Introduction à la data science

ACP – Classification - Arbres – Forêts alétoires

François Husson
husson@agrocampus-ouest.fr

Hammamet, 25 au 27 novembre de 2021

Department Statistics & Computer science, L'Institut Agro

Présentation

- Recherche : données manquantes, analyse de données, tableaux multiples
- Enseignement : cursus d'ingénieur, master *science des données*
- MOOC en analyse de données, sensométrie, plan d'expériences
- Formation continue : statistique avec R, analyse de données

2018

2nd ed: 2017

1st ed: 2011

2nd ed: 2016

1st ed: 2009

2nd ed: 2013

1st ed: 2005

2013

3rd ed: 2012

2nd ed: 2010

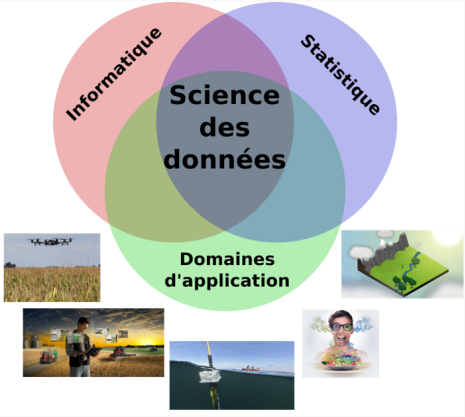
1st ed: 2008

2012

Packages: FactoMineR - missMDA - SensoMineR - Factoshiny - FactoInvestigate - RcmdrPlugin.FactoMineR

Plan

Science des données



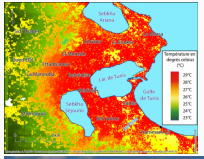
Science des données et intelligence artificielle

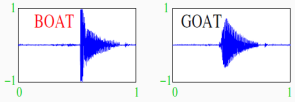
s'intéressent au management et à l'analyse des données sous toutes leurs formes (massives, complexes, hétérogènes, incertaines, etc.) et à leurs applications dans des secteurs clés (santé, environnement, transport, défense, etc.)

Deux principaux défis :

- passage à l'échelle, maîtrise de la complexité et explicabilité des algorithmes
- mise à profit de la dimension multidisciplinaire alliant méthodes, technologies et usages

Science des données : domaine d'application



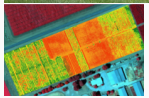


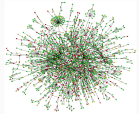
- Médecine
  - Identification de biomarqueurs en génomique
  - Classification de parcours de soin
  - Prévision du besoin de transfusion d'un patient
- Environnement
  - Modélisation de la température de surface de la mer
  - Cartographie et analyse de la qualité des eaux
- Energie
  - Suivi et pilotage en temps réel de la production, la consommation, et le stockage de l'électricité
  - Pilotage intelligent des infrastructures et bâtiments pour une optimisation énergétique
- Vie courante
  - Reconnaissance de l'écriture, de la parole, traduction (deepl)
  - Prévision de circulation et de temps de trajet (Google maps, ...)
  - Système de recommandation (Netflix, Amazon, ...)

Science des données en agronomie

Un système agronomique est complexe :

- Multi-échelle (en temps en espace) et multifactoriel
- Adaptatif et naturel





- Suivi en temps réel d'exploitations équipées de capteurs pour conduire un élevage ou une vigne
- Analyse de données de drones pour raisonner la fertilisation
- Utilisation de la spectrométrie et l'imagerie en nutrition
- Estimation de la hauteur de l'herbe par images satellites
- Détection de bactéries à partir d'images
- Analyse d'un réseau de protéines
- Représentation et analyse d'un réseau social pour comprendre le comportement du consommateur

Un peu d'histoire - voir Besse, P. (2018)

Période	Mémoire	Ordre de grandeur
1940-70	Octet	$n = 30, p \leq 10$
1970	kO	$n = 500, p \leq 10$
1980	MO	Machine Learning
1990	GO	Data-Mining
2000	TO	$p > n$, apprentissage statistique
2010	PO	n explose, cloud, cluster...
2013	??	Big data
2017	??	Intelligence artificielle...

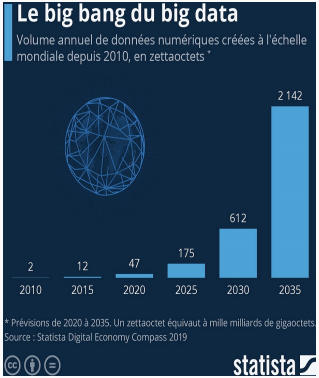
Un peu d'histoire - voir Besse, P. (2018)

Période	Mémoire	Ordre de grandeur
1940-70	Octet	$n = 30, p \leq 10$
1970	kO	$n = 500, p \leq 10$
1980	MO	Machine Learning
1990	GO	Data-Mining
2000	TO	$p > n$, apprentissage statistique
2010	PO	n explose, cloud, cluster...
2013	??	Big data
2017	??	Intelligence artificielle...

Conclusion

Capacités informatiques $\Longrightarrow$ Data Mining $\Longrightarrow$ Apprentissage statistique $\Longrightarrow$ Big Data $\Longrightarrow$ Intelligence artificielle...

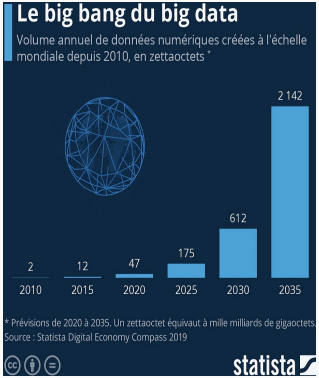
Les 3V ... ou 6V de la data science



3V de la data ...

- Volume : une avalanche de données : en 2020, 47 000 milliards gigaoctets de données crées/jour
- Variété : données de formes différentes : tweets, images, vidéos, articles, données médicales, ...
- Vélocité : traiter ces volumes et ces variétés de données en un temps réduit

Les 3V ... ou 6V de la data science



3V de la data ...

- Volume : une avalanche de données : en 2020, 47 000 milliards gigaoctets de données crées/jour
- Variété : données de formes différentes : tweets, images, vidéos, articles, données médicales, ...
- Vélocité : traiter ces volumes et ces variétés de données en un temps réduit

... voire même 6V de la data :

- Véracité : s'assurer de la fiabilité de la donnée est donc primordial
- Valeur : apporter de la valeur ajoutée en répondant aux objectifs médicaux, environnementaux, marketing, commerciaux
- Visualisation : restituer de manière lisibles et simples

Plan

Quel type de données ?

L'ACP s'intéresse à des tableaux de données rectangulaires avec des individus en lignes et des variables quantitatives en colonnes

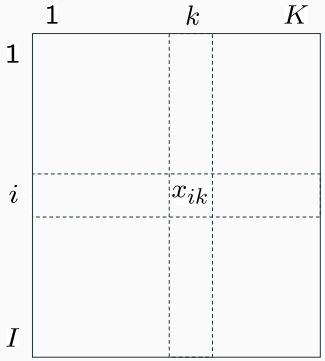


Tableau de données en ACP

Quel type de données ?

L'ACP s'intéresse à des tableaux de données rectangulaires avec des individus en lignes et des variables quantitatives en colonnes

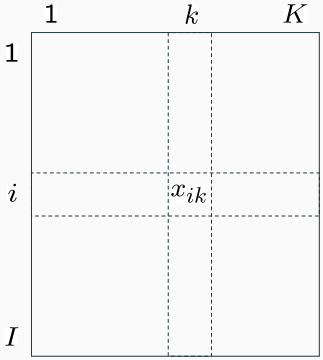


Tableau de données en ACP

- Ecologie : concentration du polluant *k* dans la rivière *i*
- Economie : valeur de l'indicateur *k* pour l'année *i*
- Génétique : expression du gène *k* pour le patient *i*
- Marketing : valeur d'indice de satisfaction *k* pour la marque *i*
- Sociologie : temps passé à l'activité *k* par les individus de la CSP *i*
- etc.

$\Rightarrow$ Il existe de très nombreux tableaux comme cela

Les données température en Tunisie

- 16 individus (lignes) : villes de Tunisie
- 12 variables (colonnes) : 12 températures mensuelles moyennes

	Janv	Fév	Mar	Avr	Mai	Juin	Juil	Août	Sept	Oct	Nov	Déc
Bizerte	12,3	11,9	13,5	15,6	18,6	22,4	25,3	25,9	23,7	21,1	16,8	13,7
Cebbala	9,2	9,8	13,1	16,6	20,7	25,5	28,9	28,2	23,9	19,8	14	10,2
Chebba	12,3	12,5	14,7	17,3	20,6	24,3	27,2	27,6	25,3	22,3	17,6	13,7
Gabès	12,3	13,1	15,9	18,9	22	25,5	28,3	28,9	26,9	23,6	18,1	13,6
Kairouan	10,8	11,4	14,4	17,6	21,6	26,2	29,3	29,1	25,4	21,3	15,9	12
Kasserine	6,7	7,3	10,8	14,5	18,7	23,6	27,1	26,3	21,6	17,5	11,4	7,7
Korba	12,3	11,9	13,6	15,7	18,8	22,8	25,7	26,3	24	21,2	17	13,7
Médenine	11,4	12,5	16	19,3	22,5	25,9	28,6	28,7	26,5	22,9	17,1	12,5
Nabeul	12,2	11,9	13,6	15,8	18,9	22,7	25,7	26,3	24,1	21,3	17,1	13,7
Sfax	11,5	12	14,6	17,4	20,7	24,5	27,4	27,7	25,3	22	16,9	12,8
Sidi Bouzid	9,5	10,2	13,5	16,9	21	25,6	28,9	28,3	24,3	20,3	14,5	10,6
Sousse	11,6	11,9	14,6	17,4	21	25,1	28,2	28,2	25,2	21,7	16,6	12,7
Tabarka	10	9,8	11,9	14,3	17,6	21,7	24,8	25,2	22,3	19,3	14,5	11,3
Tataouine	10,5	11,6	15,3	19,1	22,5	25,9	28,6	28,6	26,1	22,2	16,3	11,6
Tunis	11,4	11,3	13,5	16,1	19,6	23,9	26,9	27,1	24,2	21	16,2	12,7
Zarzis	12,6	13,2	16	18,8	21,7	24,8	27,4	28	26,6	23,4	18,3	14

Problèmes - objectifs

Tableau = ensemble de lignes ou ensemble de colonnes

Etude des individus

- construction de groupes d'individus se ressemblant du point de vue de l'ensemble des variables
- bilan des ressemblances, une partition des individus

Etude des variables

- recherche des ressemblances, liaisons (linéaires) entre variables
- bilan des liaisons : visualisation de la matrice des corrélations
- recherche d'indicateurs synthétiques résumant les variables

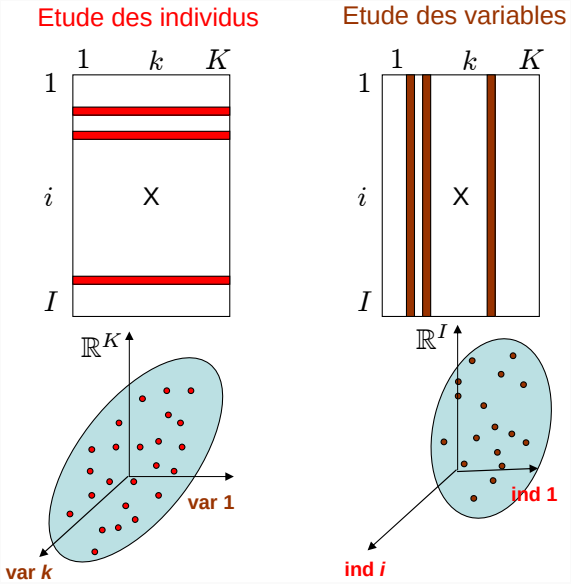
Lien entre les deux études

- caractérisation des classes d'individus par les variables
- individus spécifiques pour comprendre les liaisons entre variables

Objectifs de l'ACP :

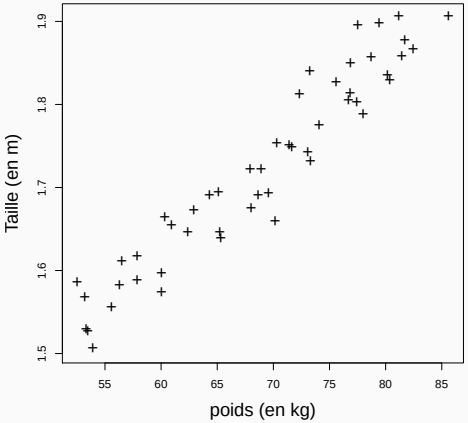
- Descriptif - exploratoire : visualisation de données
- Synthèse - résumé de grands tableaux individus $\times$ variables

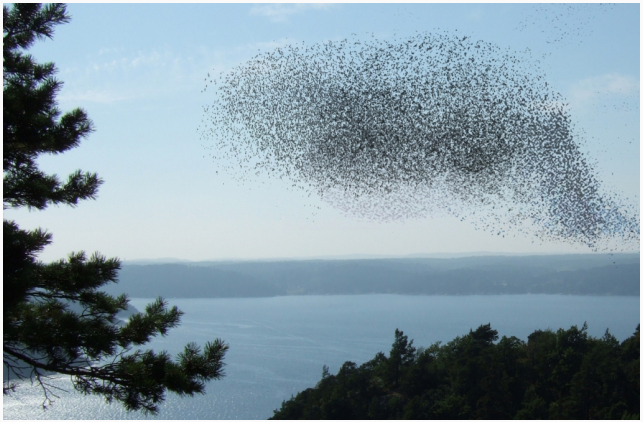
Deux nuages de points



Deux nuages de points

Nuage des individus





- Les individus vivent dans $\mathbb{R}^K$
- Etudier la forme du nuage des individus
- Notion de distance entre individus : Quelle distance ? question cruciale ! ! ! Doit-on normer les variables ? Transformer les variables (par ex. passage au log) ?

Le nuage des individus N^I

1 individu = 1 ligne du tableau $\Rightarrow$ 1 point dans un espace à K dim

- Si $K = 1$: Représentation axiale
- Si $K = 2$: Nuage de points
- Si $K = 3$: Représentation + difficile en 3D
- Si $K = 4$: Impossible à représenter MAIS le concept est simple

Notion de ressemblance : distance (au carré) entre individus i et i' :

$$d^2(i, i') = \sum_{k=1}^{K} (x_{ik} - x_{i'k})^2 \qquad \text{(merci Pythagore)}$$

Etude des individus $\equiv$ Etude de la forme du nuage N^I

Ajustement du nuage des individus

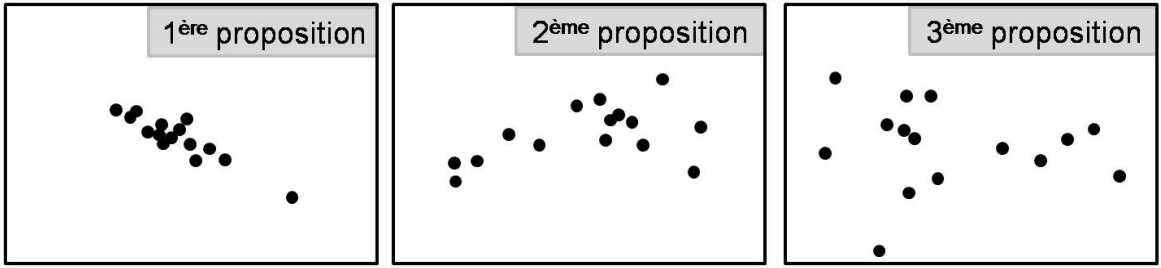
L'ACP vise à fournir une image simplifiée de N^I la + fidèle possible

$\Longleftrightarrow$ Trouver le sous-espace qui résume au mieux les données

Ajustement du nuage des individus

L'ACP vise à fournir une image simplifiée de N^I la + fidèle possible

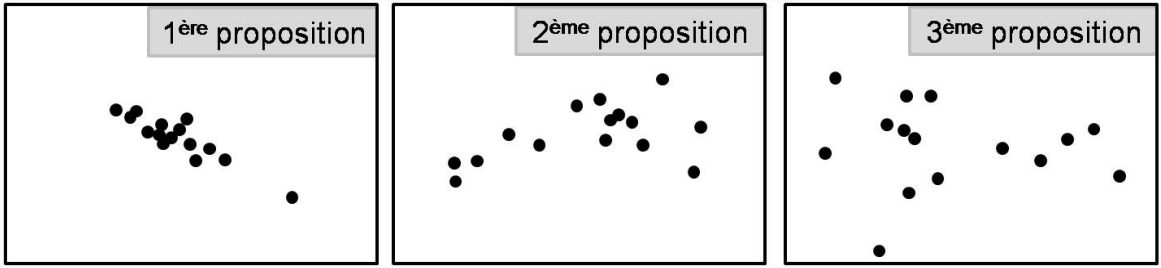
$\Longleftrightarrow$ Trouver le sous-espace qui résume au mieux les données



Ajustement du nuage des individus

L'ACP vise à fournir une image simplifiée de N^I la + fidèle possible

$\Longleftrightarrow$ Trouver le sous-espace qui résume au mieux les données

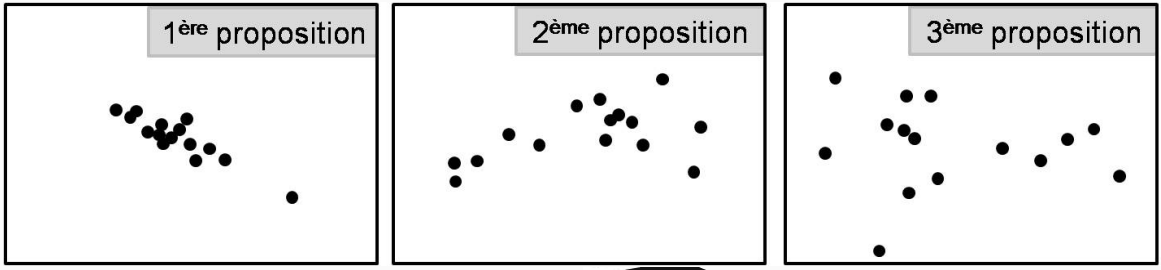


- Meilleure approximation par projection
- Meilleure représentation de la diversité, de la variabilité

Ajustement du nuage des individus

L'ACP vise à fournir une image simplifiée de N^I la + fidèle possible

$\Longleftrightarrow$ Trouver le sous-espace qui résume au mieux les données



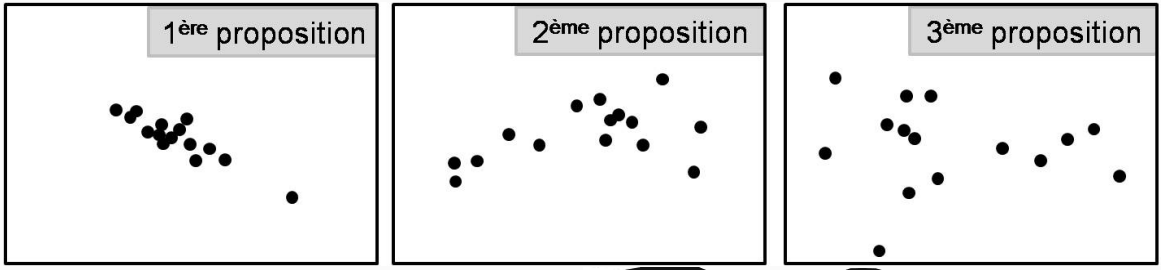
- Meilleure approximation par projection
- Meilleure représentation de la diversité, de la variabilité

Quel animal ? source J.P. Fenelon

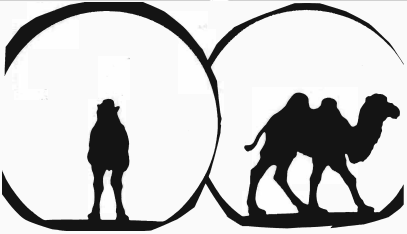
Ajustement du nuage des individus

L'ACP vise à fournir une image simplifiée de N^I la + fidèle possible

$\Longleftrightarrow$ Trouver le sous-espace qui résume au mieux les données

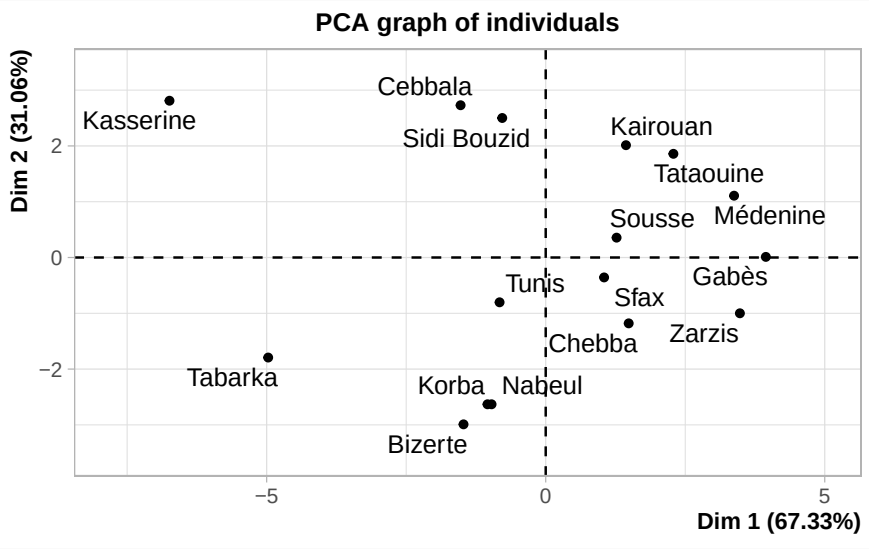


- Meilleure approximation par projection
- Meilleure représentation de la diversité, de la variabilité



Quel animal ? source J.P. Fenelon

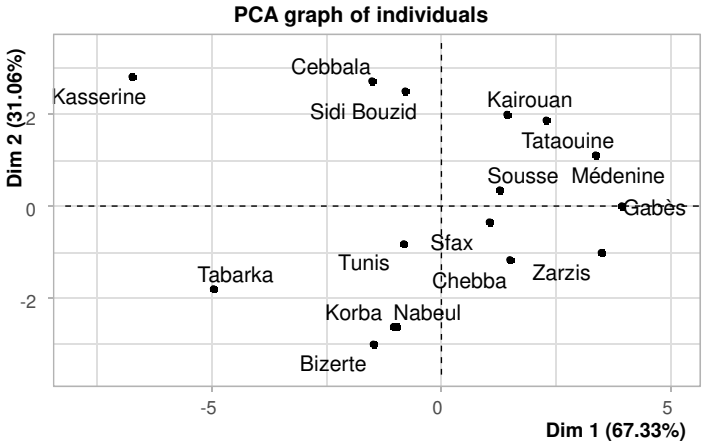
Exemple : graphe des individus



Comment interpréter les axes ? Qu'est-ce qui oppose Gabès à Kasserine ? Et Cebbala à Bizerte ?

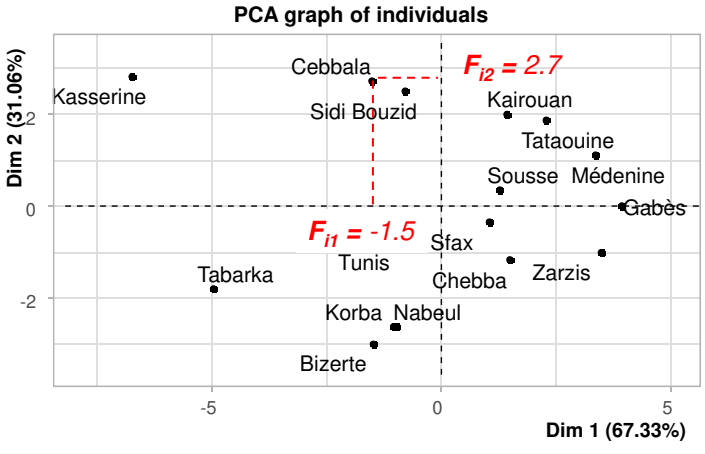
⇒ Besoin de variables pour interpréter ces dimensions de variabilité

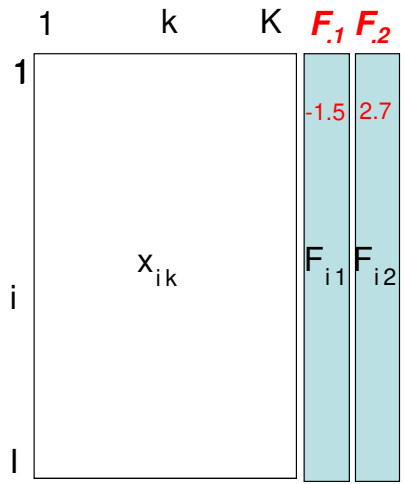
Interprétation du graphe des individus grâce aux variables



Interprétation du graphe des individus grâce aux variables

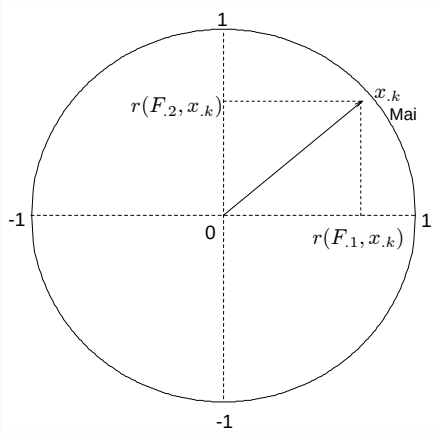
Considérons les coordonnées des individus sur les axes comme des variables





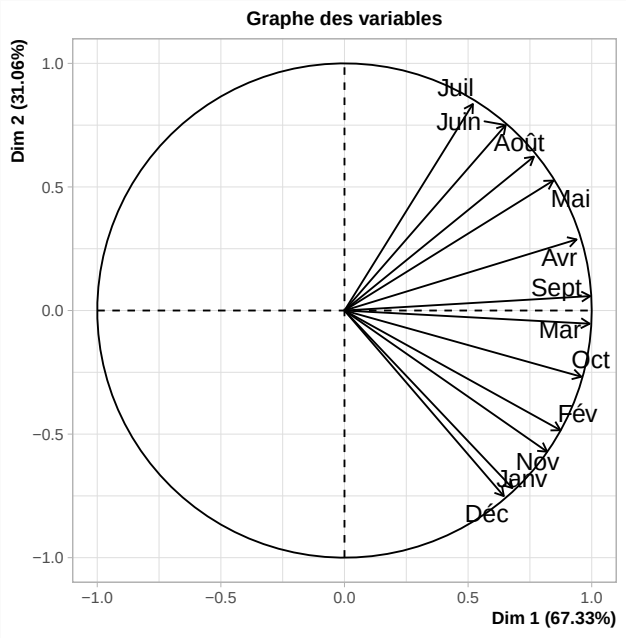
Interprétation du graphe des individus grâce aux variables

- Corrélations entre la variable $x_{.k}$ et $F_{.1}$ (et $F_{.2}$)

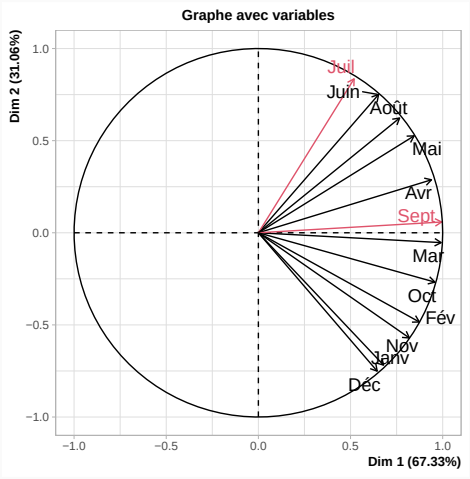


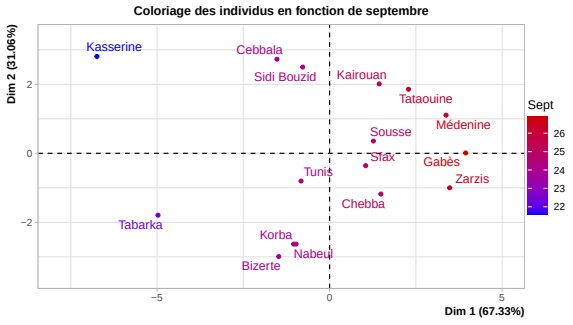
$\Rightarrow$ Cercle des corrélations

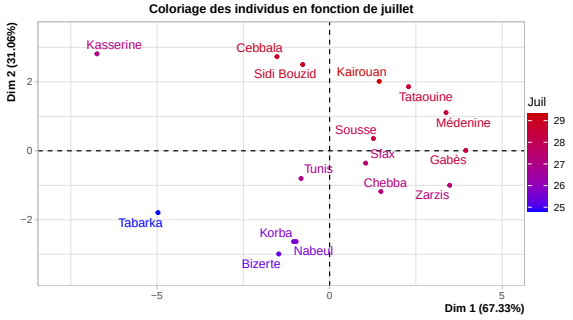
Interprétation du graphe des individus grâce aux variables



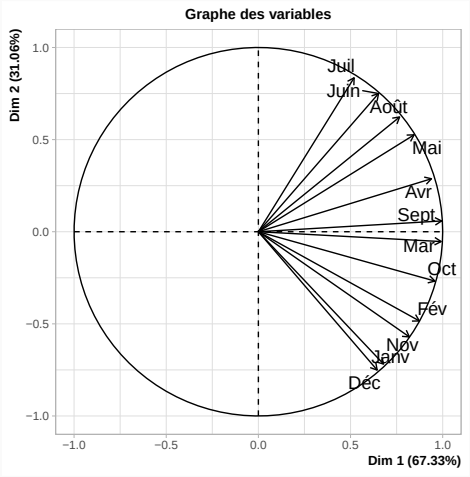
Interprétation du graphe des individus grâce aux variables



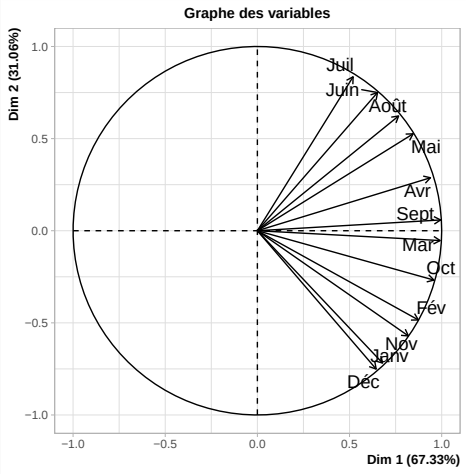




Interprétation du graphe des individus grâce aux variables



Interprétation du graphe des individus grâce aux variables



Toutes les variables sont corrélées à F_1.

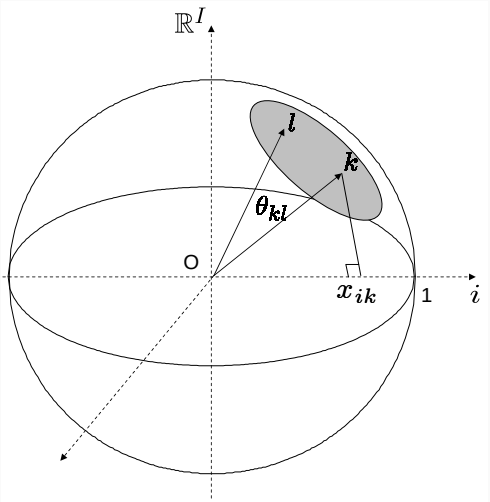
Comment interpréter le 1er axe ?
Comment interpréter le 2ème ?

Principaux facteurs de variabilité :

1 - villes chaudes et froides ;

2 - à T^{o} moyenne constante : l'amplitude thermique

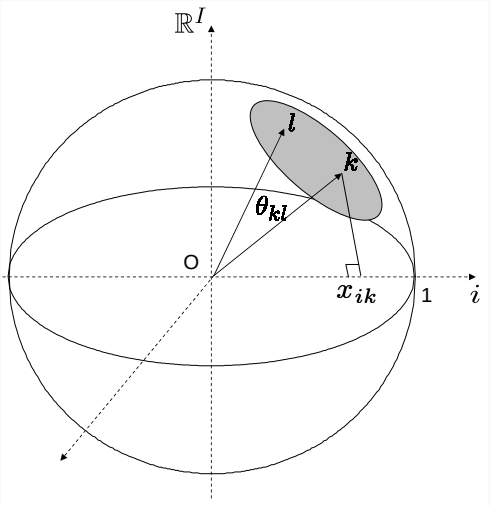
Nuage des variables N^K



1 variable = 1 point dans un espace à I dimensions

$$\begin{aligned} \cos(\theta_{kl}) &= \frac{< x_{.k}, x_{.l} >}{\|x_{.k}\| \ \|x_{.l}\|} \\ &= \frac{\sum_{i=1}^{I} x_{ik} x_{il}}{\sqrt{\sum_{i=1}^{I} x_{ik}^2} \sqrt{\sum_{i=1}^{I} x_{il}^2}} \end{aligned}$$

Nuage des variables N^K

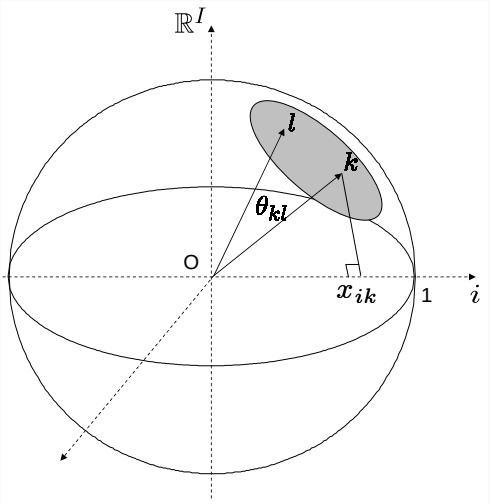


1 variable = 1 point dans un espace à I dimensions

$$\cos(\theta_{kl}) = \frac{< x_{.k}, x_{.l} >}{\|x_{.k}\| \ \|x_{.l}\|} = \frac{\sum_{i=1}^{I} x_{ik} x_{il}}{\sqrt{\sum_{i=1}^{I} x_{ik}^2} \sqrt{\sum_{i=1}^{I} x_{il}^2}}$$

Comme les variables sont centrées : $\cos(\theta_{kl}) = r(x_{.k}, x_{.l})$

Nuage des variables N^K



1 variable = 1 point dans un espace à I dimensions

$$\cos(\theta_{kl}) = \frac{< x_{.k}, x_{.l} >}{\|x_{.k}\| \ \|x_{.l}\|} = \frac{\sum_{i=1}^{I} x_{ik} x_{il}}{\sqrt{\sum_{i=1}^{I} x_{ik}^2} \sqrt{\sum_{i=1}^{I} x_{il}^2}}$$

Comme les variables sont centrées : $\cos(\theta_{kl}) = r(x_{.k}, x_{.l})$

Si variables réduites $\Rightarrow$ points sur une hypersphère de rayon 1

Ajustement du nuage des variables

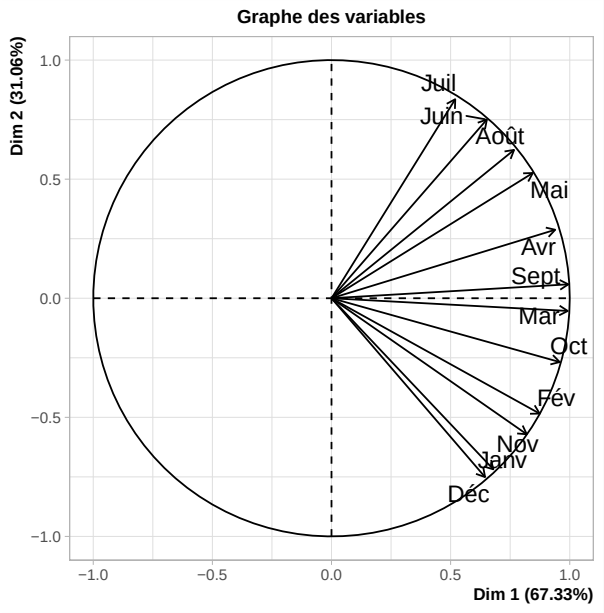
Même règle que pour les individus : recherche d'axes orthogonaux

$$\underset{v_1 \in \mathbb{R}^I}{\arg\max} \sum_{k=1}^{K} r(v_1, x_{\cdot k})^2$$

$\Rightarrow$ v_1 est la variable synthétique qui résume au mieux les variables

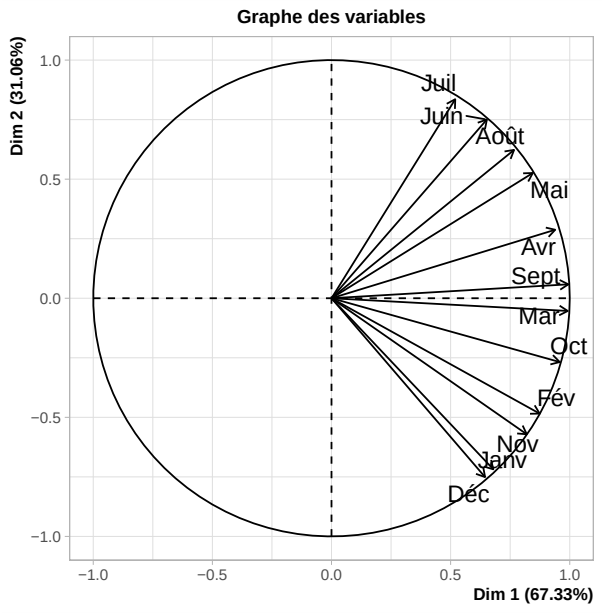
Trouver le $2^{\text{ème}}$ axe, puis le $3^{\text{ème}}$, etc.

Ajustement du nuage des variables



$\Rightarrow$ Même représentation que précédemment ! ! ! !

Ajustement du nuage des variables



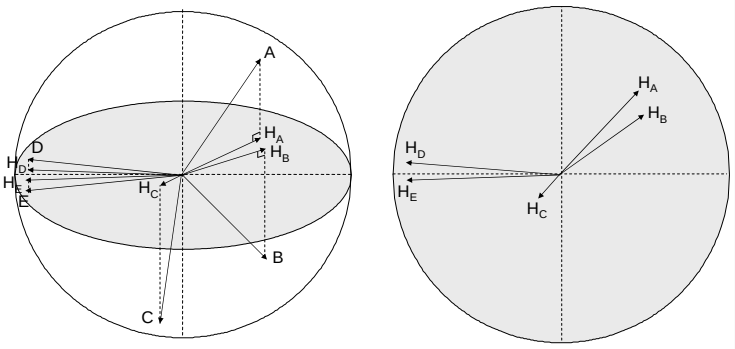
⇒ Même représentation que précédemment ! ! ! !

- aide pour interpréter les individus
- représentation optimale du nuage des variables
- visualisation de la matrice des corrélations

Projections...

$r(A, B) = cos(\theta_{A,B})$

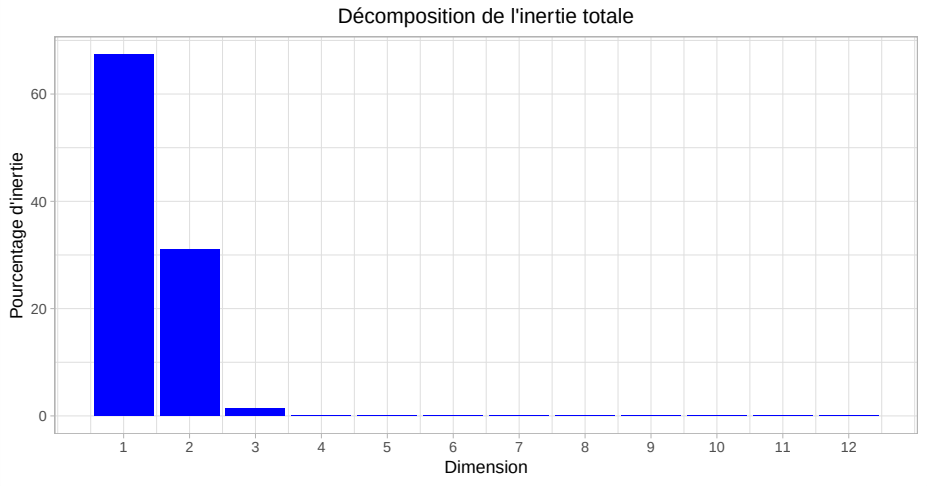
$cos(\theta_{A,B}) \approx cos(\theta_{H_A,H_B})$ si les variables sont bien projetées



Seules les variables bien projetées peuvent être interprétées !

Pourcentage d'inertie

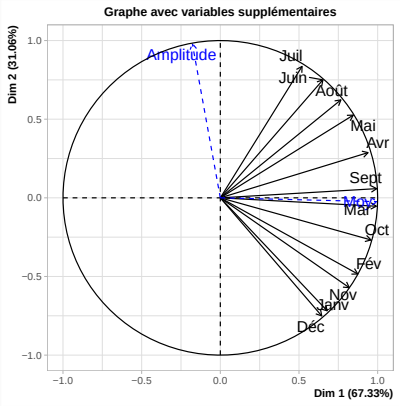
- Pourcentage d'information (d'inertie) expliqué par chaque axe

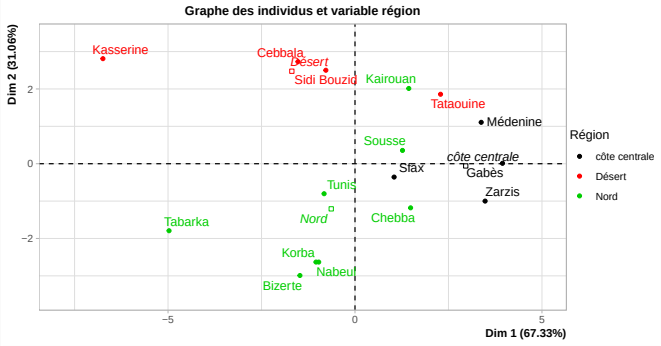


$\Rightarrow$ Choix d'un nombre de dimensions à interpréter

Information supplémentaire

- Pour les variables quantitatives : projection des variables
- Pour les modalités : projection au barycentre des individus qui prennent cette modalité





⇒ Information supp. ne participe pas à la construction des axes

Description des dimensions

Par les variables quantitatives :

- calcul des corrélations entre chaque variable et la dimension s
- tri des coefficients de corrélation (significatifs)

```
> dimdesc(res.pca)
$Dim.1$quanti                    $Dim.2$quanti
           correlation  p.value              correlation  p.value
Moy              0.999 9.26e-22  Amplitude        0.980 2.88e-11
Septembre        0.994 7.39e-15  Juillet          0.836 5.56e-05
Mars             0.992 7.17e-14  Juin             0.750 8.15e-04
Octobre          0.958 5.63e-09  Août             0.623 9.97e-03
Avril            0.939 7.14e-08  Mai              0.526 3.64e-02
Février          0.874 9.78e-06
...                              Novembre        -0.571 2.08e-02
Décembre         0.645 7.02e-03  Janvier         -0.718 1.72e-03
Juillet          0.519 3.92e-02  Décembre        -0.751 8.02e-04
```

Description des dimensions

Par les variables qualitatives :

- Analyse de variance des coordonnées des individus sur l'axe (variable Y) expliquée par la variable qualitative
  - un test F par variable
  - un test t de Student par modalité pour comparer la moyenne de la modalité avec la moyenne générale

```
> dimdesc(res.pca)
Dim.1                                    Dim.2
$quali                                   $quali
                      R2     p.value                     R2    p.value
Région          0.3847178   0.0425582   Région      0.6066451   0.002323

$category                                $category
                    Estimate  p.value                Estimate   p.value
Région=côte centrale 2.749387 0.01370   Région=Désert 2.072120 0.001046
```

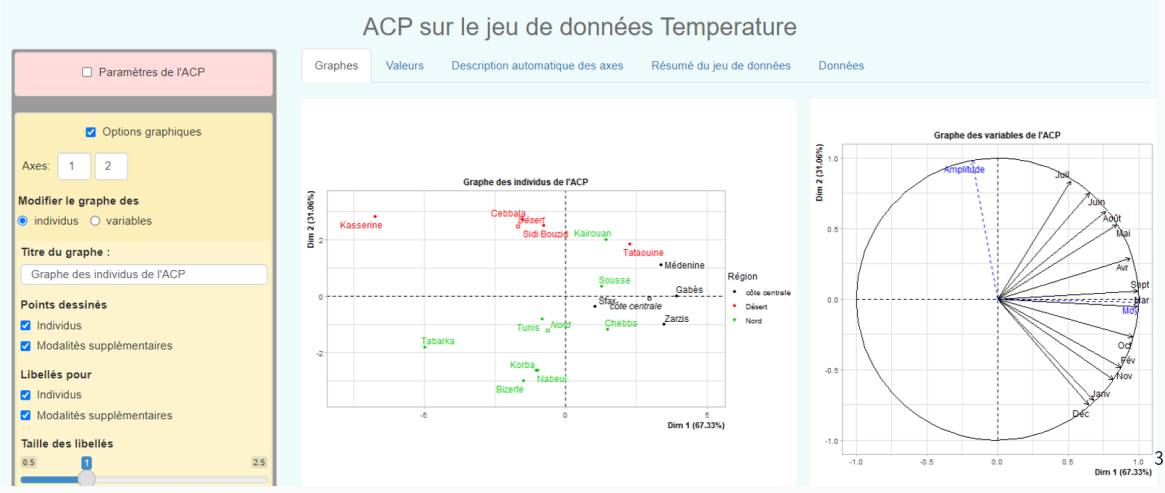
Pratique de l'ACP

1. Choisir les variables actives
2. Choisir de réduire ou non les variables
3. Réaliser l'ACP
4. Choisir le nombre de dimensions à interpréter
5. Interpréter simultanément le graphe des individus et celui des variables
6. Utiliser les indicateurs pour enrichir l'interprétation
7. Revenir aux données brutes pour interpréter

Graphiques interactifs avec le package Factoshiny

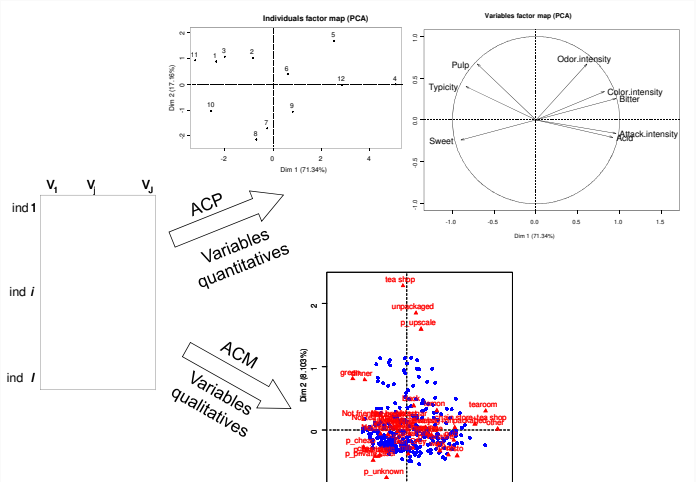
- Réaliser des analyses sans besoin de maîtriser le code
- Visualisation en temps réel des modifications apportées

```
> library(Factoshiny)
> res <- Factoshiny(don)    ## analyse factorielle sur les données
```



Les autres méthodes d'analyse de données : : l'analyse des correspondances multiples

- Tableau individu-variables quantitatives $\Longrightarrow$ Analyse en Composantes Principales
- Tableau individu-variables quantitatives $\Longrightarrow$ Analyse des correspondances Multiples



Les autres méthodes d'analyse de données : l'analyse des correspondances

- Tableau croisé $\Longrightarrow$ Analyse Factorielle des Correspondances
- Pour l'analyse textuelle

- Aragon (23 textes) :	FeuJoie, Perpétuel, Destinées, Snark, Peinture, ...
- Balzac (49 textes) :	*Chouans*, *Physiologie*, *Vendetta*, *Gobseck*, ...
- Corneille (34 textes) :	*Mélite*, *Clitandre*, *Veuve*, *Gelerie*,*Suivante*, ...
- ...	

On conserve les mots cités au moins 100 fois

accord	264	0	88	44	...
affaire	1029	2040	74	154	...
âge	545	629	92	108	
ah	219	0	0	0	
air	2093	2009	95	191	
allemagne	366	0	0	0	
allemand	476	0	0	0	
amant	303	760	566	0	
âme	478	2190	1101	240	
ami	1090	2583	307	407	
amour	1374	3286	1791	167	
an	1812	3009	112	182	
anglais	315	0	0	0	
. . .					

978 mots

Les autres méthodes d'analyse de données : l'analyse des correspondances

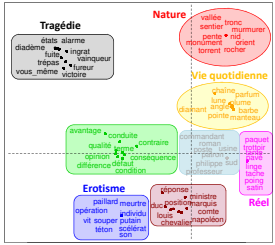
- Tableau croisé $\Longrightarrow$ Analyse Factorielle des Correspondances
- Pour l'analyse textuelle

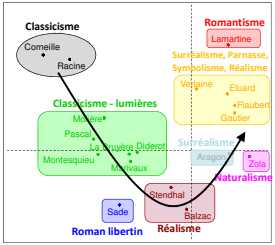
- Aragon (23 textes) : FeuJoie, Perpétuel, Destinées, Snark, Peinture, ...
- Balzac (49 textes) : *Chouans*, *Physiologie*, *Vendetta*, *Gobseck*, ...
- Corneille (34 textes) : *Mélite*, *Clitandre*, *Veuve*, *Gelerie*,*Suivante*, ...
- ...

On conserve les mots cités au moins 100 fois

accord	264	0	88	44	...
affaire	1029	2040	74	154	...
âge	545	629	92	108	
ah	219	0	0	0	
air	2093	2009	95	191	
allemagne	366	0	0	0	
allemand	476	0	0	0	
amant	303	760	566	0	
âme	478	2190	1101	240	
ami	1090	2583	307	407	
amour	1374	3286	1791	167	
an	1812	3009	112	182	
anglais	315	0	0	0	
. . .					

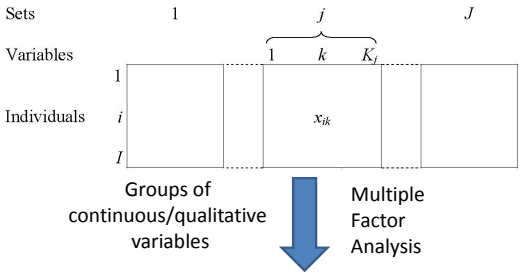
978 mots

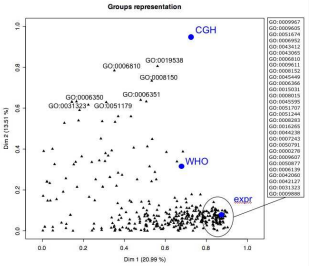


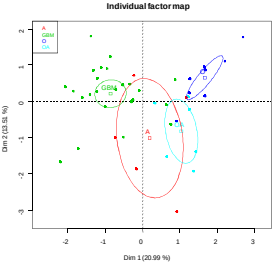


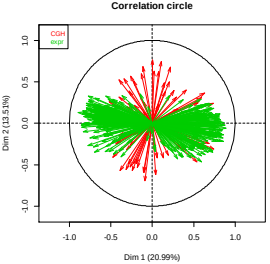
Les autres méthodes d'analyse de données : l'Analyse Factorielle Multiple

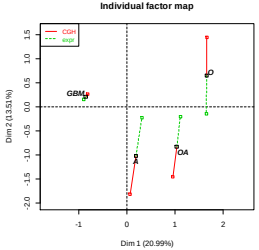
- Tableau où les variables sont structurées en groupes $\Longrightarrow$ Analyse Factorielle Multiple











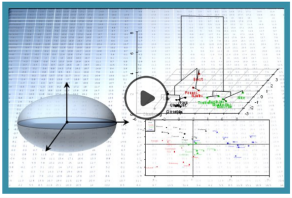
Matériel disponible

Analyse de données avec R (2e ed)



R pour la stat. et sc. des données


MOOC analyse de données multidimensionnelles



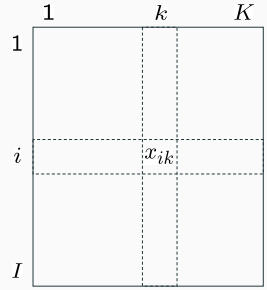
Playlists en analyse de données :

- sur l'ACP, on PCA
- sur l'AFC, on correspondence analysis,
- sur l'ACM, on multiple correspondence analysis (MCA),
- sur la classification, on clustering,
- sur l'AFM, on multiple factor analysis (MFA),
- sur la gestion de données manquantes, on handling missing values

Plan

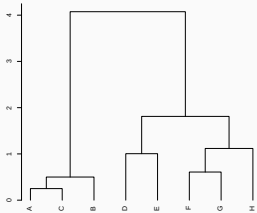
Quelles données pour quels objectifs ?

La classification s'intéresse à des tableaux de données individus $\times$ variables quantitatives



Objectifs : production d'une structure (arborescence) permettant :

- la mise en évidence de liens hiérarchiques entre individus ou groupes d'individus
- la détection d'un nb de classes « naturel » au sein de la population



Critères

Ressemblance entre individus :

- distance euclidienne
- indice de similarité
- ...

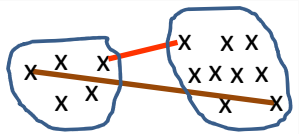
Critères

Ressemblance entre individus :

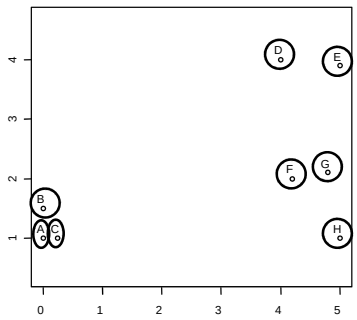
- distance euclidienne
- indice de similarité
- ...

Ressemblance entre groupes d'individus :

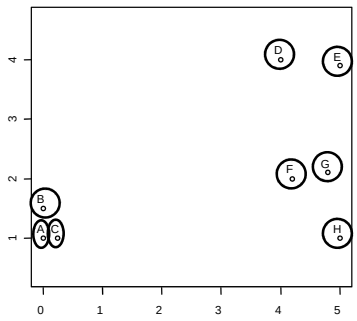
- saut minimum ou lien simple (plus petite distance)
- lien complet (plus grande distance)
- critère de Ward

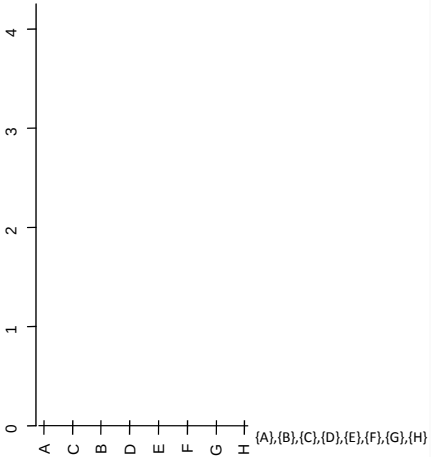


Algorithme



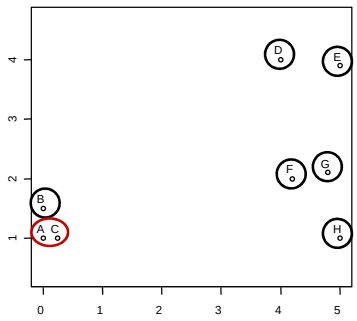
Algorithme

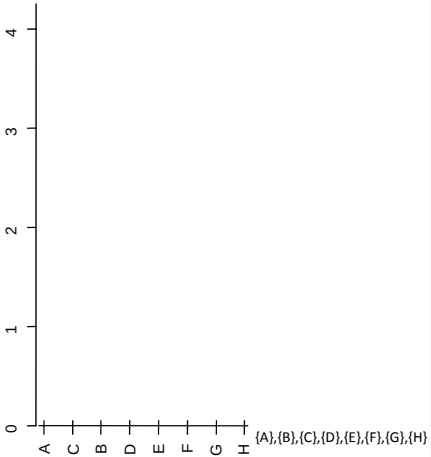




	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

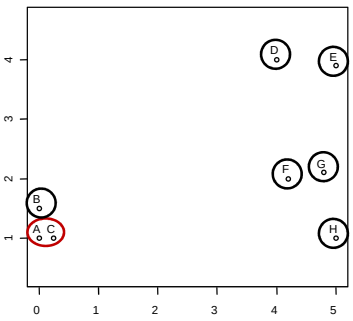
Algorithme

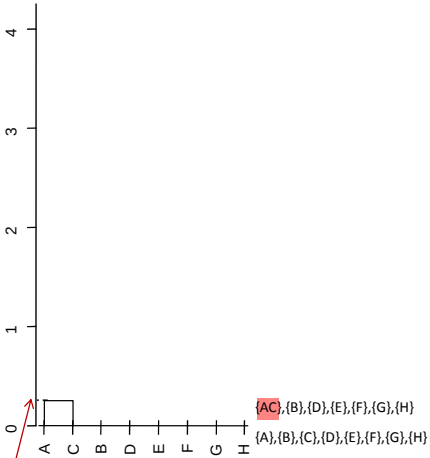




	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

Algorithme

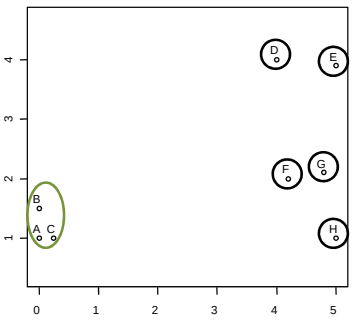


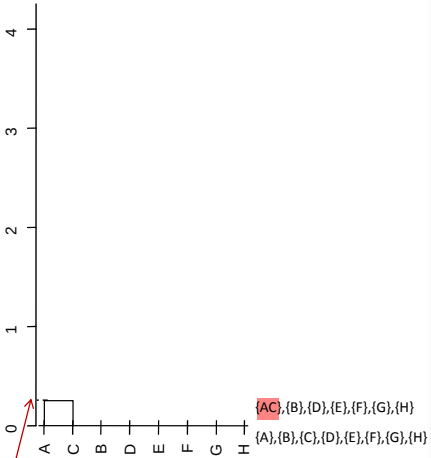


	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

1er regroupement

Algorithme



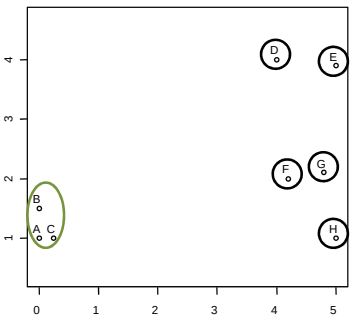


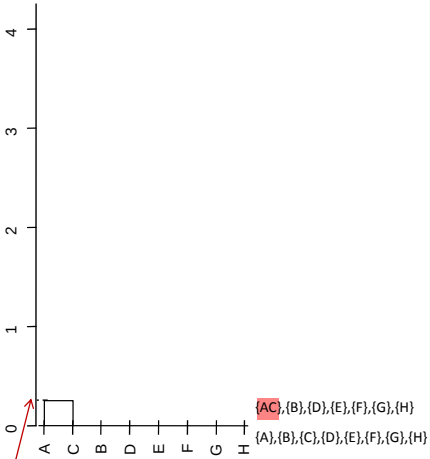
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

1er regroupement

Algorithme





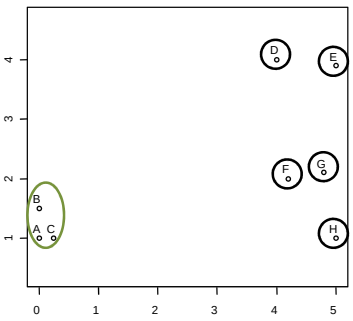
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

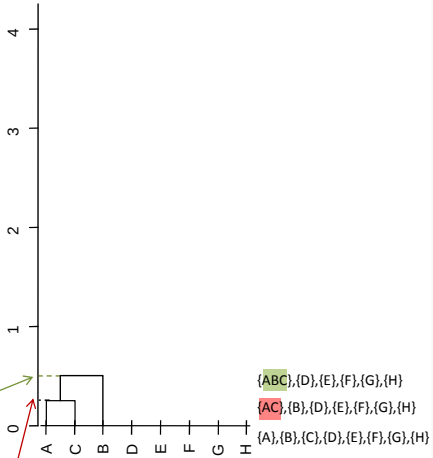
2e regroupement

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

1er regroupement

Algorithme





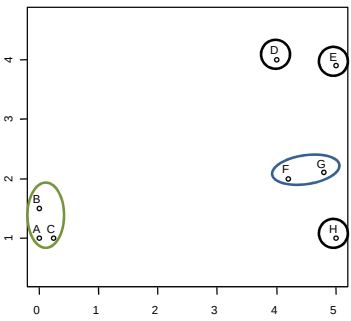
2e regroupement

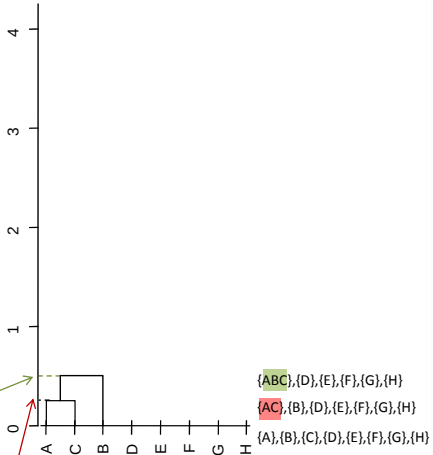
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

1er regroupement

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

Algorithme





	ABC	D	E	F	G
D	4.72				
E	5.55	1.00			
F	4.07	2.01	2.06		
G	4.68	2.06	1.81	0.61	
H	4.75	3.16	2.90	1.28	1.12

3e regroupement

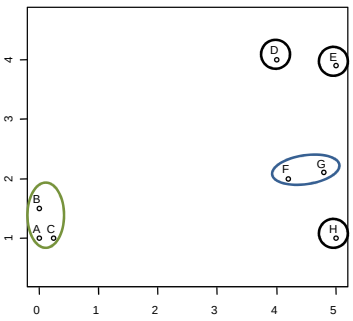
	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

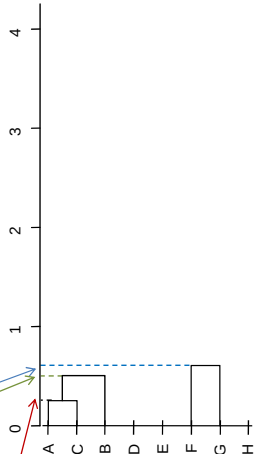
2e regroupement

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

1er regroupement

Algorithme





{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

3e regroupement

	ABC	D	E	F	G
D	4.72				
E	5.55	1.00			
F	4.07	2.01	2.06		
G	4.68	2.06	1.81	0.61	
H	4.75	3.16	2.90	1.28	1.12

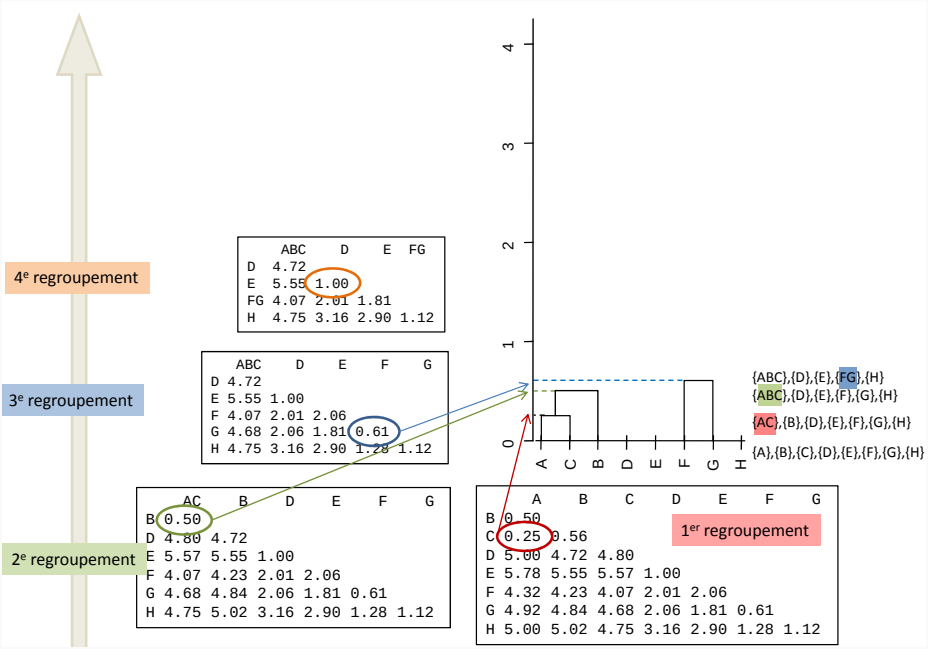
2e regroupement

	AC	B	D	E	F	G
B	0.50					
D	4.80	4.72				
E	5.57	5.55	1.00			
F	4.07	4.23	2.01	2.06		
G	4.68	4.84	2.06	1.81	0.61	
H	4.75	5.02	3.16	2.90	1.28	1.12

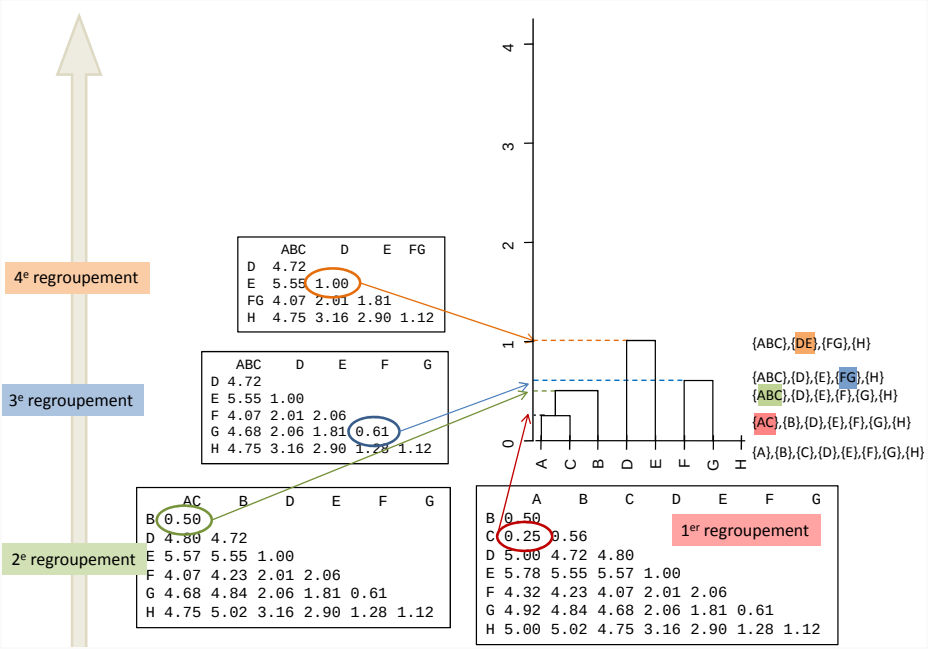
1er regroupement

	A	B	C	D	E	F	G
B	0.50						
C	0.25	0.56					
D	5.00	4.72	4.80				
E	5.78	5.55	5.57	1.00			
F	4.32	4.23	4.07	2.01	2.06		
G	4.92	4.84	4.68	2.06	1.81	0.61	
H	5.00	5.02	4.75	3.16	2.90	1.28	1.12

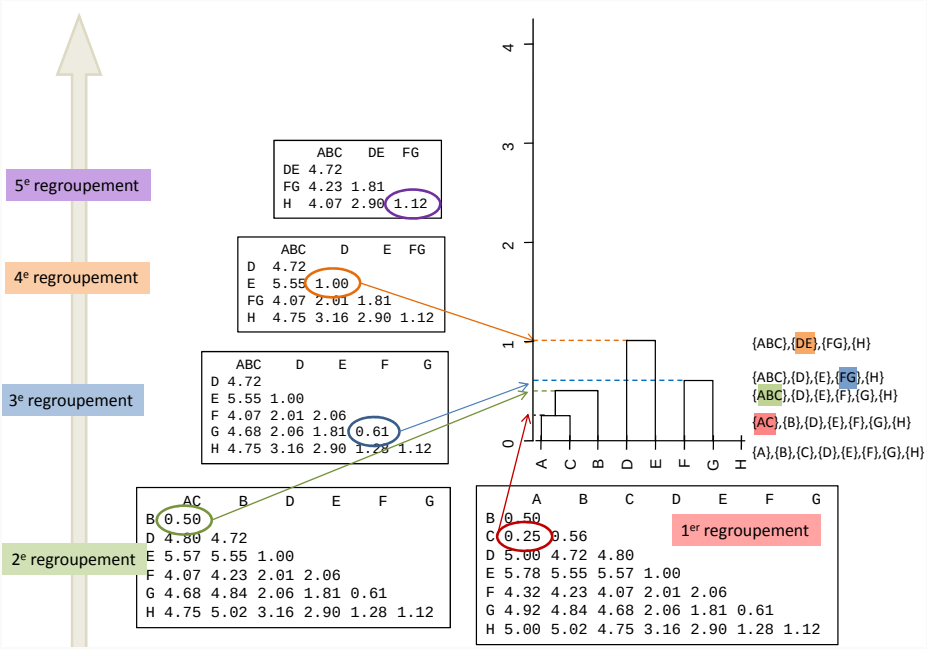
Algorithme


4e regroupement
3e regroupement
2e regroupement
1er regroupement
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
0
1
2
3
4
A C B D E F G H
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

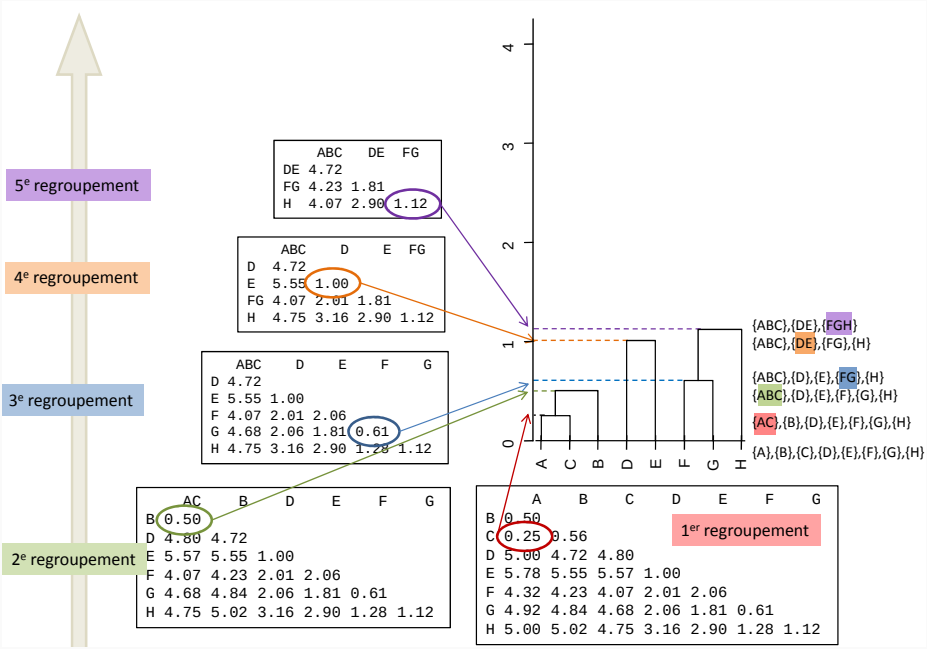
Algorithme


4e regroupement
3e regroupement
2e regroupement
1er regroupement
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
0 1 2 3 4
A C B D E F G H
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

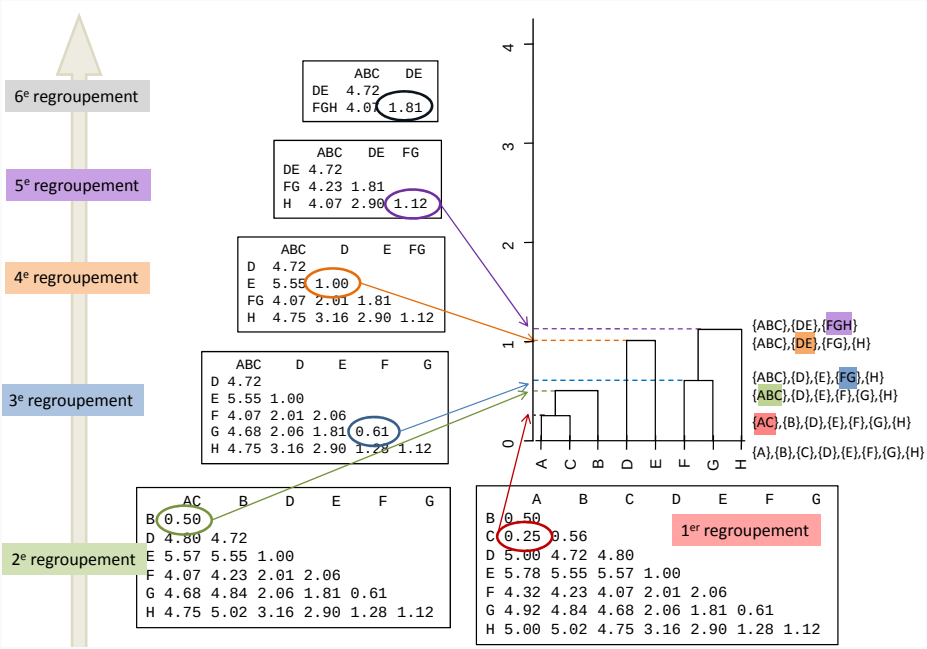
Algorithme


5e regroupement
4e regroupement
3e regroupement
2e regroupement
1er regroupement
ABC DE FG
DE 4.72
FG 4.23 1.81
H 4.07 2.90 1.12
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
0
1
2
3
4
A C B D E F G H
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

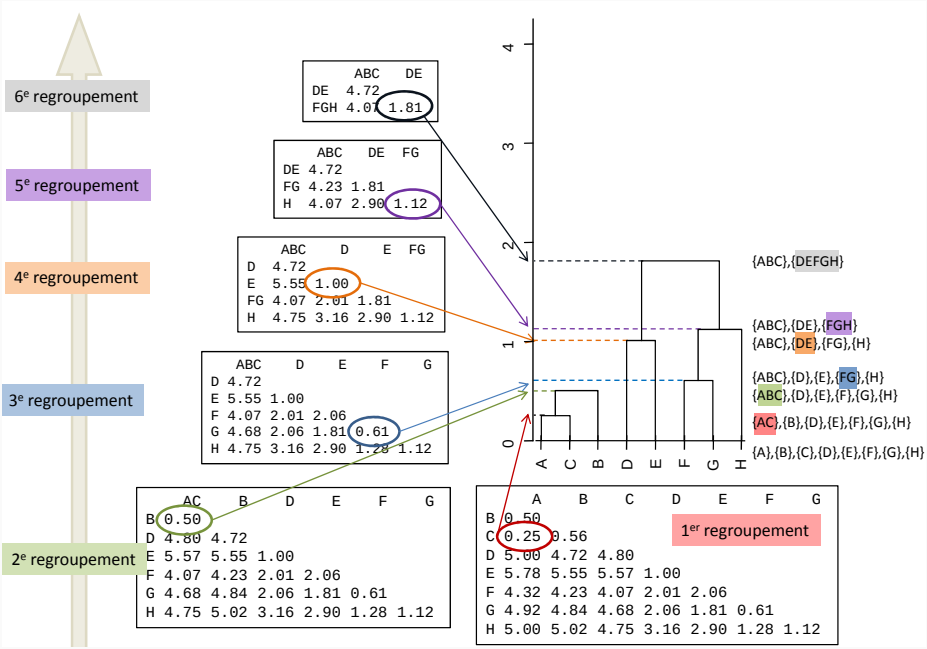
Algorithme


5e regroupement
4e regroupement
3e regroupement
2e regroupement
1er regroupement
ABC DE FG
DE 4.72
FG 4.23 1.81
H 4.07 2.90 1.12
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
0
1
2
3
4
A C B D E F G H
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

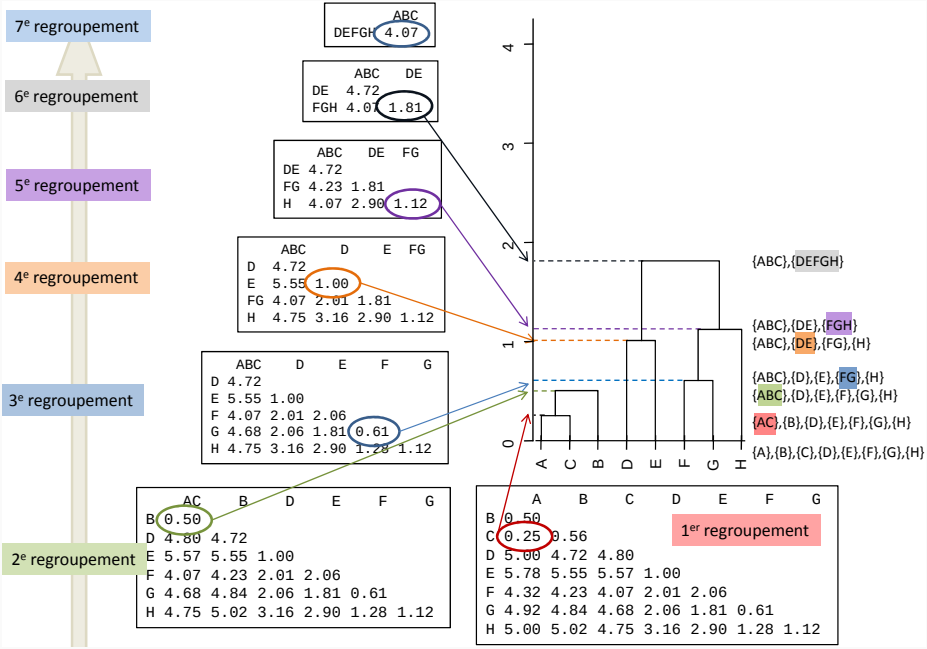
Algorithme


6e regroupement
5e regroupement
4e regroupement
3e regroupement
2e regroupement
1er regroupement
ABC DE
DE 4.72
FGH 4.07 1.81
ABC DE FG
DE 4.72
FG 4.23 1.81
H 4.07 2.90 1.12
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
0
1
2
3
4
A C B D E F G H
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

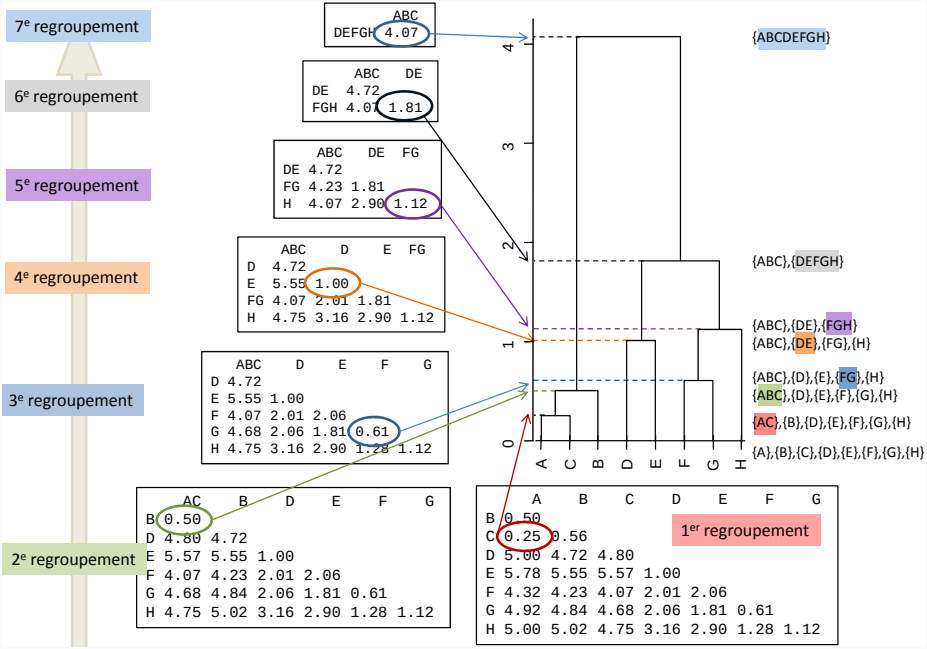
Algorithme


6e regroupement
5e regroupement
4e regroupement
3e regroupement
2e regroupement
1er regroupement
ABC DE
DE 4.72
FGH 4.07 1.81
ABC DE FG
DE 4.72
FG 4.23 1.81
H 4.07 2.90 1.12
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
0
1
2
3
4
A C B D E F G H
{ABC},{DEFGH}
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

Algorithme


7e regroupement
6e regroupement
5e regroupement
4e regroupement
3e regroupement
2e regroupement
ABC
DEFGH 4.07
ABC DE
DE 4.72
FGH 4.07 1.81
ABC DE FG
DE 4.72
FG 4.23 1.81
H 4.07 2.90 1.12
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
1er regroupement
0
1
2
3
4
A C B D E F G H
{ABC},{DEFGH}
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

Algorithme


7e regroupement
6e regroupement
5e regroupement
4e regroupement
3e regroupement
2e regroupement
1er regroupement
ABC
DEFGH 4.07
ABC DE
DE 4.72
FGH 4.07 1.81
ABC DE FG
DE 4.72
FG 4.23 1.81
H 4.07 2.90 1.12
ABC D E FG
D 4.72
E 5.55 1.00
FG 4.07 2.01 1.81
H 4.75 3.16 2.90 1.12
ABC D E F G
D 4.72
E 5.55 1.00
F 4.07 2.01 2.06
G 4.68 2.06 1.81 0.61
H 4.75 3.16 2.90 1.28 1.12
AC B D E F G
B 0.50
D 4.80 4.72
E 5.57 5.55 1.00
F 4.07 4.23 2.01 2.06
G 4.68 4.84 2.06 1.81 0.61
H 4.75 5.02 3.16 2.90 1.28 1.12
A B C D E F G
B 0.50
C 0.25 0.56
D 5.00 4.72 4.80
E 5.78 5.55 5.57 1.00
F 4.32 4.23 4.07 2.01 2.06
G 4.92 4.84 4.68 2.06 1.81 0.61
H 5.00 5.02 4.75 3.16 2.90 1.28 1.12
4
3
2
1
0
A C B D E F G H
{ABCDEFGH}
{ABC},{DEFGH}
{ABC},{DE},{FGH}
{ABC},{DE},{FG},{H}
{ABC},{D},{E},{FG},{H}
{ABC},{D},{E},{F},{G},{H}
{AC},{B},{D},{E},{F},{G},{H}
{A},{B},{C},{D},{E},{F},{G},{H}

Méthode de Ward

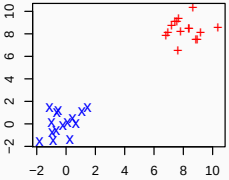
- Initialisation : 1 classe = 1 individu $\Longrightarrow$ In. inter = In. totale
- A chaque étape : agréger les classes a et b qui minimisent la diminution de l'inertie inter

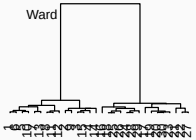
Méthode de Ward

- Initialisation : 1 classe = 1 individu $\Longrightarrow$ In. inter = In. totale
- A chaque étape : agréger les classes a et b qui minimisent la diminution de l'inertie inter

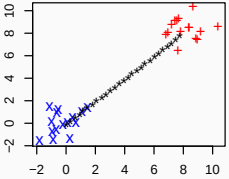
$$\text{Inertie}(a) + \text{Inertie}(b) = \text{Inertie}(a \cup b) - \underbrace{\frac{m_a m_b}{m_a + m_b}\, d^2(a,b)}_{\text{à minimiser}}$$

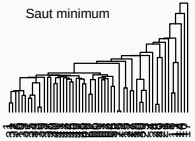
Regroupe les objets de faible poids et évite l'effet de chaîne





Regroupe des classes ayant des centres de gravité proches





Intérêt immédiat pour la classification

Les données température en Tunisie

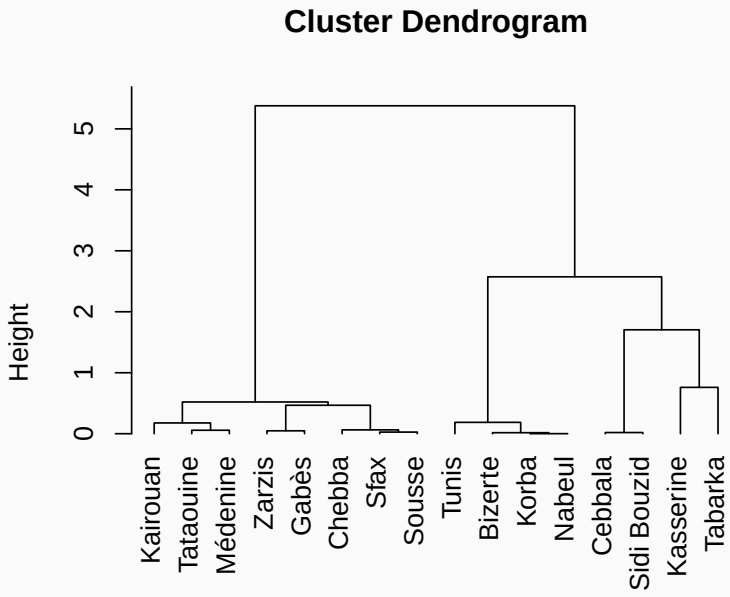
- 16 individus (lignes) : villes de Tunisie
- 12 variables (colonnes) : 12 températures mensuelles moyennes

	Janv	Fév	Mar	Avr	Mai	Juin	Juil	Août	Sept	Oct	Nov	Déc
Bizerte	12,3	11,9	13,5	15,6	18,6	22,4	25,3	25,9	23,7	21,1	16,8	13,7
Cebbala	9,2	9,8	13,1	16,6	20,7	25,5	28,9	28,2	23,9	19,8	14	10,2
Chebba	12,3	12,5	14,7	17,3	20,6	24,3	27,2	27,6	25,3	22,3	17,6	13,7
Gabès	12,3	13,1	15,9	18,9	22	25,5	28,3	28,9	26,9	23,6	18,1	13,6
Kairouan	10,8	11,4	14,4	17,6	21,6	26,2	29,3	29,1	25,4	21,3	15,9	12
Kasserine	6,7	7,3	10,8	14,5	18,7	23,6	27,1	26,3	21,6	17,5	11,4	7,7
Korba	12,3	11,9	13,6	15,7	18,8	22,8	25,7	26,3	24	21,2	17	13,7
Médenine	11,4	12,5	16	19,3	22,5	25,9	28,6	28,7	26,5	22,9	17,1	12,5
Nabeul	12,2	11,9	13,6	15,8	18,9	22,7	25,7	26,3	24,1	21,3	17,1	13,7
Sfax	11,5	12	14,6	17,4	20,7	24,5	27,4	27,7	25,3	22	16,9	12,8
Sidi Bouzid	9,5	10,2	13,5	16,9	21	25,6	28,9	28,3	24,3	20,3	14,5	10,6
Sousse	11,6	11,9	14,6	17,4	21	25,1	28,2	28,2	25,2	21,7	16,6	12,7
Tabarka	10	9,8	11,9	14,3	17,6	21,7	24,8	25,2	22,3	19,3	14,5	11,3
Tataouine	10,5	11,6	15,3	19,1	22,5	25,9	28,6	28,6	26,1	22,2	16,3	11,6
Tunis	11,4	11,3	13,5	16,1	19,6	23,9	26,9	27,1	24,2	21	16,2	12,7
Zarzis	12,6	13,2	16	18,8	21,7	24,8	27,4	28	26,6	23,4	18,3	14

Quelles villes ont des profils météo similaires ?

Comment caractériser les groupes de villes ?

Les données température : l'arbre hiérarchique



Arbres et partitions

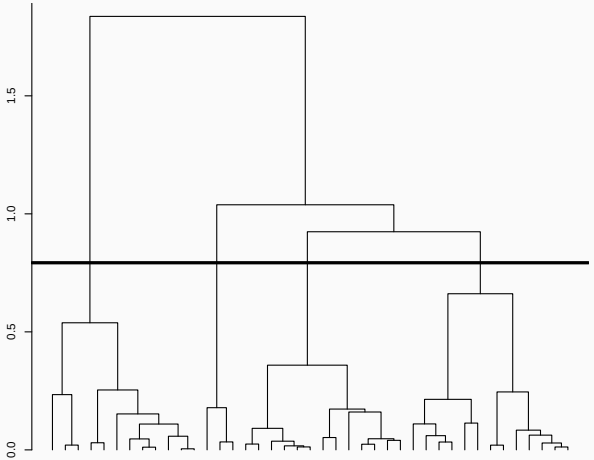
Les arbres finissent tous ...

Les arbres finissent tous ... par être coupés ! ! !

Arbres et partitions

Les arbres finissent tous ... par être coupés ! ! !

En définissant un niveau de coupure, on construit une partition

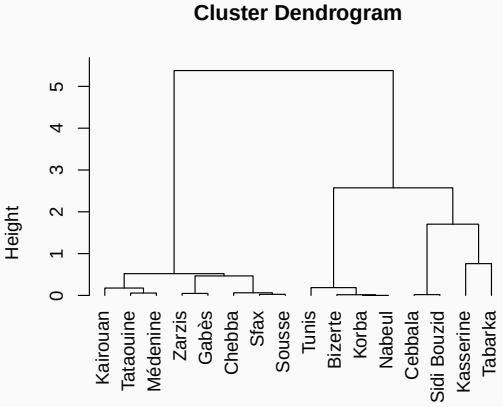


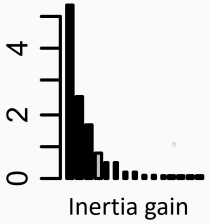
Remarque : vu le mode de construction, la partition n'est pas optimale mais est intéressante

Détermination d'un nombre de classes

- A partir de l'arbre
- Dépend de l'usage (enquête, nb individus, ...)

- A partir du diagramme des indices de niveau
- Critère ultime : interprétabilité des classes

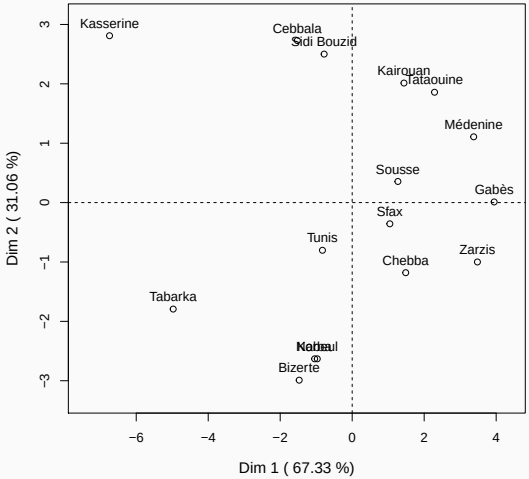




Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

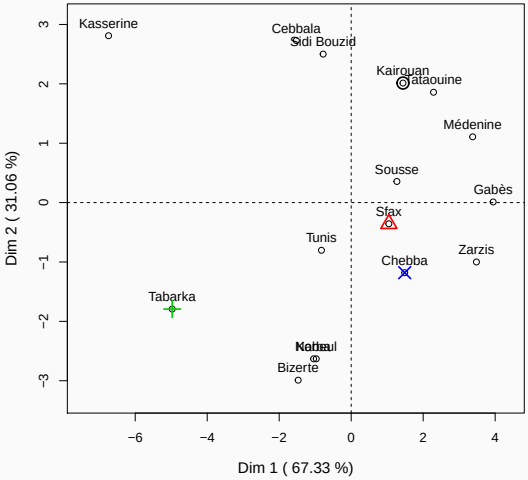
- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité



Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

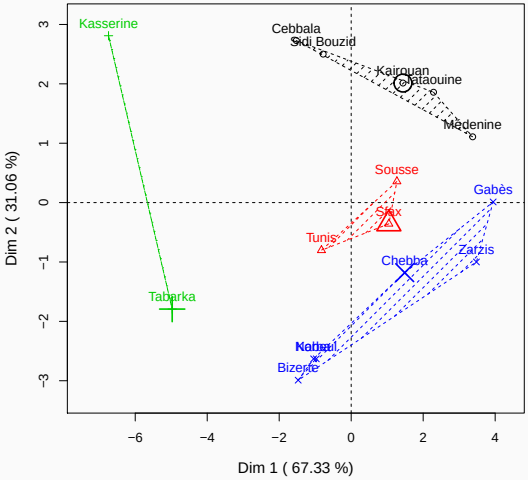
- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité



Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

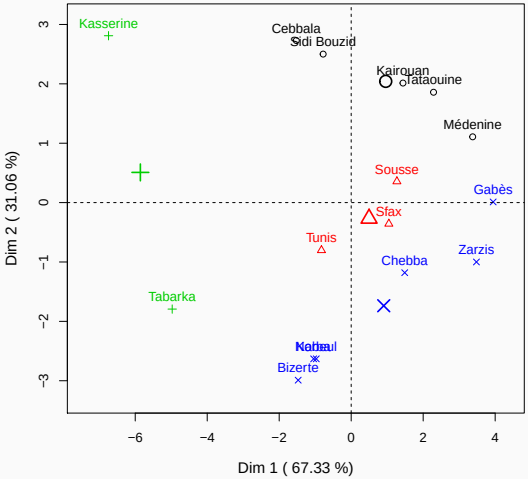
- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité



Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

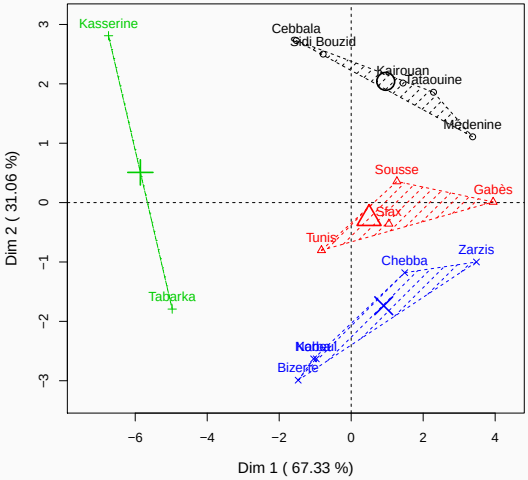
- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité



Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

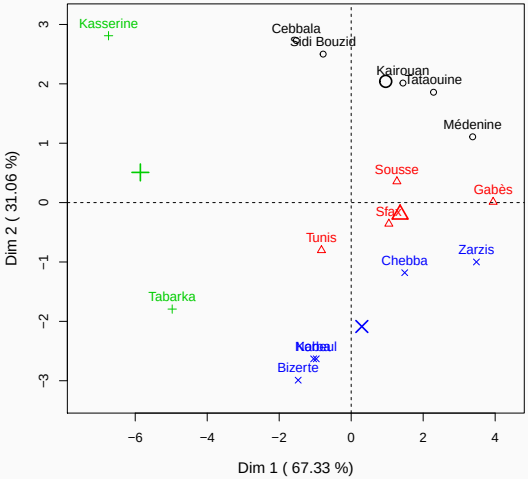
- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité



Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

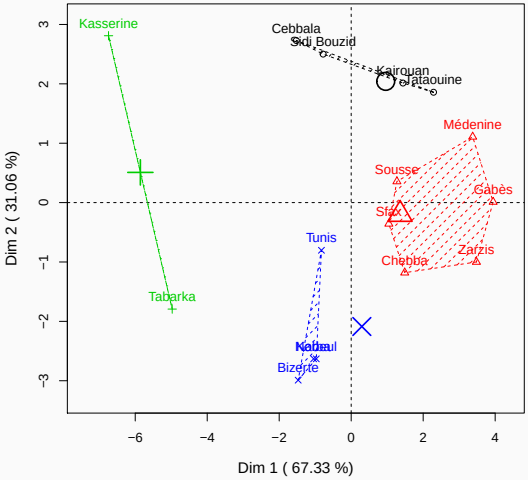
- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité



Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

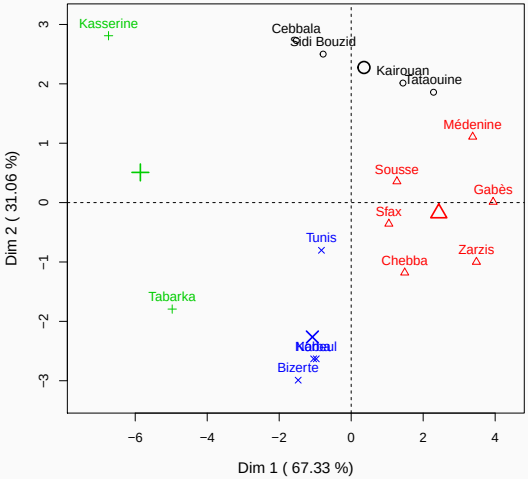
- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité



Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

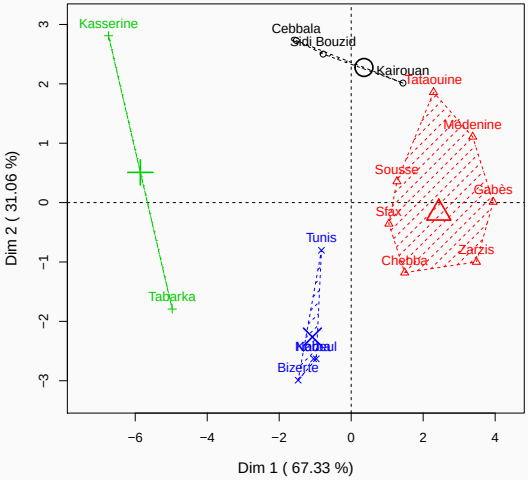
- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité



Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité

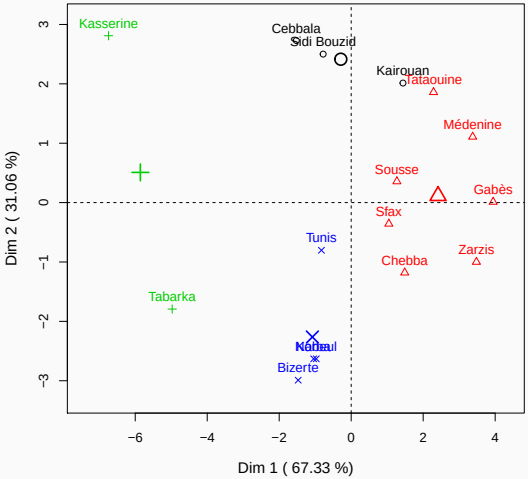


Algorithme de partitionnement : les K-means

Algorithme d'agrégation autour des centres mobiles (K-means)

- Choisir Q centres de classes au hasard
- Affecter les points au centre le plus proche
- Calculer les Q centres de gravité

L'algorithme a convergé



Consolidation d'une partition obtenue par CAH

La partition obtenue par CAH n'est pas optimale et peut être améliorée, consolidée, par les K-means

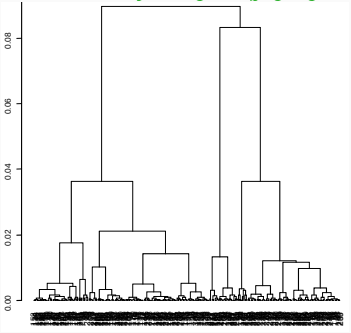
Algorithme de consolidation :

- la partition obtenue par CAH est utilisée comme initialisation de l'algorithme de partitionnement
- quelques étapes de K-means sont itérées

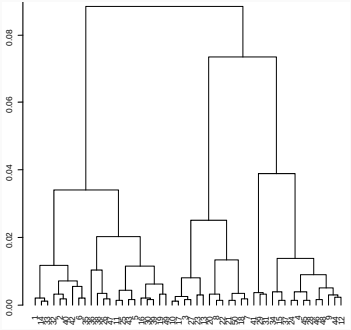
$\Longrightarrow$ amélioration de la partition (souvent non décisive)
Avantage : consolidation de la partition
Inconvénient : perte de l'info de hiérarchie

CAH en grandes dimensions

- Si beaucoup de variables : faire une ACP et ne conserver que les premières dimensions $\Longrightarrow$ on se ramène au cas classique
- Si beaucoup d'individus : algorithme de CAH trop long
  - Faire une partition (par K-means) en une centaine de classes
  - Construire la CAH à partir des classes (utiliser l'effectif des classes dans le calcul)
  - Obtention du « haut » de l'arbre de la CAH



Arbre sur données brutes

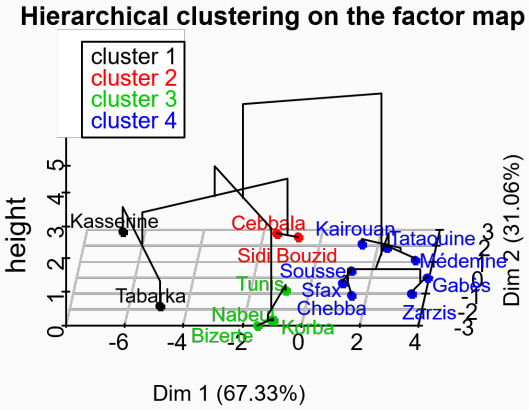


Arbre à partir de classes

Enchaînement analyse factorielle - classification

- Données qualitatives : Faire un pré-traitement par ACM
- L'analyse factorielle élimine les dernières composantes qui ne contiennent que du bruit $\Longrightarrow$ classification plus stable

- Représentation de l'arbre et des classes sur un plan factoriel



Conclusion

- La classification s'applique à des tableaux individus $\times$ variables quantitatives
  $\Rightarrow$ L'ACM transforme des variables qualitatives en variables quantitatives

Conclusion

- La classification s'applique à des tableaux individus $\times$ variables quantitatives
  $\Rightarrow$ L'ACM transforme des variables qualitatives en variables quantitatives

- CAH donne un arbre hiérarchique $\Rightarrow$ nombre de classes

Conclusion

- La classification s'applique à des tableaux individus $\times$ variables quantitatives
  $\Rightarrow$ L'ACM transforme des variables qualitatives en variables quantitatives

- CAH donne un arbre hiérarchique $\Rightarrow$ nombre de classes

- K-means consolide les classes

Conclusion

- La classification s'applique à des tableaux individus $\times$ variables quantitatives
  $\Rightarrow$ L'ACM transforme des variables qualitatives en variables quantitatives
- CAH donne un arbre hiérarchique $\Rightarrow$ nombre de classes
- K-means consolide les classes
- Caractérisation des classes par des variables actives et supplémentaires, quantitatives et qualitatives

Plan

Estimation vs apprentissage

- Estimation : objectifs explicatifs
  - notion de modèle
  - décisions prises à l'aide de tests statistiques
- Apprentissage

Estimation vs apprentissage

- Estimation : objectifs explicatifs
  - notion de modèle
  - décisions prises à l'aide de tests statistiques
- Apprentissage : objectifs prédictifs
  - complexité des modèles "peu" importante
  - décisions prises à l'aide de critères de prévisions

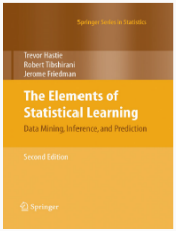
Vocabulaire

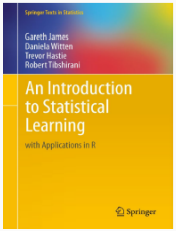
- Quand la variable à expliquer Y est quantitative, on parle de régression
- Quand Y est qualitative, on parle de discrimination ou de classification supervisée

Nombreuses applications

finance, économie, marketing, biologie, médecine...

Ressources





Disponibles (avec jeux de données, codes...) :

https://web.stanford.edu/~hastie/ElemStatLearn/

http://www-bcf.usc.edu/~gareth/ISL/

Wikistat

- Très bons cours et tutoriels sur la statistique classique et moderne
- Voir les vignettes sur la partie apprentissage :
  - http://wikistat.fr/pdf/st-m-Intro-ApprentStat.pdf
  - http://wikistat.fr/pdf/st-m-app-risque.pdf

Les supers cours de Laurent Rouvière

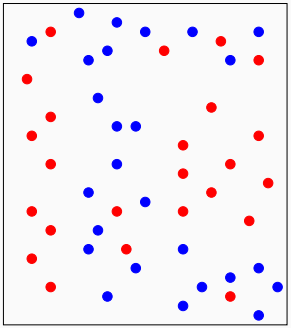
- https://lrouviere.github.io/machine_learning/cours.pdf

Présentation

- Les arbres sont des algorithmes de prédiction qui fonctionnent en régression (Y quantitative) et en discrimination (Y qualitative) ; les variables $X_1, \ldots, X_p$ peuvent être qualitatives et/ou quantitatives
- Il existe différentes variantes permettant de construire des prédicteurs par arbres.
- Nous nous focalisons dans cette partie sur la méthode CART (Breiman et al., 1984) qui est la plus utilisée.

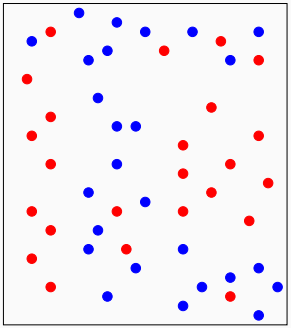
Représentation des données

- On dispose de n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ où $X_i \in \mathbb{R}^2$ et $Y_i \in \{-1, 1\}$.



Représentation des données

- On dispose de n observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ où $X_i \in \mathbb{R}^2$ et $Y_i \in \{-1, 1\}$.

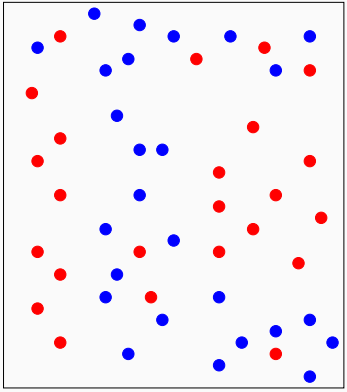


Approche par arbres

Trouver une partition des observations qui sépare "au mieux" les points rouges des points bleus.

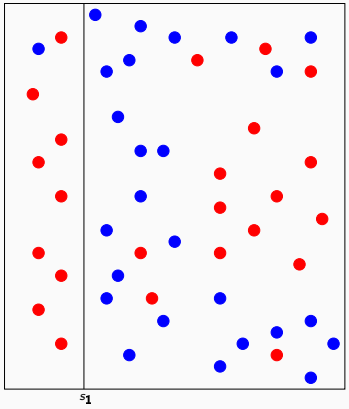
Arbre de décision CART

- C'est un algorithme qui construit une partition par divisions successives parallèles aux axes (une division = 1 variable et 1 seuil de coupure)
- et moyenne localement (moyenne ou vote à la majorité par classe)



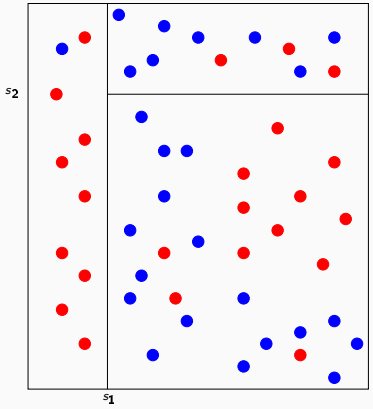
Arbre de décision CART

- C'est un algorithme qui construit une partition par divisions successives parallèles aux axes (une division = 1 variable et 1 seuil de coupure)
- et moyenne localement (moyenne ou vote à la majorité par classe)



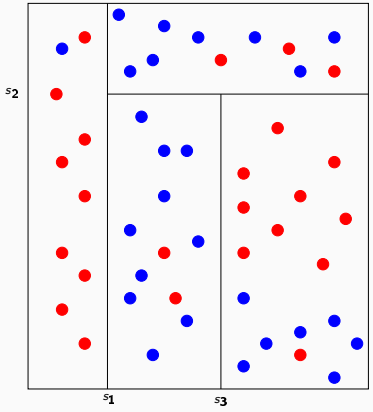
Arbre de décision CART

- C'est un algorithme qui construit une partition par divisions successives parallèles aux axes (une division = 1 variable et 1 seuil de coupure)
- et moyenne localement (moyenne ou vote à la majorité par classe)



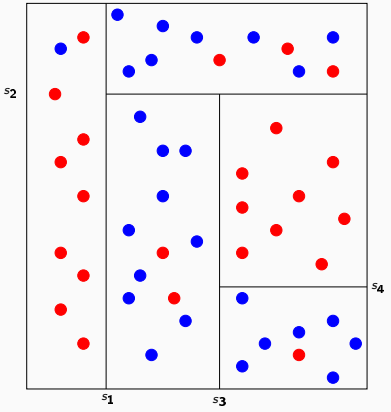
Arbre de décision CART

- C'est un algorithme qui construit une partition par divisions successives parallèles aux axes (une division = 1 variable et 1 seuil de coupure)
- et moyenne localement (moyenne ou vote à la majorité par classe)

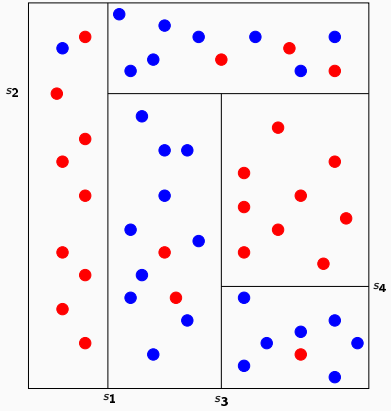


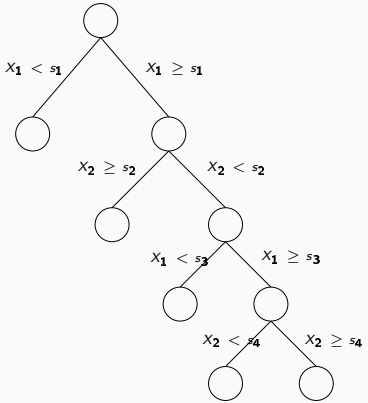
Arbre de décision CART

- C'est un algorithme qui construit une partition par divisions successives parallèles aux axes (une division = 1 variable et 1 seuil de coupure)
- et moyenne localement (moyenne ou vote à la majorité par classe)

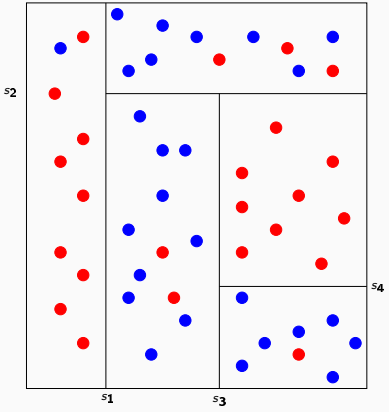


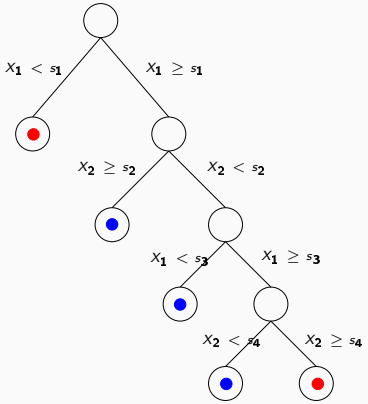
Représentation de l'arbre





Représentation de l'arbre





Règle de classification

On effectue un vote à la majorité dans les nœuds terminaux (ou feuilles) de l'arbre

Choix des découpes

Questions

- Comment choisir les découpes ?
- Faut-il stopper les découpes ? Si oui, quand ?

Choix des découpes

Questions

- Comment choisir les découpes ?
- Faut-il stopper les découpes ? Si oui, quand ?

- A chaque étape, on cherche un couple variable j - seuil s qui divise un noeud en deux nœuds fils pour optimiser l'(im)pureté ou l'hétérogénité de ces deux nœuds.

Choix des découpes

Questions

- Comment choisir les découpes ?
- Faut-il stopper les découpes ? Si oui, quand ?

- A chaque étape, on cherche un couple variable j - seuil s qui divise un noeud en deux nœuds fils pour optimiser l'(im)pureté ou l'hétérogénité de ces deux nœuds.

Critère

Régression : choisir le couple (j, s) qui maximise $n\mathbf{V}(Y) - (n_g\mathbf{V}(Y_g) + n_d\mathbf{V}(Y_d))$
Discrimination : maximiser l'indice de Gini $np(1-p) - (n_g p_g(1-p_g) + n_d p_d(1-p_d))$
avec n, $\mathbf{V}(Y)$ et p resp. le nombre d'individus, la variance de Y et la proportion de 1 dans le nœud, les indices $_g$ et $_d$ indiquent s'il s'agit du nœud gauche ou droite

Questions

- Comment construire un "bon" arbre ?

Questions

- Comment construire un "bon" arbre ?
- Construire l'arbre maximal ? (on découpe les nœuds jusqu'à ce qu'on ne puisse plus).

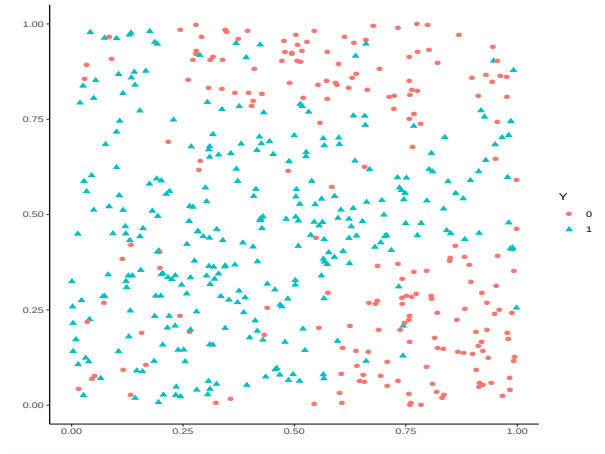
Questions

- Comment construire un "bon" arbre ?
- Construire l'arbre maximal ? (on découpe les nœuds jusqu'à ce qu'on ne puisse plus).
- Faut-il se donner un critère d'arrêt ?

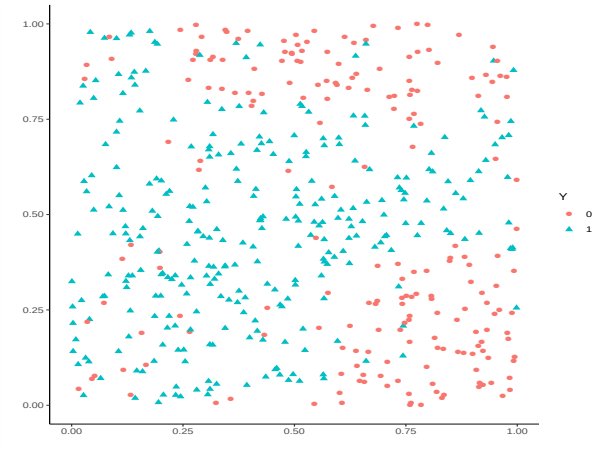
Questions

- Comment construire un "bon" arbre ?
- Construire l'arbre maximal ? (on découpe les nœuds jusqu'à ce qu'on ne puisse plus).
- Faut-il se donner un critère d'arrêt ?
- Faut-il construire un arbre grand et choisir un sous-arbre de ce dernier ?

Un exemple en discrimination



Un exemple en discrimination

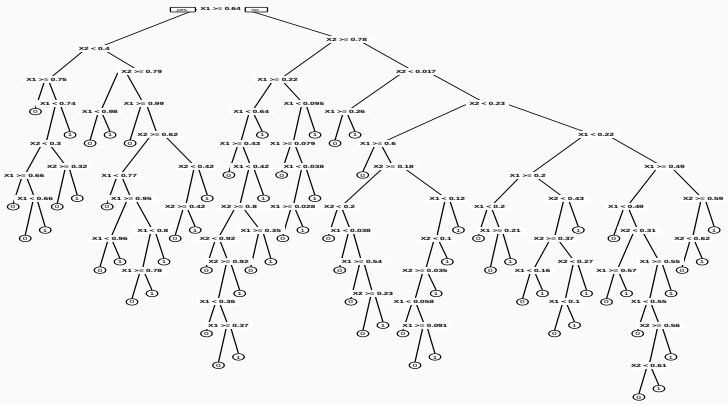


Arbre optimal ?

Intuitivement, on a envie de faire à peu près 5 classes.

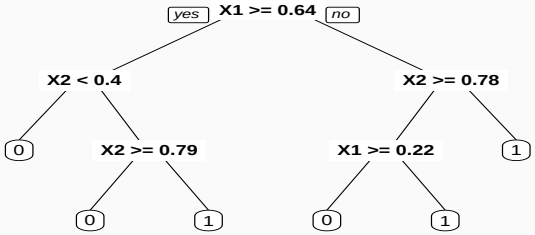
Arbre « maximal »

```
> library(rpart)
> library(rpart.plot)
> arbre1 <- rpart(Y~.,data=dta[1:350,],cp=0.0001,minsplit=2)
> prp(arbre1)
```



Un arbre plus petit

```
> arbre2 <- rpart(Y~.,data=dta[1:350,])
> prp(arbre2)
```



Comparaison des deux arbres ...

- ... par leur probabilité de mauvais classement sur un échantillon test :

```
> prev <- data.frame(arbre1=predict(arbre1,newdata=dta[351:500,],type="class"),
+                    arbre2=predict(arbre2,newdata=dta[351:500,],type="class"),
+                    obs=dta[351:500,]$Y)
> prev %>% summarize_at(1:2,funs(mean(.!=obs))) %>% round(3)
##   arbre1 arbre2
## 1  0.133  0.127
```

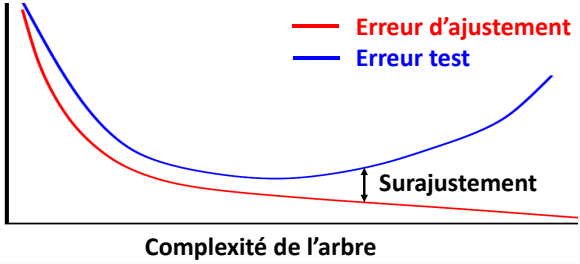
Comparaison des deux arbres ...

- ... par leur probabilité de mauvais classement sur un échantillon test :

```
> prev <- data.frame(arbre1=predict(arbre1,newdata=dta[351:500,],type="class"),
+                    arbre2=predict(arbre2,newdata=dta[351:500,],type="class"),
+                    obs=dta[351:500,]$Y)
> prev %>% summarize_at(1:2,funs(mean(.!=obs))) %>% round(3)
##   arbre1 arbre2
## 1  0.133  0.127
```

Conclusion

La performance n'augmente pas forcément avec la complexité (la profondeur) de l'arbre



Principe d'élagage

Biais et variance

La profondeur régule le compromis biais/variance :

1. Peu de découpes (arbres peu profonds) $\Longrightarrow$ arbres stables $\Longrightarrow$ peu de variance... mais... beaucoup de biais
2. Beaucoup de découpes (arbres profonds) $\Longrightarrow$ arbres instables $\Longrightarrow$ peu de biais... mais... beaucoup de variance (surapprentissage)

Principe d'élagage

Biais et variance

La profondeur régule le compromis biais/variance :

1. Peu de découpes (arbres peu profonds) $\Longrightarrow$ arbres stables $\Longrightarrow$ peu de variance... mais... beaucoup de biais
2. Beaucoup de découpes (arbres profonds) $\Longrightarrow$ arbres instables $\Longrightarrow$ peu de biais... mais... beaucoup de variance (surapprentissage)

Principe d'élagage

Plutôt que de choisir « quand couper » on raisonne en 3 temps :

1. On construit un arbre maximal (très profond) $\mathcal{T}_{max}$
2. On sélectionne une suite d'arbres emboités avec 2 feuilles, 3, 4, ...
3. On sélectionne un arbre dans cette sous-suite

Sorties printcp

- Sur R, on obtient cette sous-suite à l'aide de la fonction printcp :

```
> arbre <- rpart(Y~.,data=dta,cp=0.0001,minsplit=2)
> printcp(arbre)
## Classification tree:
## rpart(formula=Y~., data=dta, cp=1e-04, minsplit=2)
##
## Variables actually used in tree construction:
## [1] X1 X2
##
## Root node error: 204/500 = 0.408
## n= 500
##
##           CP nsplit rel error  xerror     xstd
## 1  0.2941176      0  1.000000 1.00000 0.053870
## 2  0.1225490      1  0.705882 0.72059 0.049938
## 3  0.0931373      3  0.460784 0.51471 0.044646
## 4  0.0637255      4  0.367647 0.42647 0.041555
## 5  0.0122549      5  0.303922 0.35294 0.038483
## 6  0.0098039      7  0.279412 0.35294 0.038483
## 7  0.0049020      9  0.259804 0.35784 0.038704
## 8  0.0040107     25  0.181373 0.39216 0.040184
## 9  0.0036765     41  0.112745 0.39706 0.040386
## 10 0.0032680     49  0.083333 0.40196 0.040586
## 11 0.0024510     52  0.073529 0.41667 0.041174
## 12 0.0001000     82  0.000000 0.45098 0.042473
```

- Suite de 12 arbres emboîtés
- CP : complexity parameter : CP $\searrow$ $\Longrightarrow$ complexité arbre $\nearrow$
- nsplit : nombre de coupures de l'arbre
- rel.error : erreur d'ajustement
- xerror : erreur de prévision (par validation croisée 10 blocs)
- xstd : écart-type de l'erreur de validation croisée

Sorties printcp

- Sur R, on obtient cette sous-suite à l'aide de la fonction printcp :

```
> arbre <- rpart(Y~.,data=dta,cp=0.0001,minsplit=2)
> printcp(arbre)
## Classification tree:
## rpart(formula=Y~., data=dta, cp=1e-04, minsplit=2)
##
## Variables actually used in tree construction:
## [1] X1 X2
##
## Root node error: 204/500 = 0.408
## n= 500
##
##           CP nsplit rel error  xerror     xstd
## 1  0.2941176      0  1.000000 1.00000 0.053870
## 2  0.1225490      1  0.705882 0.72059 0.049938
## 3  0.0931373      3  0.460784 0.51471 0.044646
## 4  0.0637255      4  0.367647 0.42647 0.041555
## 5  0.0122549      5  0.303922 0.35294 0.038483
## 6  0.0098039      7  0.279412 0.35294 0.038483
## 7  0.0049020      9  0.259804 0.35784 0.038704
## 8  0.0040107     25  0.181373 0.39216 0.040184
## 9  0.0036765     41  0.112745 0.39706 0.040386
## 10 0.0032680     49  0.083333 0.40196 0.040586
## 11 0.0024510     52  0.073529 0.41667 0.041174
## 12 0.0001000     82  0.000000 0.45098 0.042473
```

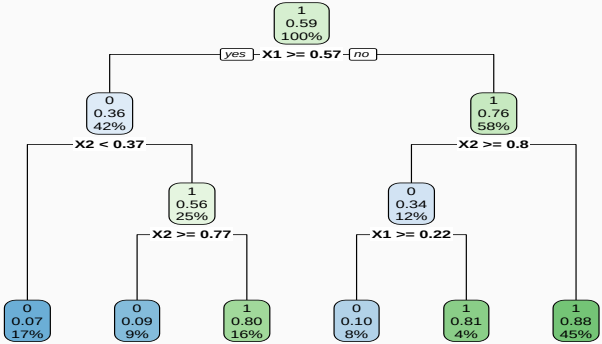
- Suite de 12 arbres emboîtés
- CP : complexity parameter : CP $\searrow$ $\Longrightarrow$ complexité arbre $\nearrow$
- nsplit : nombre de coupures de l'arbre
- rel.error : erreur d'ajustement
- xerror : erreur de prévision (par validation croisée 10 blocs)
- xstd : écart-type de l'erreur de validation croisée

Choix de l'arbre final

Choisir l'arbre qui a la plus petite erreur de prévision (calculée par validation croisée)
$\Longrightarrow$ ici arbre à 5 coupures (et donc 6 feuilles)

Tracé de l'arbre final

```
> LigneARetenir <- which.min(arbre$cptable[,"xerror"])
> cp_opt <- arbre$cptable[LigneARetenir,"CP"]
> cp_opt
## [1] 0.0122549
> arbre_final <- prune(arbre,cp = cp_opt)
> rpart.plot(arbre_final)
```



Un arbre interactif avec visNetwork

- `visTreeEditor` : fonction qui prend en entrée un jeu de données et lance une appli shiny pour construire son arbre et le récupérer ensuite
- `visTree` : fonction pour visualiser un arbre de façon interactive

```
> visTree(arbre)
```

Règle de classification et score par arbre

- Règle de classification : un nouvel individu est dirigé dans la feuille qui convient et on lui attribue la classe la plus probable pour la feuille
- Score : probabilité que $Y = 1$ quand un nouvel individu arrive dans une feuille
- Fonction predict (predict.rpart) estime la classe ou le score :

```
> x_new <- data.frame(X1=0.5,X2=0.85)
> predict(arbre_final,newdata=x_new)
##     0   1
## 1 0.9 0.1
> predict(arbre_final,newdata=x_new,type="class")
## 1
## 0
## Levels: 0 1
```

Bilan

- Méthode « simple » relativement facile à mettre en œuvre.
- Fonctionne en régression et en discrimination
- Résultats interprétables (à condition que l'arbre ne soit pas trop profond)
- Un inconvénient : méthode connue pour être instable, sensible à de légères perturbations de l'échantillon
- Cet inconvénient sera un avantage pour des agrégations bootstrap $\Longrightarrow$ forêts aléatoires

Plan

Forêts aléatoires

- Comme son nom l'indique, une forêt aléatoire est définie à partir d'un ensemble d'arbres
- Les forêts aléatoires les plus utilisées sont celles de Léo Breiman (2000)
- Elles consistent à agréger des arbres construits sur B échantillons bootstrap

Principe

Prédire par la moyenne des prévisions de nombreux arbres "indépendants"

- On pourra trouver de la doc à partir de l'url

  `http://www.stat.berkeley.edu/~breiman/RandomForests/`

  et consulter la thèse de Robin Genuer (2010)

Forêt aléatoire de Breiman – algorithme randomforest

Algorithme randomforest

Pour construire $k = 1, \ldots, B$ arbres "indépendants" :

1. Tirer un échantillon bootstrap dans $\mathcal{D}_n$
2. Construire un arbre CART sur cet échantillon bootstrap ; chaque coupure est obtenue en choisissant la meilleure variable dans un ensemble de m variables choisies au hasard parmi les p ; ne pas élaguer l'arbre

Prévision par aggrégation :
En régression (Y quantitative) : prendre la moyenne des prédictions des arbres
En discrimination (Y qualitative) : vote à la majorité des arbres

Choix des paramètres

- **B** : le plus grand possible
- Arbres "profonds" : peu d'observations dans les nœuds terminaux (par défaut dans randomForest, $n_{max} = 5$ en régression et 1 en classification)
- **m** : comparer les performances de la forêt pour plusieurs valeurs de m, le nombre de variables choisies aléatoirement à chaque coupure d'un nœud (par défaut $m = p/3$ en régression et $\sqrt{p}$ en classification)

Choix des paramètres

- **B** : le plus grand possible
- Arbres "profonds" : peu d'observations dans les nœuds terminaux (par défaut dans randomForest, $n_{max} = 5$ en régression et 1 en classification)
- **m** : comparer les performances de la forêt pour plusieurs valeurs de m, le nombre de variables choisies aléatoirement à chaque coupure d'un nœud (par défaut $m = p/3$ en régression et $\sqrt{p}$ en classification)

Commentaires

- Deux sources d'aléa : tirage bootstrap et m variables sélectionnées à chaque nœud de l'arbre
- Peu sensible au choix de ses paramètres (B, n_{max}, m...)

Intérêt de l'aggrégation ?

- Biais reste le même
- Variance :

$$\mathbf{V}(\text{forêt}) = \rho \mathbf{V}(\textit{arbre}) + \frac{1-\rho}{B}\mathbf{V}(\textit{arbre})$$

où $\rho =$ corrélation entre les prévisions des arbres, et $\mathbf{V}$(arbre) est la variance de prédiction des arbres qui dépend des 2 sources d'aléas (échantillon bootstrap, choix des variables)

Définition

Les forêts aléatoires sont une méthode de Bagging (pour Bootstrap Aggregating)

Application sur les données spam

```
> library(randomForest)
> foret1 <- randomForest(type~.,data=spam)
> foret1
##
## Call:
##  randomForest(formula = type ~ ., data = spam)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 7
##
##         OOB estimate of  error rate: 4.52%
## Confusion matrix:
##         nonspam spam class.error
## nonspam    2708   80  0.02869440
## spam        128 1685  0.07060121
```

Mesure de performance : Erreur Ouf Of Bag

Idée : Evaluer la qualité de prévision de la forêt par validation croisée sans refaire de calcul

- Pour chaque observation (X_i, Y_i) de $\mathcal{D}_n$, plusieurs arbres de la forêt ne contiennent pas cette observation dans leur échantillon bootstrap
- La prévision de la forêt $\hat{Y}_i$ au point X_i est calculée en moyennant la prévision de tous les arbres qui ne contiennent pas l'observation i

Mesure de performance : Erreur Ouf Of Bag

Idée : Evaluer la qualité de prévision de la forêt par validation croisée sans refaire de calcul

- Pour chaque observation (X_i, Y_i) de $\mathcal{D}_n$, plusieurs arbres de la forêt ne contiennent pas cette observation dans leur échantillon bootstrap
- La prévision de la forêt $\hat{Y}_i$ au point X_i est calculée en moyennant la prévision de tous les arbres qui ne contiennent pas l'observation i

Erreurs Out Of Bag

- L'erreur de prédiction de la forêt est estimée par $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$.
- La probabilité d'erreur est estimée par $\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{Y}_i \neq Y_i}$.

Exemple

3	4	6	10	3	9	10	7	7	1	A_1
2	8	6	2	10	10	2	9	5	6	A_2
2	9	4	4	7	7	2	3	6	7	A_3
6	1	3	3	9	3	8	10	10	1	A_4
3	7	10	3	2	8	6	9	10	2	A_5
7	10	3	4	9	10	10	8	6	1	A_6

Exemple

3	4	6	10	3	9	10	7	7	1	A_1
2	8	6	2	10	10	2	9	5	6	A_2
2	9	4	4	7	7	2	3	6	7	A_3
6	1	3	3	9	3	8	10	10	1	A_4
3	7	10	3	2	8	6	9	10	2	A_5
7	10	3	4	9	10	10	8	6	1	A_6

- Les échantillons 2, 3 et 5 ne contiennent pas la première observation, donc $\hat{Y}_1 = \frac{1}{3}(A_2(X_1) + A_3(X_1) + A_5(X_1))$
- On fait de même pour toutes les observations $\Longrightarrow \hat{Y}_2, \ldots, \hat{Y}_n$

Exemple

3	4	6	10	3	9	10	7	7	1	A_1
2	8	6	2	10	10	2	9	5	6	A_2
2	9	4	4	7	7	2	3	6	7	A_3
6	1	3	3	9	3	8	10	10	1	A_4
3	7	10	3	2	8	6	9	10	2	A_5
7	10	3	4	9	10	10	8	6	1	A_6

- Les échantillons 2, 3 et 5 ne contiennent pas la première observation, donc $\hat{Y}_1 = \frac{1}{3}(A_2(X_1) + A_3(X_1) + A_5(X_1))$
- On fait de même pour toutes les observations $\Longrightarrow \hat{Y}_2, \ldots, \hat{Y}_n$
- On estime l'erreur par $\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2$

Exemple

- On construit la forêt avec $m = 1$:

```
> foret2 <- randomForest(type~.,data=spam,mtry=1)
> foret2
##
## Call:
##  randomForest(formula = type ~ ., data = spam, mtry = 1)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##         OOB estimate of  error rate: 8.06%
## Confusion matrix:
##         nonspam spam class.error
## nonspam    2725   63  0.02259684
## spam        308 1505  0.16988417
```

Exemple

- On construit la forêt avec $m = 1$:

```
> foret2 <- randomForest(type~.,data=spam,mtry=1)
> foret2
##
## Call:
##  randomForest(formula = type ~ ., data = spam, mtry = 1)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 1
##
##         OOB estimate of  error rate: 8.06%
## Confusion matrix:
##         nonspam spam class.error
## nonspam    2725   63  0.02259684
## spam        308 1505  0.16988417
```

Remarque

L'erreur OOB est de 8.06%, elle est de 4.52% lorsque $m = 7$.

Conclusion sur les forêts aléatoires

Les forêts aléatoires

- donnent des prédictions précises sur données complexes (beaucoup de variables, variables quantitatives et qualitatives, données manquantes, liaison non linéaires entre variables, interactions, ...)
- sont peu sensibles au choix de ses paramètres (B, n_{max}, m...)
- sont simples à mettre en oeuvre (fonction randomForest du package randomForest)
- sont un peu boîte noire et manquent d'interprétabilité par rapport aux modèles paramétriques comme le modèle logistique même si un indicateur permet de mesurer l'importance des variables présentes dans le modèle