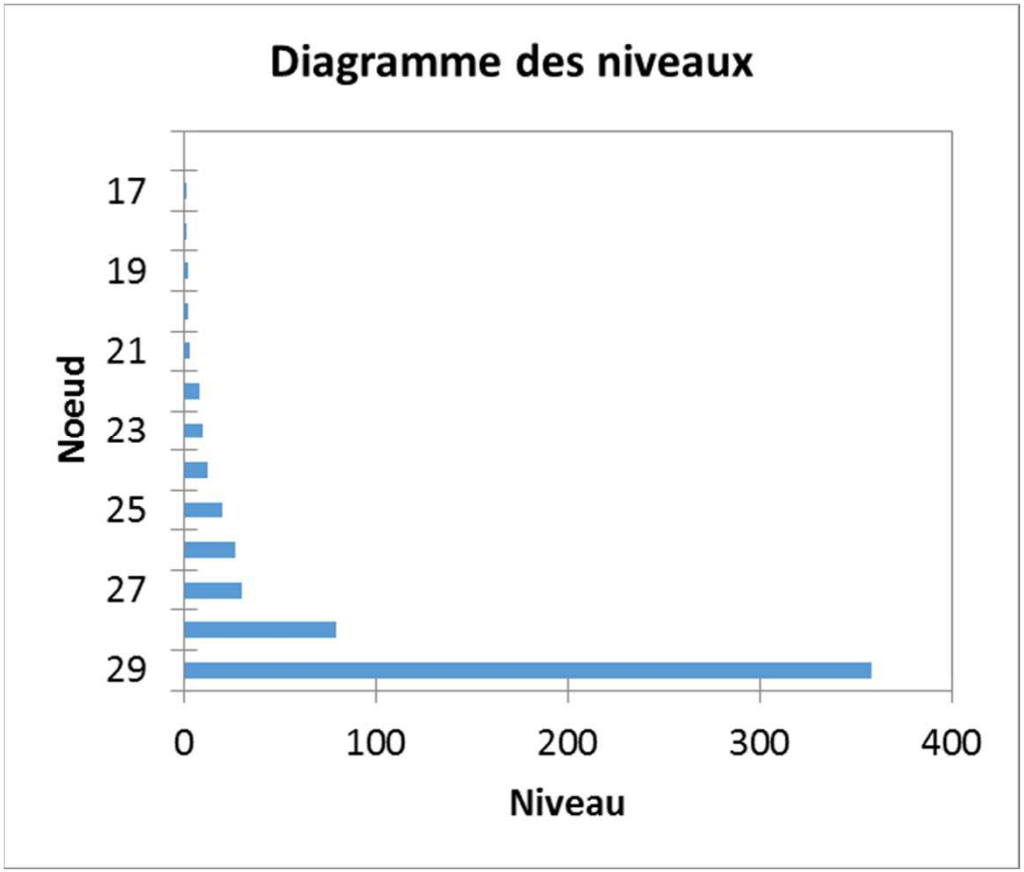
- 15 individus : 15 villes de France
- 12 variables : températures moyennes mensuelles sur 30 ans

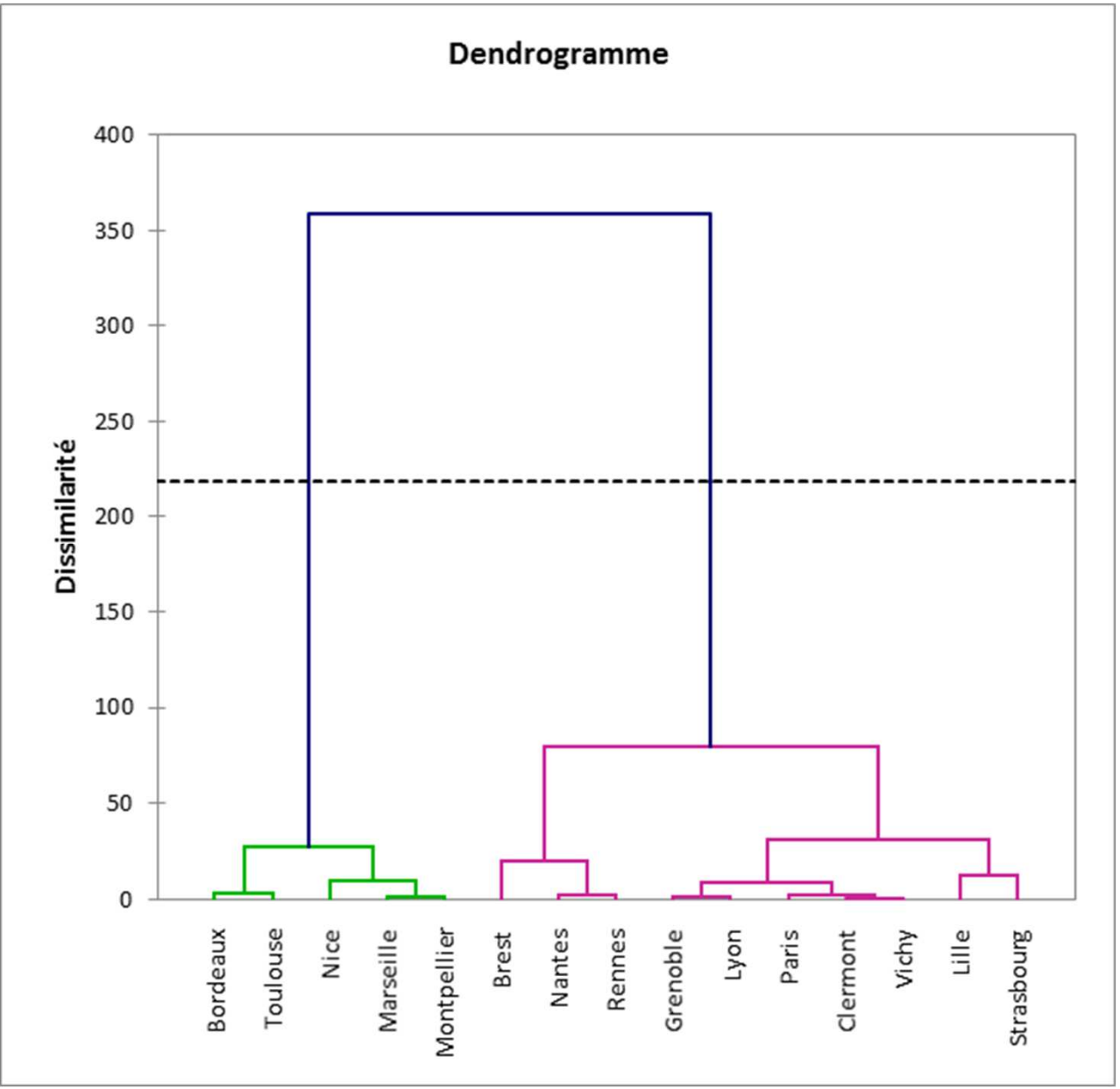
ville	janvier	février	mars	avril	mai	juin	juillet	août	septembre	octobre	novembre	décembre
Bordeaux	5.6	6.6	10.3	13	16	19	20.9	21	18.6	13.8	9.1	6.2
Brest	6.1	5.8	7.8	9.2	12	14	15.6	16	14.7	12	9	7
Clermont	2.6	3.7	7.5	10	14	17	19.4	19	16.2	11.2	6.6	3.6
Grenoble	1.5	3.2	7.7	11	15	18	20.1	20	16.7	11.4	6.5	2.3
Lille	2.4	2.9	6	8.9	12	15	17.1	17	14.7	10.4	6.1	3.5
Lyon	2.1	3.3	7.7	11	15	19	20.7	20	16.9	11.4	6.7	3.1
Marseille	5.5	6.6	10	13	17	21	23.3	23	19.9	15	10.2	6.9
Montpellier	5.6	6.7	9.9	13	16	20	22.7	22	19.3	14.6	10	6.5
Nantes	5	5.3	8.4	11	14	17	18.8	19	16.4	12.2	8.2	5.5
Nice	7.5	8.5	10.8	13	17	20	22.7	23	20.3	16	11.5	8.2
Paris	3.4	4.1	7.6	11	14	18	19.1	19	16	11.4	7.1	4.3
Rennes	4.8	5.3	7.9	10	13	16	17.9	18	15.7	11.6	7.8	5.4
Strasbourg	0.4	1.5	5.6	9.8	14	17	19	18	15.1	9.5	4.9	1.3
Toulouse	4.7	5.6	9.2	12	15	19	20.9	21	18.3	13.3	8.6	5.5
Vichy	2.4	3.4	7.1	9.9	14	17	19.3	19	16	11	6.6	3.4

Quelles villes ont des profils météo similaires ?
Comment caractériser les groupes de ville ?

Où faire la coupure ?

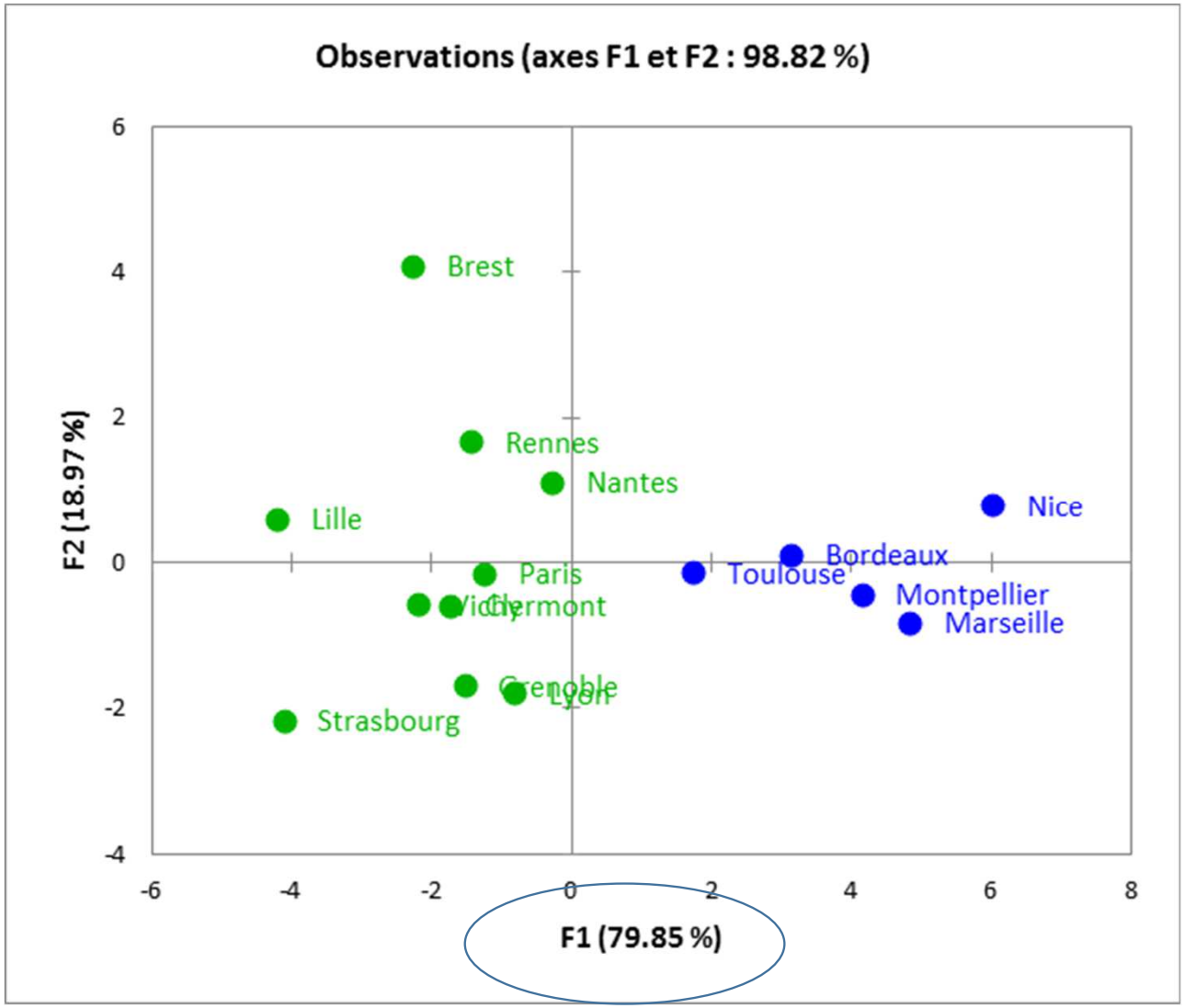






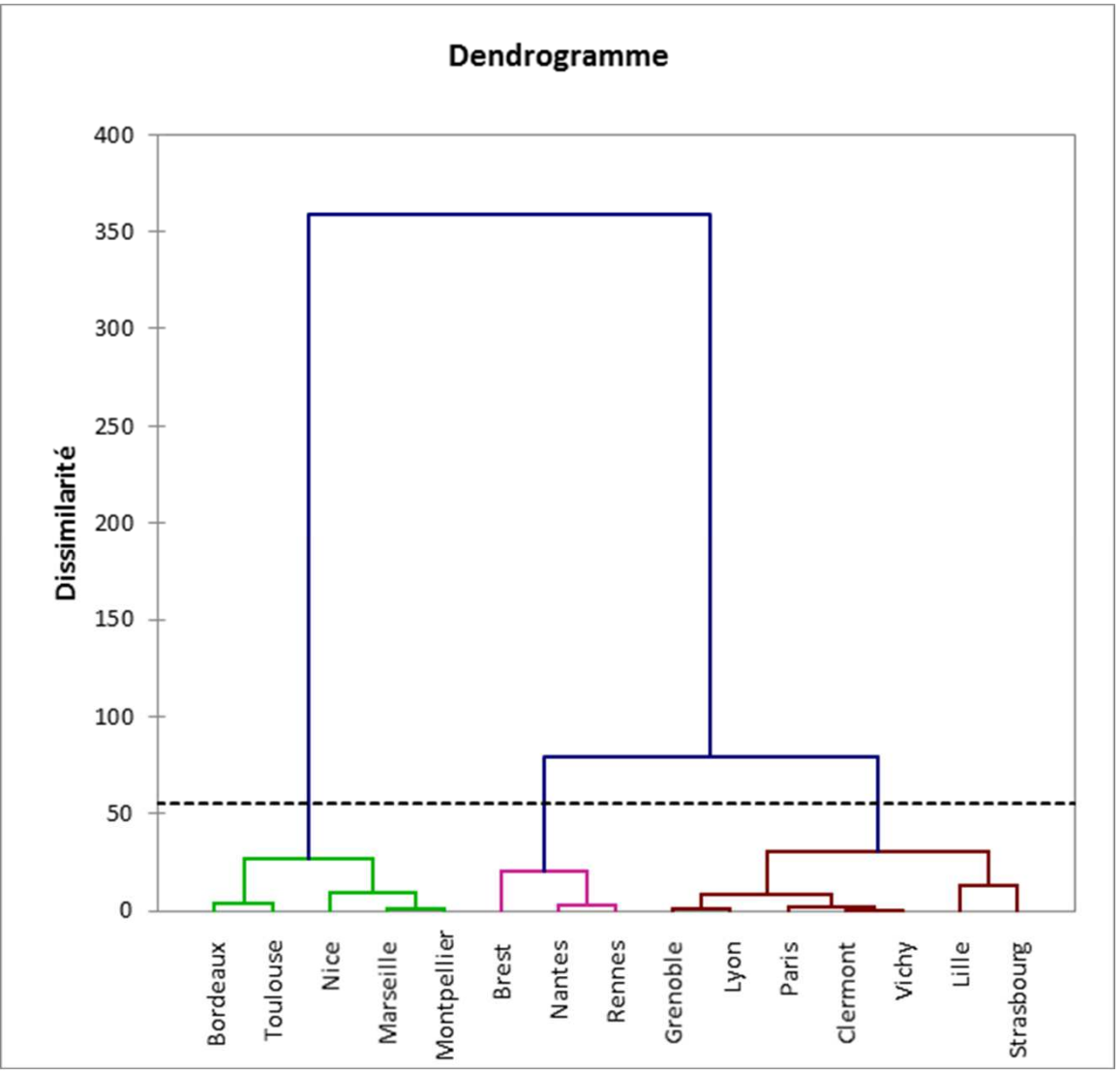
Combien de groupes ?

2 groupes :
Inertie inter-classe = 62 % de la variance



Comparaison ACP / CAH :
Discrimination selon axe F1

Axe F2 ?



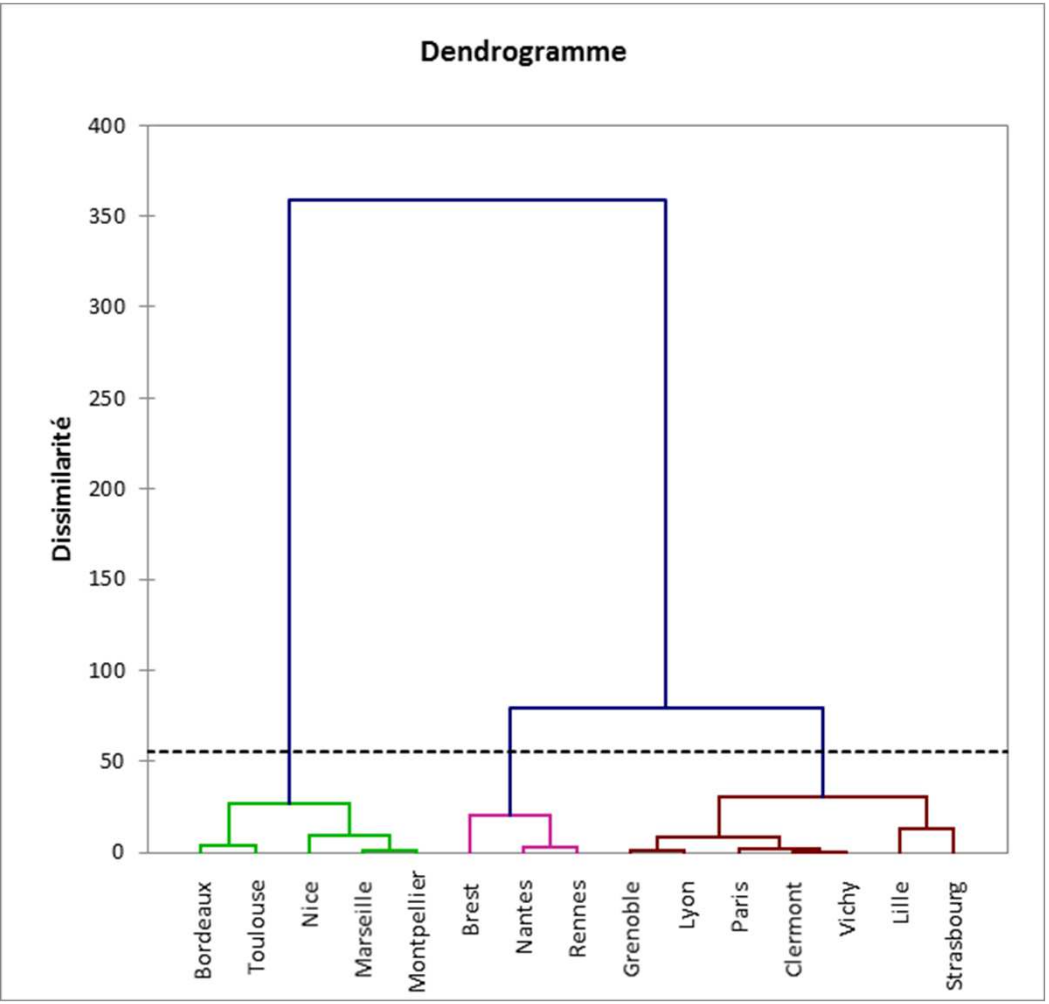
$2 \rightarrow 3$ classes

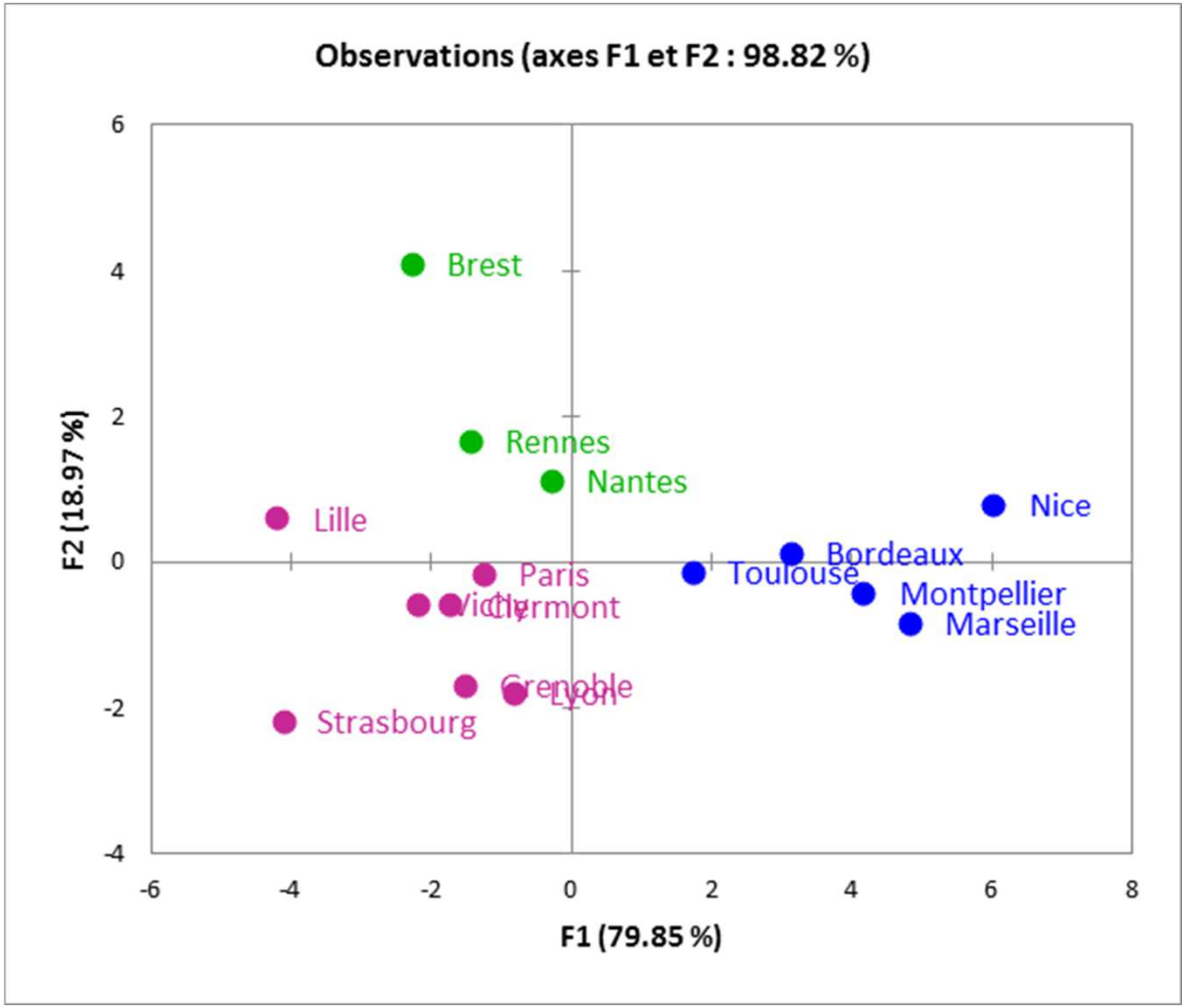
Découpage des villes « froides » en deux groupes.

Inertie inter-classe = 75 %

Algorithme des constructions des dendrogrammes
1. Construit la position dont les classes contiennent 1 élément
2. Agréger deux classes selon le critère choisi -> 1 seule classe

La partition finale s'obtient en définissant un niveau de coupure =
- correspond à un saut important de l'indice d'agrégation
- critère d'inertie

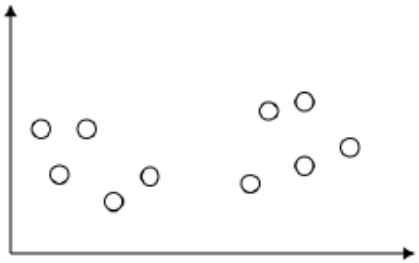




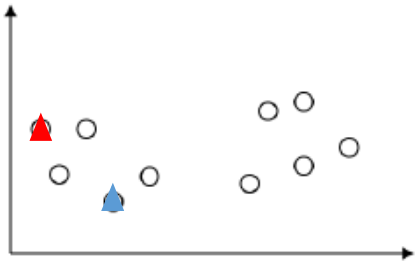
Plan du cours

1. Introduction
2. Classifications : principes
3. Classification ascendante hiérarchique
    1. Principe
    2. Dissimilarité
    3. Règles d’agrégation
    4. Inertie
    5. Construction du dendrogramme
4. Méthode de partitionnement k-means

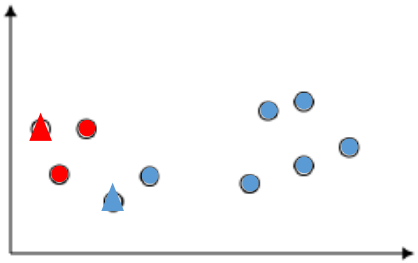
Méthodes de partitionnement



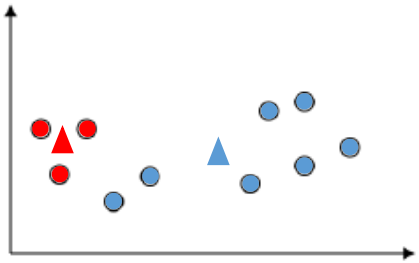
individus



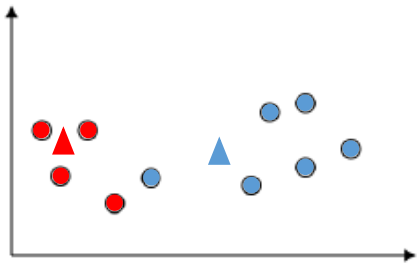
1.Tirage aléatoire des noyaux initiaux



2.Affectation de chaque individu au centre de gravité le plus proche

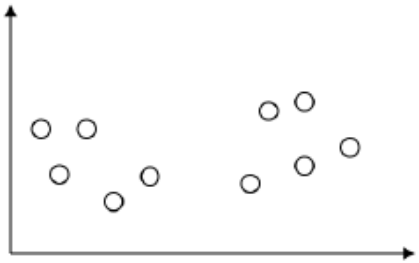


3. Calcul des nouveaux centres de gravité

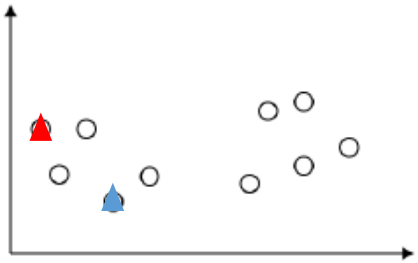


2. Affectation de chaque individu au centre de gravité le plus proche

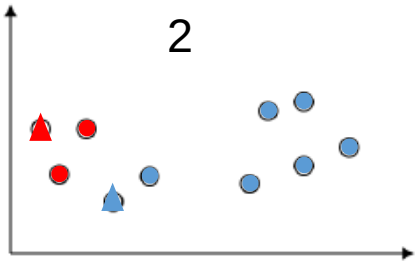
4. Principe des k-means

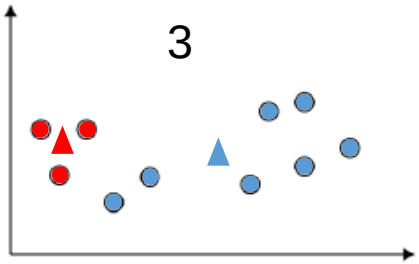


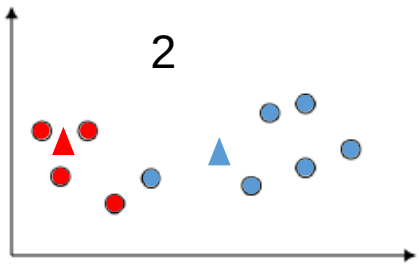
individus

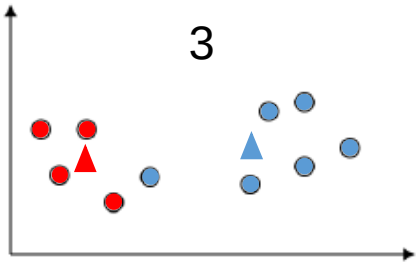


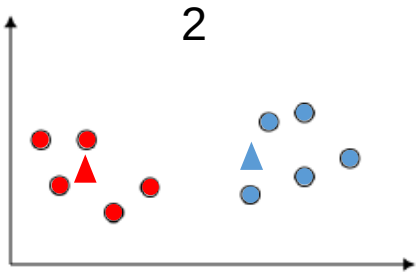
Tirage aléatoire des noyaux initiaux

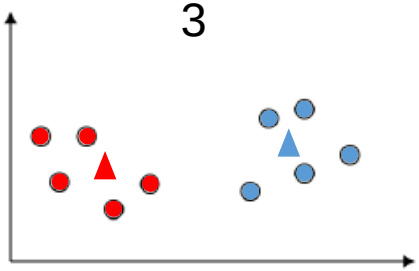


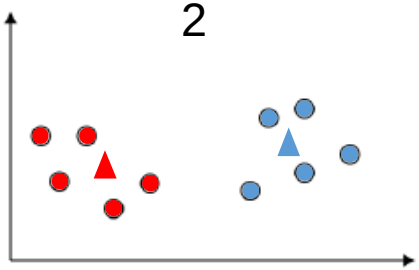












Plus de modification : arrêt de l'agorithme

Exemple

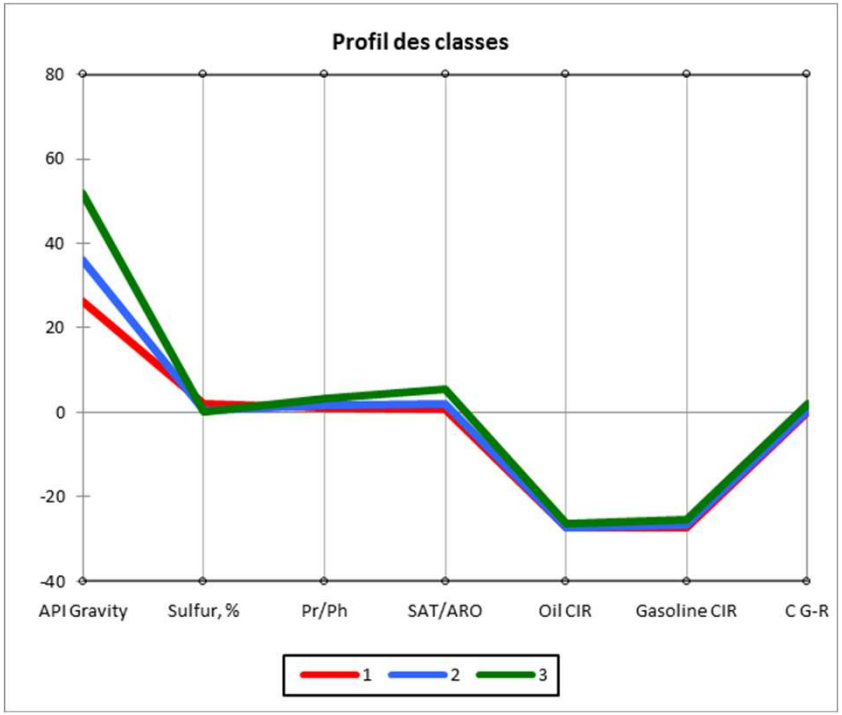
Compilation des paramètres d'une série d'échantillons de pétrole brut dont la source est connue.

No.	API Gravity	Sulfur, %	Pr/Ph	SAT/ARO	Oil CIR	Gasoline CIR	C G-R	rock	
1	24.6	1.69	1.1	1.1	-26.23	-26.3	-0.27	carbonate	Carbonate, Deltaic, Marine Shale
2	27	1.58	0.95	1.1	-26.62	-26.89	-0.33	carbonate	
3	28.1	1.53	1.02	1.2	-26.02	-26.21	-0.39	carbonate	API gravity
4	29.5	3.1	0.7	0.8	-26.1	-27.16	-1.42	carbonate	Pr/Ph = pristane/phytane ratio;
5	32.2	2.61	0.65	0.8	-26.24	-27.2	-1.09	carbonate	SAT/ARO = saturates to aromatics ratio;
6	33.6	2.27	0.75	0.7	-26.5	-27.19	-0.93	carbonate	Oil CIR = whole-oil carbon isotope ratio;
7	31.7	2.52	0.7	0.9	-26.24	-27.07	-1.12	carbonate	Gasoline CIR = carbon isotope ratio of gasoline fraction;
8	33	1.71	0.71	1.2	-26.27	-27	-0.97	carbonate	C G-R = difference in carbon isotope ratio between gasoline fraction and residuum.
9	34	1.95	0.62	1.2	-26.3	-26.95	-0.96	carbonate	From Chung, et al, 1994, Table 1.
10	28	2.78	0.67	0.7	-26.57	-27.46	-0.83	carbonate	
11	25.5	2.26	0.82	0.9	-25.59	-25.8	-0.6	carbonate	
12	35.4	1.03	0.85	1.3	-25.25	-25.65	-0.5	carbonate	
13	35.1	1.39	0.58	1.1	-25.06	-25.52	-0.54	carbonate	

63 échantillons, 3 sources, 6 variables

K-means sans centrage et réduction des données :

Les individus sont mal répartis :
Variables dans des unités différentes avec des variances et des moyennes très différentes.

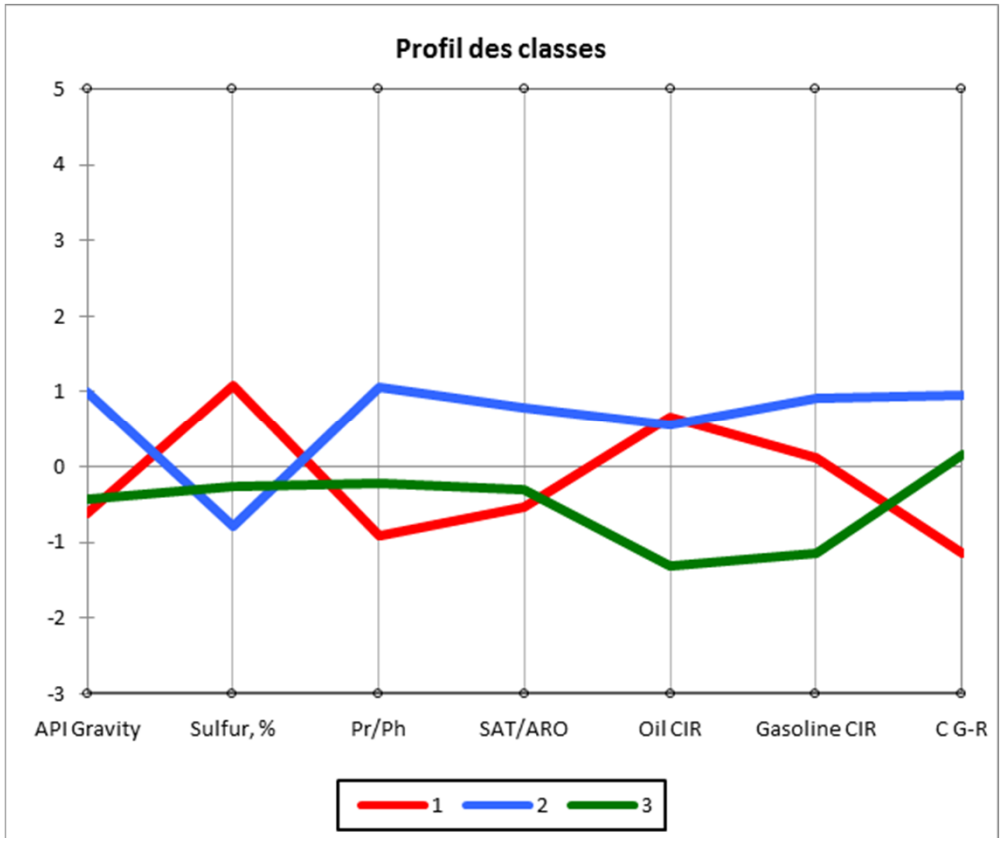


La variable « API gravity » domine.

Classe	1	2	3
Objets	22	27	14
Somme des poids	22	27	14
Variance intra-classe	20.087	18.322	44.028
Distance minimale au barycentre	1.200	1.543	3.207
Distance moyenne au barycentre	4.064	3.894	5.952
Distance maximale au barycentre	8.360	8.013	11.296
	carbonate	carbonate	deltaic
	carbonate	carbonate	deltaic
	carbonate	carbonate	deltaic
	carbonate	carbonate	deltaic
	carbonate	carbonate	deltaic
	carbonate	carbonate	deltaic
	carbonate	carbonate	deltaic
	carbonate	carbonate	deltaic
	carbonate	deltaic	deltaic
	carbonate	deltaic	deltaic
	carbonate	deltaic	deltaic
	carbonate	deltaic	deltaic
	carbonate	deltaic	deltaic
	marine shale	deltaic	marine shale
	marine shale	deltaic	
	marine shale	deltaic	
	marine shale	deltaic	
	marine shale	marine shale	
	marine shale	marine shale	
	marine shale	marine shale	
	marine shale	marine shale	
	marine shale	marine shale	
		marine shale	
		marine shale	
		marine shale	
		marine shale	
		marine shale	

K-means après centrage et réduction des données :

Décomposition de la variance pour la classification optimale :		
	Absolu	Pourcentage
Intra-classe	2.543	36.33%
Inter-classes	4.457	63.67%
Totale	7.000	100.00%



Classe	1	2	3
Objets	21	22	20
Somme des poids	21	22	20
Variance intra-classe	1.542	4.291	1.664
Distance minimale au barycentre	0.561	0.657	0.449
Distance moyenne au barycentre	1.113	1.811	1.184
Distance maximale au barycentre	2.626	4.256	2.013
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	marine shale
	carbonate	deltaic	
		deltaic	