5.5 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 :

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

5.6 CAH à partir d'indices de similarité

5.6.1 Enoncé

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.

5.6.2 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.

5.6.2.1 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.

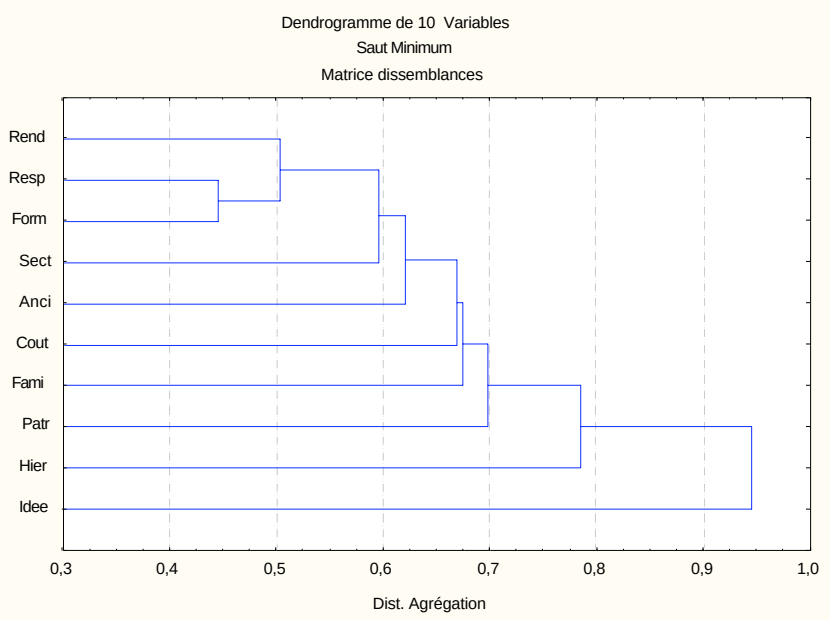
Chargez ensuite la place de cellules A29:K39 dans une feuille de données Statistica.

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.

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.

5.6.2.2 L'indice de similarité pairé

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

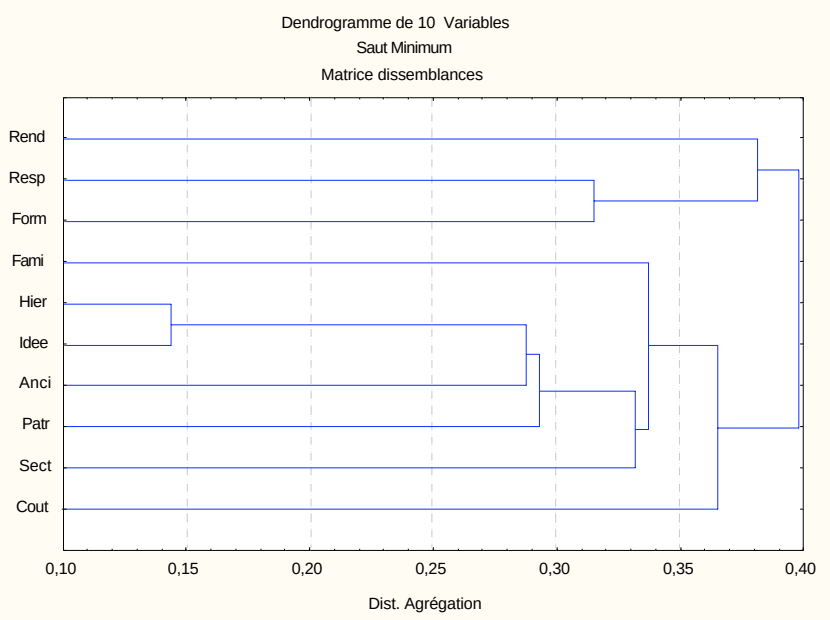
$$s'(I,J) = \frac{nombre\ \ de\ \ co-occurrences + nombre\ \ de\ \ co-absences}{nombre\ \ de\ \ répondants}$$

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 : =L13-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 :



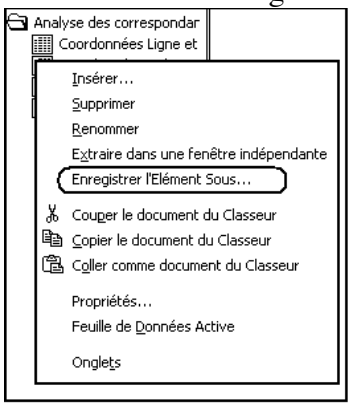
Le résultat est sensiblement différent du précédent.

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.

5.7 CAH sur les résultats d'une AFC

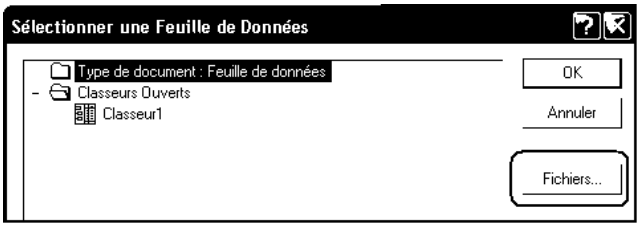
On reprend l'exemple "Elections régionales 2004 en Ile-de-France. Ouvrez le fichier idf.sta et réalisez une AFC en calculant les coordonnées lignes et colonnes sur tous les facteurs.

Enregistrer les deux tableaux de coordonnées comme feuilles de données Statistica (bouton droit sur l'icône du document dans l'onglet du classeur et menu "Enregistrer l'élément sous...")

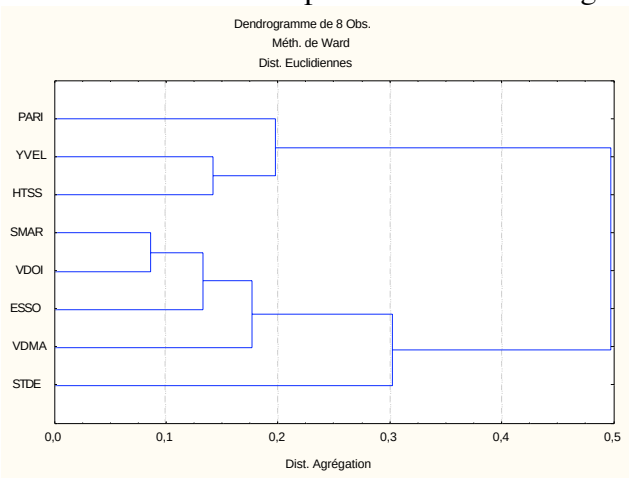


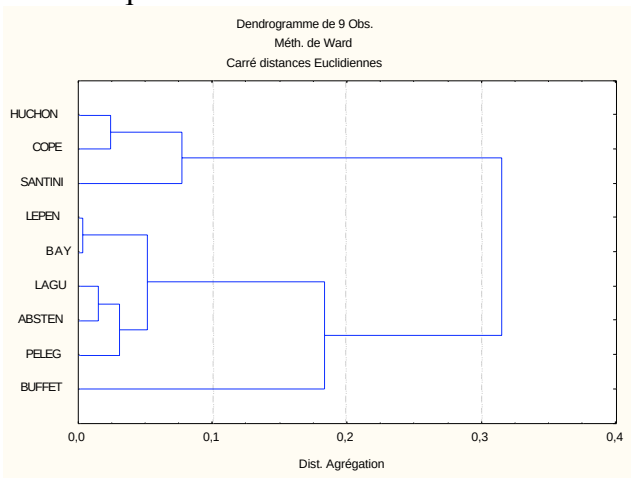
Refermez la feuille de données idf.sta

Utilisez ensuite le menu Statistiques - Techniques Exploratoires Multivariées - Classifications . Lorsque Statistica affiche la fenêtre de dialogue "Sélectionner une feuille de données", cliquer sur le bouton "Fichiers" et sélectionner l'un des fichiers de données précédents :



Réalisez ensuite la classification, en utilisant par exemple la distance euclidienne au carré et la méthode de Ward. Vous devriez parvenir à des dendrogrammes tels que :





5.8 Exercice à rendre

Reprendre le cas "Budget-temps Multimédia" (exercice à rendre pp. 22 et 23 du polycopié). Réalisez une CAH sur les individus en utilisant les 16 variables actives, sans centrage ni réduction. Faites en sorte que le dendrogramme soit étiqueté par les identificateurs des individus, et commentez le résultat obtenu en essayant de déterminer, parmi les variables nominales Age, Niveau d'éducation, Type d'agglomération, celle qui correspond le mieux à la classification obtenue.

Travail à rendre par mail à votre enseignant (carpenti@infolettres.univ-brest.fr si vous travaillez dans les salles de TD, Francois.Carpentier@univ-brest.fr si vous travaillez à l'extérieur) :

- Un classeur Statistica contenant les résultats numériques de l'ACP et les graphiques.
- Un fichier Word contenant votre interprétation des résultats.