Classification non supervisée

Résumé

Méthodes de classification non supervisée (ou clustering*). Notions de distance, classification ascendante hiérarchique et choix de distances entre classes, construction du dendrogramme, choix du nombre de classes Classification par ré-allocation dynamique (*k*-means,* partitionning around medoïds*), méthode mixte pour les grands tableaux.*

Travaux pratiques avec SAS et R pour la recherche de classes et leurs représentations.

Retour au plan du cours.

1 Introduction

1.1 Les données

Comme dans le cas du thème précédent (MDS), les données peuvent se présenter sous différentes formes ; elles concernent n individus supposés affectés, pour simplifier, du même poids :

- un tableau de distances (ou dissimilarités, ou mesures de dissemblance), $n \times n$, entre les individus pris deux à deux ;
- les observations de p variables quantitatives sur ces n individus ;
- les observations, toujours sur ces n individus, de variables qualitatives ou d'un mélange de variables quantitatives et qualitatives.

D'une façon ou d'une autre, il s'agit, dans chaque cas, de se ramener au tableau des distances deux à deux entre les individus (c'est-à-dire au premier cas). Le choix d'une matrice de produit scalaire permet de prendre en compte simplement un ensemble de variables quantitatives tandis que le troisième cas nécessite plus de développements.

1.2 Les objectifs

L'objectif d'une méthode de classification déborde le cadre strictement exploratoire. C'est la recherche d'une *typologie*, ou *segmentation*, c'est-à-dire d'une partition, ou répartition des individus en *classes* homogènes, ou catégories. Ceci est fait en optimisant un *critère* visant à regrouper les individus dans des classes, chacune le plus homogène possible et, entre elles, les plus distinctes possible. Cet objectif est à distinguer des procédures de discrimination, ou encore de classement (en anglais *classification*) pour lesquelles une typologie est *a priori* connue, au moins pour un échantillon d'apprentissage. Nous sommes dans une situation d'apprentissage *non-supervisé*, ou en anglais de *clustering* [1].

Il existe de très nombreuses méthodes de classification non supervisées, seule un sélection est décrite ci-dessous. Cette sélection est opérée en visant des méthodes fréquemment utilisées et appartenant à des types d'algorithmes différents donc complémentaires.

1.3 Les méthodes

Un calcul de combinatoire montre que le nombre de partitions possibles d'un ensemble de n éléments croît exponentiellement avec n ; le nombre de partitions de n éléments en k classes est le nombre de Stirling, le nombre total de partitions est celui de Bell. Pour $n = 20$ il est de l'ordre de 10^13. Il n'est donc pas question de chercher à optimiser le critère sur toutes les partitions possibles. Les méthodes se limitent à l'exécution d'un *algorithme itératif* convergeant vers une bonne partition et correspondant en général à un optimum local.

Plusieurs choix sont laissés à l'initiative de l'utilisateur :

- une mesure d'éloignement (dissemblance, dissimilarité ou distance) entre individus ;
- le critère d'homogénéité des classes à optimiser : il est, dans le cas de variables quantitatives, généralement défini à partir de la trace d'une matrice de variances-covariances ; soit les variances et covariances interclasses (la trace correspond alors à l'inertie de la partition), soit les variances et covariances intraclasse ;

1. Faire attention aux faux amis français / anglais : discrimination / *classification* (supervisée) et classification / *clustering* (non-supervisée)

- la méthode : classification ascendante hiérarchique, ré-allocation dynamique et DBSCAN sont les plus utilisées, seules ou combinées ;
- le nombre de classes : c'est un point délicat.

Enfin, différents outils recherchent une interprétation, ou des caractérisations, des classes obtenues.

Classification ascendante hiérarchique, ou CAH

Il s'agit de regrouper itérativement les individus, en commençant par le bas (les deux plus proches) et en construisant progressivement un arbre, ou *dendrogramme*, regroupant finalement tous les individus en une seule classe, à la racine (cf. figure 2 qui reprend les données élémentaires de la vignette sur le MDS). Ceci suppose de savoir calculer, à chaque étape ou regroupement, la distance entre un individu et un groupe ainsi que celle entre deux groupes. Ceci nécessite donc, pour l'utilisateur de cette méthode, de faire un choix supplémentaire : comment définir la distance entre deux groupes connaissant celles de tous les couples d'individus entre ces deux groupes. Différents choix, appelés *saut* en français et *linkage* en anglais, sont détaillés plus loin. Le nombre de classes est déterminé *a posteriori*, à la vue du dendrogramme ou d'un graphique représentant la décroissance de la hauteur de chaque saut, ou écart de distance, opéré à chaque regroupement.

Classification par ré-allocation dynamique

Dans ce cas, le nombre de classes, k, est fixé *a priori*. Ayant initialisé k centres de classes par tirage aléatoire (ou autre procédure), tous les individus sont affectés à la classe dont le centre est le plus proche au sens de la distance choisie (en principe, euclidienne pour cette méthode). Dans une deuxième étape, l'algorithme calcule des barycentres de ces classes qui deviennent les nouveaux centres. Le procédé (affectation de chaque individu à un centre, détermination des centres) est itéré jusqu'à convergence vers un minimum (local) ou un nombre d'itérations maximum fixé.

DBSCAN

Density-based spatial clustering of applications with noise (DBSCAN) est un algorithme plus récent (Ester et al. 1996)[2] basé sur une estimation local de la densité comme son acronyme le désigne. Basé sur deux paramètres (nombre minimum de points et rayon d'une boule, il regroupe itérativement les points par paquet sur la base de leur voisinage (nombre minimum d'individus) à l'intérieur d'une boule de rayon ϵ.

2 Mesures d'éloignement

Notons $\Omega = \{i = 1, \ldots, n\}$ l'ensemble des individus. Cette section se propose de définir sur $\Omega \times \Omega$ différentes mesures d'éloignement entre deux individus. Les hypothèses et propriétés étant de plus en plus fortes.

2.1 Indice de ressemblance, ou similarité

C'est une mesure de proximité définie de $\Omega \times \Omega$ dans $\mathbb{R}_+$ et vérifiant :

$$\begin{aligned} s(i,j) &= s(j,i), \forall (i,j) \in \Omega \times \Omega : \quad \text{symétrie} ; \\ s(i,i) &= S > 0, \forall i \in \Omega \ : \ \text{ressemblance d'un individu avec lui-même} ; \\ s(i,j) &\leq S, \forall (i,j) \in \Omega \times \Omega : \text{la ressemblance est majorée par } S. \end{aligned}$$

Un indice de ressemblance normé s^* est facilement défini à partir de s par :

$$s^*(i,j) = \frac{1}{S} s(i,j), \forall (i,j) \in \Omega \times \Omega ;$$

s^* est une application de $\Omega \times \Omega$ dans $[0,1]$.

2.2 Indice de dissemblance, ou dissimilarité

Une dissimilarité est une application d de $\Omega \times \Omega$ dans $\mathbb{R}_+$ vérifiant :

$$\begin{aligned} \forall (i,j) &\in \Omega \times \Omega \\ d(i,j) &= d(j,i), : \quad \text{symétrie} ; \\ d(i,j) = 0 &\Leftrightarrow i = j. \end{aligned}$$

Les notions de similarité et dissimilarité se correspondent de façon élémentaire. Si s est un indice de ressemblance, alors

$$d(i,j) = S - s(i,j), \forall (i,j) \in \Omega \times \Omega$$

est un indice de dissemblance. De façon réciproque, si d est un indice de dissemblance avec $D = \sup_{(i,j) \in \Omega \times \Omega} d(i,j)$, alors $s(i,j) = D - d(i,j)$ est

un indice de ressemblance. Comme s^*, un indice de dissemblance normé est défini par :

$$d^*(i,j) = \frac{1}{D} d(i,j), \forall (i,j) \in \Omega \times \Omega$$

avec $d^* = 1 - s^*$ et $s^* = 1 - d^*$. Du fait de cette correspondance immédiate, seule la notion de dissemblance, ou dissimilarité, normée est considérée par la suite.

2.3 Distance

Une distance sur Ω est, par définition, une dissimilarité vérifiant en plus la propriété d'*inégalité triangulaire*. Autrement dit, une distance d est une application de $\Omega \times \Omega$ dans $\mathbb{R}_+$ vérifiant :

$$\begin{aligned} d(i,j) &= d(j,i), \ \forall (i,j) \in \Omega \times \Omega \, ; \\ d(i,i) &= 0 \Longleftrightarrow i = j \, ; \\ d(i,j) &\leq d(i,k) + d(j,k), \ \forall (i,j,k) \in \Omega^3. \end{aligned}$$

Si Ω est fini, la distance peut être normée.

2.4 Distance euclidienne

Dans le cas où Ω est un espace vectoriel muni d'un produit scalaire, donc d'une norme, la distance définie à partir de cette norme est appelée distance euclidienne :

$$d(i,j) = < i-j, i-j >^{1/2} = \|i-j\|.$$

La condition pour qu'une matrice donnée de distances entre éléments d'un espace vectoriel soit issue d'une distance euclidienne est explicitée dans la vignette sur le positionnement multidimensionnel (MDS). Toute distance n'est pas nécessairement euclidienne ; voir, par exemple, celle construite sur la valeur absolue.

2.5 Utilisation pratique

Concrètement, il peut arriver que les données à traiter soient directement sous la forme d'une matrice d'un indice de ressemblance ou de dissemblance. Il est alors facile de la transformer en une matrice de dissemblances normées avant d'aborder une classification.

Nous précisons ci-dessous les autres cas.

Données quantitatives

Lorsque les p variables sont toutes quantitatives, il est nécessaire de définir une matrice $\mathbf{M}$ de produit scalaire sur l'espace $\mathbb{R}^P$. Le choix $\mathbf{M} = \mathbf{I}_p$, matrice identité, est un choix élémentaire et courant ; mais il est vivement conseillé de *réduire* les variables de variances hétérogènes, comme en ACP, ce qui revient à considérer, comme matrice de produit scalaire, la matrice diagonale composée des inverses des écarts-types :

$$\mathbf{M} = \Sigma^{-1} = \text{diag}\,(\frac{1}{\sigma_1} \cdots \frac{1}{\sigma_p}).$$

La métrique dite de Mahalanobis (inverse de la matrice des variances-covariances) peut aussi être utilisée pour atténuer la structure de corrélation.

Données qualitatives

Dans le cas très particulier où toutes les variables sont binaires (présence, absence de caractéristiques), de nombreux indices de ressemblances ont été proposés dans la littérature. Ils sont basés sur les quantités suivantes définis pour deux individus i et j distincts :

- a_{ij} = nombre de caractères communs à i et j sur les p considérés,
- b_{ij} = nombre de caractères possédés par i mais pas par j,
- c_{ij} = nombre de caractères possédés par j mais pas par i,
- d_{ij} = nombre de caractères que ne possèdent ni i ni j.
- bien sûr, $a_{ij} + b_{ij} + c_{ij} + d_{ij} = p$.

Les indices de ressemblance les plus courants sont :

- Concordance : $\frac{a_{ij}+d_{ij}}{p}$,
- Jaccard : $\frac{a_{ij}}{a_{ij}+b_{ij}+c_{ij}}$,
- Dice : $\frac{2a_{ij}}{2a_{ij}+b_{ij}+c_{ij}}$

Il est ensuite facile de construire un indice de dissemblance.

Dans le cas plus général de p variables qualitatives, la distance la plus utilisée est celle, euclidienne, dite du χ^2 entre profils-lignes du tableau disjonctif complet (cf. AFCM). La distance entre deux individus i et k est alors définie

par :

$$d^2_{\chi^2}(i,k) = \frac{n}{p}\sum_{j=1}^{p}\sum_{\ell=1}^{m_j}\delta_{ik}^{j\ell}\frac{1}{n_\ell^j}.$$

où m_j est le nombre de modalités de la variable qualitative Y^j, n_ℓ^j est l'effectif de la ℓ-ième modalité de Y^j et $\delta_{ik}^{j\ell}$ vaut 1 si les individus i et k présentent une discordance pour la ℓ-ième modalité de la variables Y^j et 0 sinon. L'importance donnée à une discordance est d'autant plus importante que les modalités considérées sont rares. Le coefficient n/p peut être omis.

Mélange quantitatif, qualitatif

Différentes stratégies sont envisageables dépendant de l'importance relative des nombres de variables qualitatives et quantitatives.

Rendre tout qualitatif . Les variables quantitatives sont rendues qualitatives par découpage en classes. Les classes d'une même variable sont généralement recherchées d'effectifs sensiblement égaux : bornes des classes égales à des quantiles. La métrique à utiliser est alors celle du χ^2 décrite ci-dessus.

Rendre tout quantitatif à l'aide d'une AFCM. Une AFCM est calculée sur les seules variables qualitatives ou sur l'ensemble des variables après découpage en classes des variables quantitatives. L'AFCM calculée par AFC du tableau disjonctif complet produit des *scores* (cf. chapitre 6) qui sont les composantes principales de l'ACP des profils-lignes. Dans le cas d'une AFCM partielle des seules variables qualitatives, les variables quantitatives restantes doivent être nécessairement réduites. Ces scores sont ensuite utilisés comme coordonnées quantitatives des individus en vue d'une classification.

Métrique de Gower permet de mixer les types de variables mais celle-ci reste très peu utilisée.

2.6 Bilan

Une fois ces préliminaires accomplis, nous nous retrouvons donc avec

- soit un tableau de mesures quantitatives $n \times p$, associé à une matrice de produit scalaire $p \times p$ (en général $\mathbf{I}_p$) définissant une métrique euclidienne,
- soit directement un tableau $n \times n$ de dissemblances ou de distances entre individus.

Attention, si n est grand, la deuxième solution peut se heurter rapidement à des problèmes de stockage en mémoire pour l'exécution des algorithmes.

2.7 Accord entre partitions

Une partition de n individus définit une variable qualitative dont les catégories sont les classes de la partition. Une comparaison de deux partitions est obtenue an construisant la table de contingence croisant ces deux variables. Cependant, les numéros des classes étant arbitraires, l'appréciation de cet accord est difficile aussi un indice quantitatif a été proposé en considérant toutes les paires d'individus, selon qu'ils appartiennent à la même classe dans les deux partitions, qu'ils sont dans la même classe pour l'une mais pas pour l'autre, et enfin qu'ils sont séparés dans les deux partitions.

En notant n_{kl} le terme général de la table de contingence croisant les deux partitions, l'indice dit de Rand dont une version s'écrit :

$$R = \frac{2\sum_k\sum_l n_{kl}^2 - \sum_k n_{k+}^2 - \sum_l n_{+l}^2 + n^2}{n^2}.$$

Cet indice prend ses valeurs entre 0 et 1, il est égal à 1 lorsque les deux partitions sont identiques. De nombreuses variantes de cet indice ont été proposées.

3 Classification ascendante hiérarchique

3.1 Principe

L'initialisation de cet algorithme consiste, s'il n'est déjà donné, à calculer un tableau de distances (ou de dissemblances) entre les individus à classer. L'algorithme démarre alors de la partition triviale des n singletons (chaque individu constitue une classe) et cherche, à chaque étape, à constituer des classes par agrégation des deux éléments les plus proches de la partition de l'étape précédente. L'algorithme s'arrête avec l'obtention d'une seule classe. Les regroupements successifs sont représentés sous la forme d'un arbre binaire ou *dendrogramme*.

3.2 Distance, ou dissemblance, entre deux classes

À chaque étape de l'algorithme, il est nécessaire de mettre à jour le tableau des distances (ou des dissemblances). Après chaque regroupement, de deux individus, de deux classes ou d'un individu à une classe, les distances entre ce nouvel objet et les autres sont calculées et viennent remplacer, dans la matrice, les distances des objets qui viennent d'être agrégés. Différentes approches sont possibles à ce niveau, donnant lieu à différentes CAH.

Notons A et B deux classes, ou éléments, d'une partition donnée, w_A et w_B leurs pondérations, et $d_{i,j}$ la distance entre deux individus quelconques i et j.

Le problème est de définir $d(A, B)$, distance entre deux éléments d'une partition de Ω.

Cas d'une dissemblance

Les stratégies ci-dessous s'accommodent d'un simple indice de dissemblance défini entre les individus. Elles s'appliquent également à des indices plus structurés (distance) mais n'en utilisent pas toutes les propriétés.

$$d(A,B) = \min_{i\in A, j\in B}(d_{ij}) \quad \text{(saut minimum, } \textit{single linkage}\text{)},$$

$$d(A,B) = \sup_{i\in A, j\in B}(d_{ij}) \quad \text{(saut maximum ou diamètre, } \textit{complete linkage}\text{)},$$

$$d(A,B) = \frac{1}{\text{card}(A)\text{card}(B)} \sum_{i\in A, j\in B} d_{ij} \quad \text{(saut moyen, } \textit{group average linkage}\text{)}.$$

Cas d'une distance euclidienne

Considérons que les données sont sous la forme d'une matrice $n \times p$ de variables quantitatives associée à une métrique euclidienne dans $\mathbb{R}^p$ ou directement sous la forme d'une matrice de distances euclidiennes $(n \times n)$ des individus 2 à 2. Dans le premier cas, il est facile de calculer les barycentres des classes et donc de considérer les distances suivantes entre deux groupes.

$$d(A,B) = d(g_A, g_B) \quad \text{(distance des barycentres, centroïd)},$$

$$d(A,B) = \frac{w_A w_B}{w_A + w_B} d(g_A, g_B) \quad \text{(saut de Ward)}.$$

Dans le 2ème cas, le carré de la distance entre 2 barycentres se calcule à partir de la matrice des distances entre les individus 2 à 2 :

$$d^2(g_A, g_B) = \frac{1}{\sum_{i\in A, j\in B} w_i w_j} \sum_{i\in A, j\in B} w_i w_j d_{ij}^2$$

Remarques :

- Le saut de Ward joue un rôle particulier et est la stratégie la plus courante ; c'est même souvent l'option par défaut dans le cas d'une distance euclidienne entre individus. En effet, ce critère induit, à chaque étape de regroupement, une minimisation de la décroissance de la variance interclasse.
- Même si la distance entre individus n'est pas euclidienne, la même expression est utilisée pour faire du "saut de Ward" dans le cas non-euclidien.
- Les implémentations du saut de Ward peuvent changer d'un logiciel à l'autre notamment sur le choix de la distance ou de son carré.

3.3 Algorithme

Algorithm 1 *classification ascendante hiérarchique*

Initialiser classes par les singletons
Calculer la matrice de leurs distances deux à deux
repeat
  Regrouper les deux classes les plus proches au sens de la distance entre classes choisie
  Mettre à jour le tableau de distances en remplaçant les deux classes regroupées par la nouvelle et en calculant sa distance avec chacune des autres classes.
until Agrégation en une seule classe

3.4 Résultats

Graphes

Les graphes obtenus à l'issue d'une CAH sont présentés et illustrés dans la section suivante. Il s'agit du graphique d'aide au choix du nombre de classes

et du dendrogramme, regroupant hiérarchiquement les observations et groupes par des branches dont la longueur est la distance entre les objets regroupés. Attention, la représentation du dendrogramme n'est pas unique, celui-ci est invariant par rotation d'une branche. L'ordre des observations sur l'axe horizontal est donc artificiel, il peut amener à rapprocher des observations qui sont de fait très éloignées l'une de l'autre car regroupées par de longues branches.

Qualité et choix du nombre de classes

La *corrélation cophénétique* est un indicateur de qualité d'une classification hiérarchique ou d'un dendrogramme est obtenue à partir de la notion de *distance cophénétique*. Cette distance est définie entre deux observations représentées dans un arbre par la hauteur des branches qui finissent par les réunir dans un même groupe. C'est également la distance entre les deux groupes contenant ces observations avant qu'ils ne soient réunis en un même groupe ; c'est par exemple la distance entre deux espèces dans un arbre phylogénétique. Toutes les distances ainsi définies entre les objets deux à deux sont rangées dans une matrice triangulaire de distances cophénétiques.

La qualité d'un arbre de classification peut se résumer par un coefficient de *corrélation cophénétique* entre les valeurs de la matrice de distances initiales, par exemple euclidiennes, et celle des distances cophénétiques. Évidemment, plus proche est cette valeur de 1, meilleure est la classification. *Attention* ce critère n'est néanmoins pas toujours pertinent pour opérer certains choix, notamment celui de la métrique entre les classes.

La *silhouette* (Rousseeuw, 1987)[8] d'une classification est un graphe montrant comment chaque observation appartient plus ou moins à sa classe. Supposons que n observations aient été réparties en k classes par un quelconque algorithme. Soit $a(i)$ la moyenne des dissimilarités (ou distances) de l'observation i avec toutes les autres observations au sein d'une même classe. Plus $a(i)$ est petit meilleur est l'assignation de i à sa classe ; $a(i)$ est la dissimilarité moyenne de i à cette classe.

Soit $b(i)$ la plus faible moyenne des dissimilarités (ou distances) de l'observation i à chaque autre classe dont i ne fait pas partie. La classe avec cette plus faible dissimilarité moyenne est appelé classe *voisine* de i car c'est la meilleure classe suivante pour l'observation i.

La silhouette de la $i$ème observations est alors donnée par

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

Plus ces valeurs sont proches de 1 et meilleure est la classification. La moyenne de toutes ces valeurs est un autre indicateur global de qualité.

Le choix du nombre de classes k est, comme le choix de la dimension en ACP, délicat à opérer. Plusieurs heuristiques on été proposées selon les critères précédents ou encore suivant le graphe de décroissance de la distance inter-classes qui est aussi la décroissance de la variance inter-classe dans le cas du saut de Ward. La recherche d'un "coude" dans ce graphe est une indication heuristique du choix de k ; voir l'application dans la section suivante.

La statistique du *gap* est une proposition de (Tibshirani et al.)[9] pour tenter de rationaliser cette démarche. Soit D_r la somme de toutes les distances prises entre les observations deux à deux au sein d'une même classe $r = 1, k$; W_k est la moyenne pondérée (par la taille de la classe) de ces sommes de distances. Si la distance initiale est euclidienne, W est (à un facteur 2 près) la norme carrée de la matrice de variance intra-classe. L'idée est alors de comparée le graphe de $\log(W_k)$ par rapport à celui d'une distribution de référence obtenue par simulation (Monte Carlo) selon une loi uniforme et de rechercher le plus grand écart ou *gap*. La fonction `clusGap` qui implémente ce critère dans la librairie `cluster` propose 5 méthodes ou critères ! pour rechercher le plus grand *gap*. Attention, cette fonction n'accepte que des données sous la forme d'une matrice de variables quantitatives, pas celle d'une matrice de distances ou dissimilarités.

Enfin, dans le contexte de mélanges supposés gaussiens, c'est-à-dire si l'hypothèse d'une situation gaussienne multidimensionnelle, le choix du nombre de classes s'apparente à une sélection de modèle par des critères AIC, BIC, spécifiques. Il n'est pas abordé dans ce cours ou aperçu des méthodes de classification non-supervisée.

3.5 Illustration

Les données sont celles déjà représentées à l'aide du MDS : un tableau contenant les distances kilométriques par route (Source : IGN) entre 47 grandes villes en France et dans les pays limitrophes. Toutes ces valeurs sont rangées dans le triangle inférieur d'une matrice carrée avec des 0 sur la diagonale. Il s'agit donc de regrouper au mieux ces villes, en tenant compte de leurs proximités relatives au sens de cette distance routière qui n'est pas euclidienne à cause du relief.

À l'issue de l'exécution, la classification ascendante hiérarchique fournit les deux graphiques précisés ci-dessous.

- Un graphique d'aide au choix du nombre de classes (cf. figure 1). Il représente à rebours, en fonction du nombre de classes, la décroissance de la distance interclasses. La présence d'une rupture importante dans cette décroissance aide au choix du nombre de classes comme dans le cas du choix de dimension en ACP, avec l'éboulis des valeurs propres. Dans ce cas, il faut lire le graphe de droite à gauche et s'arrêter avant le premier saut jugé significatif. Avec l'indice de Ward, cela revient à couper l'arbre avant une perte, jugée trop importante, de la variance interclasses. Dans le cas des villes repérées par leurs distances kilométriques, le choix de 5 classes semble raisonnable.
  La fonction `clusGap` ne permet pas de calculer la statistique de *gap* sur une matrice de distances. La corrélation cophénétique de l'arbre est de $0,64$ mais cela est guère utile dans l'absolu tandis que les silhouettes sont représentées dans la figure 1.
- Le *dendrogramme* (cf. figure 2) est une représentation graphique, sous forme d'arbre binaire, des agrégations successives jusqu'à la réunion en une seule classe de tous les individus. La hauteur d'une branche est proportionnelle à l'indice de dissemblance ou distance entre les deux objets regroupés. Dans le cas du saut de Ward, c'est la perte de variance interclasses.

Une fois un nombre de classes sélectionné par l'un ou l'autre des critères proposés, une coupure de l'arbre fournit, dans chaque sous-arbre, la répartition des individus en classes. Ces classes peuvent ensuite être représentées dans les axes d'une analyse factorielle :

- ACP si la classification a été opérée sur des variables quantitatives assorties d'une métrique euclidienne,
- AFCM si la classification a été opérée sur les composantes d'une AFCM de variables qualitatives,
- MDS dans le cas de l'exemple (figure 3) car la classification est directement calculée sur un tableau de distance.

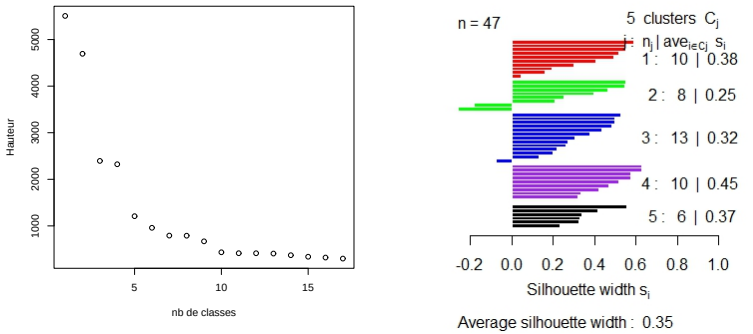

FIGURE 1 – *Villes : Décroissance de la variance interclasses à chaque regroupement dans le cas du saut de Ward (à gauche) et à droite silhouettes des observations dans leur classe respective.*

Signalons qu'il est courant, dans la pratique, de mettre en œuvre, à l'issue d'une CAH, une méthode de ré-allocation dynamique avec pour nombre de classes celui choisi par CAH et pour centres initiaux les barycentres des classes obtenues : on stabilise ainsi les classes.

Notons également que l'exemple présenté ici est relativement simple et bien structuré. Modifier le critère de saut ne change pas grand chose dans ce cas. Mais, attention, il est facile de vérifier expérimentalement qu'une classification ascendante est un objet très sensible. En effet, il suffit de modifier une distance dans le tableau, par exemple de réduire sensiblement la distance de Grenoble à Brest, pour que la classification (nombre de classes, organisation) devienne très sensible au choix du critère de saut. En revanche, la structure des données fait que la représentation factorielle de l'ACP du tableau de distance (MDS) est très robuste à ce type d'"erreur de mesure"; il est recommandé de systématiquement compléter une classification par une représentation facto-

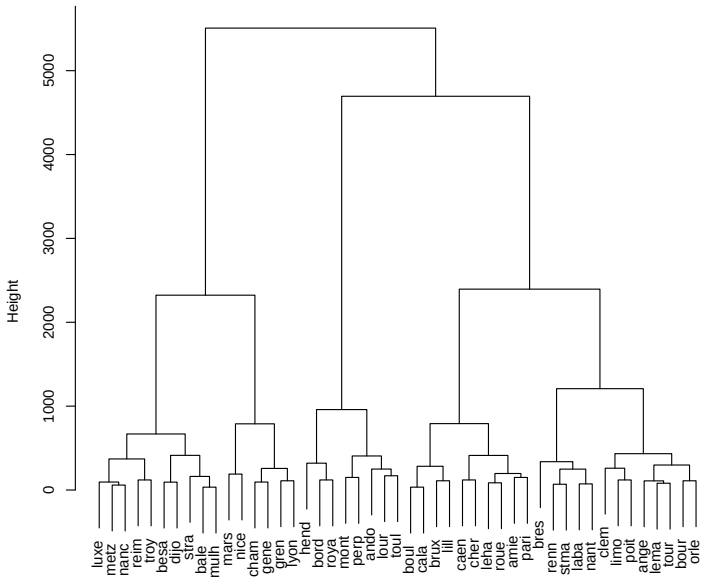

FIGURE 2 – *Villes : Exemple d'un dendrogramme issu de la classification des données par CAH et saut de Ward.*

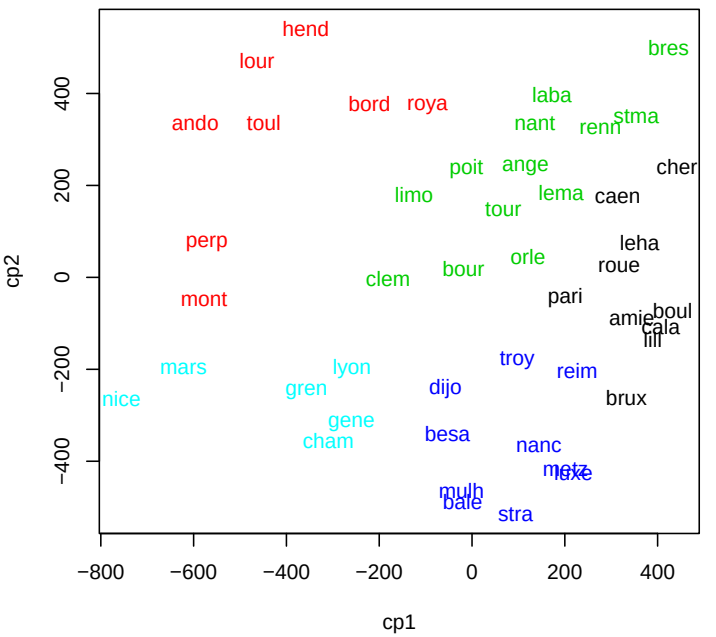

FIGURE 3 – *Villes : Représentation des classes (couleurs) obtenues par CAH dans les coordonnées du MDS.*

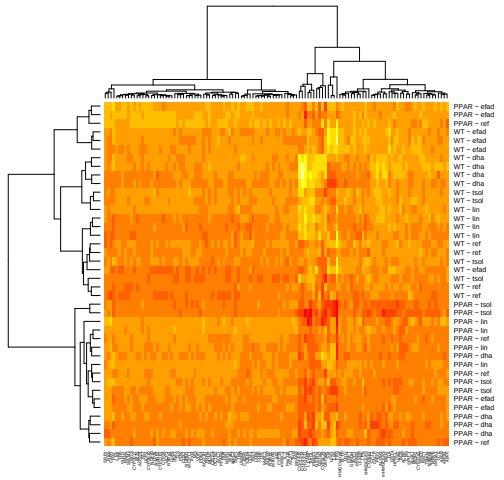

FIGURE 4 – *Souris : double classification ascendante hiérarchique des individus-souris et des variables-gènes selon la méthode de Ward, avec la distance euclidienne.*

rielle.

3.6 Double classification

Lors de l'analyse de données génomiques (transcriptome, protéome...), les biologistes apprécient de construire une double classification hiérarchique opérant à la fois sur les lignes et sur les colonnes (gènes et échantillons). La double classification hiérarchique induit une réorganisation des lignes et colonnes avant représentation en fausses couleurs. Cette représentation fournit une lecture susceptible de prendre en compte les distances respectives des lignes (gènes) d'une part et des colonnes (échantillons biologiques) d'autre part, et de se faire ainsi une idée des gènes pouvant influencer la hiérarchie obtenue pour les échantillons. Néanmoins, cette lecture, même en se limitant à une sélection des gènes, n'est pas très aisée (figure 4).

4 Agrégation autour de centres mobiles

4.1 Principes

Différents types d'algorithmes ont été définis autour du même principe de *ré-allocation dynamique* des individus à des centres de classes, eux-mêmes recalculés à chaque itération. Ces algorithmes requièrent une représentation vectorielle des individus dans $\mathbb{R}^p$ muni d'une métrique, généralement euclidienne. Il est important de noter que, contrairement à la méthode hiérarchique précédente, le nombre de classes k doit être déterminé *a priori*.

Ces méthodes sont itératives : après une initialisation des centres consistant, par exemple, à tirer aléatoirement k individus, l'algorithme répète deux opérations jusqu'à la convergence d'un critère :

1. Chaque individu est affecté à la *classe* dont le centre est le plus proche au sens d'une métrique.
2. Calcul des k *centres* des classes ainsi constituées.

4.2 Principale méthode

Il s'agit de la version proposé par Forgy (1965)[3] des algorithmes de type *k-means*.

Algorithm 2 Algorithme de Forgy

Initialisation Tirer au hasard, ou sélectionner pour des raisons extérieures à la méthode, k points dans l'espace des individus, en général k individus de l'ensemble, appelés centres ou noyaux.
repeat
Allouer chaque individu au centre (c'est-à-dire à la classe) le plus proche au sens de la métrique euclidienne choisie ; on obtient ainsi, à chaque étape, une classification en k classes, ou moins si, finalement, une des classes devient vide.
Calculer le centre de gravité de chaque classe : il devient le nouveau noyau ; si une classe s'est vidée, on peut éventuellement retirer aléatoirement un noyau complémentaire.
until Le critère de variance interclasses ne croisse plus de manière significative, c'est-à-dire jusqu'à la stabilisation des classes.

4.3 Propriétés

Convergence Le critère (la variance interclasses) est majoré par la variance totale. Il est simple de montrer qu'il ne peut que croître à chaque étape de l'algorithme, ce qui en assure la convergence. Il est équivalent de maximiser la variance interclasses ou de minimiser la variance intraclasse. Cette dernière est alors décroissante et minorée par 0. Concrètement, une dizaine d'itérations suffit généralement pour atteindre la convergence.

Optimum local La solution obtenue est un optimum local, c'est-à-dire que la répartition en classes dépend du choix initial des noyaux. Plusieurs exécutions de l'algorithme permettent de s'assurer de la présence de *formes fortes*, c'est-à-dire de classes, ou partie de classes, présentes de manière stable dans la majorité des partitions obtenues.

4.4 Variantes

k-means

Toujours sous la même appellation (une option de la commande `kmeans` de R) Mac Queen (1967)[7] a proposé une modification de l'algorithme précédent. Les noyaux des classes, ici les barycentres des classes concernées, sont recalculés à chaque allocation d'un individu à une classe. L'algorithme est ainsi plus efficace, mais la solution dépend de l'ordre des individus dans le fichier.

Nuées dynamiques

La variante proposée par Diday (1973)[1] et parallèlement par Hartigan et Wong (1979)[4] consiste à remplacer chaque centre de classe par un noyau constitué d'éléments représentatifs de cette classe. Cela permet de corriger l'influence d'éventuelles valeurs extrêmes sur le calcul du barycentre. Diday (1973) a également proposé la recherche de formes fortes communes à plusieurs partitions issues d'initialisations différentes.

Partitionning Around Medoïds

Cet algorithme (PAM), proposé par Kaufman & Rousseeuw (1990)[6], permet de classifier des données de façon plus robuste, c'est-à-dire moins sensible à des valeurs atypiques. Le noyau d'une classe est alors un médoïd c'est-à-dire l'observations d'une classe qui minimise la moyenne des distances ou dissimilarités aux autres observations de la classes. Une différence majeur avec l'algorithme *kmeans* est qu'un médoïd fait partie des données et permet donc de partitionner des matrices de dissimilarités. En contre-partie, il est limité par le nombre d'observations (matrice de dissimilarités à stocker) et en temps de calcul (algorithme en $\mathcal{O}(n^2)$). Il fonctionne de manière analogue à celui de Mac Queen. À chaque itération, un médoïd est mis en concurrence avec un autre individu aléatoire. Si l'échange améliore le critère, cet individu devient le nouveau médoïd.

D'autres algorithmes ont été proposés pour des types de données spécifiques : k-modes (Huang, 1998)[5] pour des variables qualitatives et k-prototypes pour des variables mixtes. Ils sont disponibles dans R.

La classification des villes par partitionnement autour de médoïds est fournie dans la figure 5 ; le nombre de classes est fixé *a priori* à 5 comme le suggère la CAH alors que les classes obtenues sont sensiblement différentes.

4.5 Combinaison

Chaque méthode précédente peut être plus ou moins adaptée à la situation rencontrée. La classification hiérarchique, qui construit nécessairement la matrice des distances, n'accepte qu'un nombre limité d'individus ; de son côté, la ré-allocation dynamique nécessite de fixer *a priori* le nombre de classes. La stratégie suivante, adaptée aux grands ensembles de données, permet de contourner ces difficultés.

1. Exécuter une méthode de ré-allocation dynamique en demandant un grand nombre de classes, de l'ordre de 10% de n.
2. Sur les barycentres des classes précédentes, exécuter une classification hiérarchique puis déterminer un nombre "optimal" k de classes.
3. Exécuter une méthode de ré-allocation dynamique sur tout l'ensemble en fixant à k le nombre de classes. Pour initialiser l'algorithme, il est habituel de choisir pour noyaux les barycentres (calculés en pondérant par les effectifs de classes) des classes de l'étape précédente.

5 *DBSCAN*

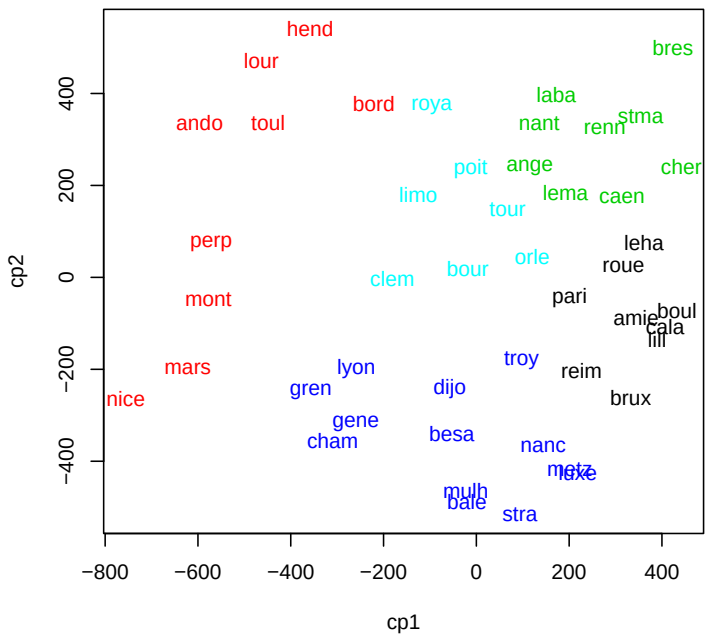

FIGURE 5 – *Villes : Représentation des classes (couleurs) obtenues par PAM dans les coordonnées du MDS.*

5.1 Principe

Le principe de DBSCAN (*Density-based spatial clustering of applications with noise*; Ester et al. 1996[2]) repose sur la notion de ε-voisinage d'un individu ou point défini comme l'ensemble des points appartenant à la boule de rayon ϵ centrée sur ce point. En plus du rayon ε, un autre paramètre est considéré : `MinPts` qui précise un nombre minimum de points à prendre en compte dans cette boule.

L'ensemble des points ou individus se répartit en trois catégories illustrées dans la figure 6 : les cœurs (*core points*), les points atteignables (*reachable*), les points atypiques (*ouliers*).

- Un point p est un *cœur* (*core*) si son ε-voisinage contient au moins `MinPts` points en l'incluant. Un point du ε-voisinage de p est dit *densément atteignable* depuis p. Par construction, aucun point ne peut être densément atteignable par un point qui n'est pas un cœur.
- Un point q est dit *atteignable* depuis p s'il existe un chemin $p = p_1, p_2, \ldots, p_n = q$ de sorte que chaque p_{i+1} est *densément atteignable* depuis p_i. Sous entendu, tous les points du chemin doivent être des cœurs à l'exception de q.
- Tous les points *non atteignables* de tout autre point, donc isolés, sont considérés comme atypiques (*outliers*) ou du bruit.

Ainsi, si p est un cœur, il forme une classe avec tous les points (cœur ou pas) qui sont atteignables à partir de lui. Chaque classe contient au moins un cœur. Des points qui ne sont pas des cœurs peuvent appartenir à une classe mais à sa frontière car ils ne peuvent servir à atteindre d'autres points.

L'*atteignabilité* n'est pas une relation symétrique car un point qui n'est pas un cœur peut être atteint mais aucun point ne peut être atteint à partir de lui.

5.2 Algorithme

DBSCAN nécessite de fixer les valeurs de deux paramètres : ε et `minPts` : le nombre minimum de points requis pour constituer une région dense.

L'algorithme démarre d'un point arbitrairement fixé qui n'a pas encore été visité. Si le ε-voisinage de ce point contient suffisamment de points, la construction d'une classe démarre. Sinon ce point est, au moins temporairement, étiqueté comme atypique mais peut être reconsidéré s'il apparaît par la

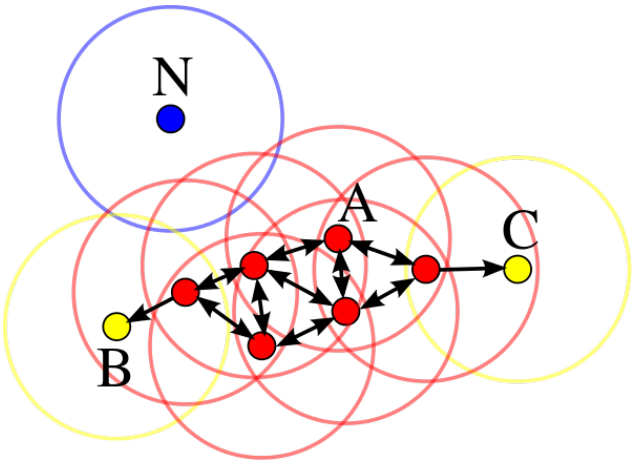

FIGURE 6 – *Exemple avec* `MinPts=4`. *A et tous les autres points rouges sont des cœurs ; chacun de leur ε-voisinage, eux compris, contient au moins* `MinPts` *points. Comme ils sont densément atteignables les uns les autres, ils forment une seule classe. B et C, ne sont pas des cœurs mais, atteignables de A via d'autres cœurs, ils appartiennent à la même classe. R n'est ni un cœur ni directement atteignable, c'est un atypique.*

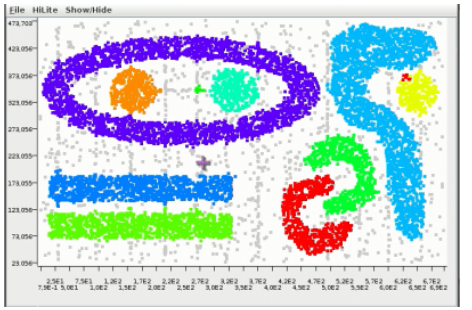

FIGURE 7 – *Exemple de classes discernables uniquement par un algorithme de type DBSCAN. Situation très rarement rencontrée en pratique sauf peut-être dans le cas d'une recherche de formes en analyse d'image.*

suite dans le ε-voisinage du cœur d'une autre classe.

Si un point est un cœur, tous les points de son ε-voisinage sont affectés à sa classe puis chacun considéré par l'algorithme. Si l'un de ces points est un cœur son ε-voisinage vient compléter la classe et le processus continue jusqu'à la complétion de la classe des points atteignables.

Puis, l'algorithme considère autre un point pas encore visité pour renouvelé le procédé jusqu'à ce que tous les points soient étiquetés.

5.3 Propriétés

DBSCAN présente quelques avantages :

- Il n'est pas nécessaire de définir *a priori* le nombre de classes, celui-ci est une conséquence du choix des paramètres ε et `minPts`.
- Il peut distinguer des classes avec des formes pathologiques ou même imbriquées entre elles. Il est même renommé et activement expérimenté sur des situations comme celui de la figure 7.
- Il est robuste aux observations atypiques et même propose de les détecter. Il ne classe pas les points isolés.

Et des inconvénients :

- Le choix des valeurs des paramètres ε et `minPts` doit être opéré de façon experte à partir de la compréhension des données et de leur en-

vironnement ; `minPts` apparaît comme une borne inférieure pour la taille des classes. Pour fixer ε il est conseillé (librairie `dbscan` de R) de tracer le graphe des distances des k plus proches voisins (`kNNdistplot`) avec `minPts` pour valeur de k. La recherche d'un genou dans ce graphe est une indication pour le choix de ε.

- DBSCAN est largement mis en défaut si des classes présentent des densités locales hétérogènes et, contrairement aux autres algorithmes, il ne met pas en évidence des classes si des groupes de points ne sont pas suffisamment distincts les uns des autres. Dans le cas des données contenant les distances kilométriques de cette vignette, il fournit une classe et une ou deux villes atypiques. C'est aussi le cas pour la plupart des exemples du tutoriel associé à cette vignette.

En résumé, DBSCAN est un algorithme de classification non supervisée très différent des autres vus et dont l'emploi est justifié lorsque des classes sont bien distinctes même si celles-ci ont des frontières ou des enveloppes avec des formes pathologiques. C'est-à-dire lorsque les autres algorithmes sont mis en défaut.

6 Conclusion

Une méthode de classification non supervisée produit une variable qualitative dont les modalités précisent la classe retenue pour chaque individu. Chaque méthode, algorithme, chaque choix de critère, dont le nombre de classes, conduit à des solutions différentes. Les critères proposés dits de qualité : corrélation cophénétique, silhouette... n'aident pas spécialement à un choix entre les solutions. Une représentation par ACP, AFD, AFCM ou MDS aide mieux à ce choix en faisant également intervenir, ensuite, des compétences métier pour cerner au mieux l'objectif.

Ces représentations aident également à l'interprétation des classes de même que l'ajustement d'un modèle de type arbre binaire de discrimination.

Références

[1] E. Diday, *The dynamic clusters method in nonhierarchical clustering*, International Journal of Computer & Information Sciences **2** (1973), n° 1, 61–88.

[2] M. Ester, H. P. Kriegel, J. Sander et X. Xu, *A density-based algorithm for discovering clusters in large spatial databases with noise*, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD-96, 1996, p. 226–231.

[3] R. Forgy, *Cluster Analysis of Multivariate Data : Efficiency versus Interpretability of Classification*, Biometrics (1965), n° 21, 768–769.

[4] J. A. Hartigan et M. A. Wong, *Algorithm AS 136 : a k-means clustering algorithm*, Applied Statistics **28** (1979), 100–108.

[5] Zhexue Huang, *Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values*, Data Min. Knowl. Discov. **2** (1998), n° 3, 283–304.

[6] Leonard Kaufman et Peter J. Rousseeuw, *Finding Groups in Data – An Introduction to Cluster Analysis*, John Wiley & Sons, 1990.

[7] J. Macqueen, *Some methods for classification and analysis of multivariate observations*, In 5-th Berkeley Symposium on Mathematical Statistics and Probability, 1967, p. 281–297.

[8] Peter J. Rousseeuw, *Silhouettes : A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics **20** (1987), n° 0, 53 – 65.

[9] Robert Tibshirani, Guenther Walther et Trevor Hastie, *Estimating the number of clusters in a data set via the gap statistic*, Journal of the Royal Statistical Society : Series B (Statistical Methodology) **63** (2001), n° 2, 411–423.