4 Classification Ascendante Hiérarchique

4.1 Introduction

Classifier, c'est regrouper entre eux des objets similaires selon tel ou tel critère. Les diverses techniques de classification (ou d'"analyse typologique", de "taxonomie", ou "taxinomie" ou encore "analyse en clusters" (amas)) visent toutes à répartir *n* individus, caractérisés par *p* variables X_1, X_2, ..., X_p en un certain nombre m de sous-groupes aussi homogènes que possible.
On distingue deux grandes familles de techniques de classification :

- Les classifications non hiérarchiques ou partitionnements, aboutissant à la décomposition de l'ensemble de tous les individus en m ensembles disjoints ou classes d'équivalence ; le nombre m de classes est fixé. Le résultat obtenu est alors une partition de l'ensemble des individus, un ensemble de parties, ou classes de l'ensemble I des individus telles que :
  - toute classe soit non vide
  - deux classes distinctes sont disjointes
  - tout individu appartient à une classe.

- Les classifications hiérarchiques : pour un niveau de précision donné, deux individus peuvent être confondus dans un même groupe, alors qu'à un niveau de précision plus élevé, ils seront distingués et appartiendront à deux sous-groupes différents. Le résultat d'une classification hiérarchique n'est pas une partition de l'ensemble des individus. C'est une hiérarchie de classes telles que :
  - toute classe est non vide
  - tout individu appartient à une (et même plusieurs) classes
  - deux classes distinctes sont disjointes, ou vérifient une relation d'inclusion (l'une d'elles est incluse dans l'autre)
  - toute classe est la réunion des classes qui sont incluses dans elle.

Remarques. Ces méthodes jouent un rôle un peu à part dans l'univers des méthodes statistiques. En effet :
- L'aspect inférentiel est ici inexistant ;
- Il existe un grand nombre de variantes de ces méthodes, et on peut être amené à appliquer plusieurs de ces méthodes sur un même jeu de données, jusqu'à obtenir une classification "qui fasse sens" ;
- Au contraire des méthodes factorielles, l'accent est souvent mis sur les n individus et non sur les p variables qui les décrivent.

4.2 Exemple

4.2.1 Enoncé

On reprend le cas "Basket", qui a été présenté au paragraphe 2.6 page 26.

4.2.2 Choix des variables représentant les individus

Une première étape consiste à choisir une mesure de la "dissimilarité" ou "distance" entre les sujets. Mais, les variables de départ (Taille en cm, Poids, ...) s'expriment avec des unités différentes et prennent leurs valeurs sur des échelles difficilement comparables. Nous choisissons donc de représenter les individus à l'aide des variables centrées réduites associées aux variables de départ (en utilisant l'écart type corrigé comme dénominateur) :

	TAI	VIT	DET	PAS	LEG	STA
I1	-1,1125	1,3473	1,5025	0,9702	1,1665	0,5535
I2	-0,0056	-0,7615	-0,7446	0,9702	-1,0845	-0,9793
I3	1,1013	-0,9724	-0,6643	1,5023	-0,9514	0,0426
I4	-0,8106	1,1364	0,9407	1,2120	1,1423	0,5535
I5	-1,1125	1,3473	0,9407	-0,2392	1,2391	1,0644

I6	-0,6093	0,7146	1,1012	-0,2876	0,9124	0,8090
I7	-1,1125	0,5038	0,9407	-0,4810	1,3964	-1,2347
I8	-1,3137	1,3473	1,4222	-0,9648	1,2028	-2,0011
I9	-1,5150	1,3473	1,4222	-1,4485	1,2028	-1,7456
I10	-0,0056	-1,3941	-0,8248	1,0669	-0,4673	-1,2347
I11	0,4975	-0,1289	-0,2630	1,2120	-0,3705	-0,2129
I12	-0,1062	0,0820	-0,6643	-0,4810	-0,2857	0,2980
I13	0,3969	-0,3398	-0,6643	-0,0941	-0,5278	1,0644
I14	1,1013	-0,7615	-0,8248	-0,7229	-1,0240	0,5535
I15	1,0007	-0,5506	-1,0656	-0,8197	-0,9877	0,2980
I16	1,1013	-0,5506	-1,2261	-1,2067	-0,6246	0,8090
I17	1,1013	-0,9724	-0,6643	-1,2067	-1,0966	0,2980
I18	1,4032	-1,3941	-0,6643	1,0185	-0,8424	1,0644

4.2.3 Choix d'un indice de dissimilarité ou distance entre individus

Chaque individu statistique est ici représenté par 6 "coordonnées", à savoir les valeurs des variables centrées réduites associées aux 6 variables TAI, VIT, DET, PAS, LEG, STA. Pour évaluer la dissimilarité entre les individus, nous utiliserons la distance euclidienne. Autrement dit, si les coordonnées des individus I_i et I_j sont données par : $(x_{i1}, x_{i2}, x_{i3}, x_{i4}, x_{i5}, x_{i6})$ et $(x_{j1}, x_{j2}, x_{j3}, x_{j4}, x_{j5}, x_{j6})$, on a :

$$d(I_i,I_j) = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + (x_{i3} - x_{j3})^2 + (x_{i4} - x_{j4})^2 + (x_{i5} - x_{j5})^2 + (x_{i6} - x_{j6})^2}$$

Ainsi, la distance entre les sujets I1 et I2 est donnée par :

$$d(I_1,I_2) = \sqrt{(-1,1125 + 0,0056)^2 + (1,3473 + 0,7615)^2 + ... + (0,5535 + 0,9793)^2} = 4,2588$$

Le tableau des distances mutuelles entre sujets est ainsi donné par :

Dist. Euclidiennes (Basket-CR.sta)

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	I16	I17	I18
I1	0,00	4,26	4,47	0,71	1,43	1,59	2,53	3,21	3,36	4,48	3,30	3,40	3,75	4,74	4,75	4,89	4,99	4,78
I2	4,26	0,00	1,62	3,80	4,42	3,84	3,74	4,57	4,81	0,93	1,43	2,26	2,44	2,54	2,45	3,11	2,77	2,57
I3	4,47	1,62	0,00	3,93	4,66	4,02	4,55	5,52	5,76	1,87	1,31	2,65	2,16	2,30	2,41	2,92	2,72	1,25
I4	0,71	3,80	3,93	0,00	1,58	1,62	2,57	3,43	3,63	4,00	2,76	3,03	3,31	4,34	4,35	4,50	4,65	4,26
I5	1,43	4,42	4,66	1,58	0,00	0,92	2,47	3,19	3,12	4,66	3,54	2,86	3,29	4,25	4,24	4,20	4,45	4,73
I6	1,59	3,84	4,02	1,62	0,92	0,00	2,18	3,07	3,05	4,05	2,96	2,35	2,72	3,58	3,61	3,63	3,75	4,06
I7	2,53	3,74	4,55	2,57	2,47	2,18	0,00	1,36	1,53	3,72	3,39	2,99	3,83	4,33	4,21	4,42	4,33	5,01
I8	3,21	4,57	5,52	3,43	3,19	3,07	1,36	0,00	0,58	4,67	4,33	3,89	4,82	5,18	5,03	5,27	5,12	6,06
I9	3,36	4,81	5,76	3,63	3,12	3,05	1,53	0,58	0,00	4,92	4,58	3,91	4,86	5,21	5,05	5,23	5,11	6,21
I10	4,48	0,93	1,87	4,00	4,66	4,05	3,72	4,67	4,92	0,00	1,80	2,64	2,82	2,89	2,82	3,39	3,06	2,73
I11	3,30	1,43	1,31	2,76	3,54	2,96	3,39	4,33	4,58	1,80	0,00	1,92	1,89	2,42	2,42	2,90	2,81	2,12
I12	3,40	2,26	2,65	3,03	2,86	2,35	2,99	3,89	3,91	2,64	1,92	0,00	1,11	1,69	1,55	1,75	1,94	2,76
I13	3,75	2,44	2,16	3,31	3,29	2,72	3,83	4,82	4,86	2,82	1,89	1,11	0,00	1,27	1,38	1,47	1,75	1,86
I14	4,74	2,54	2,30	4,34	4,25	3,58	4,33	5,18	5,21	2,89	2,42	1,69	1,27	0,00	**0,43**	0,82	0,61	1,96
I15	4,75	2,45	2,41	4,35	4,24	3,61	4,21	5,03	5,05	2,82	2,42	1,55	1,38	**0,43**	0,00	0,76	0,71	2,24
I16	4,89	3,11	2,92	4,50	4,20	3,63	4,42	5,27	5,23	3,39	2,90	1,75	1,47	0,82	0,76	0,00	0,99	2,49
I17	4,99	2,77	2,72	4,65	4,45	3,75	4,33	5,12	5,11	3,06	2,81	1,94	1,75	0,61	0,71	0,99	0,00	2,42
I18	4,78	2,57	1,25	4,26	4,73	4,06	5,01	6,06	6,21	2,73	2,12	2,76	1,86	1,96	2,24	2,49	2,42	0,00

Le minimum non nul de ce tableau est 0,43 et correspond à la distance entre les sujets I14 et I15. Le premier groupe sera donc formé en réunissant ces deux sujets.

4.2.4 Choix d'un indice d'agrégation et algorithme de classification

Pour poursuivre la méthode, il faut maintenant faire le choix d'une "distance" entre groupes (ou entre un individu et un groupe). Choisissons, par exemple, l'indice d'agrégation défini par la méthode du "saut minimal". La distance D entre deux groupes A et B est alors définie par :

$$D(A,B) = \min_{I \in A} \min_{J \in B} d(I,J)$$

Le tableau des distances mutuelles entre les 17 objets restant après regroupement de I14 et I15 est donné par :

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14 I15	I16	I17	I18
I1	0	4,26	4,47	0,71	1,43	1,59	2,53	3,21	3,36	4,48	3,3	3,4	3,75	4,74	4,89	4,99	4,78
I2	4,26	0	1,62	3,8	4,42	3,84	3,74	4,57	4,81	0,93	1,43	2,26	2,44	2,45	3,11	2,77	2,57
I3	4,47	1,62	0	3,93	4,66	4,02	4,55	5,52	5,76	1,87	1,31	2,65	2,16	2,3	2,92	2,72	1,25
I4	0,71	3,8	3,93	0	1,58	1,62	2,57	3,43	3,63	4	2,76	3,03	3,31	4,34	4,5	4,65	4,26
I5	1,43	4,42	4,66	1,58	0	0,92	2,47	3,19	3,12	4,66	3,54	2,86	3,29	4,24	4,2	4,45	4,73
I6	1,59	3,84	4,02	1,62	0,92	0	2,18	3,07	3,05	4,05	2,96	2,35	2,72	3,58	3,63	3,75	4,06
I7	2,53	3,74	4,55	2,57	2,47	2,18	0	1,36	1,53	3,72	3,39	2,99	3,83	4,21	4,42	4,33	5,01
I8	3,21	4,57	5,52	3,43	3,19	3,07	1,36	0	**0,58**	4,67	4,33	3,89	4,82	5,03	5,27	5,12	6,06
I9	3,36	4,81	5,76	3,63	3,12	3,05	1,53	**0,58**	0	4,92	4,58	3,91	4,86	5,05	5,23	5,11	6,21
I10	4,48	0,93	1,87	4	4,66	4,05	3,72	4,67	4,92	0	1,8	2,64	2,82	2,82	3,39	3,06	2,73
I11	3,3	1,43	1,31	2,76	3,54	2,96	3,39	4,33	4,58	1,8	0	1,92	1,89	2,42	2,9	2,81	2,12
I12	3,4	2,26	2,65	3,03	2,86	2,35	2,99	3,89	3,91	2,64	1,92	0	1,11	1,55	1,75	1,94	2,76
I13	3,75	2,44	2,16	3,31	3,29	2,72	3,83	4,82	4,86	2,82	1,89	1,11	0	1,27	1,47	1,75	1,86
I14 I15	4,74	2,45	2,3	4,34	4,24	3,58	4,21	5,03	5,05	2,82	2,42	1,55	1,27	0	0,76	0,61	1,96
I16	4,89	3,11	2,92	4,5	4,2	3,63	4,42	5,27	5,23	3,39	2,9	1,75	1,47	0,76	0	0,99	2,49
I17	4,99	2,77	2,72	4,65	4,45	3,75	4,33	5,12	5,11	3,06	2,81	1,94	1,75	0,61	0,99	0	2,42
I18	4,78	2,57	1,25	4,26	4,73	4,06	5,01	6,06	6,21	2,73	2,12	2,76	1,86	1,96	2,49	2,42	0

Le minimum non nul de ce tableau est 0,58, distance entre les sujets I8 et I9. Ce sont donc ces deux individus qui seront regroupés à cette étape, ce qui conduit au tableau de distances suivant :

	I1	I2	I3	I4	I5	I6	I7	I8 I9	I10	I11	I12	I13	I14 I15	I16	I17	I18
I1	0	4,26	4,47	0,71	1,43	1,59	2,53	3,21	4,48	3,3	3,4	3,75	4,74	4,89	4,99	4,78
I2	4,26	0	1,62	3,8	4,42	3,84	3,74	4,57	0,93	1,43	2,26	2,44	2,45	3,11	2,77	2,57
I3	4,47	1,62	0	3,93	4,66	4,02	4,55	5,52	1,87	1,31	2,65	2,16	2,3	2,92	2,72	1,25
I4	0,71	3,8	3,93	0	1,58	1,62	2,57	3,43	4	2,76	3,03	3,31	4,34	4,5	4,65	4,26
I5	1,43	4,42	4,66	1,58	0	0,92	2,47	3,12	4,66	3,54	2,86	3,29	4,24	4,2	4,45	4,73
I6	1,59	3,84	4,02	1,62	0,92	0	2,18	3,05	4,05	2,96	2,35	2,72	3,58	3,63	3,75	4,06
I7	2,53	3,74	4,55	2,57	2,47	2,18	0	1,36	3,72	3,39	2,99	3,83	4,21	4,42	4,33	5,01
I8 I9	3,21	4,57	5,52	3,43	3,12	3,05	1,36	0	4,67	4,33	3,89	4,82	5,03	5,23	5,11	6,06
I10	4,48	0,93	1,87	4	4,66	4,05	3,72	4,67	0	1,8	2,64	2,82	2,82	3,39	3,06	2,73
I11	3,3	1,43	1,31	2,76	3,54	2,96	3,39	4,33	1,8	0	1,92	1,89	2,42	2,9	2,81	2,12
I12	3,4	2,26	2,65	3,03	2,86	2,35	2,99	3,89	2,64	1,92	0	1,11	1,55	1,75	1,94	2,76
I13	3,75	2,44	2,16	3,31	3,29	2,72	3,83	4,82	2,82	1,89	1,11	0	1,27	1,47	1,75	1,86
I14 I15	4,74	2,45	2,3	4,34	4,24	3,58	4,21	5,03	2,82	2,42	1,55	1,27	0	0,76	**0,61**	1,96
I16	4,89	3,11	2,92	4,5	4,2	3,63	4,42	5,23	3,39	2,9	1,75	1,47	0,76	0	0,99	2,49
I17	4,99	2,77	2,72	4,65	4,45	3,75	4,33	5,11	3,06	2,81	1,94	1,75	**0,61**	0,99	0	2,42
I18	4,78	2,57	1,25	4,26	4,73	4,06	5,01	6,06	2,73	2,12	2,76	1,86	1,96	2,49	2,42	0

Le minimum non nul de ce tableau est 0,61 et correspond à la distance entre le groupe {I14, I15} et le sujet I17. Un nouveau groupe, rassemblant I14, I15 et I17 est donc formé.

En poursuivant la méthode, on obtient la suite d'objets suivante :

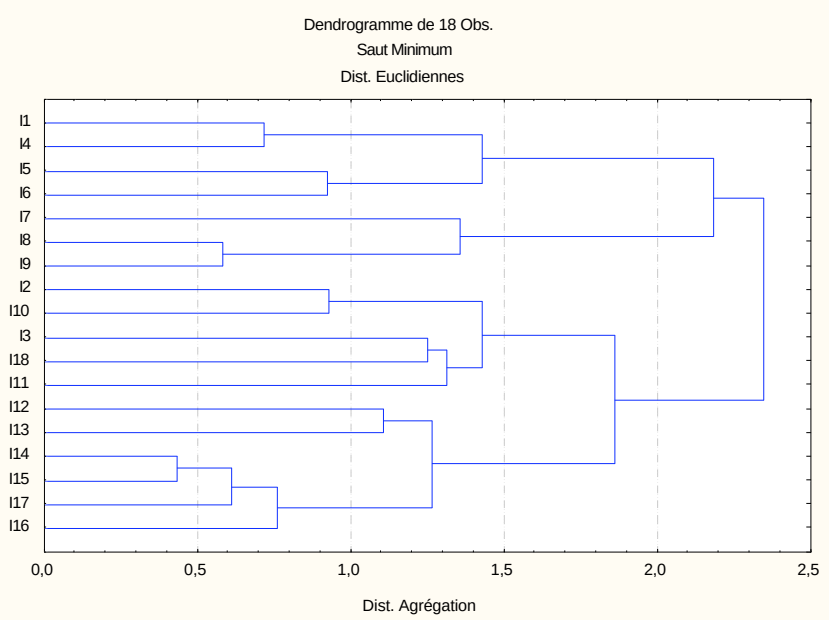
	Objet # 1	Objet # 2	Objet # 3	Objet # 4	Objet # 5	Objet # 6	Objet # 7	Objet # 8	Objet # 9	Objet # 10	Objet # 11	Objet # 12	Objet # 13	Objet # 14	Objet # 15	Objet # 16	Objet # 17	Objet # 18
,4341613	I14	I15																
,5828903	I8	I9																
,6121795	I14	I15	I17															
,7143260	I1	I4																
,7605890	I14	I15	I17	I16														
,9238485	I5	I6																
,9285629	I2	I10																
1,107561	I12	I13																
1,248607	I3	I18																
1,265891	I12	I13	I14	I15	I17	I16												
1,313007	I3	I18	I11															
1,357462	I7	I8	I9															
1,428591	I2	I10	I3	I18	I11													
1,429821	I1	I4	I5	I6														
1,860421	I2	I10	I3	I18	I11	I12	I13	I14	I15	I17	I16							
2,184427	I1	I4	I5	I6	I7	I8	I9											
2,346147	I1	I4	I5	I6	I7	I8	I9	I2	I10	I3	I18	I11	I12	I13	I14	I15	I17	I16

A l'inverse, on peut aussi indiquer les classes successives auxquelles appartiennent les objets élémentaires. Les classes sont ici numérotées de C1 à C35 ; C1 à C18 désignent les classes formées d'un

seul objet, C35 désigne la classe finale, formée de tous les objets. Chaque colonne du tableau correspond alors à une partition.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
I1	C1	C1	C1	C1	C22	C22	C22	C22	C22	C22	C22	C22	C22	C22	C32	C32	C34	C35
I2	C2	C2	C2	C2	C2	C2	C2	C25	C25	C25	C25	C25	C25	C31	C31	C33	C33	C35
I3	C3	C3	C3	C3	C3	C3	C3	C3	C3	C27	C27	C29	C29	C31	C31	C33	C33	C35
I4	C4	C4	C4	C4	C22	C22	C22	C22	C22	C22	C22	C22	C22	C22	C32	C32	C34	C35
I5	C5	C5	C5	C5	C5	C5	C24	C24	C24	C24	C24	C24	C24	C24	C32	C32	C34	C35
I6	C6	C6	C6	C6	C6	C6	C24	C24	C24	C24	C24	C24	C24	C24	C32	C32	C34	C35
I7	C7	C7	C7	C7	C7	C7	C7	C7	C7	C7	C7	C7	C30	C30	C30	C30	C34	C35
I8	C8	C8	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C30	C30	C30	C30	C34	C35
I9	C9	C9	C20	C20	C20	C20	C20	C20	C20	C20	C20	C20	C30	C30	C30	C30	C34	C35
I10	C10	C10	C10	C10	C10	C10	C10	C25	C25	C25	C25	C25	C25	C31	C31	C33	C33	C35
I11	C11	C11	C11	C11	C11	C11	C11	C11	C11	C11	C11	C29	C29	C31	C31	C33	C33	C35
I12	C12	C12	C12	C12	C12	C12	C12	C12	C26	C26	C28	C28	C28	C28	C28	C33	C33	C35
I13	C13	C13	C13	C13	C13	C13	C13	C13	C26	C26	C28	C28	C28	C28	C28	C33	C33	C35
I14	C14	C19	C19	C21	C21	C23	C23	C23	C23	C23	C28	C28	C28	C28	C28	C33	C33	C35
I15	C15	C19	C19	C21	C21	C23	C23	C23	C23	C23	C28	C28	C28	C28	C28	C33	C33	C35
I16	C16	C16	C16	C16	C16	C23	C23	C23	C23	C23	C28	C28	C28	C28	C28	C33	C33	C35
I17	C17	C17	C17	C21	C21	C23	C23	C23	C23	C23	C28	C28	C28	C28	C28	C33	C33	C35
I18	C18	C18	C18	C18	C18	C18	C18	C18	C18	C27	C27	C29	C29	C31	C31	C33	C33	C35

Le résultat est cependant plus lisible lorsqu'il est représenté sous forme de dendrogramme :



Le dendrogramme nous donne la composition des différentes classes, ainsi que l'ordre dans lequel elles ont été formées. Il nous indique également, sur l'axe horizontal, quelle était la valeur de l'indice entre les deux classes qui ont été agrégées à une étape donnée. Sur l'exemple proposé, on voit un "saut" de l'indice entre la partition en 4 classes et les partitions en 3, 2, 1 classes. Il pourrait être intéressant d'étudier plus précisément cette partition en 4 classes.

4.2.5 Lecture du résultat

On choisit par exemple de conserver 4 classes. La partition correspondante est alors :

Classe 1 = {I1, I4, I5, I6}
Classe 2 = {I7, I8, I9}
Classe 3 = {I2, I10, I3, I18, I11}
Classe 4 = {I12, I13, I14, I15, I17, I16}

4.3 Les 4 étapes de la méthode

4.3.1 Choix des variables représentant les individus

Dans le cas où les données observées sont les valeurs de p variables numériques sur n individus, on pourra choisir d'effectuer une classification des individus, ou une classification des variables. On peut choisir, par exemple, de retenir certains "traits" des individus (autrement dit certaines variables qui ont servi à les décrire) et réaliser la classification sur les individus décrits par ce choix de variables.
On peut noter qu'il revient au même par exemple :

- de réaliser la CAH des individus à partir de p variables centrées réduites ;
- de réaliser la CAH des individus à partir des p facteurs obtenus à l'aide d'une ACP normée sur les variables précédentes.

Toutefois, il peut être intéressant de réaliser la CAH à partir des q premiers facteurs (q<p). Cela a pour effet d'éliminer une partie des variations entre individus, qui correspond en général à des fluctuations aléatoires, c'est-à-dire à un "bruit statistique".

Dans le cas où les données observées sont représentées par un tableau de contingence, c'est-à-dire sont les valeurs de 2 variables nominales sur n individus, on pourra effectuer une CAH des modalités-lignes par exemple, à partir des coordonnées lignes obtenues par une AFC. On pourra, de même, réaliser une CAH des modalités-colonnes.

Enfin, si les données observées sont les valeurs de p variables nominales sur n individus, on pourra effectuer une CAH des individus en partant du tableau disjonctif complet, ou en utilisant les coordonnées des individus obtenues par une ACM. On pourra également traiter les modalités comme dans le cas d'une AFC.

4.3.2 Choix d'un indice de dissimilarité

De nombreuses mesures de la "distance" entre individus ont été proposées. Le choix d'une (ou plusieurs) d'entre elles dépend des données étudiées. Statistica nous propose les mesures suivantes :

- Distance Euclidienne. C'est probablement le type de distance le plus couramment utilisé. Il s'agit simplement d'une distance géométrique dans un espace multidimensionnel.
$$d(I_i,I_j) = \sqrt{\sum_k (x_{ik} - x_{jk})^2}$$
- Distance Euclidienne au carré. On peut élever la distance euclidienne standard au carré afin de "sur-pondérer" les objets atypiques (éloignés).
$$d(I_i,I_j) = \sum_k (x_{ik} - x_{jk})^2$$
- Distance du City-block (Manhattan) :
$$d(I_i,I_j) = \sum_k \left| x_{ik} - x_{jk} \right|$$
- Distance de Tchebychev :
$$d(I_i,I_j) = Max \left| x_{ik} - x_{jk} \right|$$
- Distance à la puissance.
$$d(I_i,I_j) = \left(\sum_k \left| x_{ik} - x_{jk} \right|^p \right)^{1/r}$$
- Percent disagreement. Cette mesure est particulièrement utile si les données des dimensions utilisées dans l'analyse sont de nature catégorielle.
$$d(I_i,I_j) = \frac{Nombre \;\; de \;\; x_{ik} \neq x_{jk}}{K}$$

- 1- r de Pearson : calculée à partir du coefficient de corrélation, à l'aide de la formule :

$$d(I_i, I_j) = 1 - r_{ij}$$

4.3.3 Choix d'un indice d'agrégation

L'application de la méthode suppose également que nous fassions le choix d'une "distance" entre classes. Là encore, de nombreuses solutions existent. Il faut noter que ces solutions permettent toutes de calculer la distance entre deux classes quelconques sans avoir à recalculer celles qui existent entre les individus composant chaque classe.

Les choix proposés par Statistica sont les suivants :

- Saut minimum ou "single linkage" (distance minimum). C'est celle que nous avons utilisée ci-dessus :

$$D(A,B) = \min_{I \in A} \min_{J \in B} d(I,J)$$

- Diamètre ou "complete linkage" (distance maximum). Dans cette méthode, les distances entre classes sont déterminées par la plus grande distance existant entre deux objets de classes différentes (c'est-à-dire les "voisins les plus éloignés").

$$D(A,B) = \max_{I \in A} \max_{J \in B} d(I,J)$$

- Moyenne non pondérée des groupes associés. Ici, la distance entre deux classes est calculée comme la moyenne des distances entre tous les objets pris dans l'une et l'autre des deux classes différentes.

$$D(A,B) = \frac{1}{n_A n_B} \sum_{I \in A, J \in B} d(I,J)$$

- Moyenne pondérée des groupes associés. La moyenne précédente est étendue à l'ensemble des paires d'objets trouvées dans la réunion des deux classes.

$$D(A,B) = \frac{1}{(n_A + n_B)(n_A + n_B - 1)} \sum_{I,J \in A \cup B} d(I,J)$$

- Centroïde non pondéré des groupes associés. Le centroïde d'une classe est le point moyen d'un espace multidimensionnel, défini par les dimensions. Dans cette méthode, la distance entre deux classes est déterminée par la distance entre les centroïdes respectifs.

- Centroïde pondéré des groupes associés (médiane). Cette méthode est identique à la précédente, à la différence près qu'une pondération est introduite dans les calculs afin de prendre en compte les tailles des classes (c'est-à-dire le nombre d'objets contenu dans chacune).

- Méthode de Ward (méthode du moment d'ordre 2). Cette méthode se distingue de toutes les autres en ce sens qu'elle utilise une analyse de la variance approchée afin d'évaluer les distances entre classes. En résumé, cette méthode tente de minimiser la somme des carrés (SC) de tous les couples (hypothétiques) de classes pouvant être formés à chaque étape. Les indices d'agrégation sont recalculés à chaque étape à l'aide de la règle suivante : si une classe M est obtenue en regroupant les classes K et L, sa distance à la classe J est donnée par :

$$D(M,J) = \frac{(N_J + N_K)D(K,J) + (N_J + N_L)D(L,J) - N_J D(K,L)}{N_J + N_K + N_L}$$

La méthode de Ward se justifie bien lorsque la "distance" entre les individus est le carré de la distance euclidienne. Choisir de regrouper les deux individus les plus proches revient alors à choisir la paire de points dont l'agrégation entraîne la diminution minimale de l'inertie du nuage. Le calcul des nouveaux

indices entre la paire regroupée et les points restants revient alors à remplacer les deux points formant la paire par leur point moyen, affecté du poids 2.

4.3.4 Quelle partie du dendrogramme faut-il conserver ?

Le dendrogramme nous indique l'ordre dans lequel les agrégations successives ont été opérées. Il nous indique également la valeur de l'indice d'agrégation à chaque niveau d'agrégation. Il est généralement pertinent d'effectuer la coupure après les agrégations correspondant à des valeurs peu élevées de l'indice et avant les agrégations correspondant à des valeurs élevées. En coupant l'arbre au niveau d'un saut important de cet indice, on peut espérer obtenir une partition de bonne qualité car les individus regroupés en-dessous de la coupure étaient proches, et ceux regroupés après la coupure sont éloignés.

4.4 La CAH avec Statistica

4.4.1 Classification à partir d'un tableau Individus x Variables Numériques

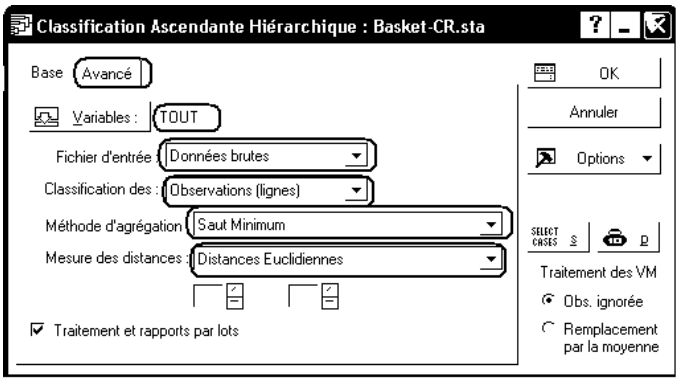
4.4.1.1 Classification des individus

On se propose de refaire la classification qui a été donnée comme exemple au paragraphe 1 (CAH sur les sujets, dans le cas "Basket").

Ouvrez le classeur Statistica Basket.stw et rendez active la feuille de données Basket-CR.sta, qui contient les variables centrées réduites correspondant au protocole observé.

Utilisez le menu Statitiques - Techniques Exploratoires Multivariées - Classifications et choisissez l'item "Classification Hiérarchique"

Sélectionnez ensuite l'onglet "Avancé", et complétez la fenêtre de dialogue comme suit :



En cochant la boîte "Traitement et rapports par lots", on obtient directement dans un classeur les trois principaux résultats, à savoir : le dendrogramme, le tableau intitulé "agrégation finale" et le tableau des distances entre objets.
Eventuellement, on pourra aussi afficher le graphique des distances d'agrégation par étape.

On pourra également essayer plusieurs distances, plusieurs indices d'agrégation et comparer les dendrogrammes obtenus.
Il est également intéressant de comparer les résultats de la classification à ceux de l'ACP, voire de représenter les classes sur le graphique de l'ACP.

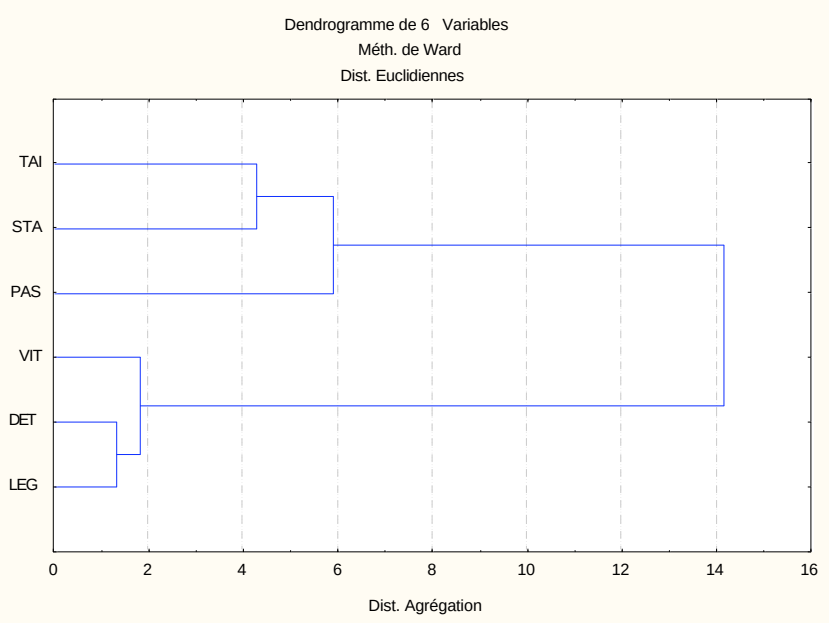
Remarques.

1. Contrairement à d'autres logiciels (Statgraphics par exemple), Statistica ne nous propose pas d'opérer le centrage-réduction des données avant de procéder à la classification. Il nous appartient donc de le faire nous-mêmes.

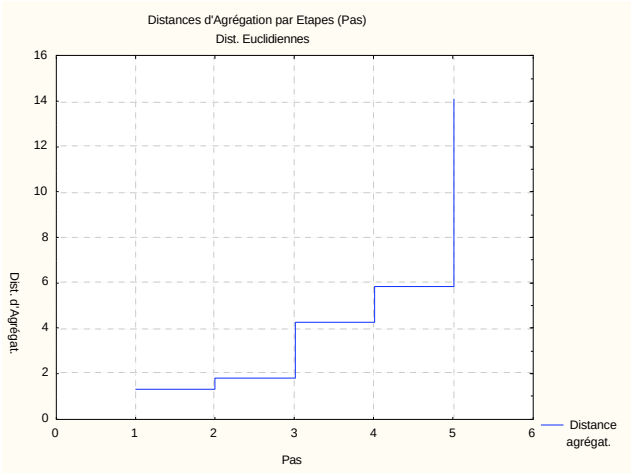
2. Le dendrogramme est souvent plus facile à lire lorsqu'on ne représente pas les dernières classes obtenues. Avec Statistica, il suffit de modifier l'échelle de l'axe horizontal pour parvenir à ce résultat.

4.4.1.2 Classification des variables

La classification peut porter aussi bien sur les individus que sur les variables. En modifiant la zone de dialogue "Classification des :...." de la fenêtre de dialogue ci-dessus, on peut réaliser une classification des 6 variables. Pour la distance euclidienne et l'agrégation selon la méthode de Ward, on obtient le dendrogramme suivant :



On voit sur ce diagramme, deux groupes bien distincts de variables : d'une part, Tai, Sta et Pas, d'autre part Vit, Det et Leg. Ce résultat rejoint celui qui avait été obtenu par l'AFC.
Les indices d'agrégation entre les classes qui ont été associées peuvent également être représentés par le graphique suivant :



Essayer plusieurs distances, plusieurs indices d'agrégation. Comparer les dendrogrammes obtenus. Comparer les résultats de la classification à ceux de l'ACP.

4.5 Indice d'agrégation et distances ultramétriques

Pour réaliser une CAH, nous devons faire le choix d'une distance entre les individus, et d'un indice d'agrégation mesurant la distance entre les classes. A chaque classe *H* est associé un nombre *v(H)* : la valeur de l'indice d'agrégation entre les deux objets qui ont été réunis pour former cette classe.

Par exemple, pour la classification des sujets dans le cas "Basket", l'indice correspondant à la classe *H1={I14, I15}* est *v(H1)=0,4342*, pendant que l'indice correspondant à la classe *H2={I14, I15,I16,I17}* est *v(H2)=0,7606* (cf. page 74).

Cet indice est croissant pour la relation d'inclusion : si une classe H est incluse dans une classe H', l'indice de H est inférieur à l'indice de H'. On dit que l'ensemble des classes forme une *hiérarchie indicée*.

L'indice *v(H)* nous fournit à son tour une nouvelle distance entre individus, définie par :

La distance δ(I,J) entre les individus I et J est l'indice correspondant à la plus petite classe contenant à la fois I et J.

Ainsi, sur l'exemple précédent :

$$\delta(I14,I15) = 0,4342 \quad ; \quad \delta(I14,I17)=0,7606.$$

Cette distance possède des propriétés mathématiques intéressantes. Elle vérifie une relation plus forte que l'inégalité triangulaire :

$$\delta(I_i,I_j) \leq Max(\delta(I_i,I_k),\ \delta(I_k,I_j))$$

Une distance vérifiant cette propriété est appelée *distance ultramétrique*.

Comme conséquence remarquable de cette propriété, on pourra noter que, dans un espace muni d'une distance ultramétrique, tout triangle est isocèle, la base étant le plus petit côté.
Ainsi, dans le triangle formé par les individus I14, I15 et I17 de l'exemple précédent, on a :

$$\delta(I14,I17) = \delta(I15,I17) = 0,7606 \quad ; \quad \delta(I14,I15)=0,4342$$

4.6 CAH sur les résultats d'une AFC

Réf. Examen de Statistiques de mai 2004, Module MULT, Maîtrise de Psychologie, Université René Descartes. Site Web : http://piaget.psycho.univ-paris5.fr/Statistiques/

Les données sont constituées par les résultats du premier tour des élections régionales de 2004 pour la région Ile de France. Pour chacun des huit départements de l'Ile de France (en lignes), on a les effectifs de suffrages pour chacune des huit listes candidates ainsi que les effectifs d'abstentions (en colonnes). Voici les codes de désignation des départements et des listes :

Départements	Code
Paris (75)	PARI
Seine et Marne (77)	SMAR
Yvelines (78)	YVEL
Essonne (91)	ESSO
Hauts de Seine (92)	HTSS
Seine Saint-Denis (93)	STDE
Val de Marne (94)	VDMA
Val d'Oise (95)	VDOI

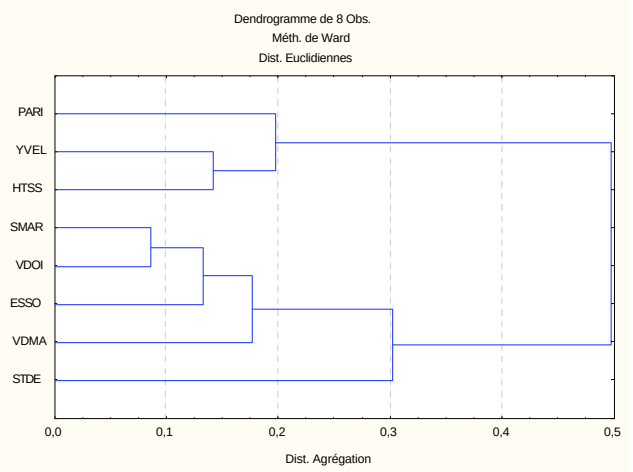
Listes	Tête de liste	Code
PS-Verts-MRG-MRC	Huchon	HUCH
UMP	Copé	COPE
UDF	Santini	SANT
FN	Le Pen	LEPE
PC-AGR-AC	Buffet	BUFF
LO-LCR	Laguiller	LAGU
GE-Les Bleus	Pelegrin	PELE
MNR	Bay	BAY
Abstentions		ABST

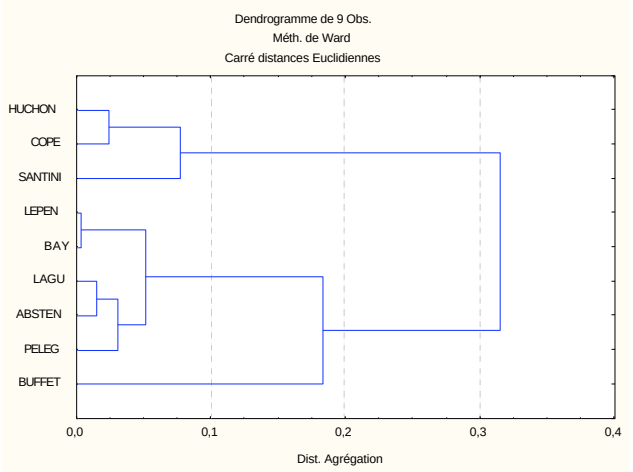
Chargez le classeur Statistica Regionales-2004-idf.stw et réalisez une AFC en calculant les coordonnées lignes et colonnes sur tous les facteurs et faites calculer à Statistica les coordonnées lignes et colonnes.

Rendez active la feuille "Coordonnées lignes". Utilisez ensuite le menu Statistiques - Techniques Exploratoires Multivariées - Classifications et réalisez ensuite une classification portant sur les variables nommées "coord.", en utilisant par exemple la distance euclidienne au carré et la méthode de Ward.

Procédez de même pour la feuille "Coordonnées colonnes".

Vous devriez parvenir à des dendrogrammes tels que :





4.7 CAH à partir d'indices de (dis)similarité

4.7.1 Indices de dissimilarité et distances

On peut également utiliser d'autres indices de dissimilarité puisque Statistica permet d'effectuer la classification à partir du tableau des scores de dissimilarités entre individus. En fait, un indice de dissimilarité doit simplement satisfaire les conditions suivantes :

- non-négativité : $d(I_i,I_j) \geq 0$
- symétrie : $d(I_i,I_j) = d(I_j,I_i)$
- normalisation : $d(I_i,I_i) = 0$

Un indice de dissimilarité est une "vraie" distance, s'il vérifie également l'inégalité triangulaire :

$$d(I_i,I_j) \leq d(I_i,I_k) + d(I_k,I_j).$$

La plupart des "distances" proposées par Statistica sont de véritables distances.

De nombreux indices de dissimilarité (ou au contraire de similarité) ont été proposés dans le cas de variables qualitatives (à deux modalités, ou après codage disjonctif). Par exemple, si les individus sont décrits par K variables dichotomiques (oui/non), on peut introduire :

$$a_{ij} = Nombre\ co-occurrences\ entre\ les\ individus\ i\ et\ j$$

$$d_{ij} = Nombre\ co-absences\ entre\ les\ individus\ i\ et\ j$$

$$b_{ij} = Nombre\ d'attributs\ présents\ chez\ i\ et\ absents\ chez\ j$$

$$c_{ij} = Nombre\ d'attributs\ absents\ chez\ i\ et\ présents\ chez\ j$$

On peut proposer par exemple, comme indice de dissimilarité :

$$d(I_i,I_j) = \sqrt{b_{ij} + c_{ij}}$$

ou au contraire, comme indice de similarité :

$$s(I_i,I_j) = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}$$

Un indice de similarité peut être converti en distance par la relation :

$$d(I_i,I_j) = s_{\max} - s(I_i,I_j)$$

4.7.2 Exemple

L'exemple qui suit est extrait de :
Doise W., Clemence A., Lorenzi-Cioldi F., Représentations Sociales et Analyses de Données, Presses Universitaires de Grenoble, 1992.

On demandait aux sujets interrogés d'indiquer *de quoi dépend la paie d'un travailleur*, en cochant la (ou les) réponse(s) qui correspondai(en)t le mieux à leur opinion. Les items proposés étaient les suivants : *de son rendement, de sa situation familiale, des responsabilités qu'il exerce, de sa formation, du coût de la vie, de son niveau hiérarchique, de son patron, de son ancienneté, de l'entreprise, du secteur où il travaille, de ses idées politiques.* Le nombre de répondants est égal à 181.

On donne ci-dessous le tableau des co-occurrences (nombre de sujets ayant accepté simultanément les deux items).

	Rend	Fami	Resp	Form	Coût	Hier	Patr	Anci	Sect	Idee	Univ
Rend	105	40	68	60	43	18	33	44	42	3	105
Fami	40	62	35	34	32	10	17	22	25	2	62
Resp	68	35	100	71	40	17	32	39	45	3	100
Form	60	34	71	99	38	18	36	39	49	2	99
Coût	43	32	40	38	68	7	23	24	26	1	68
Hier	18	10	17	18	7	25	11	14	17	1	25
Patr	33	17	32	36	23	11	56	20	27	3	56
Anci	44	22	39	39	24	14	20	55	33	3	55
Sect	42	25	45	49	26	17	27	33	71	3	71
Idee	3	2	3	2	1	1	3	3	3	3	3
Univarié	105	62	100	99	68	25	56	55	71	3	181

N.B. Les marges du tableau donnent le nombre de sujets ayant choisi l'item correspondant, pris isolément.

4.7.3 Exploration de divers indices de similarité entre les items.

Une première idée est de calculer la proportion que chaque co-occurrence représente par rapport au nombre total de répondants. Mais on s'aperçoit rapidement que cet indice donne trop de poids aux items les plus fréquemment cités. Les liens entre les items les moins cités sont donc sous-estimés.

4.7.3.1 L'indice de similarité de Jaccard

On décide de mesurer la similarité entre deux items à l'aide de l'indice dit de Jaccard :

$$s(I,J) = \frac{nombre\ \ de\ \ co-occurrences}{nombre\ \ de\ \ choix\ \ de\ \ I\ \ ou\ \ de\ \ J}$$

Ouvrez dans Excel le fichier Determinants-salaire.xls et calculez l'indice de Jaccard entre les items dans la plage B17:K26.

N.B. On pourra utiliser en B17 la formule : =B3/(B$13+$L3-B3). Réfléchissez aux différents éléments de cette formule avant de l'utiliser.

Calculez ensuite l'indice de dissimilarité *d(I,J)=1-s(I,J)* dans la plage B30:K39.

Chargez ensuite la plage de cellules A29:K39 dans une feuille de données Statistica.

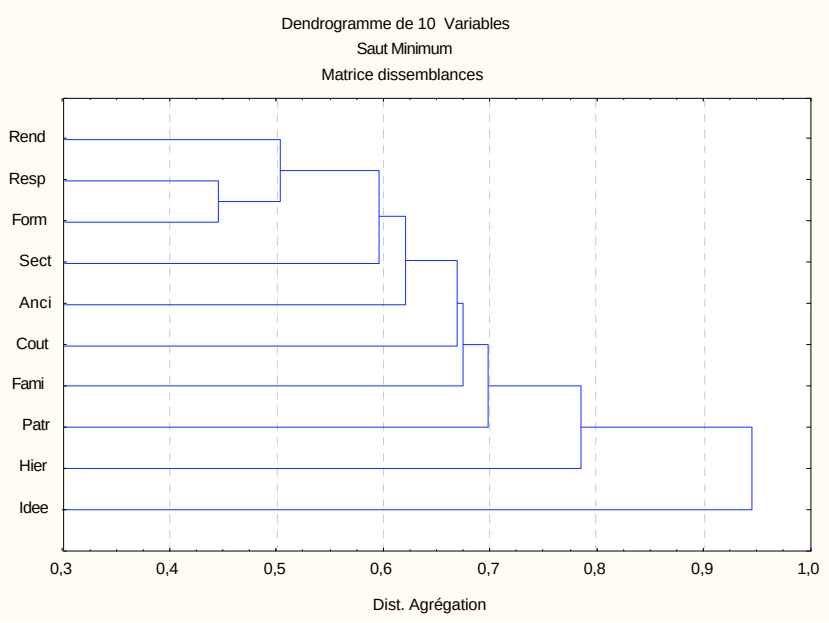
Pour que Statistica accepte ce fichier comme "matrice de dissimilarités", il faut ajouter les 4 observations suivantes après les 10 observations existantes (la première colonne est celle des noms d'observations) :

Moyennes										
Ec-Types										
Nb Obs.										
Matrix	3									

Notez que c'est la dernière ligne qui est la plus importante. Le nom d'observation "matrix" est reconnu comme mot clé, et francisé en "matrice" ; le code 3 indique à Statistica qu'il s'agit d'une matrice de dissimilarités. Lorsqu'on essaiera d'enregistrer ce fichier, Statistica proposera l'extension .smx, caractéristique des fichiers de matrices du logiciel.

Remarque : Statistica utilise les codes suivants pour les fichiers de matrices :
1 : matrice de corrélations
2 : matrice de similarités
3 : matrice de dissimilarités
4 : matrice de covariances.

Utiliser ces données comme matrice de distances pour effectuer une CAH, en utilisant la méthode d'agrégation du saut minimal. Vous devriez aboutir au dendrogramme suivant :



Le résultat obtenu est plutôt décevant. A chaque étape, c'est un élément supplémentaire qui s'ajoute à la classe déjà formée, et aucune "coupure" du dendrogramme ne semble satisfaisante. Ce résultat est dû à deux effets qui s'additionnent :
- l'indice de similarité retenu donne un poids important aux fréquences des items, pris individuellement
- L'agrégation par la méthode du saut minimal induit un effet de chaînage.

4.7.3.2 L'indice de similarité pairé

On choisit de mesurer la similarité entre deux items à l'aide de la formule :

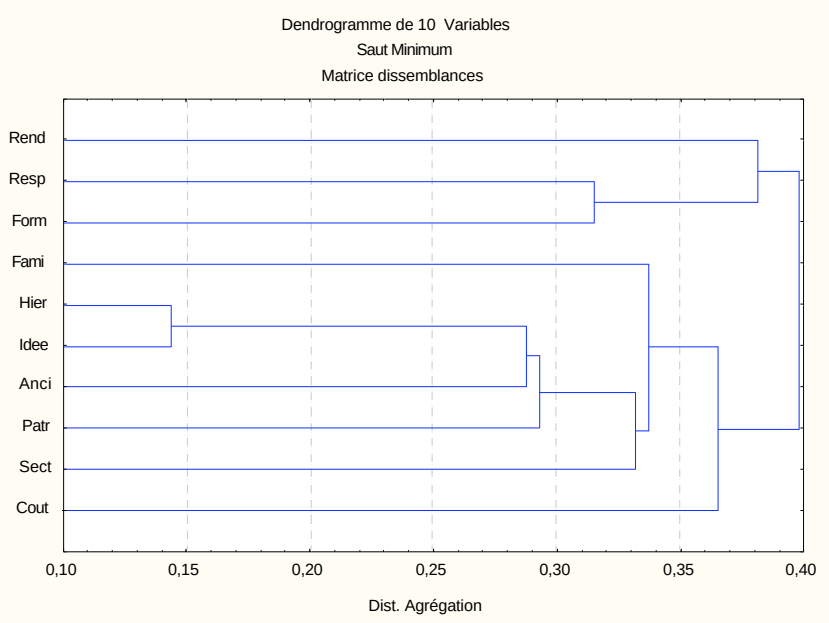
$$s'(I,J) = \frac{nombre\ \ de\ \ co-occurrences + nombre\ \ de\ \ co-absences}{nombre\ \ de\ \ répondants}$$

Calculer le nombre de co-absences (nombre de sujets n'ayant choisi ni l'un ni l'autre des deux items) dans la plage O3:X12 de la feuille Excel. On pourra pour cela indiquer en O3 la formule : =$L$13-B$13-$L3+B3, et réfléchir à la signification de cette formule...

Calculer l'indice de similarité s' dans la plage O17:X26.

Calculez ensuite l'indice de dissimilarité *d'(I,J)=1-s'(I,J)* dans la plage O30:X39.

Après avoir importé la plage N29:X39 dans une feuille de données Statistica, réalisez comme précédemment une CAH, en utilisant la méthode d'agrégation du saut minimal. Vous devriez aboutir au dendrogramme suivant :



Le résultat est sensiblement différent du précédent.

Réalisez ensuite la CAH à partir du tableau des dissimilarités d', mais en utilisant la méthode d'agrégation du diamètre. Comparer avec le résultat précédent.

4.7.3.3 Khi2 d'association, khi-2 signé, Phi de contingence

On peut également construire, pour chaque paire d'items, un tableau de contingence croisant la présence (codée 1) ou l'absence (codée 0) de chacun d'eux dans les réponses des sujets. Par exemple, pour rendement et famille, on obtient :

		Fami		
		1	0	Total
Rend	1	40	65	105
	0	22	54	76
	Total	62	119	181

On peut alors mesurer le lien entre les deux items en calculant un khi-2 sur ce tableau de contingence. Les effectifs théoriques sous hypothèse d'indépendance sont donnés par :

		Fami	
		1	0
Rend	1	36	69
	0	26	50

et la valeur du khi-2 observé est 1,64.

Certains auteurs désignent cette mesure sous le nom de **khi-2 d'association**, et affectent ce khi-2 du signe de la différence (co-occurrences observées - co-occurrences théoriques). Sur l'exemple précédent, la valeur du khi-2 est ainsi affectée du signe +.

Il est ainsi possible de faire un test statistique relatif à l'association entre deux items. Le khi-2 présente l'inconvénient d'être lié à l'effectif de l'échantillon considéré, et on lui préfère souvent le phi de contingence, mesure de similarité liée au khi-2 d'association par :

$$\Phi^2 = \frac{\chi^2}{N}$$

La valeur du phi de contingence peut aussi être calculée directement. Avec les notations suivantes pour le tableau de contingence des présences / absences :

	1	0	Total
1	a	b	n_3
0	c	d	n_4
Total	n_1	n_2	N

on a :

$$\Phi = \frac{ad - bc}{\sqrt{n_1 n_2 n_3 n_4}}$$

La valeur $1 - \Phi$ est alors une mesure de dissimilarité à partir de laquelle on peut faire une classification.

4.7.3.4 Exemple

On reprend le fichier Determinants-Salaires.xls. Activez la feuille "Phi de Contingence".

Calculez à partir de O17 les co-absences des différents items. La formule correspond à : Nombre total de répondants - Nombre d'occurrences de l'item colonne - Nombre d'occurrences de l'item ligne + Nombre de co-occurrences.
Formule en O17 : =$L$13-B$13-$L3+B3

Calculez à partir de O3 les présences de l'item ligne associées à des absences de l'item colonne. La formule correspond à : Nombre d'occurrences de l'item ligne - Nombre de co-occurrences
Formule en O3 : =$L3-B3

Calculez à partir de B17 les présences de l'item colonne associées à des absences de l'item ligne. La formule correspond à : Nombre d'occurrences de l'item colonne - Nombre de co-occurrences.
Formule en B17 : =B$13-B3

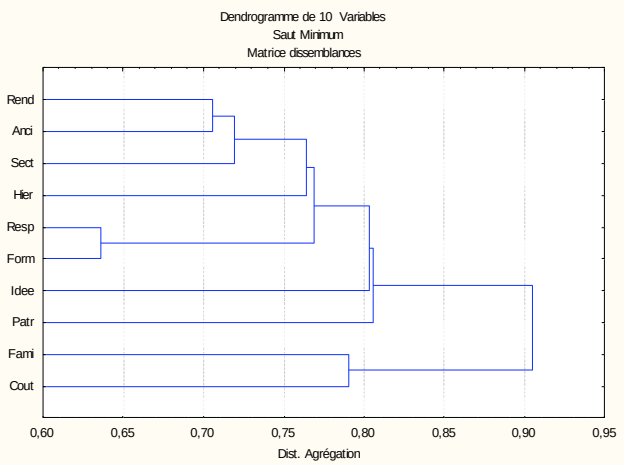
Calculez les valeurs du Phi de contingence à partir de B30. Les présences n_1 et n_3 des items ligne et colonne sont reprises du tableau de données, les absences n_2 et n_4 sont calculées par des formules du type : Nombre de répondants - Nombre de présences.
Formule en B30 : =(B3*O17-O3*B17)/RACINE(B$13*($L$13-B$13)*$L3*($L$13-$L3))

Enfin, calculez la mesure de dissimilarité à partir de O30
Formule en O30 : = 1 - B30.

Vous devriez obtenir :

	Rend	Fami	Resp	Form	Cout	Hier	Patr	Anci	Sect	Idee
Rend	0,00	0,90	0,78	0,94	0,92	0,89	0,99	0,71	0,98	0,89
Fami	0,90	0,00	0,98	1,00	0,79	0,95	1,05	0,92	0,98	0,91
Resp	0,78	0,98	0,00	0,64	0,94	0,90	0,97	0,79	0,87	0,88
Form	0,94	1,00	0,64	0,00	0,98	0,86	0,87	0,78	0,77	0,97
Cout	0,92	0,79	0,94	0,98	0,00	1,08	0,95	0,92	1,02	1,01
Hier	0,89	0,95	0,90	0,86	1,08	0,00	0,89	0,78	0,76	0,93
Patr	0,99	1,05	0,97	0,87	0,95	0,89	0,00	0,92	0,88	0,81
Anci	0,71	0,92	0,79	0,78	0,92	0,78	0,92	0,00	0,72	0,80
Sect	0,98	0,98	0,87	0,77	1,02	0,76	0,88	0,72	0,00	0,84
Idee	0,89	0,91	0,88	0,97	1,01	0,93	0,81	0,80	0,84	0,00

Importez ensuite cette matrice de dissimilarités dans Statistica et réalisez une CAH sur ces données. On obtient par exemple :



4.8 Exercice à rendre

Source : Pierre Vergès et Boumedienne Bouriche, L'analyse des données par les graphes de similitude, Sciences Humaines, Juin 2001.

Exemple tiré de : Enquête Eric Tafani , 1999, Laboratoire de Psychologie Sociale de l'Université de Provence ; et Beauvois, L., (ed) La construction sociale de la personne vol.4, P.U.G.

On mène une enquête sur les "valeurs" à partir d'un questionnaire de Schwartz passé auprès de 268 sujets. Les réponses au questionnaire ont permis de construire une série de scores pour chaque sujet : chaque score reflète l'opinion d'un sujet à propos d'une valeur. Cette méthode identifie dix valeurs :

B-Accomplissement

A-Pouvoir
K-Sécurité
J-Conformisme
H-Tradition
G-Bienveillance
F-Universalisme
D-Stimulation
E-Centration sur soi
C-Hédonisme

Ouvrez le fichier Valeurs-Schwartz.stw.

N.B. Les données présentées dans ce fichier ont été générées artificiellement afin de retrouver la matrice de corrélation indiquée dans la publication citée supra.

On mesure ici la similarité entre deux valeurs à l'aide du coefficient de corrélation entre les variables. Réalisez une CAH sur les variables en utilisant la "distance" 1 - r de Pearson et la distance d'agrégation "saut minimum".

Donnez la composition des classes dans le cas d'une partition en 3 classes, puis en 2 classes.

Générer la matrice des corrélations entre variables et commentez la classification obtenue en utilisant les valeurs de ces corrélations.

Travail à rendre par mail à votre enseignant (francois.carpentier@univ-brest.fr) :
- Un classeur Statistica contenant les résultats numériques et les graphiques.
- Un fichier Word ou un rapport Statistica contenant votre interprétation des résultats.