DISTANCE DU KHI 2 ET ALGORITHMES DE CLASSIFICATION HIERARCHIQUE

Georges Laplace

On se propose de classer selon leurs ressemblances :

a - soit des séries ou ensembles industriels considérés à un niveau structural donné ;
b - soit les catégories composantes du niveau structural considéré dans les séries ou ensembles industriels donnés.

La distance choisie entre les séries ou entre les catégories est appelée "distance du Khi 2". Deux algorithmes, ou procédés de calcul, de classification hiérarchique indicée sont proposés : l'un consistant à rendre ultramétrique la distance d, l'autre procédant par réduction.

A. DISTANCE DU KHI 2.

1. Formules de la distance du Khi 2.

Soit le tableau de contingence r x k, c'est-à-dire comportant r lignes et k colonnes :

		Catégories			
		j	j'		
Séries	i	n_{ij}	$n_{ij'}$		$n_{i.}$
	i'	$n_{i'j}$	$n_{i'j'}$		$n_{i'.}$
		$n_{.j}$	$n_{.j'}$		$n_{..}$

(A.1.1)

Dans ce tableau de données brutes, comportant k catégories et r séries :

a) n_{ij} exprime le nombre d'observations présentant à la fois la modalité i de la variable "séries" et la modalité j de la variable "catégories" ;

b) $n_{i.}$ représente l'effectif de la série i, c'est-à-dire la somme le long de i, j variant de 1 à k, soit :

$$n_{i.} = \sum_{j=1}^{k} n_{ij}$$

c) $n_{.j}$ représente l'effectif de la catégorie j, c'est-à-dire la somme le long de j, i variant de 1 à r, soit :

$$n_{.j} = \sum_{i=1}^{r} n_{ij}$$

d) $n_{..}$ représente l'effectif total, soit :

$$n_{..} = \sum_{j=1}^{k} \sum_{i=1}^{r} n_{ij}$$

A partir de ce tableau on construit le tableau des fréquences :

Catégories

Séries		j	j'		
i		f_{ij}	$f_{ij'}$		$f_{i.}$
i'		$f_{i'j}$	$f_{i'j'}$		$f_{i'.}$
		$f_{.j}$	$f_{.j'}$		1

(A.1.2)

Le fait d'avoir choisi des "profils", ou fréquences conditionnelles, conduit à adopter une distance différente de la distance euclidienne usuelle, laquelle se calcule comme suit :

$$d^2(i, i') = \sum_{j=1}^{k} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}} \right)^2$$

ou

$$d^2(i, i') = \sum_{j=1}^{k} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}} \right)^2$$

car $f_{ij} = \frac{n_{ij}}{n_{..}}$, $f_{i'j} = \frac{n_{i'j}}{n_{..}}$, $f_{i.} = \frac{n_{i.}}{n_{..}}$ et $f_{i'.} = \frac{n_{i'.}}{n_{..}}$

et, de façon symétrique :

$$d^2(j, j') = \sum_{i=1}^{r} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}}\right)^2$$

ou $$d^2(j, j') = \sum_{i=1}^{r} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2$$

La distance du Khi 2 ne diffère en fait de la métrique euclidienne usuelle que par la pondération de chaque carré par les inverses des fréquences correspondant à chaque terme.

La distance du Khi 2 entre deux modalités sérielles i et i' sera donnée par les formules :

$$d^2(i, i') = \sum_{j=1}^{k} \frac{n_{..}}{n_{.j}} \left(\frac{n_{ij}}{n_{i.}} - \frac{n_{i'j}}{n_{i'.}}\right)^2 \qquad (A.1.3)$$

ou $$d^2(i, i') = \sum_{j=1}^{k} \frac{1}{f_{.j}} \left(\frac{f_{ij}}{f_{i.}} - \frac{f_{i'j}}{f_{i'.}}\right)^2 \qquad (A.1.4)$$

De façon symétrique, la distance du Khi 2 entre les deux modalités catégorielles j et j' vaut :

$$d^2(j, j') = \sum_{i=1}^{r} \frac{n_{..}}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}}\right)^2 \qquad (A.1.5)$$

ou $$d^2(j, j') = \sum_{i=1}^{r} \frac{1}{f_{i.}} \left(\frac{f_{ij}}{f_{.j}} - \frac{f_{ij'}}{f_{.j'}}\right)^2 \qquad (A.1.6)$$

2. Exemple d'application des formules de la distance du Khi 2.

Prenons pour exemple d'application de la formule de la distance du Khi 2 les effectifs des trois classes de Grattoirs (G.f = grattoirs frontaux ; G.m = grattoirs à museau ; G.c = grattoirs carénés) dans les séries ou ensembles industriels Gjn2, Cjnl, Cbf et Cb de la grotte Gatzarria. On construit le tableau de contingence (A.1.1) :

	G.f	G.m	G.c	
Cjn2	5	4	14	23
Cjnl	10	8	20	38
Cbf	33	21	61	115
Cb	41	27	72	140
	89	60	167	316

Si l'on se propose de calculer la distance entre les quatre séries ou ensembles industriels, on doit utiliser la formule (A.1.3). Par contre, si l'on avait voulu calculer les distances entre les trois catégories de grattoirs on aurait utilisé la formule symétrique (A.1.5).

Calculons la distance du Khi 2 entre Cjn2 et Cjnl :

$$d^2(Cjn2,Cjnl) = \frac{316}{89}\left(\frac{5}{23} - \frac{10}{38}\right)^2 + \frac{316}{60}\left(\frac{4}{23} - \frac{8}{38}\right)^2 + \frac{316}{167}\left(\frac{14}{23} - \frac{20}{38}\right)^2$$

$$= 3{,}551\ (.217 - .263)^2 + 5{,}267\ (.174 - .211)^2 + 1{,}892\ (.609 - .526)^2$$

$$= 0{,}028$$

d'où $d(Cjn2, Cjnl) = \sqrt{0{,}028} = 0{,}167$

Pour faciliter les calculs on dresse le tableau pertinent des fréquences utilisées dans la formule (A.1.3) :

		j	j'		
i		$\frac{n_{ij}}{n_{i\circ}}$	$\frac{n_{ij'}}{n_{i\circ}}$		1
i'		$\frac{n_{i'j}}{n_{i'\circ}}$	$\frac{n_{i'j'}}{n_{i'\circ}}$		1
1/f		$\frac{n_{\circ\circ}}{n_{\circ j}}$	$\frac{n_{\circ\circ}}{n_{\circ j'}}$		

et on obtient :

	G.f	G.m	G.c	
Cjn2	.217	.174	.609	1
Cjnl	.263	.211	.526	1
Cbf	.287	.183	.530	1
Cb	.293	.193	.514	1
1/f	3,551	5,267	1,892	

A partir de ce tableau on calcule itérativement les distances entre les quatre séries ou ensembles industriels :

$d^2(Cjn2, Cjnl) = 0{,}028$ d'où $d(Cjn2, Cjnl) = 0{,}167$
$d_2^2(Cjn2, Cbf) = 0{,}030$ d'où $d(Cjn2, Cbf) = 0{,}172$
$d_2^2(Cjn2, Cb) = 0{,}039$ d'où $d(Cjn2, Cb) = 0{,}199$
$d_2^2(Cjnl, Cbf) = 0{,}006$ d'où $d(Cjnl, Cbf) = 0{,}079$
$d_2^2(Cjnl, Cb) = 0{,}005$ d'où $d(Cjnl, Cb) = 0{,}072$
$d^2(Cbf, Cb) = 0{,}001$ d'où $d(Cbf, Cb) = 0{,}034$

B. ALGORITHME UTILISANT UNE DISTANCE ULTRAMETRIQUE.

Cet algorithme concerne le passage d'une matrice de distance à une matrice ultramétrique. Plusieurs procédés peuvent être utilisés. Nous en présentons trois :

a - celui de l'ultramétrique supérieure minimale ;
b - celui de l'ultramétrique inférieure maximale ;
c - celui de l'ultramétrique moyenne.

Soit E un ensemble de n objets, muni d'un indice de "distance" d, ici la distance du Khi 2. On aurait pu prendre un indice de similarité mais, pour rester homogène avec le sens de variation d'une distance, on a préféré considérer un indice qui varie en sens contraire d'une similarité. Un tel indice est appelé quelquefois indice de dissimilarité. Si d n'est pas ultramétrique, il n'induit pas une seule hiérarchie indicée. Le choix d'une formule pour calculer à chaque étape les dissimilarités entre parties de E s'impose. Quand on choisit une telle formule, on construit très simplement une hiérarchie indicée correspondante, c'est-à-dire un dendrogramme niveau par niveau. En haut, on place les n points de E. Au niveau 1, on agrège les deux points les plus proches et on calcule la dissimilarité des n - 2 autres points à cette partie à l'aide de la formule, et ainsi de suite. On s'arrête au niveau n - 1.

Remarque : cet algorithme est ascendant puisq'il procède par regroupement des séries ou des catégories.

1. Construction d'une ultramétrique supérieure minimale.

L'algorithme proposé revient à rendre isocèle tout triangle de l'indice initial en lui donnant pour base son plus petit côté et pour longueur des deux côtés égaux celle de son plus grand côté. On construit ainsi une hiérarchie indicée, et l'ultramétrique correspondante u_1 est "supérieure" à la dissimilarité initiale. On démontre que cette ultramétrique est minimale dans l'ensemble des ultramétriques supérieures à d.

On dispose les résultats obtenus par le calcul de la distance du Khi 2 de manière à former une matrice des distances entre les quatre séries ou ensembles industriels :

d	Cjn2	Cjnl	Cbf	Cb
Cjn2	0	0,167	0,177	0,199
Cjnl		0	0,079	0,072
Cbf			0	0,034
Cb				0

On constate que la distance la plus courte est de 0,034 entre

les séries Cb et Cbf. En conséquence, on agrège Cb et Cbf et on calcule, pour obtenir une nouvelle matrice des distances ou tableau de dissimilarité, les distances entre la nouvelle série (Cb,Cbf) et les séries Cjn2 et Cjnl.

Soit le triangle Cb-Cbf-Cjn2. On obtient les longueurs : 0,034 pour le côté Cb-Cbf, 0,199 pour le côté Cb-Cjn2 et 0,177 pour le côté Cbf-Cjn2. On choisira donc 0,034, longueur du plus petit côté, pour longueur de la base et 0,199, longueur du plus grand côté, pour longueur des deux côtés égaux. Ainsi, la distance ultramétrique entre (Cb,Cbf) et Cjn2 sera de 0,199.

De même, si l'on considère le triangle Cb-Cbf-Cjnl, on obtient les longueurs : 0,034 pour le côté Cb-Cbf, 0,072 pour le côté Cb-Cjnl et 0,079 pour le côté Cbf-Cjnl. Par conséquent, le triangle "isocèle pointu" aura une base de longueur 0,034 et deux côtés égaux de longueur 0,079. Ainsi, la distance ultramétrique entre (Cb,Cbf) et Cjnl sera de 0,079. La distance entre Cjn2 et Cjnl, soit 0,167, ne variant pas on construit sur l'ensemble de ces données la nouvelle matrice des distances:

1re étape	(Cb,Cbf)	Cjn2	Cjnl
(Cb,Cbf)	0	0,199	0,079
Cjn2		0	0,167
Cjnl			0

Procédant comme précédemment, on constate que la distance la plus courte est de 0,079 entre la série (Cb,Cbf) et la série Cjnl.

On agrège (Cb,Cbf) et Cjnl et on calcule, pour obtenir une nouvelle matrice des distances, la distance entre la nouvelle série (Cb,Cbf,Cjnl) et la série Cjn2.

On rend "isocèle pointu" le triangle (Cb,Cbf)-Cjnl-Cjn2 de côtés (Cb,Cbf)-Cjnl = 0,079, (Cb,Cbf)-Cjn2 = 0,199 et Cjnl-Cjn2 = 0,167. On en déduit la distance ultramétrique entre la série (Cb, Cbf,Cjnl) et la série Cjn2, soit 0,199. On construit la dernière matrice des distances :

2e étape	(Cb,Cbf, Cjnl)	Cjn2
(Cb,Cbf,Cjnl)	0	0,199
Cjn2		0

Nous passons donc de la matrice des distances à une matrice ultramétrique u_1 :

u_1	Cjn2	Cjn1	Cbf	Cb
Cjn2	0	0,199	0,199	0,199
Cjn1		0	0,079	0,079
Cbf			0	0,034
Cb				0

Remarque : on constate que toutes les distances figurant dans ce tableau sont "ultramétriques", c'est-à-dire qu'elles vérifient la condition dite de Bourbaki :

"Pour tout triplet $(X_h, X_i, X_e,)$ on a :

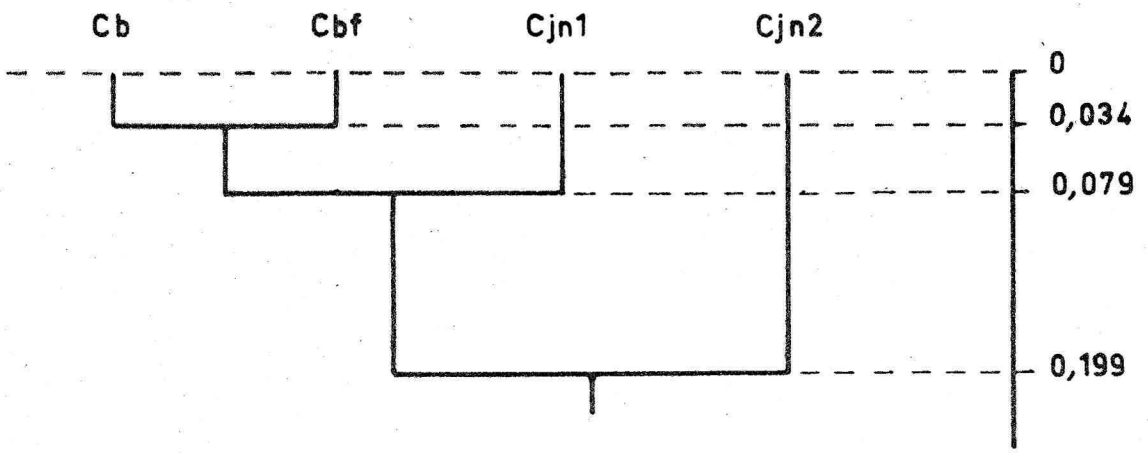
$$d(X_h, X_i) \leqslant \text{SUP}\ d(X_h, X_e), d(X_e, X_i)"$$

Chaque fois que l'on considère trois séries (ou catégories) la distance entre deux séries (ou catégories) est toujours inférieure ou égale à la plus grande des distances de ces deux séries (ou catégories) à la troisième, ce qui revient à dire que tous les triangles ayant pour sommets les séries (ou catégories) sont isocèles ou équilatéraux.

On peut maintenant dessiner le dendrogramme. Il se présente comme un arbre de classification associé à une échelle de distance ultramétrique et figure, de ce fait, une hiérarchie stratifiée indicée.

feuilles

Il est évident que deux séries (ou catégories) figurant les branches se ressemblent d'autant plus que le premier noeud qui les relie est plus élevé dans l'arbre.



2. Construction d'une ultramétrique inférieure maximale.

L'algorithme proposé revient à rendre isocèle tout triangle de l'indice initial en lui donnant pour base son plus petit côté et en réduisant le plus grand côté à la longueur du côté immédiatement inférieur. On obtient une hiérarchie différente de la précédente et l'ultramétrique correspondante u_2 n'est donc pas équivalente à u_1 : elle est inférieure à d et on démontre qu'elle est maximale dans l'ensemble des ultramétriques inférieures à d.

Reprenons la matrice des distances :

d	Cjn2	Cjnl	Cbf	Cb
Cjn2	0	0,167	0,177	0,199
Cjnl		0	0,079	0,072
Cbf			0	0,034
Cb				0

On obtient successivement les matrices des distances :

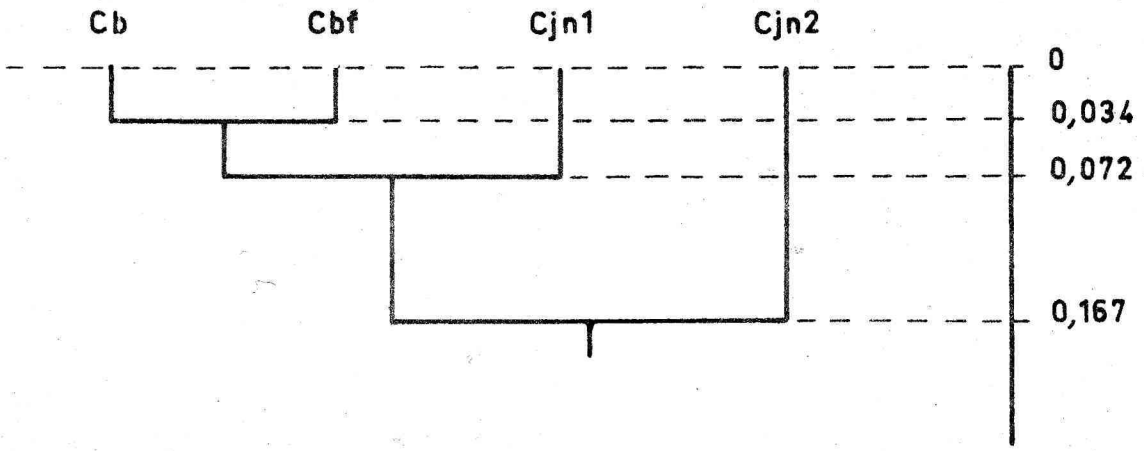
1re. étape	(Cb,Cbf)	Cjn2	Cjnl
(Cb,Cbf)	0	0,177	0,072
Cjn2		0	0,167
Cjnl			0

2e étape	(Cb,Cbf, Cjnl)	Cjn2
(Cb,Cbf,Cjnl)	0	0,167
Cjn2		0

la matrice ultramétrique :

u_2	Cjn2	Cjnl	Cbf	Cb
Cjn2	0	0,167	0,167	0,167
Cjnl		0	0,072	0,072
Cbf			0	0,034
Cb				0

et, enfin, le dendrogramme :



3. Construction d'une ultramétrique moyenne.

L'algorithme proposé revient à rendre isocèle tout triangle de l'indice initial en lui donnant pour base son plus petit côté et pour longueur des deux côtés égaux la moyenne des longueurs des deux autres côtés.

Reprenons la matrice des distances :

d	Cjn2	Cjnl	Cbf	Cb
Cjn2	0	0,167	0,177	0,199
Cjnl		0	0,079	0,072
Cbf			0	0,034
Cb				0

On obtient successivement les matrices des distances :

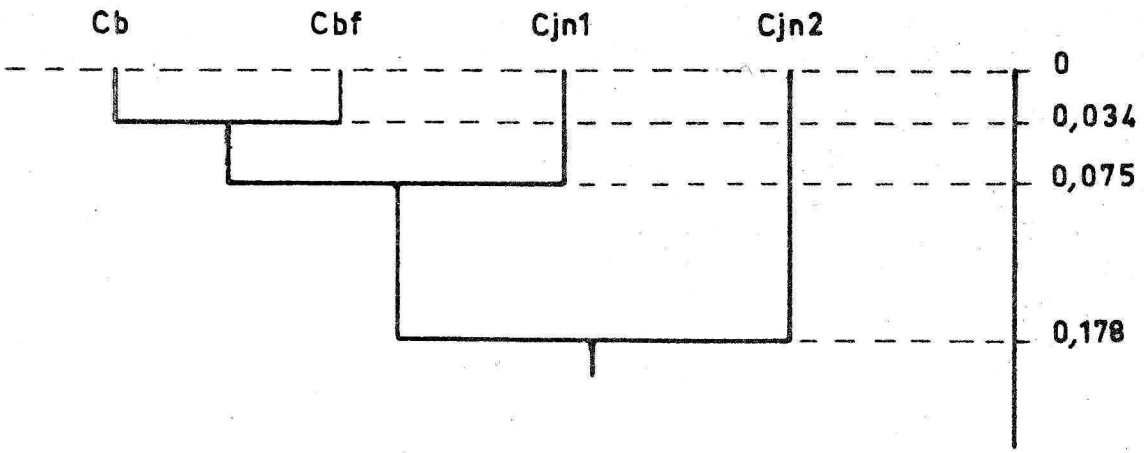
1re étape	(Cb,Cbf	Cjn2	Cjnl
(Cb,Cbf)	0	0,188	0,075
Cjn2		0	0,167
Cjnl			0

2e étape	(Cb,Cbf, Cjnl)	Cjn2
(Cb,Cbf,Cjnl)	0	0,178
Cjn2		0

la matrice ultramétrique :

u_3	Cjn2	Cjnl	Cbf	Cb
Cjn2	0	0,178	0,178	0,178
Cjnl		0	0,075	0,075
Cbf			0	0,034
Cb				0

et, enfin, le dendrogramme :



C. ALGORITHME PROCEDANT PAR REDUCTIONS.

Cet algorithme procède par regroupements successifs de façon à réduire progressivement les dimensions de la matrice de départ.

Remarque. Cet algorithme est ascendant puisqu'il procède par regroupement des séries ou des catégories.

1. L'algorithme.

La démarche est exposée à partir du tableau pertinent des fréquences (A.1.7) utilisées dans la formule (A.1.3) pour les séries ou ensembles industriels Cjn2, Cjnl, Cbf et Cb de la grotte Gatzarria considérés au niveau structural des classes de Grattoirs :

	G.f	G.m	G.c	
Cjn2	.217	.174	.609	1
Cjnl	.263	.211	.526	1
Cbf	.287	.183	.530	1
Cb	.293	.193	.514	1
1/f	3,551	5,267	1,892	

Les distances du Khi 2 calculées entre les quatre séries ou ensembles industriels sont reportées dans une matrice des distances ou tableau de dissimilarité :

d	Cjn2	Cjnl	Cbf	Cb
Cjn2	0	0,167	0,177	0,199
Cjnl		0	0,079	0,072
Cbf			0	0,034
Cb				0

On constate que la distance la plus courte est de 0,034 entre les séries Cb et Cbf. En conséquence, on agrège les séries Cb et Cbf, soit (Cb,Cbf), pour former avec les séries Cjn2 et Cjnl un nouveau tableau des fréquences :

	G.f	G.m	G.c	
(Cb,Cbf)	.290	.188	.522	1
Cjn2	.217	.174	.609	1
Cjnl	.263	.211	.526	1
1/f	3,551	5,267	1,892	

Calculons les distances du Khi 2 entre les trois séries Cjn2, Cjnl et (Cb,Cbf) :

d^2 (Cjn2,Cjnl) = 0,028 d'où d (Cjn2,Cjnl) = 0,167

d^2 ((Cb,Cbf),Cjn2) = 0,034 d'où d ((Cb,Cbf),Cjn2) = 0,185

d^2 ((Cb,Cbf),Cjnl) = 0,005 d'où d ((Cb,Cbf),Cjnl) = 0,074

La matrice se réduit comme suit :

1re étape	(Cb,Cbf)	Cjn2	Cjnl
(Cb,Cbf)	0	0,185	0,074
Cjn2		0	0,167
Cjnl			0

On constate que la distance la plus courte est de 0,074 entre (Cb,Cbf) et Cjnl. En conséquence, on agrège les séries (Cb,Cbf) et

Cjnl, soit (Cb, Cbf, Cjnl), pour former avec la série Cjn2 un nouveau tableau de fréquences :

	G.f	G.m	G.c	
(Cb,Cbf,Cjnl)	.287	.191	.522	1
Cjn2	.217	.174	.609	1
1/f	3,551	5,267	1,892	

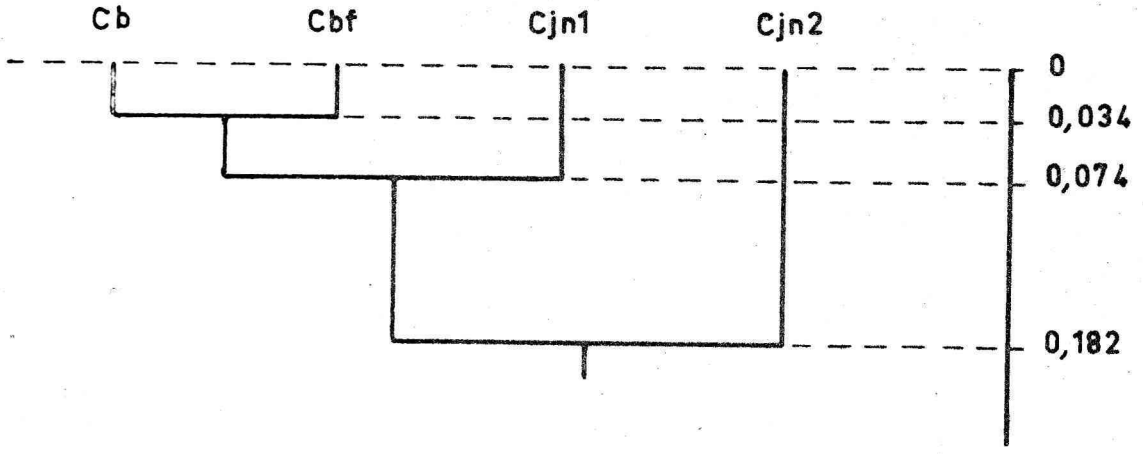
Calculons les distances du Khi 2 entre les deux séries (Cb,Cbf, Cjnl) et Cjn2 :

$d^2((Cb,Cbf,Cjnl),Cjn2) = 0,033$ d'où $d((Cb,Cbf,Cjnl),Cjn2) = 0,182$

La matrice se réduit enfin comme suit :

2e étape	(Cb,Cbf,Cjnl)	Cjn2
(Cb,Cbf,Cjnl)	0	0,182
Cjn2		0

Nous pouvons maintenant dessiner le dendrogramme, c'est-à-dire un arbre de classification qui, muni d'une échelle de distance, figure une hiérarchie indicée.



2. Autre exemple d'application de l'algorithme.

On se propose de classer hiérarchiquement les catégories d'Ordres (S = Simples; A = Abrupts; SE = Surélevés; E = Ecaillés; B = Burins et P = Plans) dans les trois séries ou ensembles industriels Cjn2, Cjnl et Cbf de la grotte Gatzarria.

On compose avec ces données le tableau de contingence à r séries et k catégories (A.1.1) :

	S	A	SE	E	B	P	
Cjn2	132	93	19	0	26	0	270
Cjn1	70	25	21	3	6	0	125
Cbf	216	42	76	52	8	0	394
	418	160	116	55	40	0	789

Pour faciliter le calcul de la distance du Khi 2 on dresse le tableau pertinent des fréquences utilisées dans la formule (A.1.5):

$$d^2(j,j') = \sum_{i=1}^{k} \frac{n_{..}}{n_{i.}} \left(\frac{n_{ij}}{n_{.j}} - \frac{n_{ij'}}{n_{.j'}} \right)^2$$

		j	j'		
i		$\frac{n_{ij}}{n_{.j}}$	$\frac{n_{ij'}}{n_{.j'}}$		$\frac{n_{..}}{n_{i.}}$
i'		$\frac{n_{i'j}}{n_{.j}}$	$\frac{n_{i'j'}}{n_{.j'}}$		$\frac{n_{..}}{n_{i'.}}$
		1	1		

c'est-à-dire :

	S	A	SE	E	B	1/f
Cjn2	.316	.581	.164	0	.650	2,922
Cjn1	.167	.156	.181	.055	.150	6,312
Cbf	.517	.263	.655	.945	.200	2,003
	1.	1.	1.	1.	1.	

La catégorie P disparaît car le quotient $\frac{n_{ij}}{n_{.j}} = \frac{f_{ij}}{f_{.j}} = \frac{0}{0}$

est une forme indéterminée. Par contre, il en serait tout autrement si le numérateur étant égal à zéro le dénominateur était différent de zéro.

Les distances du Khi 2 calculées entre les cinq catégories sont reportées dans une matrice des distances ou tableau de dissimilarité:

	S	A	SE	E	B
S	0	0,579	0,327	0,859	0,727
A		0	0,905	1,408	0,149
SE			0	0,584	1,054
E				0	1,550
B					0

On constate que la distance la plus courte est de 0,149 entre les catégories A et B. En conséquence, on agrège ces deux catégories, soit (A,B), pour former avec les catégories S,SE,E et B un nouveau tableau des fréquences :

	(A,B)	S	SE	E	1/f
Cjn2	.595	.316	.164	.	2,922
Cjn1	.155	.167	.181	.055	6,312
Cbf	.250	.517	.655	.945	2,003
	1.	1.	1.	1.	

On calcule les nouvelles distances du Khi 2 et on les reporte dans une nouvelle matrice :

	(A,B)	S	SE	E
(A,B)	0	0,609	0,936	1,437
S		0	0,327	0,859
SE			0	0,589
E				0

On constate que la distance la plus courte est de 0,327 entre les catégories S et SE. En conséquence, on agrège ces catégories,

soit (S,SE), pour former avec les catégories (A,B) et E un nouveau tableau des fréquences :

	(A,B)	(S,SE)	E	1/f
Cjn2	.595	.282	.	2,922
Cjnl	.155	.170	.055	6,312
Cbf	.250	.547	.945	2,003
	1.	1.	1.	

On calcule les nouvelles distances du Khi 2 et on les reporte dans une nouvelle matrice des distances :

	(A,B)	(S,SE)	E
(A,B)	0	0,681	1,437
(S,SE)		0	0,797
E			0

On constate que la distance la plus courte est de 0,681 entre (S,SE) et (A,B). On agrège donc ces catégories, soit (A,B,S,SE) pour former avec E un nouveau tableau des fréquences :

	(A,B,S,SE)	E	1/f
Cjn2	.368	.	2,922
Cjnl	.166	.055	6,312
Cbf	.466	.945	2,003
	1.	1.	

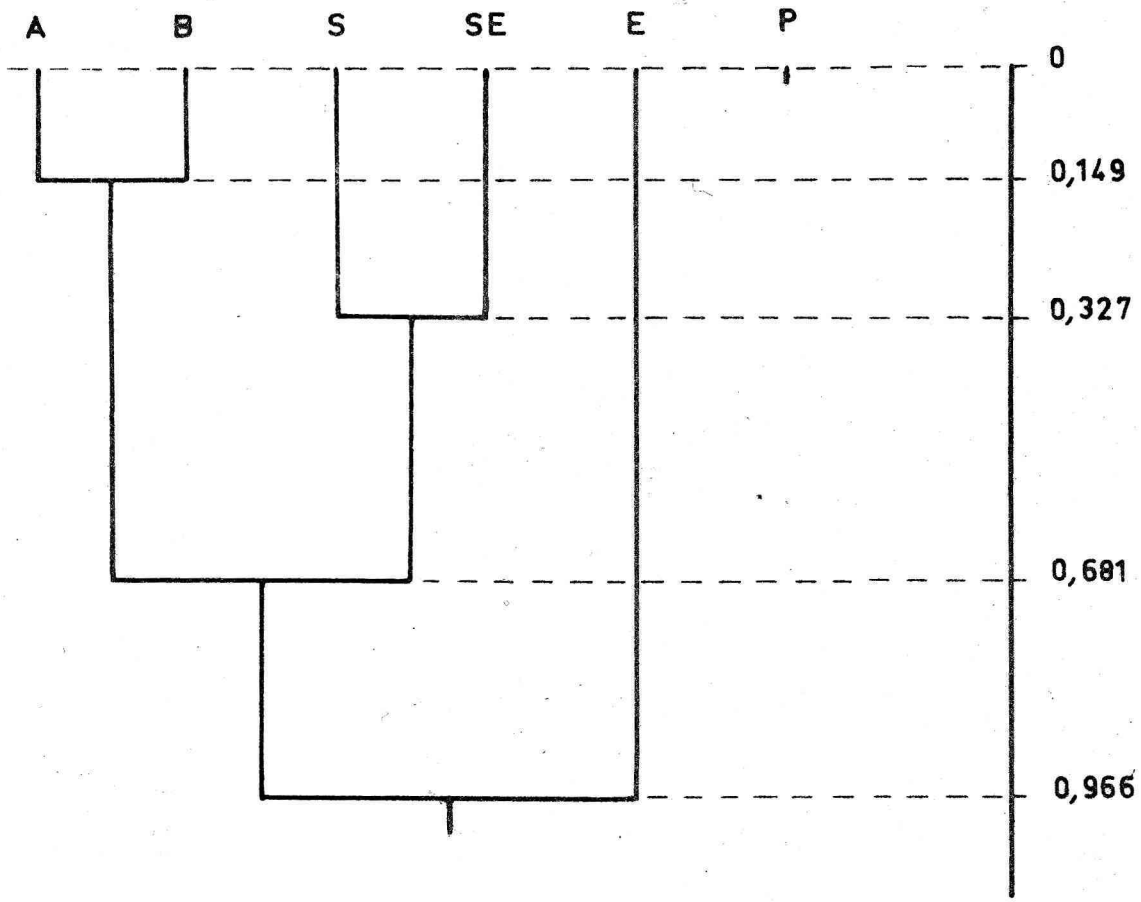
On calcule les nouvelles distances du Khi 2 et nous les reportons dans une dernière matrice des distances :

	(A,B,S,SE)	E
(A,B,S,SE)	0	0,966
E		0

L'ultime distance est de 0,966, entre les catégories (A,B,S,SE) et E.

Nous pouvons maintenant dessiner le dendrogramme qui exprime les analogies évolutives des catégories en une classification hié-

rarchisée et indicée.



BIBLIOGRAPHIE

P. BERTIER, J.M. BOUROCHE - Analyse des données multidimensionnelles. Presses Universitaires de France, 1975.

N.B. Cette étude résulte des recherches les plus récentes. Elle fera l'objet d'un exposé au cours du séminaire du mois d'Août 1976.