HAX711X - Analyse des Données Multidimensionnelles
DM2 Classification automatique

SCAIA Matteo et MARIAC Damien

28 novembre 2024

Table des matières

1 Treillis de Galois

1.1 Introduction

Le treillis de Galois est une structure mathématique utilisée en analyse de données pour extraire des règles d'implication. Il est construit à partir de données décrites par des propriétés booléennes, permettant de représenter les relations entre ces propriétés et les ensembles d'objets associés. Le treillis de Galois peut également intégrer des relations liant les données entre elles.

1.2 Interprétation et création du treillis de Galois

Dans cette question, nous allons analyser le treillis de Galois construit à partir des données fournies par le sujet afin de modéliser les relations entre les films et leurs caractéristiques.

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
Default N...	E_O	E_N	A_O	A_N	L<30	30<L<60	60<L<180	L>180	D_O	D_N	Edu_O	Edu_N	Ir_o	Ir_N	Is_O	Is_N
Série polic...	0	X	X	0	0	0	0	X	X	0	0	X	X	0	0	X
Série hum...	X	0	X	0	0	0	0	X	X	0	0	X	X	0	0	X
Long métr...	X	0	X	0	0	0	X	0	X	0	0	X	0	X	X	0
Court mét...	X	0	0	X	X	0	0	0	X	0	0	X	0	X	X	0
Clip chans...	X	0	X	0	X	0	0	0	X	0	0	X	X	0	X	0
Document...	X	0	X	0	0	X	0	0	X	0	X	0	X	0	0	X
Document...	0	X	X	0	0	X	0	0	0	X	X	0	X	0	X	0
Document...	0	X	X	0	0	X	0	0	X	0	X	0	X	0	X	0
Document...	0	X	X	0	0	X	0	0	X	0	X	0	X	0	X	0
Film de fa...	X	0	X	0	0	0	X	0	X	0	0	X	X	0	X	0
Film horre...	0	X	X	0	0	0	X	0	X	0	0	X	X	0	X	0
Film dram...	0	X	X	0	0	0	X	0	X	0	0	X	X	0	0	X
Film polici...	0	X	X	0	0	0	X	0	X	0	0	X	X	0	0	X
Film comi...	X	0	X	0	0	0	X	0	X	0	0	X	X	0	0	X

Figure 1 – Tableau des relations binaires

À l'aide du logiciel Galicia, nous obtenons le treillis de Galois suivant :

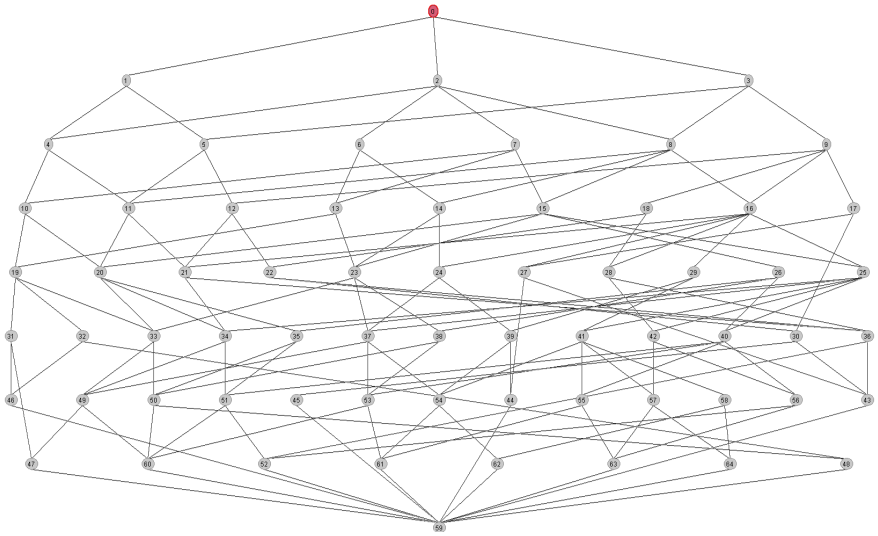

Figure 2 – Treillis de Galois de notre tableau

Dans un treillis de Galois, les nœuds situés aux extrémités correspondent soit à l'ensemble de tous les individus (en haut), soit à l'ensemble de toutes les caractéristiques (en bas). Ces nœuds étant trop généraux ou trop spécifiques, leur analyse n'est pas nécessaire.

Le nœud 57 regroupe les films dramatiques (FD), les films policiers (FP) et les séries policières (SP), caractérisés par les propriétés suivantes : ils s'adressent aux adolescents et aux adultes (A = Oui), ont un objectif distractif (D = Oui), ne sont pas éducatifs (ED = Non), ne ciblent pas les enfants (E = Non), utilisent des images réelles (IR = Oui) et n'incluent pas d'images de synthèse (IS = Non). Ce regroupement correspond à des œuvres qui partagent des thématiques et des intentions narratives du genre "thriller".
Par ailleurs, ce nœud est lié à la classe 42, qui partage les mêmes caractéristiques mais inclut également des œuvres utilisant des images de synthèse. Cela permet d'y intégrer des films d'horreur (FH).
Cette classe peut être interprétée comme un regroupement de film conçues pour provoquer du suspense ou des "frissons".

Le nœud 16 représente une classe large regroupant plusieurs types de films et séries partageant diverses caractéristiques : un public adulte, une vocation distractive, et des images réelles. Ce nœud est particulièrement intéressant, car il se connecte à plusieurs autres classes. Ces caractéristiques sont partagées par des genres variés. De ce fait, cette classe regroupe de nombreux types de films et séries, tels que les clips musicaux, les documentaires artistiques, les documentaires sur la nature, les documentaires scientifiques, les films comiques, les films dramatiques, les films de fantasy, les films d'horreur, ainsi que les séries humoristiques et policières.

On remarque également le nœud 17, qui regroupe tous les documentaires. Les caractéristiques de cette classe sont qu'elle s'adresse aux adultes (A = Oui), a une vocation éducative (ED = Oui), utilise des images réelles (IR = Oui) et correspond à des productions de courte durée (30-60 minutes). Ces caractéristiques décrivent ce que sont les documentaires.

2 Classification hiérarchique de parcelles forestières tropicales

Charger dans le logiciel les données relatives au peuplement arboré de la forêt du bassin du Congo (Datagenus.csv). Inspectez le fichier et corrigez-en les erreurs triviales s'il en est. Ces données fournissent sur 1000 parcelles de cette forêt : les variables de comptage de 27 espèces d'arbres (gen1, ..., gen27), la surface de la parcelle, le type forestier (forest) tel qu'identifié par les écologues . On ne tiendra pas compte des autres variables. Calculer la densité de peuplement de chaque espèce par unité de surface pour les 1000 parcelles. Les parcelles seront traduits en nuage dans l'espace des 27 densités de peuplement.

L4.2 -¿ justification a l'oral : 1 min

2.1 Préparation des données

Nous traduisons les observations dans l'espace euclidien, où la distance servira de mesure de dissimilarité globale entre les parcelles. Cette mesure repose sur la comparaison des écarts sur les différentes dimensions (densités des espèces). Puisque toutes les dimensions (espèces) sont supposées avoir une importance équivalente, aucune ne doit dominer le calcul.

On note x_i la i-ième parcelle. La distance euclidienne entre deux parcelles x_1 et x_2 est :

$$d(x_1, x_2) = \sqrt{\sum_{j=1}^{27} (x_1^j - x_2^j)^2}$$

où x_1^j et x_2^j représentent respectivement la densité de l'espèce j dans les parcelles 1 et 2.

En analysant la contribution de chaque espèce, nous obtenons le tableau suivant :

Table 1 – Contribution des espèces (gen1 à gen11)

	gen1	gen2	gen3	gen4	gen5	gen6	gen7	gen8	gen9	gen10	gen11
Pourcentage	0.23	0.00	0.03	0.00	0.01	0.05	0.00	64.90	0.07	0.48	1.03

Nous voyons que les variables ne contribuent pas tous de la même manière. Les contributions des espèces gen1 à gen27 à la distance euclidienne ne sont pas uniformes. Par exemple, gen8 représente 64.90 % de la distance, ce qui montre que les variations dans cette dimension dominent les calculs de dissimilarité.

Nous standardisons les densités de peuplement et on note $\mathbf{Z}$ la matrice associée :

$$z_i^j = \frac{x_i^j - \overline{x}^j}{\sigma_{x^j}}$$

avec x_i^j la densité de l'espèce j sur la parcelle i, $\overline{x}^j$ la moyenne des densités pour l'espèce j, et σ_{x^j} l'écart-type de l'espèce j.

La standardisation assure que chaque dimension contribue de manière équivalente à la mesure de la dissimilarité, indépendamment de son échelle initiale ou de sa variabilité.

2.2 CAH des parcelles sur les densités de peuplement

2.2.1 Classification ascendante hiérarchique

Dans cette section, nous procéderons à la classification des parcelles en fonction de leur peuplement arboré. L'objectif est de déterminer le nombre optimal de classes pertinentes pour cette partition. À partir des données standardisées représentant les densités de peuplement obtenues à la question 2.1, nous utiliserons le code en langage R fourni dans le sujet pour effectuer une classification ascendante hiérarchique. Cette méthode nous permettra d'identifier les classes les plus adaptées à la caractérisation des parcelles.

• Indice de Ward
Nous obtenons le dendogramme suivant :

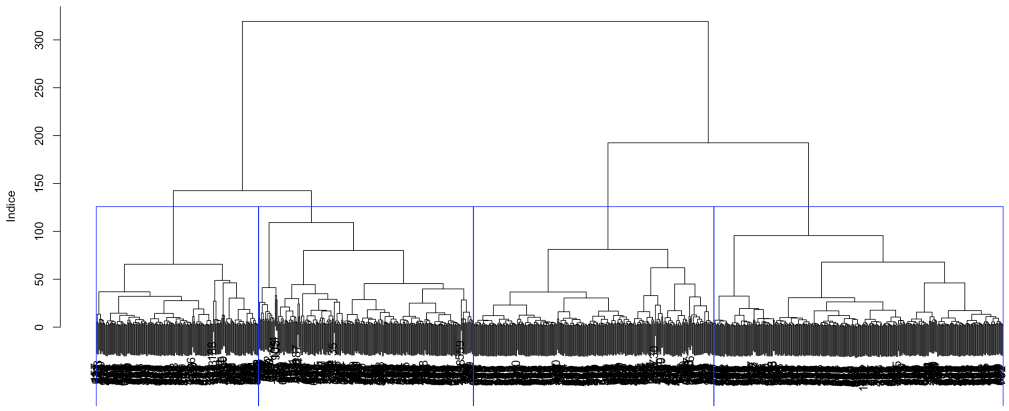

Figure 3 – Dendogramme (indice de Ward)

Le choix de couper en 4 classes signifie que les données ont été divisées en 4 groupes distincts. Pour confirmer notre choix d'utiliser 4 classes, nous pouvons utiliser l'histogramme des niveaux de μ_{ward}.

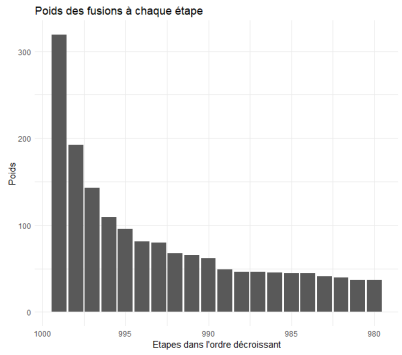

Figure 4 – Histogramme des niveaux μ_{ward}

On observe sur la figure que la différence de coût d'agrégation entre quatres classes et cinq classes n'est pas élevé, on peut donc partitionner en 4 classes.

• Indice du saut maximum

Avec l'indice de saut maximum, on obtient le dendogramme suivant :

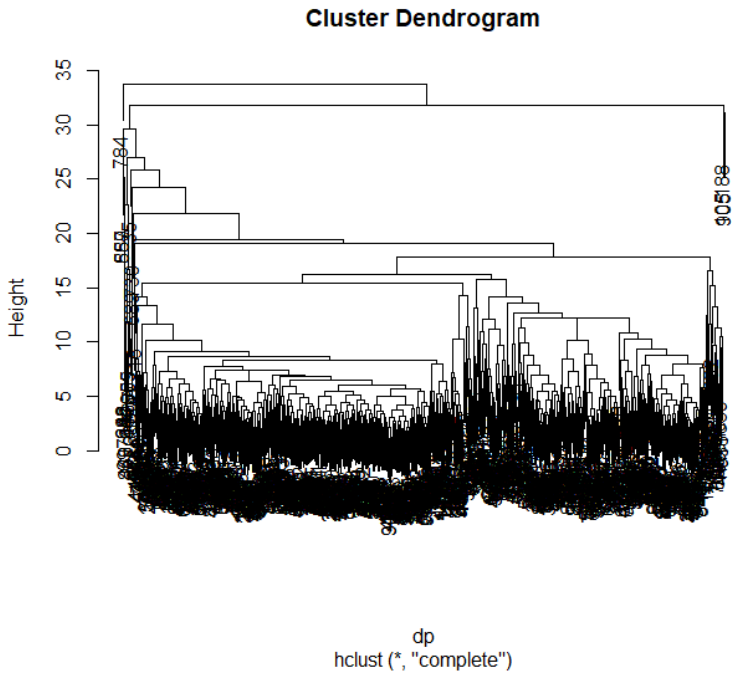

Figure 5 – Dendrogramme indice du saut max

Avec son histogramme associé

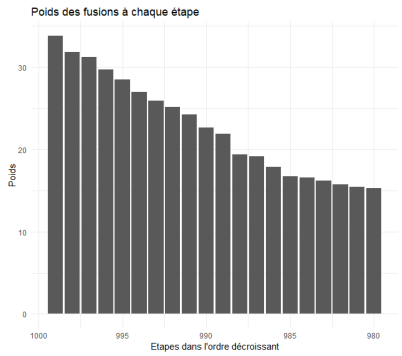

Figure 6 – Histogramme des niveaux saut maximum

- Indice du saut minimum

Bien que cette indice ne soit efficace que pour étudier les continuité/chaines de points, on peut l'afficher. Il donne le dendogramme suivant :

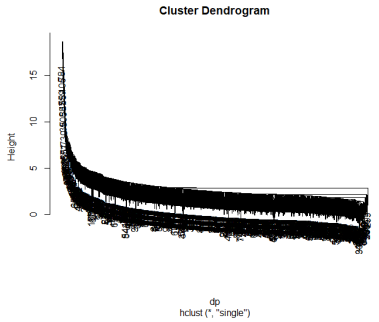

Figure 7 – Dendrogramme des niveaux saut minimum

Avec son histogramme associé :

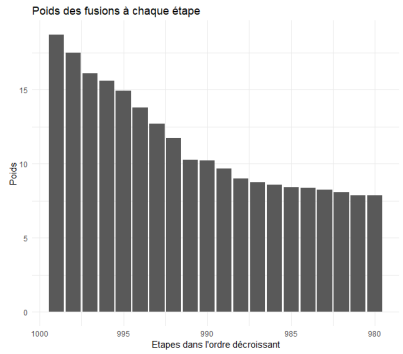

Figure 8 – Histogramme des niveaux saut minimum

2.2.2 Autour du R^2

2.3 Optimisation d'une partition avec les K-means

Notons :
— $X \in \mathbb{R}^{1000\times 27}$ la matrice des variables quantitatives.
— $W = \frac{1}{1000} I_{1000}$ la matrice des poids des individus.
— $C \in \mathbb{R}^{1000\times 1000}$ la matrice des centres de gravité.