1 Objectif

Classification automatique de variables (de modalités de variables) catégorielles.

La classification de variables vise à les regrouper en paquets homogènes. Les variables fortement liées entre elles sont classées dans un même groupe ; celles qui sont faiblement liées, voire orthogonales, sont classées dans des groupes différents.

Nous avions présenté dans un précédent tutoriel les techniques de classification autour de variables latentes issues d'une analyse en composantes principales (ACP), dédiées aux descripteurs quantitatifs (VARHAC, VARKMEANS, VARCLUS)[1]. Elles sont très séduisantes parce que les informations exploitées par l'algorithme – les corrélations entre les variables – sont parfaitement comprises par les praticiens du data mining.

A priori, le passage aux variables qualitatives ne devrait pas poser problème. De fait, des techniques de classification basées sur le regroupement autour de variables latentes issues d'une analyse des correspondances multiples (ACM) permettent de généraliser les approches ci-dessus. Il reste quand même une petite insatisfaction je trouve s'agissant du traitement spécifique des variables catégorielles. En effet, savoir que deux variables qualitatives sont liées n'indique pas à quel titre elles le sont. A l'instar de l'étude du KHI-2 d'indépendance qui doit être complétée par l'analyse des contributions au KHI-2, plutôt que le regroupement en classes des variables, il paraît plus judicieux de réaliser la typologie des modalités des variables qualitatives pour analyser la structure des relations entre les variables.

Dans ce tutoriel, nous décrivons dans un premier temps une technique de classification ascendante hiérarchique (CAH) des modalités de variables qualitatives basée sur l'indice de Dice (Abdallah et Saporta, 1998)[2]. Nous montrons sa mise en œuvre à l'aide du composant **CATVARHCA** introduit dans la version **1.4.50** de **Tanagra**. Ce dernier intègre la possibilité d'utiliser des variables illustratives, permettant ainsi d'enrichir l'interprétation des résultats. Dans un second temps, pour bien cerner les tenants et aboutissants de l'approche, nous détaillons les différentes étapes des calculs sous le **logiciel R 3.0.1**.

[1] Tutoriel Tanagra, « Classification de variables », http://tutoriels-data-mining.blogspot.fr/2008/03/classification-de-variables.html, Mars 2008.

[2] H. Abdallah, G. Saporta, « Classification d'un ensemble de variables qualitatives », in Revue de Statistique Appliquée, Tome 46, N°4, pp. 5-26, 1998.

Enfin, dans une troisième et dernière partie, nous présenterons des approches alternatives de catégorisation des modalités des variables qualitatives. L'une, implémentée dans le package 'Hmisc', s'appuie sur une autre mesure de dissimilarité. L'autre, de type « tandem analysis », procède par une CAH sur les coordonnées factorielles des modalités issues d'une ACM. Nous comparerons les résultats obtenus avec ceux de CATVARHCA. Nous constaterons que ces approches sont tout à fait valables et proposent des résultats exploitables.

2 Données

Nous utilisons une fraction de la base VOTE AU CONGRES issue du serveur UCI. Elle décrit les votes effectués par **(n = 435)** députés sur différents thèmes soumis au parlement US en 1984. Nous avons retenu **(p = 6)** variables actives : budget (adoption of the budget resolution), physician (physician fee freeze), salvador (el salvador aid), nicaraguan (aid to nicaraguan contras), missile (mx missile), education (education spending).

Pour chaque thème, nous avons 3 réponses possibles, recouvrant en réalité 9 situations différentes : « **yes** » correspond à « voted for », « paired for » et « announced for » ; « **no** » reflète « voted against », « paired against » et « announced against » ; enfin, « **neither** », codée « ? » dans la base originelle, recouvre « voted present », « vote present to avoid conflict of interest », et « did not vote or otherwise make a position known ». La modalité « neither » ne correspond pas à une donnée manquante comme on pourrait le croire à la lecture de la page de présentation de la base sur le serveur UCI[3] (Missing Values ? = Yes)[4], mais plutôt à un comportement particulier que je résumerais par « le député ne veut pas se mouiller ». Cette nuance est très importante pour la bonne compréhension des résultats fournis par la classification des modalités que nous présenterons dans ce document.

Enfin, chaque député est affilié à un groupe politique (démocrate ou républicain). Nous utilisons AFFILIATION comme variable illustrative (variable supplémentaire) pour caractériser les classes de modalités issues de la typologie.

Voici les 5 premières lignes du fichier « **vote_catvarclus.xls** ».

[3] http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records

[4] L'information est précisée d'ailleurs lors de la description détaillée de la base http://archive.ics.uci.edu/ml/machine-learning-databases/voting-records/house-votes-84.names : « *It is important to recognize that "?" in this database does not mean that the value of the attribute is unknown. It means simply, that the value is not "yea" or "nay"* ».

affiliation	budget	physician	salvador	nicaraguan	missile	education
republican	n	y	y	n	n	y
republican	n	y	y	n	n	y
democrat	y	neither	y	n	n	n
democrat	y	n	neither	n	n	n
democrat	y	n	y	n	n	neither

Ci-dessous la distribution de fréquences de chaque variable.

```
    affiliation      budget      physician      salvador      nicaraguan      missile      education
democrat  :267   n       :171   n       :247   n       :208   n       :178   n       :206   n       :233
republican:168   neither: 11   neither: 11   neither: 15   neither: 15   neither: 22   neither: 31
                 y       :253   y       :177   y       :212   y       :242   y       :207   y       :171
```

3 Classification de variables qualitatives

Le package « ClustOfVar » pour R propose plusieurs techniques de classification de variables (Chavent et al., 2011)[5]. Elles reposent sur la même idée fondatrice : durant le processus de classification, un groupe de variables est représentée par une composante latente issue du premier facteur d'une analyse factorielle. Lorsque les variables sont toutes qualitatives, nous utilisons l'analyse des correspondances multiples (ACM). Lorsqu'elles sont composées d'un mélange de variables quantitatives et qualitatives, on utilise l'analyse factorielle des données mixtes (AFDM)[6] qui se révèle être à la fois une généralisation de l'ACM et de l'ACP[7]. Ces méthodes constituent une extension de l'approche de Vigneau et Qanari (2003)[8] définie pour la classification des variables quantitatives, où la composante représentative d'un groupe est fournie par une analyse en composantes principales (ACP).

La procédure **hclustvar()** réalise une classification ascendante hiérarchique des variables actives. Voici le code source du programme sous R.

```
#chargement des données
library(xlsx)
vote.data <- read.xlsx(file="vote_catvarclus.xls",header=T,sheetIndex=1)
summary(vote.data)
#spécification des variables actives (votes réalisés)
```

[5] M. Chavent, V. Kuentz, B. Liquet, J. Saracco, « Classification de variables : la package ClustOfVar », in Actes des 43èmes journées de Statistique (SFDS), Tunisie, 2011.

[6] J. Pagès, « Analyse Factorielle de Données Mixtes », in Revue de Statistique Appliquée, 52(4), pp.93-111, 2004.

[7] Voir le support de cours : R. Rakotomalala, « Analyse Factorielle de Données Mixtes – Diapos », Août 2013.

[8] E. Vigneau, E. Qannari, « Clustering of Variables around Latent Components », in Communications in Statistics Simulation and Computation, 32(4), pp. 1131-1150, 2003. Nous décrivons l'approche dans un de nos supports de cours : R. Rakotomalala, « Classification de Variables : Classification autour de variables latentes ».

```
vote.active <- subset(vote.data,select=2:7)
#clustering des variables avec la procédure hclustvar
#du package ClustOfVar (Chavent et al., 2011)[9]
library(ClustOfVar)
arbre <- hclustvar(X.quali=vote.active)
plot(arbre)
```

Nous obtenons un dendrogramme. Il met en évidence les proximités entre les variables (Figure 1). Si l'on en juge au « saut » dans les distances d'agrégation, il semblerait qu'un partitionnement en 2 groupes des variables originelles soit pertinent avec (missile, salvador et nicaraguan) d'une part, (education, budget, et physician) d'autre part.

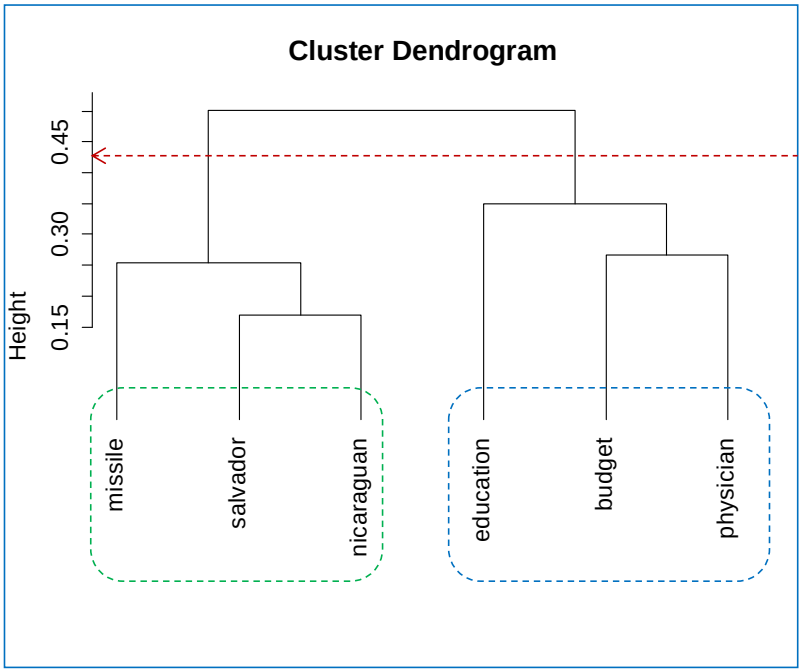


Figure 1 - Dendrogramme fournie par hclustvar() - Traitement du fichier "Vote"

Ces résultats sont intéressants en soi. Mais lorsque nous essayons d'établir une interprétation, nous constatons qu'ils ne donnent qu'une vision parcellaire des relations entre les variables.

Association entre salvador et nicaraguan. On note que salvador (el salvador aid) et nicaraguan (aid to nicaraguan contras) sont très proches. La hauteur d'agrégation est la plus basse. Mais nous ne savons pas dans quel sens lire la relation. Est-ce que les députés qui votent « yes » à la première question (salvador) font de même pour la seconde (nicaraguan) ? Ou inversement, y a-t-il une conjonction des votes « yes » et « no » ? Comment se positionnent ceux qui votent « neither » ? La lecture du dendrogramme ne permet pas de

[9] Package 'ClustOfVar', December 2013 ; http://cran.r-project.org/web/packages/ClustOfVar/ClustOfVar.pdf.

répondre à ces interrogations. Nous avons formé le tableau de contingence entre ces variables pour compléter l'analyse.

Row (Y)	Column (X)
salvador	nicaraguan

Statistical indicator	
Stat	Value
d.f.	4
Tschuprow's t	0.611119
Cramer's v	0.611119
Phi²	0.746933
Chi² (p-value)	324.92 (0.0000)
Lambda	0.775785
Tau (p-value)	0.6297 (0.0000)
U(R/C) (p-value)	0.5305 (0.0000)

Cross-tab

	n	y	neither	Sum
y	172 (+ 26 %)	31 (- 20 %)	9 (+ 0 %)	212
neither	4 (- 0 %)	7 (- 0 %)	4 (+ 7 %)	15
n	2 (- 25 %)	204 (+ 21 %)	2 (- 1 %)	208
Sum	178	242	15	435 *100%*

La situation se précise : ceux qui votent « yes » à « salvador aid » choisissent « no » à « aid to nicaraguan contras », et inversement. Ces deux thèmes différencient véritablement les députés. Les deux variables sont effectivement très liées (V de Cramer = 0.61), mais il était impossible de détecter la nature de la relation sans effectuer cette analyse complémentaire. Nous noterons également que le vote « neither » sur un des thèmes ne correspond à aucun choix particulier sur le second thème, et réciproquement.

Interprétation des clusters. Il semble que 2 groupes se détachent à l'issue de l'analyse. Nous avons calculé le V de Cramer entre les variables.

Cramer's V	missile	salvador	nicaraguan	education	budget	physician
missile	1	0.56	0.55	0.43	0.44	0.47
salvador		1	0.61	0.47	0.51	0.58
nicaraguan			1	0.47	0.52	0.52
education				1	0.48	0.51
budget					1	0.64
physician						1

Les clusters ne sont pas aussi marqués que l'on pourrait le penser finalement. Et surtout, leur interprétation n'est pas plus évidente pour autant. Il est impossible par exemple de situer le pourquoi de la formation de la classe (missile, salvador, nicaraguan) en l'absence d'informations sur les valeurs prises conjointement par les modalités.

Positionnement de la variable supplémentaire AFFILIATION. La variable AFFILIATION (démocrate ou républicain) nous permet d'apprécier la coloration politique des votes. Nous l'introduisons dans notre analyse en calculant le V de Cramer avec les variables de chaque cluster.

Cramer's V	affiliation (republican or democrat)
missile	0.63
salvador	0.71
nicaraguan	0.66
education	0.69
budget	0.74
physician	0.91

La variable AFFILIATION indiquant l'appartenance politique est en moyenne quasiment autant liée aux variables du premier que du second cluster. Identifier les votes qui différencient les démocrates des républicains n'est pas possible.

Nécessité du traitement des modalités pour l'interprétation. La classification des variables qualitatives est intéressante pour détecter les redondances. Il n'y a pas à revenir dessus. Cependant, lorsque nous cherchons à interpréter dans le détail les résultats et comprendre la nature des regroupements, contrairement aux variables quantitatives où la corrélation permet de traduire l'essence des relations, il est nécessaire de descendre d'un cran et s'intéresser au positionnement relatif des modalités (attraction, répulsion).

4 Classification des modalités des variables qualitatives

Dans cette section, nous présentons très brièvement les principaux éléments de la méthode de classification ascendante hiérarchique des modalités. La CAH est bien connue et maîtrisée[10,11]. Deux aspects surtout retiennent notre attention concernant sa transposition à la classification des modalités de variables qualitatives :

1. **La mesure de dissimilarité entre modalités**. Elle est essentielle car définit la proximité entre les objets. Elle pèse très fortement sur les résultats obtenus. Pour le traitement des modalités des variables qualitatives, nous nous sommes basés sur l'article d'Abdallah et Saporta (1998). Elle décrit les différentes mesures possibles et s'attache à détailler leurs caractéristiques. C'est la lecture de ce document qui m'a persuadé d'une part de l'intérêt de la classification des modalités et, d'autre part, permis de la programmer relativement facilement dans Tanagra avec une mesure présentant de bonnes propriétés.

[10] http://fr.wikipedia.org/wiki/Regroupement_hi%C3%A9rarchique

[11] http://en.wikipedia.org/wiki/Hierarchical_clustering

2. **La stratégie d'agrégation** que l'on peut également considérer sous l'angle de la mesure de **dissimilarité entre clusters** (saut mininum [single linkage], saut maximum [complete linkage], lien moyen [average linkage], méthode de Ward). Ce point ne pose pas de problèmes particuliers pour ce qui est de la transposition de la CAH à la classification des modalités. Nous le reconsidérerons plus en détail lorsqu'il s'agira d'aborder le traitement des variables supplémentaires.

Ainsi, la véritable particularité de la classification des modalités des variables qualitatives réside dans le choix de la mesure de dissimilarité.

4.1 Mesure de dissimilarité entre modalités – Indice de Dice

L'indice de Dice s'appuie sur les cooccurrences des modalités des variables qualitatives chez les individus de l'échantillon de données. Nous reprenons l'exemple de la base « Races Canines » de notre article de référence (Abdallah et Saporta, 1998). Nous disposons d'un échantillon de n = 27 canidés, décrits par plusieurs variables. Intéressons nous au « poids » (poids-, poids+, poids++) et à la « taille » (taille-, taille+, taille++). Nous transformons la base en tableau d'indicatrices pour calculer les dissimilarités entre les modalités.

	Variables initiales		Indicatrices des modalités					
Chien	Taille	Poids	Taille = Taille-	Taille = Taille+	Taille = Taille++	Poids = Poids-	Poids = Poids+	Poids = Poids++
Beauceron	Taille++	Poids+	0	0	1	0	1	0
Basset	Taille-	Poids-	1	0	0	1	0	0
Berger All	Taille++	Poids+	0	0	1	0	1	0
Boxer	Taille+	Poids+	0	1	0	0	1	0
Bull-Dog	Taille-	Poids-	1	0	0	1	0	0
Bull-Mastif	Taille++	Poids++	0	0	1	0	0	1
Caniche	Taille-	Poids-	1	0	0	1	0	0
Chihuahua	Taille-	Poids-	1	0	0	1	0	0
Cocker	Taille+	Poids-	**0**	1	0	**1**	0	0
Colley	Taille++	Poids+	0	0	1	0	1	0
Dalmatien	Taille+	Poids+	0	1	0	0	1	0
Doberman	Taille++	Poids+	0	0	1	0	1	0
Dogue All	Taille++	Poids++	0	0	1	0	0	1
Epag. Breton	Taille+	Poids+	0	1	0	0	1	0
Epag. Français	Taille++	Poids+	0	0	1	0	1	0
Fox-Hound	Taille++	Poids+	0	0	1	0	1	0
Fox-Terrier	Taille-	Poids-	1	0	0	1	0	0
Gd Bleu Gasc	Taille++	Poids+	0	0	1	0	1	0
Labrador	Taille+	Poids+	0	1	0	0	1	0
Levrier	Taille++	Poids+	0	0	1	0	1	0
Mastiff	Taille++	Poids++	0	0	1	0	0	1
Pekinois	Taille-	Poids-	1	0	0	1	0	0
Pointer	Taille++	Poids+	0	0	1	0	1	0
St-Bernard	Taille++	Poids++	0	0	1	0	0	1
Setter	Taille++	Poids+	0	0	1	0	1	0
Teckel	Taille-	Poids-	1	0	0	1	0	0
Terre-Neuve	Taille++	Poids++	0	0	1	0	0	1

Figure 2 - Exemple des canidés - Variables Taille et Poids - Indicatrices associées

Soit k_{ij} la valeur de l'indicatrice de la modalité j pour l'individu i. L'indice de Dice matérialisant la dissimilarité entre les modalités « j » et « j' » s'écrit :

$$\delta^2_{jj'} = \frac{1}{2}\sum_{i=1}^{n}(k_{ij} - k_{ij'})^2 = \frac{1}{2}\sum_{i=1}^{n}(k_{ij}\bar{k}_{ij'} + \bar{k}_{ij}k_{ij'})$$

Où $\bar{k}_{ij} = 1 - k_{ij}$

L'indice de Dice correspond au carré d'une distance euclidienne entre les indicatrices de modalités des variables qualitatives. Elle est parfaitement adaptée à la CAH. Calculons l'indice pour les modalités « Poids- » et «Taille- » sur notre échantillon de n = 27 observations :

$$\delta^2_{1,4} = \frac{1}{2}\{[0 \times (1-0) + (1-0) \times 0] + [1 \times (1-1) + (1-1) \times 1] + \cdots + [0 \times (1-0) + (1-0) \times 0]\} = 0.5$$

Ces deux modalités semblent très proches. Elles concordent systématiquement, excepté pour l'observation « Cocker » (Figure 2).

Nous pouvons ainsi calculer la matrice des distances entre les modalités prises deux à deux. Elle sert de point de départ à la CAH.

Distance	Taille = Taille-	Taille = Taille+	Taille = Taille++	Poids = Poids-	Poids = Poids+	Poids = Poids++
Taille = Taille-	0	6	11	0.5	10.5	6
Taille = Taille+		0	10	5.5	5.5	5
Taille = Taille++			0	11.5	4.5	5
Poids = Poids-				0	11	6.5
Poids = Poids+					0	9.5
Poids = Poids++						0

Par la suite, durant le processus de regroupement hiérarchique, cette matrice sera mise à jour selon la stratégie d'agrégation choisie (saut minimum, saut maximum, lien moyen, Ward) jusqu'à l'obtention d'une seule et unique classe qui est le point final de l'algorithme.

4.2 Traitement des variables supplémentaires qualitatives

L'usage des **objets supplémentaires** ou **illustratifs** – par opposition aux **objets actifs** qui ont servi à construire la typologie – est courant en analyse de données.

Soit il s'agit d'un élément supplémentaire que l'on cherche à positionner par rapport à ceux qui ont servi aux calculs, dans le cadre du déploiement de la solution ; soit il s'agit d'un élément illustratif qui permet de proposer une lecture différenciée des résultats, de préciser leur interprétation. Dans la CAH, il s'agit de déterminer la classe qui est la plus proche de l'objet additionnel en respectant leur mécanisme de formation, à savoir la distance utilisée et la stratégie d'agrégation.

Nous transformons la variable supplémentaire en indicatrices. Puis nous calculons leurs distances aux clusters à l'aide de l'indice de Dice en se conformant à la stratégie d'agrégation utilisée durant la construction de la partition c.-à-d. si la CAH a été réalisée à l'aide de la

méthode du saut minimum, le calcul de la distance au cluster de la modalité supplémentaire doit également s'appuyer sur le même principe, etc.

Chien	Taille	Poids	Fonction
Beauceron	Taille++	Poids+	utilite
Basset	Taille-	Poids-	chasse
Berger All	Taille++	Poids+	utilite
Boxer	Taille+	Poids+	compagnie
Bull-Dog	Taille-	Poids-	compagnie
Bull-Mastif	Taille++	Poids++	utilite
Caniche	Taille-	Poids-	compagnie
Chihuahua	Taille-	Poids-	compagnie
Cocker	Taille+	Poids-	compagnie
Colley	Taille++	Poids+	compagnie
Dalmatien	Taille+	Poids+	compagnie
Doberman	Taille++	Poids+	utilite
Dogue All	Taille++	Poids++	utilite
Epag. Breton	Taille+	Poids+	chasse
Epag. Français	Taille++	Poids+	chasse
Fox-Hound	Taille++	Poids+	chasse
Fox-Terrier	Taille-	Poids-	compagnie
Gd Bleu Gasc	Taille++	Poids+	chasse
Labrador	Taille+	Poids+	chasse
Levrier	Taille++	Poids+	chasse
Mastiff	Taille++	Poids++	utilite
Pekinois	Taille-	Poids-	compagnie
Pointer	Taille++	Poids+	chasse
St-Bernard	Taille++	Poids++	utilite
Setter	Taille++	Poids+	chasse
Teckel	Taille-	Poids-	compagnie
Terre-Neuve	Taille++	Poids++	utilite

Figure 3 - Exemple des canidés - Variables actives (taille, poids) et illustrative (fonction)

Mettons que nous obtenons une partition en 3 clusters : C1 = (Poids-, Taille), C2 = (Poids+, Taille++) et C3 = (Poids++ et Taille+). La typologie s'appuie sur les particularités physiques des canidés. Par ailleurs, ces derniers peuvent également se distinguer selon leur fonction (variable à 3 modalités : chasse, utilité, compagnie). Une idée intéressante serait d'associer les fonctions des chiens à leurs caractéristiques physiques, si cela est possible (Figure 3).

Nous souhaitons positionner le « chien de compagnie » par rapport aux différentes classes en calculant sa distance avec les indicatrices des variables actives. Nous résumons les résultats dans le tableau suivant (Figure 4) :

			Distance au cluster		
		Dist(indicatrice)	Single linkage	Complete linkage	Average Linkage
Cluster 1	Poids-	2	**2**	**2.5**	**2.25**
	Taille-	2.5			
Cluster 2	Poids+	9	9	11.5	10.25
	Taille++	11.5			
Cluster 3	Poids++	7.5	4.5	7.5	6
	Taille+	4.5			

Figure 4 - Distance aux clusters de la modalité supplémentaire "Fonction = compagnie"

Plusieurs commentaires viennent à l'esprit :

1. Le calcul de la distance d'une modalité supplémentaire avec les indicatrices des variables actives se fait tout à fait naturellement. Illustrons cela à l'aide de la modalité « Taille- », nous utilisons le tableau de calcul suivant :

Taille = Taille-	Fonction = compagnie
0	0
1	0
0	0
0	1
1	1
0	0
1	1
1	1
0	1
0	1
0	1
0	0
0	0
0	0
0	0
0	0
1	1
0	0
0	0
0	0
0	0
1	1
0	0
0	0
0	0
1	1
0	0

Indice de Dice	**2.5**

L'indice de Dice est obtenu avec

$$\delta^2 = \tfrac{1}{2} \times \{ (0 \times 1 + 1 \times 0) + (1 \times 1 + 0 \times 0) + \ldots (1 \times 0 + 0 \times 1) + (0 \times 1 + 1 \times 0) \} = 2.5$$

2. Il est tout à fait possible d'en dériver une distance aux clusters issus de la typologie.
3. Selon la stratégie d'agrégation, nous pouvons obtenir des valeurs différentes, et possiblement des conclusions différentes.

Prenons l'exemple du lien moyen (average linkage) pour détailler la démarche (Figure 4). La distance de la modalité supplémentaire au premier cluster (C1) est égale à (2 + 2.5)/2 = 2.25, par rapport au second (C2) (9 + 11.5)/2 = 10.25, par rapport au troisième (C3) (7.5 + 4.5)/2 = 6. Il paraît logique d'associer la modalité supplémentaire à la première classe réunissant les modalités (Poids-, Taille) : les chiens de compagnie sont plutôt petits et de faible poids.

Pour ce cas particulier, le type d'agrégation ne modifie pas l'affectation aux clusters (« compagnie » est toujours associée au cluster 1). Il peut en être autrement de manière générale. **Le plus important est d'être parfaitement cohérent avec la démarche de**

classification : la stratégie d'agrégation utilisée durant l'affectation de la modalité supplémentaire doit être identique à celle utilisée durant la construction des classes.

4.3 Quelques remarques

Traitement des variables supplémentaires quantitatives. Est-il possible de traiter des variables supplémentaires quantitatives ? A priori oui puisqu'il est tout à fait possible de mesurer le degré de liaison entre une variable quantitative et une indicatrice[12]. J'avoue avoir quelques réticences pour l'instant. Il faudrait analyser plus en détail le comportement de l'indicateur que l'on utilisera, surtout lorsqu'il s'agira de reproduire la stratégie d'agrégation (saut minimum, saut maximum, etc.).

Méthodes des centroïdes ou Méthode de Ward. La méthode de Ward est disponible dans Tanagra. Les distances entre clusters candidats sont calculées de proche en proche à l'aide de la formule de Lance et Williams (1967) durant la phase d'apprentissage. La moyenne des indicatrices représente chaque cluster lors du déploiement. Cette notion de barycentre d'un ensemble de modalités ne pose aucun problème géométriquement. En revanche j'ai un peu du mal à concevoir ce qu'elle représente réellement (intellectuellement tout du moins). De fait, les techniques basées sur les regroupements autour des variables latentes me semblent plus intéressantes de ce point de vue. Le 1er facteur issu de l'analyse factorielle fait office justement de variable « moyenne » représentative du cluster.

5 Classification des modalités avec Tanagra

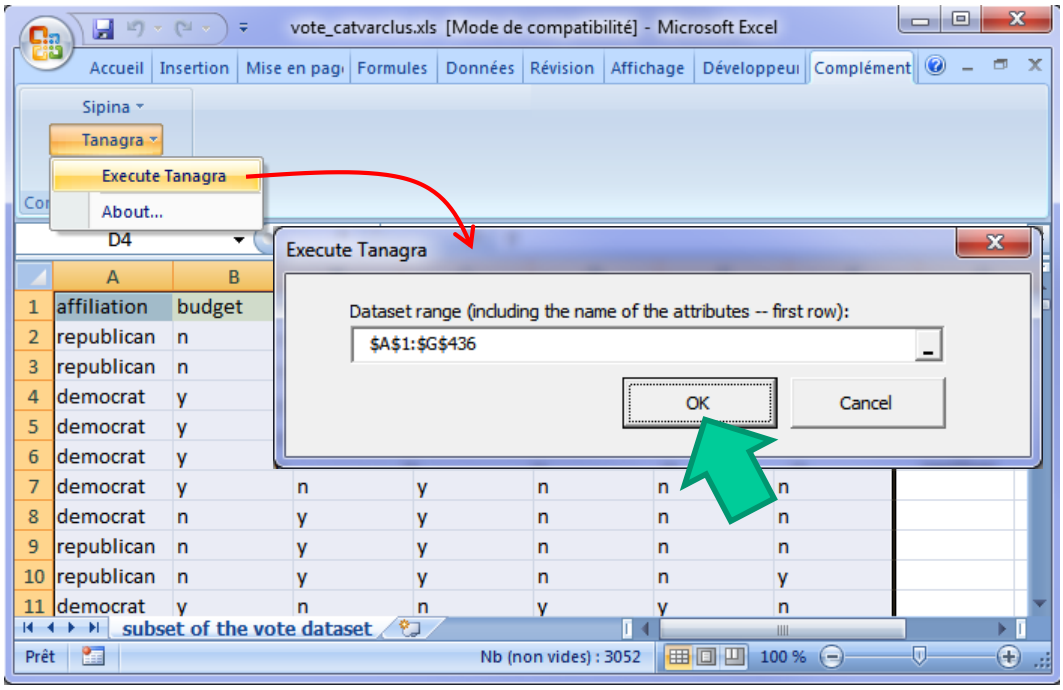
Le composant **CatVARHCA** de Tanagra implémente la classification ascendante hiérarchique de modalités basée sur la métrique de Dice. Il propose plusieurs stratégies d'agrégation : saut minimum, saut maximum, lien moyen et Ward. Dans cette section, nous montrons sa mise en œuvre sur les données « vote », nous mettrons l'accent sur la lecture des résultats.

5.1 Importation des données

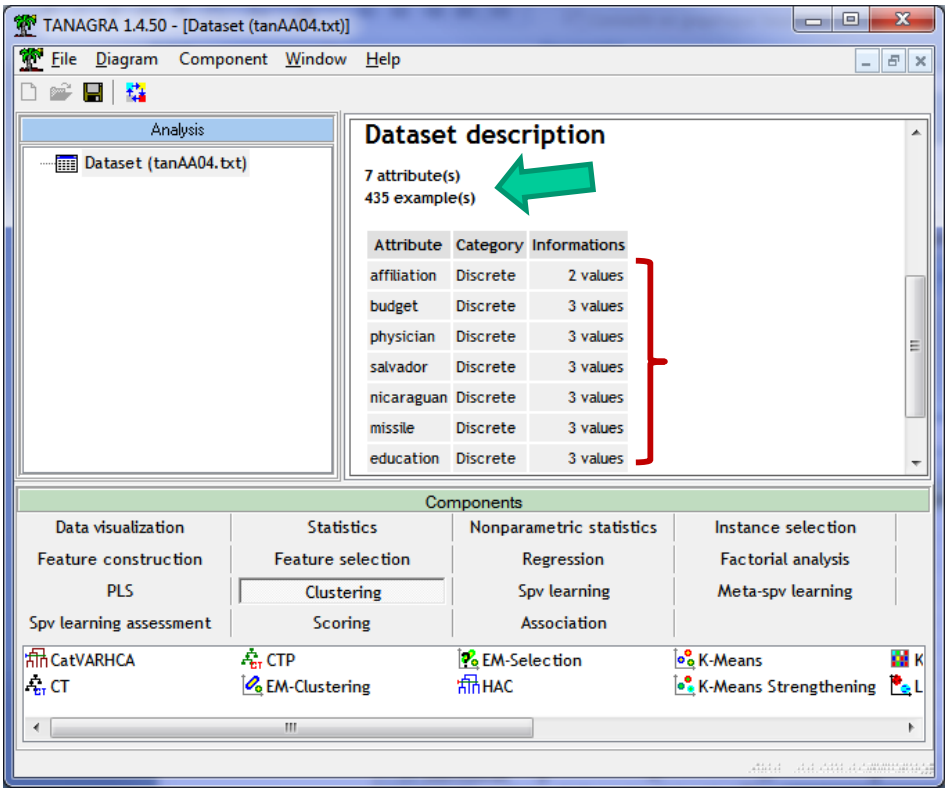
Nous chargeons le fichier « **vote_catvarclus.xls** » dans Excel. Nous sélectionnons les données puis, via la macro complémentaire « **tanagra.xla** »[13], nous les envoyons vers Tanagra.

[12] Avec le rapport de corrélation par exemple, etc. Voir R. Rakotomalala, « Analyse de Corrélation – Etude des dépendances - Variables quantitatives », Janvier 2012.

[13] Voir http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html pour l'installation et l'utilisation de la macro-complémentaire. Ce type de dispositif existe également pour les versions



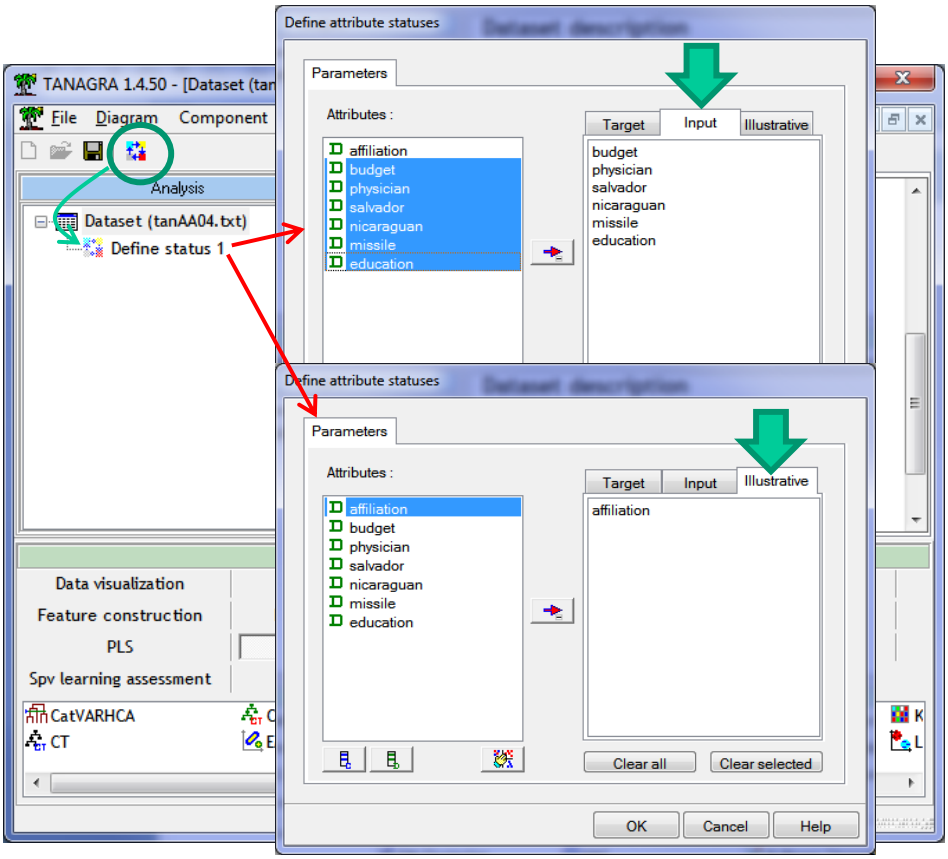
Tanagra est automatiquement démarré. Nous vérifions que nous avons bien 435 observations et 7 variables, toutes qualitatives.



précédentes d'Excel (2003 à 1997, http://tutoriels-data-mining.blogspot.fr/2008/03/importation-fichier-xls-excel-macro.html) ou encore pour le tableur Calc des suites « Libre Office » et « Open Office » (http://tutoriels-data-mining.blogspot.fr/2011/07/tanagra-addon-pour-openoffice-33.html).

5.2 Spécification des variables

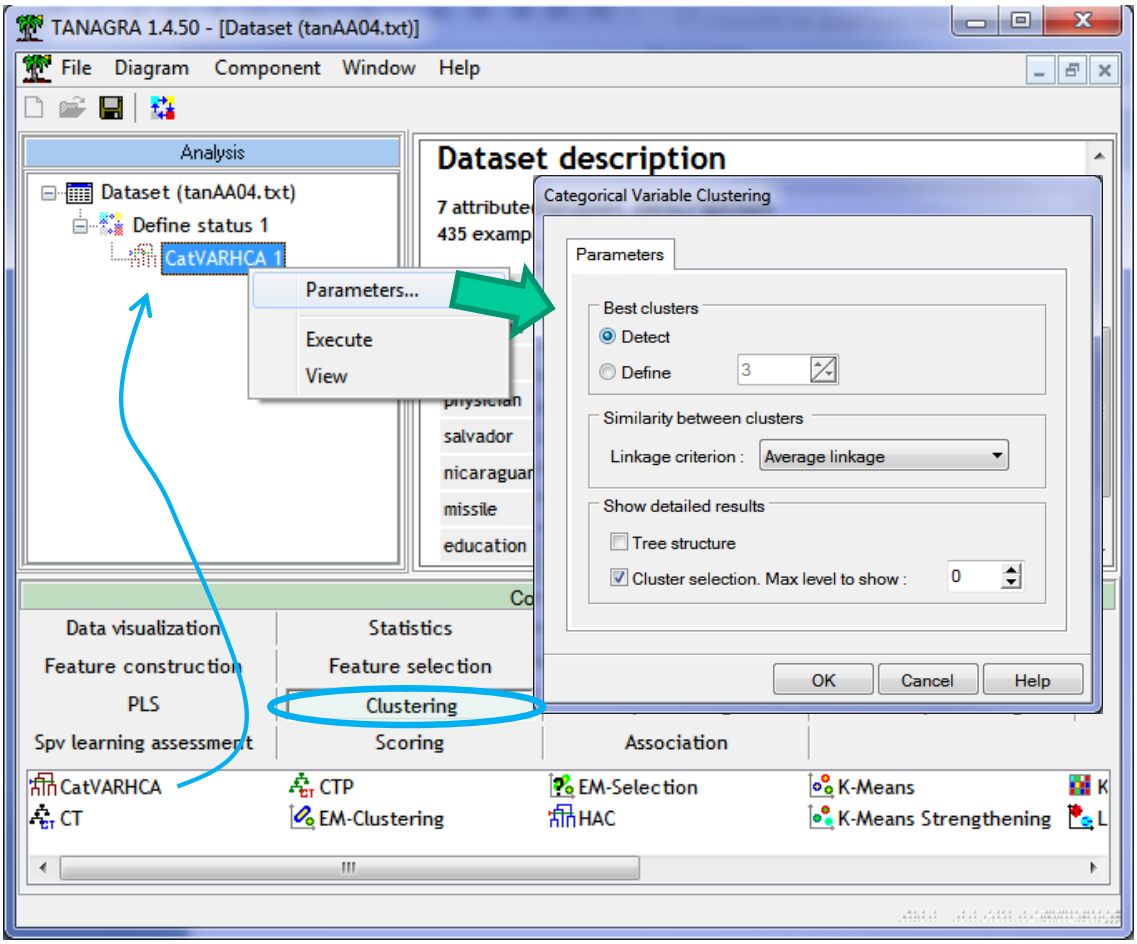
A l'aide du raccourci de la barre d'outils, nous insérons le composant DEFINE STATUS.



Nous plaçons les variables actives en INPUT (budget, ..., éducation), la variable supplémentaire en ILLUSTRATIVE (AFFILIATION).

5.3 CAH sur les modalités

Nous insérons ensuite le composant CATVARHCA (onglet CLUSTERING) dans le diagramme. Nous actionnons le menu PARAMETERS pour spécifier les paramètres de l'étude.



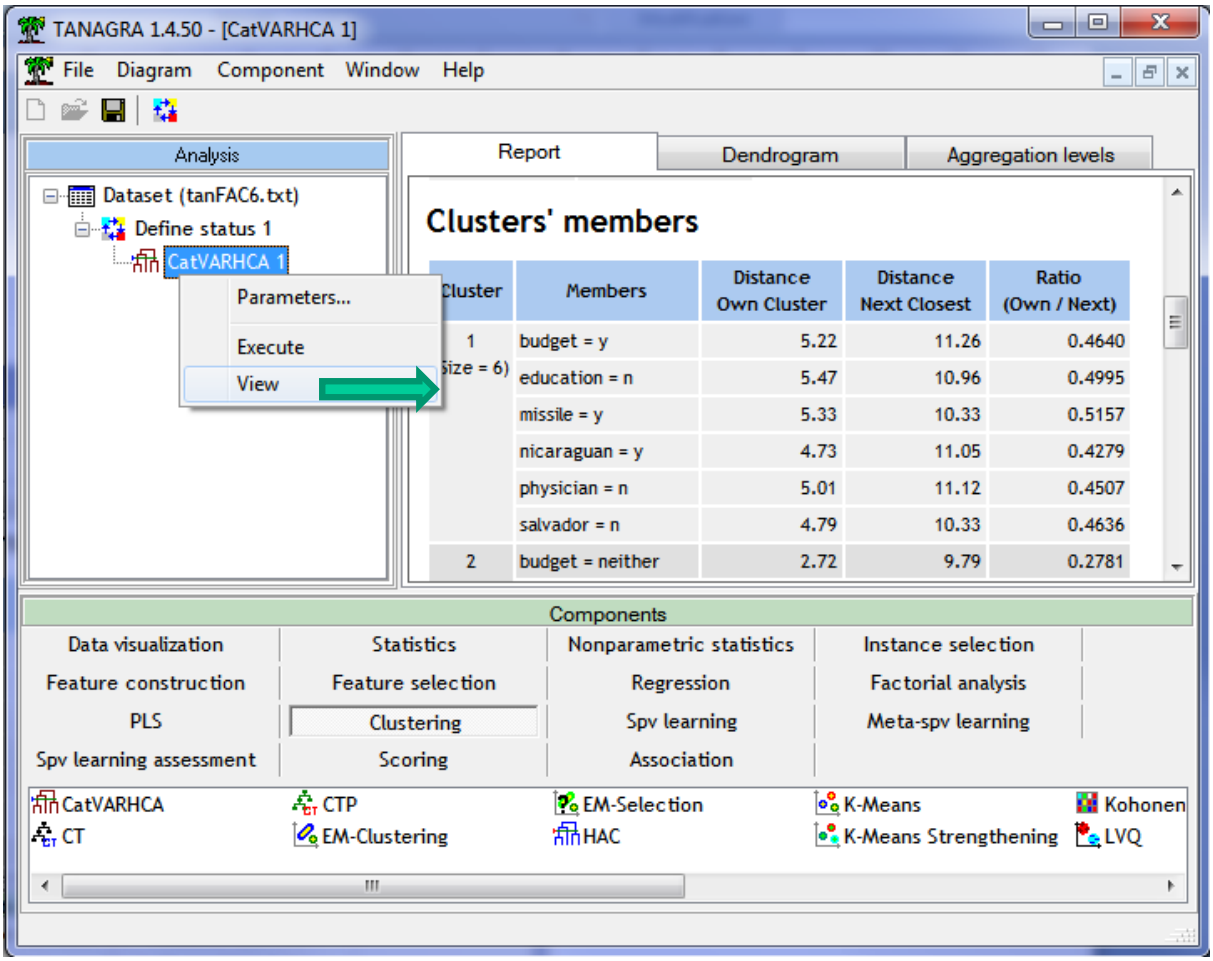
Best clusters – Detect. Le composant se charge de détecter automatiquement le nombre adéquat de clusters. Nous y reviendrons en détail plus bas. Si l'utilisateur le souhaite, il peut spécifier lui-même le nombre de clusters à produire (option « Define »).

Similarity between clusters – Average linkage. Il indique la stratégie d'agrégation utilisée durant la construction de la hiérarchie et le traitement des modalités supplémentaires. Les autres choix possibles sont « single linkage », « complete linkage » et « Ward ».

Show detailed results – Tree structure. Si elle est cochée, cette option permet d'énumérer dans un tableau les étapes des fusions effectuées lors de la construction de la hiérarchie. Comme nous disposons du dendrogramme par ailleurs, on peut s'en passer souvent.

Show detailed results – Cluster selection. Quand cette option est cochée, Tanagra affiche les valeurs utilisées pour la sélection du nombre « optimal » de clusters. Si « **Max level to show** = 0 », toutes les valeurs sont affichées. Attention, le tableau peut être très encombré lorsque le nombre initial de modalités est élevé.

Une fois les paramètres validés, nous actionnons le menu VIEW.



5.3.1 Onglet REPORT

L'onglet « **Report** » décrit l'essentiel des résultats. Plusieurs sections sont disponibles.

Clusters' characteristics. Cette section décrit les clusters avec leurs effectifs respectifs. Nous obtenons une partition en 3 groupes à 6 modalités.

Clusters' characteristics

Cluster number	Cluster size
1	6
2	6
3	6

Clusters' members. Ce tableau fournit le détail des clusters.

Clusters' members

Cluster	Members	Distance Own Cluster	Distance Next Closest	Ratio (Own / Next)
1 (Size = 6)	budget = y	5.22	11.26	0.4640
	education = n	5.47	10.96	0.4995
	missile = y	5.33	10.33	0.5157
	nicaraguan = y	4.73	11.05	0.4279
	physician = n	5.01	11.12	0.4507
	salvador = n	4.79	10.33	0.4636
2 (Size = 6)	budget = neither	2.72	9.79	0.2781
	education = neither	3.54	9.92	0.3569
	missile = neither	3.35	10.00	0.3353
	nicaraguan = neither	2.93	9.67	0.3027
	physician = neither	2.72	9.84	0.2766
	salvador = neither	2.94	9.88	0.2973
3 (Size = 6)	budget = n	5.01	9.50	0.5273
	education = y	5.31	9.54	0.5562
	missile = n	5.00	10.29	0.4861
	nicaraguan = n	4.76	9.68	0.4913
	physician = y	4.70	9.65	0.4865
	salvador = y	4.61	10.44	0.4419

Figure 5 - Cluster de rattachement pour chaque modalité - Qualité de l'association

1. Le premier groupe (Cluster 1) est constitué des votes (budget = y, education = y, missile = y, nicaraguan = y, physician = n, salvador = n). L'énorme intérêt de cette approche est que nous pouvons directement observer les conjonctions des votes. Ainsi, si l'on se réfère à l'association (nicaraguan, salvador), nous savons que ceux qui votent « y » au premier thème, votent « n » au second.
2. Le second groupe (Cluster 2) traduit l'absence implication d'un groupe de députés. Le vote « neither » correspond à une attitude particulière. Peut être que l'analyse des appartenances politiques permettra de mieux comprendre ce comportement ? Nous y reviendrons dans la section consacrée à la variable supplémentaire AFFILIATION.
3. Le troisième groupe (Cluster 3) est exactement à l'opposé du 1er. Il correspond à la conjonction des votes (budget = n, education = n, missile = n, nicaraguan = n, physician = y, salvador = y).

Indicateurs de Clusters' members. Des indicateurs permettent de situer la fiabilité de la typologie. « **Distance to Own Cluster** » indique la distance de chaque modalité par rapport à

son propre cluster. Elle est raccord avec le mode d'agrégation utilisé. Ainsi, elle serait systématiquement nulle si nous nous basons sur la méthode du saut minimum. « **Distance to Next Closest** » indique la distance de la modalité au cluster qui lui est le plus proche, en excluant son groupe d'appartenance. Le « **ratio** » est le rapport entre ces deux distances. Plus le ratio est faible, plus l'affectation de la modalité au cluster est crédible. S'il est supérieur à 1, cela voudrait dire que la modalité est plus proche d'un autre cluster que le sien. Il y a un problème dans la partition proposée. Il faudrait reconsidérer peut être le nombre de clusters à produire dans ce cas.

Distance to clusters – Supplementary variables. Ce tableau positionne les modalités de la variable supplémentaire AFFILIATION.

Distance to clusters - Supplementary variables

Variable = level	Cluster 1	Cluster 2	Cluster 3
affiliation = republican	13.558	9.303	5.445
affiliation = democrat	5.683	11.441	13.654

Pour chaque modalité, la valeur minimale est surlignée, indiquant le cluster qui lui est le plus proche. La distance est cohérente avec la stratégie d'agrégation utilisée durant la CAH (moyenne des indices de Dice dans notre exemple). Nous constatons que les **républicains** sont indiscutablement associés au 3ème cluster avec les votes (budget = n, education = n, missile = n, nicaraguan = n, physician = y, salvador = y) ; les **démocrates** au premier cluster c.-à-d. les votes (budget = y, education = y, missile = y, nicaraguan = y, physician = n, salvador = n). Nous notons également que le vote « neither » (2nd cluster) n'a pas vraiment de couleur politique.

Best cluster selection. Ce tableau décrit les indices d'agrégation et indique les calculs effectués pour détecter la meilleure partition. Nous le détaillerons plus loin (section 5.5).

5.4 Onglet dendrogram

Le dendrogramme est visible dans l'onglet du même nom. Il est possible de cliquer sur chaque nœud de l'arborescence pour visualiser les modalités présentes. Il n'y a qu'une modalité dans les feuilles ; toutes les modalités se retrouvent dans la racine. Les nœuds sont de couleur blanche lorsqu'ils correspondent aux clusters solutions. Nous en distinguons 3 dans notre exemple.

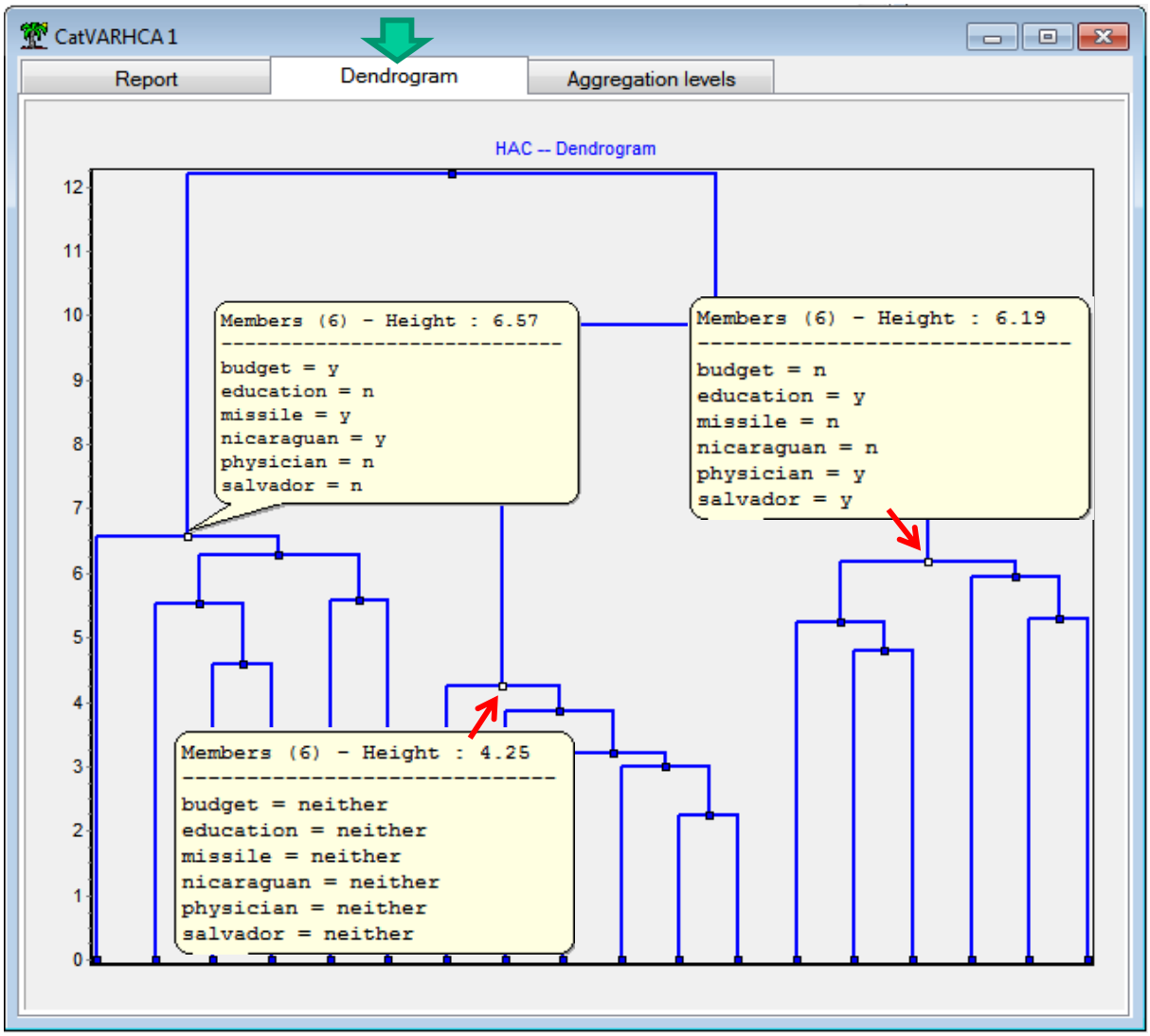
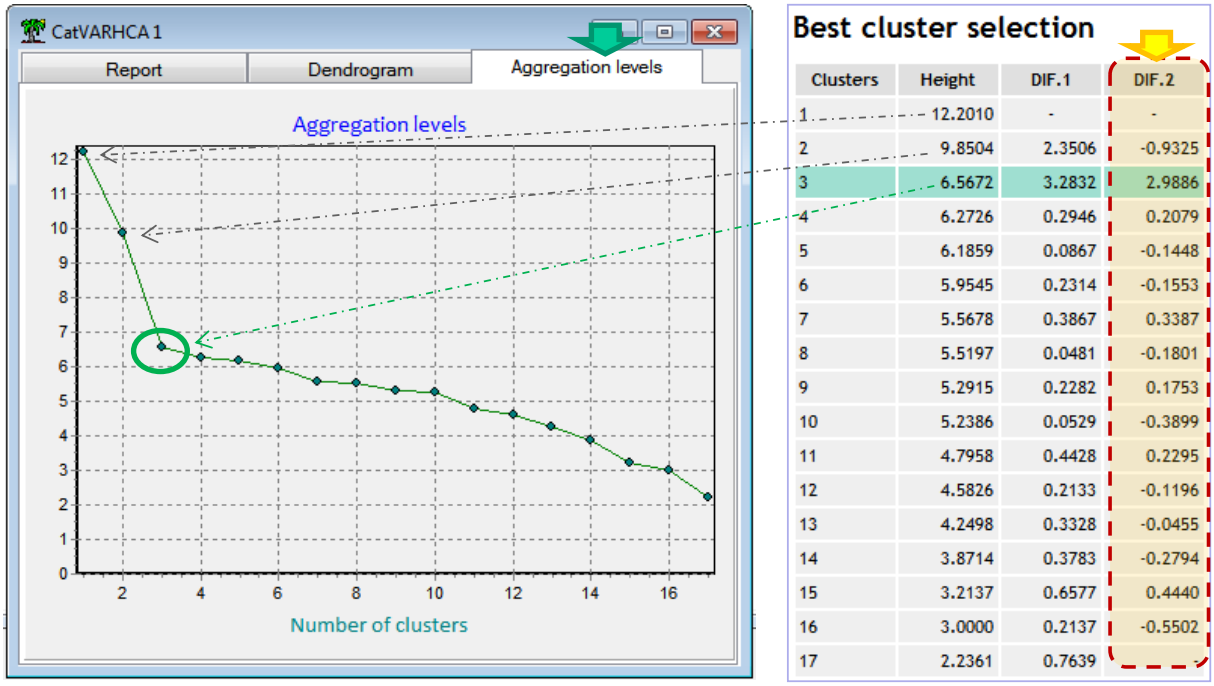


Figure 6 - Dendrogramme - Fichier "vote" - Tanagra

5.5 Détection automatique du nombre de clusters (Aggregation levels)

Dans l'onglet « Aggregation levels », nous disposons des indices d'agrégation (ou distance d'agrégation) en fonction du nombre de clusters formés. Ce graphique permet d'identifier visuellement la « cassure » dans l'évolution de l'indice, laissant présager d'une modification de structure forte dans les données. C'est la fameuse « règle du coude ». On s'appuie souvent sur cette information visuelle pour choisir le bon nombre de clusters.

Dans la Figure 7, nous couplons la visualisation graphique avec le tableau « Best cluster Selection ». Manifestement, une partition en 3 classes semble la plus judicieuse concernant le fichier VOTE.



Clusters	Height	DIF.1	DIF.2
1	12.2010	-	-
2	9.8504	2.3506	-0.9325
3	6.5672	3.2832	2.9886
4	6.2726	0.2946	0.2079
5	6.1859	0.0867	-0.1448
6	5.9545	0.2314	-0.1553
7	5.5678	0.3867	0.3387
8	5.5197	0.0481	-0.1801
9	5.2915	0.2282	0.1753
10	5.2386	0.0529	-0.3899
11	4.7958	0.4428	0.2295
12	4.5826	0.2133	-0.1196
13	4.2498	0.3328	-0.0455
14	3.8714	0.3783	-0.2794
15	3.2137	0.6577	0.4440
16	3.0000	0.2137	-0.5502
17	2.2361	0.7639	

Figure 7 - Indices d'agrégation en fonction du nombre de clusters

Le tableau recense les hauteurs d'agrégation (HEIGHT), les différences premières entre ces valeurs (DIF.1), puis les différences secondes (DIF.2). Cette dernière permet *d'estimer* la dérivée seconde de la fonction sous-jacente à l'évolution des valeurs, traduisant sa courbure. On choisir alors le point de courbure maximum qui correspond au « coude » de la courbe.

Pourquoi les différences secondes sont-elles pertinentes pour détecter le coude ? Les étudiants sont souvent sceptiques face à cette procédure qui leur paraît plus qu'hasardeuse. Ce tutoriel me donne l'occasion d'expliciter les calculs.

L'objectif est de calculer les dérivées secondes d'une fonction dont on ne dispose pas de l'expression analytique, mais d'une série de points (x_i, y_i). L'idée consiste à approximer localement la courbe avec un polynôme. Ainsi, avec 3 points, nous pouvons produire un polynôme de degré 2 de la forme « $y = P(x) = ax^2 + bx + c$ » qui passe par ces 3 points, c'est pour cela que l'on parle d'interpolation. La dérivée seconde approximée dans le voisinage étudié correspond alors à la dérivée seconde du polynôme, soit « 2a ».

Dans notre cas, les (x_i) en abscisse correspondent aux nombres de clusters candidats dans la hiérarchie, ils sont équidistants (x_i = 1, 2, 3, ...) c.-à-d. l'écart entre chaque point x_i est constant (**h = 1**). Cet élément est très important car il nous donne accès à des formules

d'interpolation simplifiées. Mettons que nous disposons de 3 couples de points (x_{-1},y_{-1}), (x_0,y_0) et (x_1,y_1), la formule d'interpolation de Stirling s'écrit[14] :

$$P(x) = y_0 + \frac{q}{1!} \times \frac{\Delta y_{-1} + \Delta y_0}{2} + \frac{q^2}{2!} \times \Delta^2 y_{-1}$$

Où $q = \frac{x - x_0}{h}$; Δy_0correspond à la différence première ; $\Delta^2 y_{-1}$ la différence seconde.

Prenons tout de suite un exemple pour expliciter la démarche. Mettons que nous voulons produire le polynôme d'interpolation au voisinage (x_0 = 3). Nous formons les calculs suivants à partir des hauteurs d'agrégation (Figure 7).

x_{-1} = 2	y_{-1} = 9.8504		
		Δy_{-1} = 6.5672 – 9.8504 = - 3.2832	
x_0 = 3	y_0 = 6.5672		$\Delta^2 y_{-1}$ = -0.2946 – (- 3.2832) = 2.9886
		Δy_0 = 6.2726 – 6.5672 = - 0.2946	
x_1 = 4	y_1 = 6.2726		

Le polynôme s'écrit :

$$P(x) = 6.5672 + q\frac{-3.2872 + (-0.2946)}{2} + \frac{q^2}{2}(2.9886)$$

Dans ce voisinage, le coefficient « a » du polynôme d'interpolation P (x) est égal à (2.9886/2), et la dérivée seconde correspond bien à $\Delta^2 y_{-1} = 2.9886$.

En scrutant la colonne des différences secondes (DIF.2) du tableau « Best Cluster Selection », nous constatons que la courbure la plus forte correspond à une partition en 3 groupes. Elle est surlignée en vert (Figure 7). C'est la solution proposée automatiquement par Tanagra[15].

6 Classification des modalités avec R

A chaque fois que j'introduis une nouvelle méthode dans Tanagra, je regarde si elle n'existe pas par ailleurs et, si ce n'est pas le cas, je la programme dans R. L'objectif est de vérifier la cohérence des résultats. Dans cette section, je décris le code source que j'ai mis au point sous R. La très bonne nouvelle est que nous obtenons exactement les mêmes résultats avec les deux logiciels. C'est toujours rassurant.

[14] D'autres méthodes d'interpolation existent (Gauss, Bessel, etc.). Voir B. Démidovitch I. Maron, « Eléments de calcul numérique », Editions MIR, Moscou, 1979 ; page 554. Cet ouvrage est un vrai bijou ! Il fait partie des quelques bouquins qui ont forgé ma vision des statistiques et des mathématiques appliquées !

[15] On notera que l'écriture de la différence seconde peut être simplifiée :

$$\Delta^2 y_{-1} = y_{-1} - 2 \times y_0 + y_1$$

6.1 Calcul de la matrice des indices de Dice

Après avoir importé les données, nous sélectionnons les variables actives et nous produisons les indicatrices. Nous calculons la matrice des distances à l'aide de deux boucles imbriquées.

```
#chargement des données
library(xlsx)
vote.data <- read.xlsx(file="vote_catvarclus.xls",header=T,sheetIndex=1)
summary(vote.data)
#variables actives
vote.active <- subset(vote.data,select=2:7)
#codage disjonctif complet
#utilisation du package ade4
library(ade4)
disj <- acm.disjonctif(vote.active)
#matrice indice de Dice
#Abdallah & Saporta, page 79
m <- matrix(0,ncol(disj),ncol(disj))
for (j in 1:ncol(disj)){
  for (jprim in 1:ncol(disj)){
    m[j,jprim] <- 0.5 * sum(disj[,j]*(1-disj[,jprim])+ (1-disj[,j])*disj[,jprim])
  }
}
#nommage des lignes et des colonnes
colnames(m) <- colnames(disj)
rownames(m) <- colnames(disj)
```

Nous obtenons la matrice suivante (Figure 8) :

Dice index	budget.n	budget.neither	budget.y	physician.n	physician.neither	physician.y	salvador.n	salvador.neither	salvador.y	nicaraguan.n	nicaraguan.neithe	nicaraguan.y	missile.n
budget.n	0	91	212	184	91	28	173.5	91	38.5	34.5	85	183.5	42.5
budget.neither	91	0	132	126	5	92	105.5	9	108.5	90.5	10	122.5	105.5
budget.y	212	132	0	31	127	186	42.5	125	176.5	181.5	130	32.5	172.5
physician.n	184	126	31	0	129	212	32.5	121	187.5	179.5	128	33.5	178.5
physician.neither	91	5	127	129	0	94	104.5	9	109.5	92.5	10	120.5	104.5
physician.y	28	92	186	212	94	0	184.5	95	26.5	34.5	87	184.5	37.5
salvador.n	173.5	105.5	42.5	32.5	104.5	184.5	0	111.5	210	191	109.5	21	191
salvador.neither	91	9	125	121	9	95	111.5	0	113.5	92.5	11	121.5	104.5
salvador.y	38.5	108.5	176.5	187.5	109.5	26.5	210	113.5	0	23	104.5	196	25
nicaraguan.n	34.5	90.5	181.5	179.5	92.5	34.5	191	92.5	23	0	96.5	210	30
nicaraguan.neither	85	10	130	128	10	87	109.5	11	104.5	96.5	0	128.5	102.5
nicaraguan.y	183.5	122.5	32.5	33.5	120.5	184.5	21	121.5	196	210	128.5	0	188
missile.n	42.5	105.5	172.5	178.5	104.5	37.5	191	104.5	25	30	102.5	188	0
missile.neither	92.5	14.5	121.5	115.5	14.5	98.5	102	16.5	110	96	14.5	118	114
missile.y	168	103	50	47	104	170	28.5	104	188.5	180.5	108	32.5	206.5
education.n	174	118	42	38	116	180	43.5	117	173.5	169.5	121	43.5	168.5
education.neither	91	17	125	123	17	93	107.5	18	107.5	94.5	17	121.5	104.5
education.y	38	88	177	180	90	33	170.5	90	42.5	42.5	87	173.5	47.5

Figure 8 - Indices de Dice - Modalités des variables actives du fichier "Vote"

Nous constatons entres autres que les modalités « budget = neither » et « physician = neither » sont les plus proches avec un indice $\delta^2 = 5$. La fusion commencera par ce couple de modalités durant la construction de la hiérarchie indicée (Figure 9).

6.2 CAH sur les modalités et indices d'agrégation

Après avoir transformé la matrice en type interne 'dist', nous l'envoyons à la méthode **hclust**() de R avec la stratégie des liens moyens (average linkage).

```
#transformation de type[16]
d <- as.dist(sqrt(m))
#clustering des modalités
arbre.moda <- hclust(d,method="average")
plot(arbre.moda)
```

Le dendrogramme correspond en tous points à celui de Tanagra.

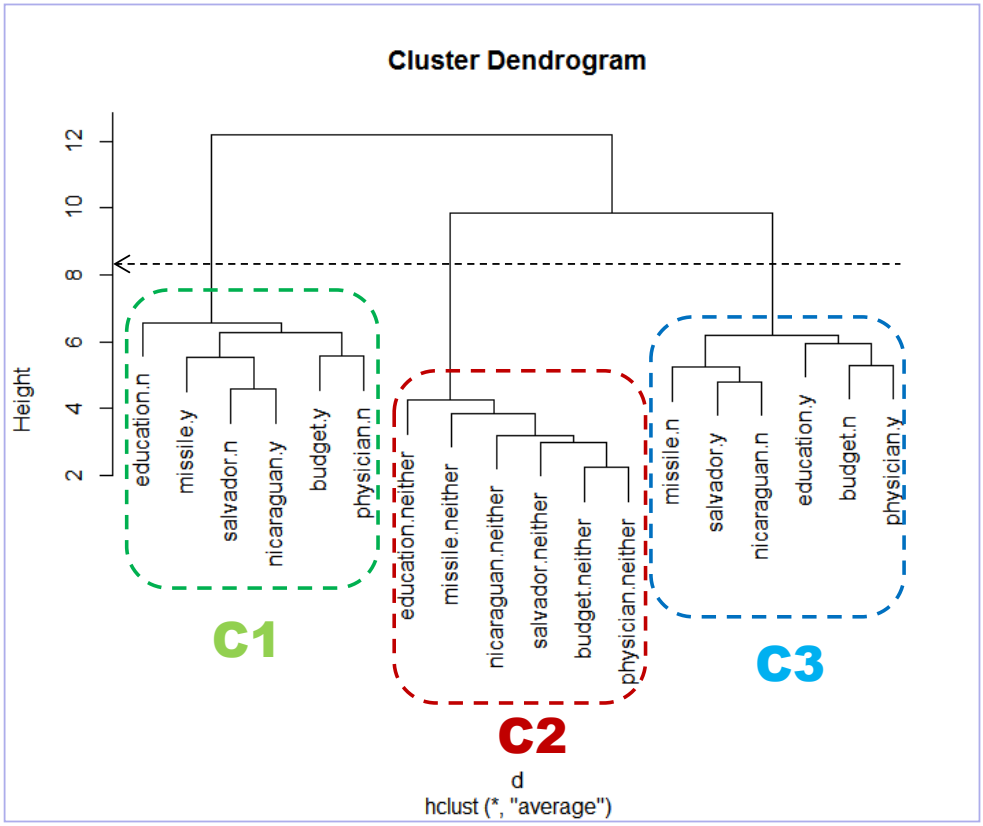


Figure 9 - Dendrogramme - CAH sur les modalités - Fichier "vote" - R

Il en est de même en ce qui concerne les hauteurs d'agrégation.

[16] En toute rigueur, il faut passer par δ (la racine carrée de l'indice) pour les stratégies « saut minimim », « saut maximum » et « lien moyen ». En pratique, utiliser δ^2 ne change en rien la nature des résultats. Pour la méthode de Ward en revanche, il est impératif d'utiliser δ^2 pour quantifier les dissimilarités entre les modalités afin que l'interprétation sous forme de perte d'inertie minimale à chaque fusion soit valide.

```
#tri des indices d'agrégation de manière décroissante
v <- sort(arbre.moda$height,decreasing=T)
#graphique
plot(1:length(v),v,type="b")
```

La courbe (Figure 10) est cohérente avec celle de Tanagra (Figure 7).

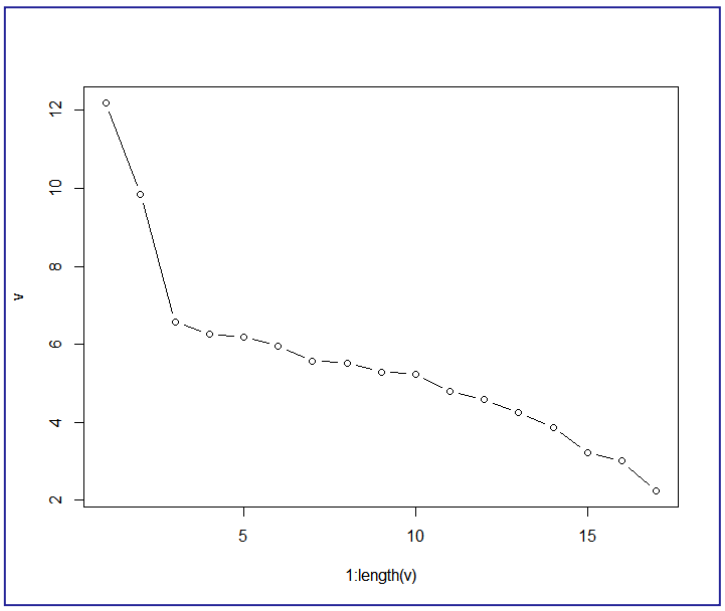


Figure 10 : Indices d'agrégation selon le nombre de clusters - Fichier "vote" - R

En nous référant à la démarche décrite précédemment (section 4.2), il est tout aussi aisé de positionner les modalités supplémentaires sous R.

7 Autres approches pour la classification des modalités

7.1 La procédure « varclus » du package 'Hmisc'

Il me paraissait extraordinaire que personne n'ait proposé un outil pour la classification de modalités jusqu'à présent. En cherchant un peu (*vive Google*), j'ai fini par trouver la procédure « varclus » du package Hmisc[17]. En lisant attentivement la documentation, je me suis rendu compte qu'il pouvait traiter des indicatrices avec une mesure de dissimilarité particulière. Voici le code utilisé sous R,

```
#chargement des données
library(xlsx)
vote.data <- read.xlsx(file="vote_catvarclus.xls",header=T,sheetIndex=1)
summary(vote.data)
#variables actives
```

[17] http://cran.r-project.org/web/packages/Hmisc/index.html

```
vote.active <- subset(vote.data,select=2:7)
#codage disjonctif complet
library(ade4)
disj <- acm.disjonctif(vote.active)
#Varclus de Hmisc
library(Hmisc)
v <- varclus(as.matrix(disj),type="data.matrix",similarity="bothpos",method="average")
plot(v)
```

Nous obtenons le dendrogramme suivant :

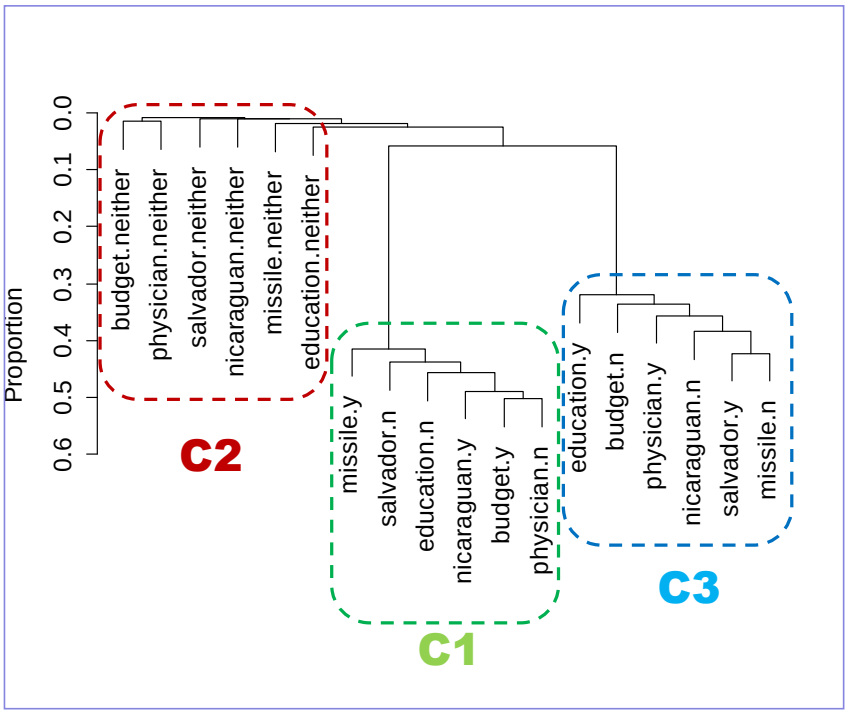


Figure 11 - Regroupement des modalités avec la procédure varclus() du package 'Hmisc' – Lien moyen

Nous disposons d'une autre vision des cooccurrences des votes. Si on coupait l'arbre comme on le fait habituellement, nous n'obtiendrions pas la bonne solution. Il est plus indiqué dans cette configuration d'isoler les 2 blocs très compacts C1 et C3, et de mettre les autres modalités dans un groupe à part (C2). Le vote « neither » correspond à un comportement distinct mais disparate des députés, nous avons une sorte de groupe « autre que C1 et C3 ».

Intrigué par cette conclusion qui n'était pas perceptible avec l'indice de Dice, j'ai calculé la distribution de fréquences de « neither » sur les 6 votes.

Étiquettes de lignes	Nombre de neither
0	360
1	58
2	12
3	2
5	1
6	2
Total général	**435**

Il semble effectivement que le vote « neither » soit le fruit d'un comportement ponctuel et ne résulte pas d'une volonté délibérée. Ca aurait été le cas si un député votant « neither » sur un des thèmes le fait sur la majorité des thèmes. Ce n'est pas du tout ce que l'on constate ici (2 députés seulement ont voté « neither » sur l'ensemble des thèmes).

Mesure de similarité. Ce résultat est tributaire de la mesure de dissimilarité utilisée par varclus(). L'option « similarity = ''bothpos'' » permet de calculer dans un premier temps la similarité entre les indicatrices à l'aide d'une estimation de la probabilité conjointe c.-à-d.

$$P_{jj'} = \frac{1}{n}\sum_{i=1}^{n} k_{ij} \times k_{ij'}$$

Nous pouvons accéder à cette matrice en faisant appel à la propriété « sim » de l'objet varclus() [**print(v$sim)**].

similarity=bothpos	budget.n	budget.neithe	budget.y	physician.n	physician.ne	physician.y	salvador.n	salvador.neit	salvador.y	nicaraguan.n	nicaraguan.n	nicaraguan.y	missile.n	missile.neith	missile.y	education.n	education.ne	education.y
budget.n	0.393	0.000	0.000	0.057	0.000	0.336	0.037	0.005	0.352	0.322	0.018	0.053	0.336	0.009	0.048	0.064	0.023	0.306
budget.neither	0.000	0.025	0.000	0.007	0.014	0.005	0.009	0.009	0.007	0.009	0.007	0.009	0.007	0.005	0.014	0.009	0.009	0.007
budget.y	0.000	0.000	0.582	0.503	0.011	0.067	0.432	0.021	0.129	0.078	0.009	0.494	0.131	0.037	0.414	0.462	0.039	0.080
physician.n	0.057	0.007	0.503	0.568	0.000	0.000	0.448	0.023	0.097	0.076	0.007	0.485	0.110	0.044	0.414	0.464	0.037	0.067
physician.neither	0.000	0.014	0.011	0.000	0.025	0.000	0.011	0.009	0.005	0.005	0.007	0.014	0.009	0.005	0.011	0.014	0.009	0.002
physician.y	0.336	0.005	0.067	0.000	0.000	0.407	0.018	0.002	0.386	0.329	0.021	0.057	0.354	0.002	0.051	0.057	0.025	0.324
salvador.n	0.037	0.009	0.432	0.448	0.011	0.018	0.478	0.000	0.000	0.005	0.005	0.469	0.037	0.030	0.411	0.407	0.028	0.044
salvador.neither	0.005	0.009	0.021	0.023	0.009	0.002	0.000	0.034	0.000	0.009	0.009	0.016	0.014	0.005	0.016	0.016	0.011	0.007
salvador.y	0.352	0.007	0.129	0.097	0.005	0.386	0.000	0.000	0.487	0.395	0.021	0.071	0.423	0.016	0.048	0.113	0.032	0.343
nicaraguan.n	0.322	0.009	0.078	0.076	0.005	0.329	0.005	0.009	0.395	0.409	0.000	0.000	0.372	0.009	0.028	0.083	0.023	0.303
nicaraguan.neither	0.018	0.007	0.009	0.007	0.007	0.021	0.005	0.009	0.021	0.000	0.034	0.000	0.018	0.009	0.007	0.007	0.014	0.014
nicaraguan.y	0.053	0.009	0.494	0.485	0.014	0.057	0.469	0.016	0.071	0.000	0.000	0.556	0.083	0.032	0.441	0.446	0.034	0.076
missile.n	0.336	0.007	0.131	0.110	0.009	0.354	0.037	0.014	0.423	0.372	0.018	0.083	0.474	0.000	0.000	0.117	0.032	0.324
missile.neither	0.009	0.005	0.037	0.044	0.005	0.002	0.030	0.005	0.016	0.009	0.009	0.032	0.000	0.051	0.000	0.025	0.011	0.014
missile.y	0.048	0.014	0.414	0.414	0.011	0.051	0.411	0.016	0.048	0.028	0.007	0.441	0.000	0.000	0.476	0.393	0.028	0.055
education.n	0.064	0.009	0.462	0.464	0.014	0.057	0.407	0.016	0.113	0.083	0.007	0.446	0.117	0.025	0.393	0.536	0.000	0.000
education.neither	0.023	0.009	0.039	0.037	0.009	0.025	0.028	0.011	0.032	0.023	0.014	0.034	0.032	0.011	0.028	0.000	0.071	0.000
education.y	0.306	0.007	0.080	0.067	0.002	0.324	0.044	0.007	0.343	0.303	0.014	0.076	0.324	0.014	0.055	0.000	0.000	0.393

Figure 12 - Indice de similarité entre modalités utilisé par varclus() de 'Hmisc'

Prenons 2 valeurs pour bien comprendre cette matrice :

1. La proportion des députés qui ont voté « Budget = n » est P(Budget = n) = 0.393 ;
2. La proportion des députés qui ont voté à la fois « Budget = n » ET « physician = n » est P (Budget = n & Physician = n) = 0.057.

Mesure de dissimilarité et àppel à hclust() de R. La procédure varclus() fait appel en interne à hclust() avec l'instruction suivante :

```
> v$hclust

Call:
hclust(d = as.dist(1 - x), method = method)

Cluster method   : average
Number of objects: 18
```

L'indice de dissimilarité entre modalités utilisée durant la CAH est donc tout simplement

$$D_{jj'} = 1 - P_{jj'}$$

Deux informations importantes sautent aux yeux :

1. Cet indice n'est pas une distance puisque que ($D_{jj} \neq 0$). Mais ça ne pose pas de problème à hclust() parce qu'elle ignore tout simplement cette singularité dans ses calculs internes.
2. La dissimilarité entre 2 modalités distinctes issues de la même variable prend la valeur la plus élevée possible (D = 1), rendant difficile leur réunion dans la même classe. Cette caractéristique ne pesait pas aussi fortement avec l'indice de Dice.

Ici, comme précédemment, l'utilisation des variables qualitatives supplémentaires ne présente aucune difficulté. Le calcul des probabilités conjointes avec les modalités existantes est aisé.

7.2 ACM + CAH sur coordonnées factorielles

Une autre approche très simple basée sur des outils existants permet de résoudre un problème de classification de modalités. Elle s'appuie sur la démarche « tandem analysis ». A savoir : nous réalisons dans un premier temps une analyse factorielle sur les données ; puis, dans un second temps, nous effectuons la classification sur les objets projetés dans l'espace intermédiaire. Dans notre cas, il s'agirait d'effectuer tout d'abord une ACM (analyse des correspondances multiples ; on parle aussi d'AFCM, analyse factorielle des correspondances multiples) sur les données originelles, puis réaliser **une classification à partir des coordonnées factorielles des modalités**. Je vois plusieurs avantages ici :

1. L'ACM est une technique connue, bien maîtrisée.
2. Nous projetons les objets (les modalités) dans un espace numérique.
3. En ne conservant que les facteurs pertinents, nous éliminons le bruit des données, les fluctuations d'échantillonnage qui ne sont pas représentatives des « vraies » relations entres les variables dans la population.

4. Une fois la typologie construite, les centres de classes ont un sens. Il est aisé de les obtenir à partir des coordonnées factorielles.
5. Cela ouvre la porte à la CAH basée sur la méthode des centroïdes ou la méthode de Ward, les autres stratégies d'agrégation restant opérationnelles.
6. Ainsi, le traitement des variables qualitatives supplémentaires est facilité. Il suffit de rattacher chaque modalité au centre de classe le plus proche.

Dans ce qui suit, nous réalisons sous R les calculs pour le fichier « vote ».

7.2.1 ACM et récupération des coordonnées factorielles des modalités

Nous effectuons une ACM en ne conservant que les 2 premiers facteurs. Nous projetons les modalités dans le premier plan factoriel.

```
#chargement des données
library(xlsx)
vote.data <- read.xlsx(file="vote_catvarclus.xls",header=T,sheetIndex=1)
summary(vote.data)
#variables actives
vote.active <- subset(vote.data,select=2:7)
#codage disjonctif complet
library(ade4)
disj <- acm.disjonctif(vote.active)
#acm - dudi.coa du package ade4
acm <- dudi.coa(disj,scannf=F,nf=2)
acm.coord <- data.frame(acm$co)
rownames(acm.coord) <- colnames(disj)
#projection des modalités dans le premier plan factoriel
plot(acm.coord[,1],acm.coord[,2],type="n",xlim=c(-1.25,1.25),ylim=c(-0.5,4.5),xlab="F1",ylab="F2")
text(acm.coord[,1],acm.coord[,2],labels=rownames(acm.coord),cex=0.75,col="blue")
abline(h=0,v=0)
```

Leurs positions relatives dans le plan factoriel (Figure 13) laissent déjà présager du résultat de la CAH à venir.

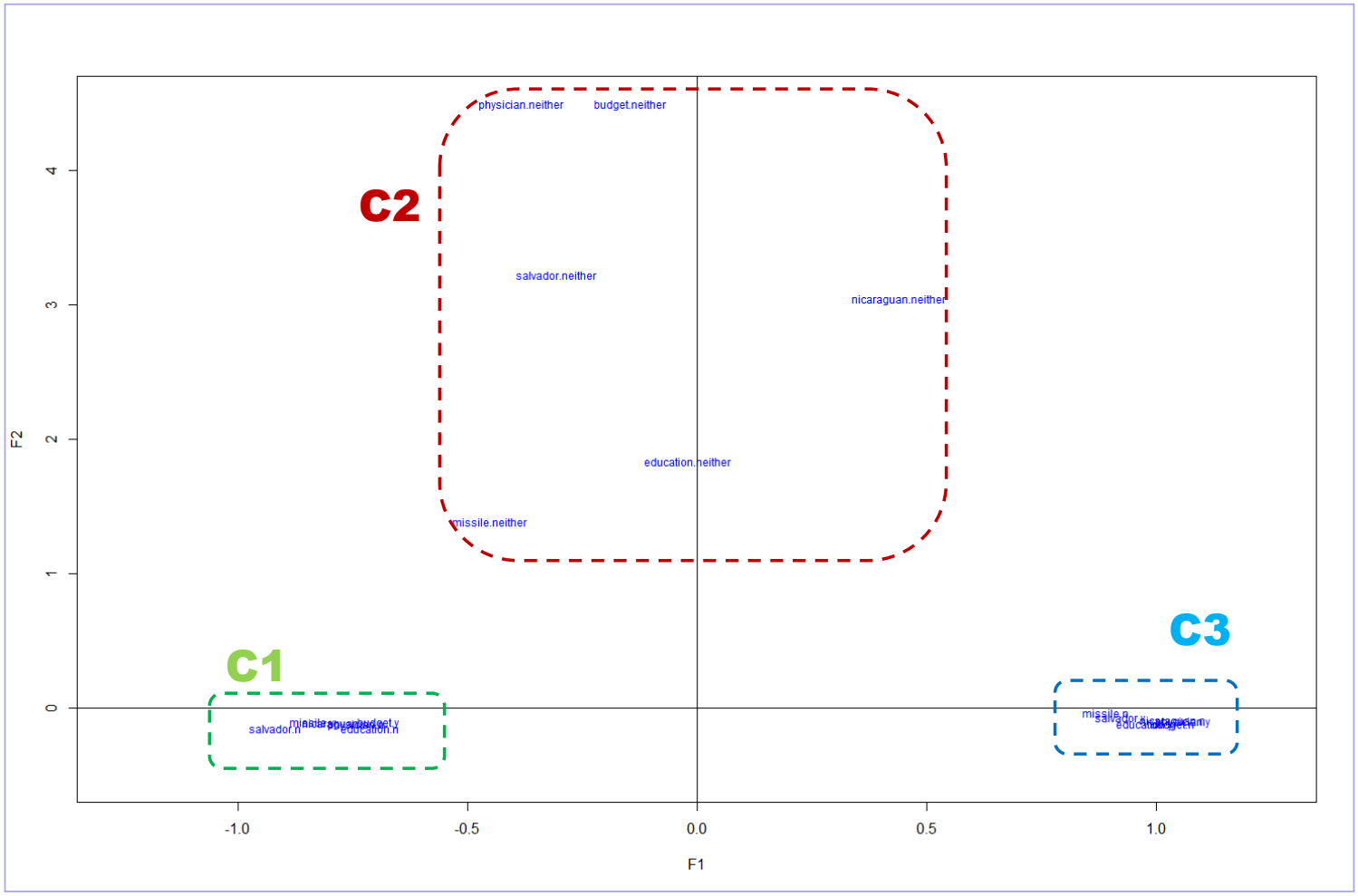


Figure 13 - Projection des modalités dans le premier plan factoriel de l'ACM - Fichier "vote"

7.2.2 CAH sur les coordonnées factorielles

Le code suivant permet de réaliser une CAH des modalités basée sur distance euclidienne dans l'espace factoriel, avec la méthode du lien moyen (average linkage)[18].

```
#distance entre modalités
m.acm <- dist(acm.coord,method="euclidian")
#CAH, méthode du lien moyen
arbre.acm <- hclust(m.acm,method="average")
plot(arbre.acm)
```

Nous distinguons bien les 3 types de comportement des députés soulignés dans les analyses précédentes. Mais les groupes, en accord avec les coordonnées des modalités dans le plan factoriel (Figure 13), sont analogues à ceux obtenus avec la procédure varclus() de Hmisc (section 7.1), à savoir : C1 et C3 forment des classes compactes, C2 décrit une attitude à part.

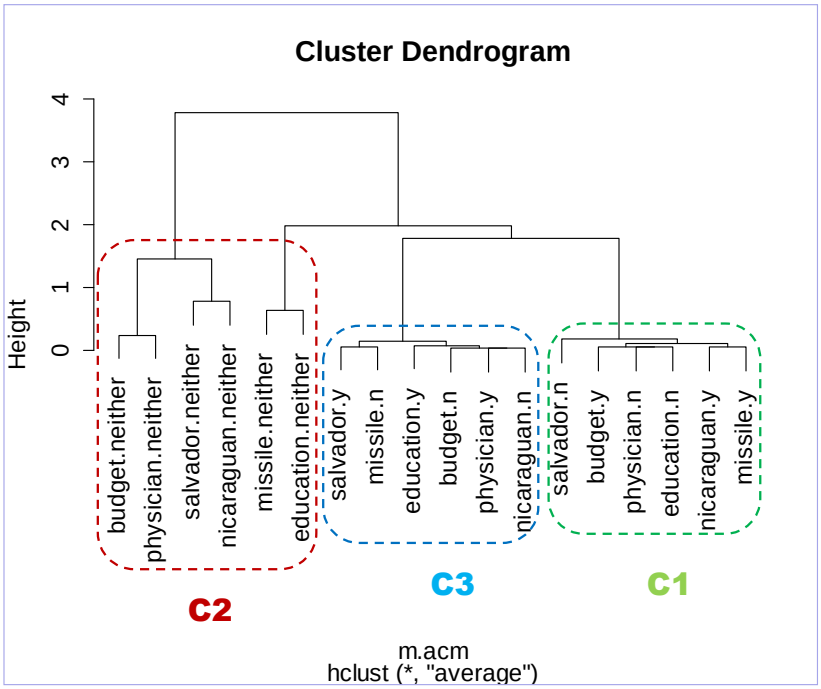


Figure 14 - Regroupement des modalités à partir de leurs coordonnées factorielles (ACM) - Fichier Vote

Remarque : Les modalités représentent un certain nombre d'observations. En toute rigueur, elles devraient être pondérées par leur fréquence dans hclust(), en utilisant l'option « members ». Cette pondération sera d'autant plus déterminante que les fréquences des modalités sont fortement disparates.

[18] Si l'on souhaite mettre en œuvre la **méthode Ward** avec **hclust()**, il faut utiliser le carré de la distance euclidienne (cf. http://en.wikipedia.org/wiki/Ward%27s_method).

A partir de cette partition, nous pouvons calculer les centres de classes dans le premier plan factoriel.

	Centres de classes	
Cluster	F1	F2
1	-0.7775	-0.1277
2	-0.1363	2.6701
3	0.9821	-0.0923

Figure 15 - Centres de classes dans le repère factoriel (ACM) - Fichier "vote"

Attention, il faut prendre en considération le poids de chaque modalité dans les calculs. Détaillons cela pour le cluster C1. Nous prenons comme point de départ le tableau contenant le poids des modalités (n_j/n pour la modalité « j »), leurs coordonnées sur les 2 composantes, et leur classe de rattachement.

row.names	weight(n_j/n)	F1	F2	cluster
budget.y	0.582	-0.6941	-0.1123	1
physician.n	0.568	-0.7422	-0.1238	1
salvador.n	0.478	-0.9184	-0.1523	1
nicaraguan.y	0.556	-0.7900	-0.1160	1
missile.y	0.476	-0.8384	-0.1117	1
education.n	0.536	-0.7124	-0.1527	1
budget.neither	0.025	-0.1444	4.4933	2
physician.neither	0.025	-0.3829	4.4897	2
salvador.neither	0.034	-0.3057	3.2224	2
nicaraguan.neither	0.034	0.4412	3.0428	2
missile.neither	0.051	-0.4521	1.3898	2
education.neither	0.071	-0.0194	1.8384	2
budget.n	0.393	1.0363	-0.1228	3
physician.y	0.407	1.0596	-0.1062	3
salvador.y	0.487	0.9227	-0.0786	3
nicaraguan.n	0.409	1.0368	-0.0987	3
missile.n	0.474	0.8908	-0.0362	3
education.y	0.393	0.9742	-0.1253	3

Figure 16 : Poids, coordonnées factorielles et classes d'appartenance des modalités

Nous avons pour le 1er facteur :

$$C_1(F_1) = \frac{0.582 \times (-0.6941) + 0.568 \times (-0.7422) + \cdots + 0.536 \times (-0.7124)}{0.582 + 0.568 + 0.478 + 0.556 + 0.476 + 0.536} = -0.7775$$

Et pour le 2nd :

$$C_1(F_2) = \frac{0.582 \times (-0.1123) + 0.568 \times (-0.1238) + \cdots + 0.536 \times (-0.1527)}{0.582 + 0.568 + 0.478 + 0.556 + 0.476 + 0.536} = -0.1277$$

Le barycentre est bien placé au sein du nuage des modalités du 1er cluster (Figure 13).

8 Conclusion

La classification des modalités des variables qualitatives vient en complément de la classification des variables qualitatives. Elle précise la nature des relations entre les variables. Elle fournit des résultats interprétables et permet l'utilisation des variables supplémentaires.

De manière plus anecdotique, nous avons également profité de ce tutoriel pour justifier l'usage des différences secondes pour la détection du « coude » dans la courbe des indices d'agrégation (section 5.5), l'objectif étant de proposer une réponse plausible au véritable serpent de mer de la classification automatique : identifier le nombre de clusters le plus approprié dans le processus de partitionnement. Cette astuce est implémentée dans le composant de classification des modalités CATVARHCA de Tanagra. Bien sûr, elle n'est pas exempte de reproches. Elle laisse entendre entres autres que la courbe de décroissance des indices d'agrégation est convexe en tous points. Ce n'est pas forcément le cas. Bref, quitte à rabâcher, il est évident qu'on ne peut pas se passer de l'interprétation pour valider les résultats d'une classification automatique.