

Classification hiérarchique ascendante (CAH)

Objectif de la méthode

Cette méthode de classification automatique a pour objectif ;

- de répartir des individus dans un certain nombre de classes selon une règle d'agrégation

Données manipulées

Variables quantitatives uniquement

(Remarque¹ : il est possible d'utiliser des variables qualitatives en utilisant au préalable une analyse des correspondances multiples.)

Exemple

ID	Fidélité Marque	Sensibilité au prix	Acheteur en ligne	Amer	Glacé	Croquant
Bobby	7	4	7	6	5	9
Muriel	5	5	5	4	6	5
Shelia	7	5	9	6	4	1
Juana	5	6	3	4	4	3
Tami	2	6	7	4	6	7
Frank	4	5	4	6	4	9
Sam	1	7	1	4	6	10
Marsha	4	7	5	3	7	10
Dominic	8	5	4	6	4	3
Kevin	5	3	3	9	0	0
Melinda	4	6	1	2	6	7
Candice	5	6	2	4	5	2
Sherri	2	5	4	3	6	0
Jordan	1	6	1	2	5	2
Anita	9	7	3	5	4	2
Alfredo	9	4	1	6	4	5

Figure extraite de l'article de XLStat, voir annexe

Description

L'analyse hiérarchique ascendante fait partie de la famille des méthodes de regroupement hiérarchique de l'analyse de donnée. Cette famille recouvre différentes méthodes de clustering et se distingue en deux catégories : les méthodes ascendantes et les méthodes descendantes.

Les méthodes dites "descendantes" partent d'une solution générale vers une autre plus spécifique. Les méthodes de cette catégorie démarrent avec un seul cluster contenant la totalité puis se divisent à chaque étape selon un critère jusqu'à l'obtention d'un ensemble de clusters différents.

A l'inverse des méthodes dites "descendantes", la classification ascendante hiérarchique est dite "ascendante" car elle part d'une situation où tous les individus sont seuls dans une classe, puis sont rassemblés en classes de plus en plus grandes.

Définition de "cluster" : Ensemble de données ou individus partageant une caractéristique commune.

Base théorique

Le qualificatif *hiérarchique* vient du fait qu'elle produit une hiérarchie H , l'ensemble des classes à toutes les étapes de l'algorithme, qui vérifie les propriétés suivantes :

- $\Omega \in H$: au sommet de la hiérarchie, lorsqu'on groupe de manière à obtenir une seule classe, tous les individus sont regroupés ;
- $\forall \omega \in \Omega, \{ \omega \} \in H$: en bas de la hiérarchie, tous les individus se trouvent seuls ;
- $\forall (h, h') \in H^2, h \cap h' = \emptyset$ ou $h \subset h'$ ou $h' \subset h$: si l'on considère deux classes du regroupement, alors soit elles n'ont pas d'individu en commun, soit l'une est incluse dans l'autre.

(Ω constitue l'ensemble des n individus.)

La méthode suppose qu'on dispose d'une mesure de dissimilarité entre les individus; dans le cas de points situés dans un espace euclidien, on peut utiliser la *distance* comme mesure de dissimilarité. La dissimilarité entre des individus x et y sera notée $\text{dissim}(x,y)$..

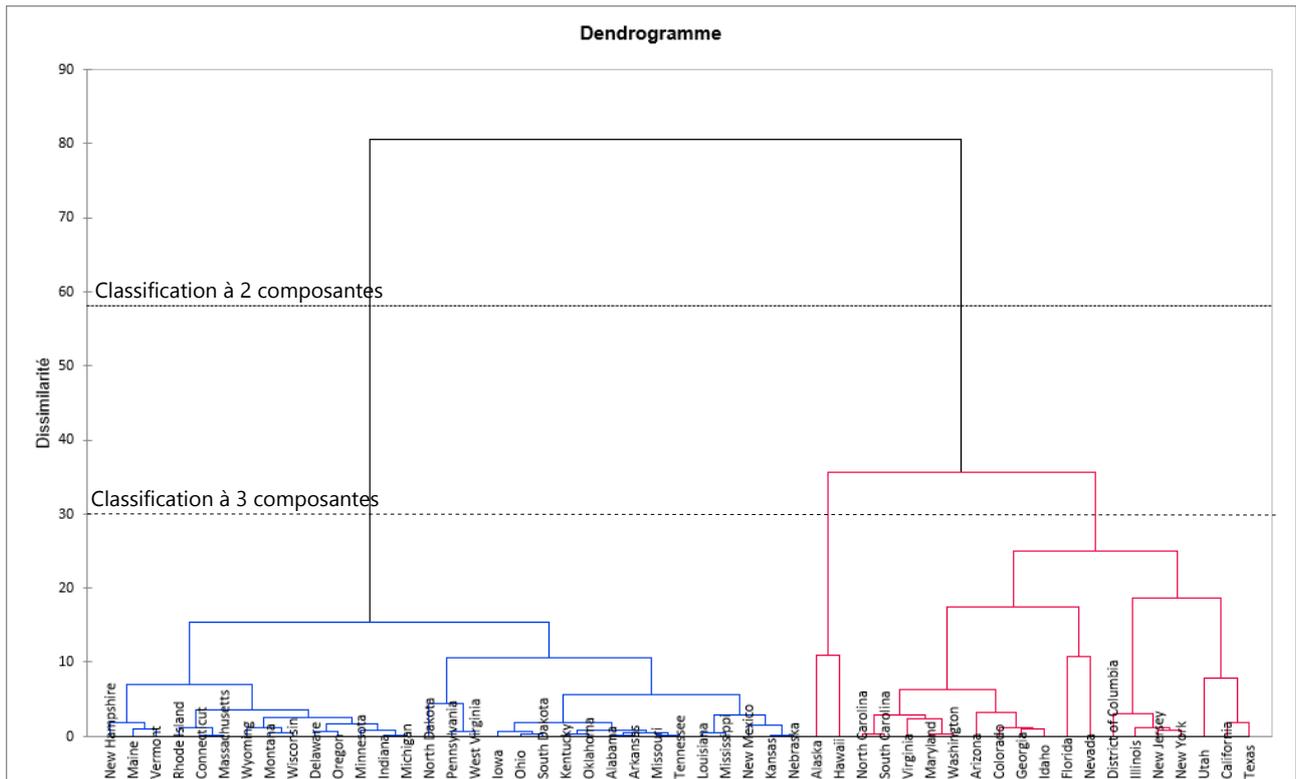
Principe algorithmique :

Initialement, chaque individu forme une classe, soit n classes. On cherche à réduire le nombre de classes à $n_{\text{classes}} < n$, ceci se fait itérativement. À chaque étape, on fusionne deux classes, réduisant ainsi le nombre de classes. Les deux classes choisies pour être fusionnées sont celles qui sont les plus « proches », en d'autres termes, celles dont la dissimilarité entre elles est minimale, cette valeur de dissimilarité est appelée *indice d'agrégation*. Comme on rassemble d'abord les individus les plus proches, la première itération a un indice d'agrégation faible, mais celui-ci va croître d'itération en itération.

Il existe de multiples critères qui permettent de calculer la dissimilarité, ils sont majoritairement basés sur des calculs de distances et d'inertie des données.

Interprétation des résultats d'une CAH

Un dendrogramme est la représentation graphique usuelle d'une classification ascendante hiérarchique. Il se présente souvent comme un arbre binaire dont les feuilles sont les individus alignés sur l'axe des abscisses. Lorsque deux classes ou deux individus se rejoignent avec l'indice d'agrégation τ , des traits verticaux sont dessinés de l'abscisse des deux classes jusqu'à l'ordonnée τ , puis ils sont reliés par un segment horizontal. À partir d'un indice d'agrégation τ , on peut tracer une droite d'ordonnée τ qui permet de voir une classification sur le dendrogramme.



Exemple de dendrogramme, figure extraite de l'article de XLStat, voir annexe.

¹Bien que la CAH s'applique uniquement sur des variables quantitatives, il est tout à fait possible de segmenter une base de données constitués de variables qualitatives. Il suffit de construire une analyse des correspondances multiples (ACM) sur les données en question puis d'exécuter une CAH sur les coordonnées des observations générées par l'ACM.

Annexe ;

Analyse-R : <http://larmarange.github.io/analyse-R/classification-ascendante-hierarchique.html>

Wikipédia : https://fr.wikipedia.org/wiki/Regroupement_h%C3%A9rarchique

Article de XLStat sur comment choisir sa méthode de classification :

<https://help.xlstat.com/s/article/choisir-une-methode-de-classification-avec-xlstat?language=fr>

XLStat CAH : <https://help.xlstat.com/s/article/classification-ascendante-hierarchique-cah-dans-excel?language=fr>