

J. P. BENZÉCRI

F. BENZÉCRI

Y. L. CHEUNG

S. MAIZA

**Aides à l'interprétation et étiquetage des arbres
en classification ascendante hiérarchique :
listages FACOR, VACOR et INFSUP**

Les cahiers de l'analyse des données, tome 10, n° 3 (1985),
p. 311-338

http://www.numdam.org/item?id=CAD_1985__10_3_311_0

© Les cahiers de l'analyse des données, Dunod, 1985, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

AIDES A L'INTERPRÉTATION ET ÉTIQUETAGE DES ARBRES EN CLASSIFICATION ASCENDANTE HIÉRARCHIQUE : LISTAGES FACOR, VACOR ET INFSUP

[INTERPRET CAH]

par J.P. Benzécri *, F. Benzécri **, Y.L. Cheung ***, S. Maiza ****

Nous avons exposé dans [CORR. ECO.] le cadre géométrique et les principes de l'analyse des correspondances. Sans sortir de ce cadre, nous avons, dans [CAH ECO] défini ce qu'est une classification ascendante hiérarchique en expliquant l'algorithme de base et en décrivant la présentation des résultats de CAH. Le présent exposé est consacré à l'interprétation des résultats de CAH utilisant les aides fournies par les programmes FACOR, VACOR et INFSUP. Ici encore nous utiliserons un exemple extrait des données du commerce extérieur des phosphates, consignées dans le tableau ternaire $\{k(i,j,t) | i \in I ; j \in J ; t \in T\}$ présenté en [CORR ECO] § 1. Mais alors que jusqu'ici nous avons considéré un tableau aussi réduit que possible (le tableau de marge $k_{I,J}$ à 14 lignes et 8 colonnes), ici, le sujet nécessite un exemple de tableau recelant des faits plus complexes : nous considérerons le tableau $k_{I \times T, J}$ à 112 lignes (14×8) et 8 colonnes, sans nous interdire de regarder l'arbre de classification sur I issu du petit tableau $k_{I,J}$.

Le présent exposé comprend trois parties dont voici les thèmes : § 1 choix d'une partition terminale ; § 2 : définition des classes et des noeuds retenus en termes de facteurs et de variables ; § 3 : étiquetage de l'arbre. Les listages utilisés sont principalement au § 1, le listage de CAH déjà commenté dans [CAH ECO] ; § 2, les listages FACOR et VACOR ; au § 3 le listage INFSUP. Le § 1 est d'une lecture facile ; les §§ 2.1 et 2.2 (listages FACOR et VACOR) reprennent avec quelques développements nouveaux des notions familières à ceux qui ont étudié [CORR. ECO.] et [CAH ECO]. Nous pensons qu'un lecteur peu mathématicien assimilera bien les nouveautés d'après l'usage qui en est fait au § 2.3 ; et surtout au § 3.2, dans l'étiquetage de l'arbre (auquel sert le listage INFSUP, qui n'est en bref qu'un extrait de VACOR, propre à suggérer des rapprochements utiles).

1 Lecture descendante de l'arbre et choix d'une partition terminale

Rappelons que l'algorithme de CAH construit l'arbre hiérarchique à partir des individus : il agrège en priorité les deux individus qui se ressemblent le plus (i.e. pour lesquels le critère de l'inertie a la plus faible valeur) ; puis il agrège deux autres individus ou la classe déjà créée et un autre individu selon le même critère ; critère qui équivaut à ceci : à chaque agrégation, l'inertie intraclasse de la partition qui en résulte doit croître le moins possible : cf. [CLAS ECO] §§ 2.3.3 et 3.5. La construction se poursuit jusqu'au sommet de l'arbre où est créée la classe comprenant tous les individus.

(*) Professeur de statistique. Université P. et M. Curie.

(**) Docteur-ès-Sciences.

(***) Docteur d'université. "Lecturer" au City polytechnic de Hong-Kong.

(****) Docteur 3° cycle.

On aspire à lire cet arbre de haut en bas en comprenant quels faits déterminent les dichotomies successives rencontrées dans cette lecture descendante : ce sera l'objet de l'interprétation de la CAH. En général l'intérêt se concentre à un certain niveau de généralité : on ne suivra pas les divers chemins descendants jusqu'aux individus ; mais on s'arrêtera aux classes d'une partition choisie pour les qualités de généralité et d'homogénéité de ses classes ; qualités que toutes les classes de cette partition ne possèdent pas nécessairement à un même degré : cf. *infra* § 1.2. Formellement, le plus simple est de couper l'arbre à un niveau déterminé d'après l'histogramme des niveaux (§ 1.1) ; mais en examinant le contenu des classes de la CAH, on découvrirait souvent que la partition la plus satisfaisante ne peut être définie par une telle coupure (§ 1.2). Le choix définitif dépend d'ailleurs d'une interprétation précise des classes tenant compte des variables (§ 1.3).

1.1 Coupure horizontale de l'arbre de CAH : Nous reproduisons ici, à titre d'exemple, l'arbre de CAH sur l'ensemble I des 14 pays importateurs, d'après le tableau de marge $k_{I,J}$ (14x8) ; ainsi que l'histogramme des niveaux des noeuds ([CAH ECO] §§ 1.2 et 4.1.

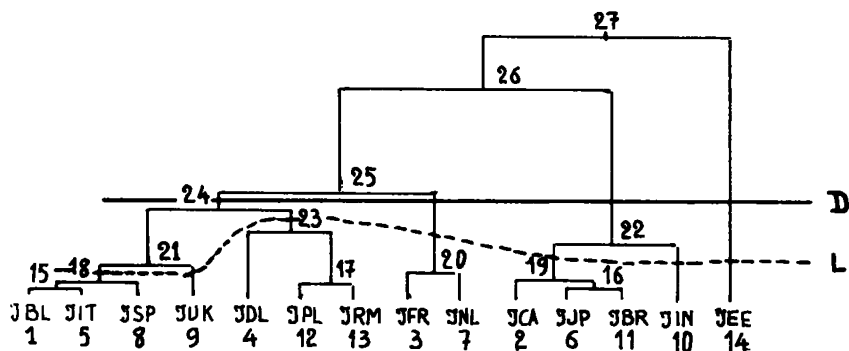


Fig. 1 : Arbre de CAH sur l'ensemble I des 14 pays importateurs et histogramme des niveaux des noeuds.

N	D(N)
27	34
26	27
25	11
24	10
23	6
22	5
21	4
20	2
19	2
18	2
17	1
16	1
15	0

En coupant l'arbre par une droite D horizontale au-dessus de laquelle se trouvent exactement n noeuds, on définit une partition S en n+1 classes ; les classes de S étant caractérisées par la propriété d'être constituées au-dessous de D et d'être des descendants immédiats d'un noeud situé au-dessus de D. Ainsi, les noeuds situés au-dessus de D définissent sur S une structure hiérarchique, les classes étant délimitées par n dichotomies successives ; comme on le voit sur la Figure 1 qui montre le cas n = 3 pour la CAH sur l'ensemble I des 14 pays importateurs de phosphate. On dira qu'on a retenu une hiérarchie

restreinte de $(2n+1)$ classes, dont les noeuds sont les n noeuds de la CAH situés au-dessus de D et dont les terminaux sont les $(n+1)$ classes de la partition S. De ce point de vue, choisir une partition c'est

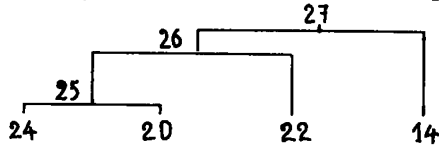


Fig. 2 : Hiérarchie restreinte de 4 classes définie par la droite D de la Fig. 1.

choisir le nombre n , ou ce qui est équivalent, choisir un niveau de coupure. Dans la pratique, ayant décidé d'un ordre de grandeur pour n (e.g. : on désire une dizaine de classes...) on fixera le niveau d'après l'histogramme (cf. [CAH ECO] § 4.1) l'idéal étant qu'entre $v[n]$ et $v[n+1]$ la variation relative soit aussi forte que possible. Pour la CAH sur IxT (14 pays $\times$ 8 années), on songe à une partition idéale en 14 classes dont chacune contiendrait les 8 profils annuels d'un pays importateur : même si le contenu des classes ne confirme pas ce modèle, on peut chercher une valeur de n voisine de 13. L'histogramme (où les valeurs de T sont en millièmes) montre une nette dénivellation entre les noeuds 216 et 215 ; (ce qui correspond à $n = 8$; donc à une

N°ED A(N):B(N) TX(=) *				HISTOGRAMME DES INDICES DE NIVEAU DE LA HIÉRARCHIE	
'203	'198	'222	'271	'#'	*****
'222	'221	'218	'214	'#'	*****
'221	'220	'219	'87	'#'	*****
'220	'216	'217	'83	'#'	*****
'219	'213	'218	'55	'#'	*****
'218	'215	'206	'52	'#'	*****
'217	'212	'214	'75	'#'	*****
'216	'211	'205	'54	'#'	*****
'215	'195	'208	'21	'#'	*****
'214	'175	'284	'18	'#'	*****
'213	'280	'209	'15	'#'	*****
'212	'187	'287	'13	'#'	*****
'211	'180	'283	'10	'#'	*****
'210	'192	'181	'9	'#'	*****
'209	'178	'282	'8	'#'	*****
'208	'194	'183	'8	'#'	*****
'207	'196	'201	'7	'#'	*****
'206	'171	'123	'7	'#'	*****
'205	'163	'167	'5	'#'	*****
'204	'189	'193	'5	'#'	*****
'203	'190	'199	'3	'#'	*****
'202	'197	'RM79	'2	'#'	*****
'201	'FR77'	'FR76	'2	'#'	*****
'200	'184	'182	'2	'#'	*****
'199	'179	'148	'2	'#'	*****
'198	'191	'176	'2	'#'	*****
'197	'174	'186	'2	'#'	*****
'196	'1178'	'185	'1	'#'	*****
'195	'IN74'	'IN73	'1	'#'	*****
'194	'157	'149	'1	'#'	*****
'193	'152	'FR75	'1	'#'	*****
'192	'155	'188	'1	'#'	*****

Fig. 3 : Histogramme des niveaux de la CAH sur I $\times$ T (14 pays $\times$ 8 années).

partition en 9 classes) et une autre moins nette entre 212 et 211 ($n = 12$; partition en 13 classes).

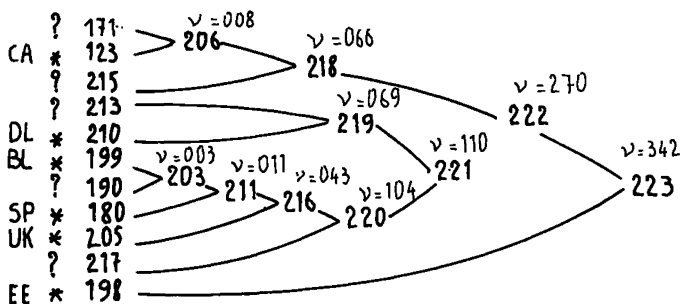
1.2 Partition générale extraite de la CAH : Plus généralement (cf. [CAH ECO] § 1.2) une partition C peut se définir par une ligne de coupure L : ici encore, les classes de C sont caractérisées par la propriété d'être constituées au-dessous de L et d'être des descendants immédiats de noeuds situés au-dessus de L ; ces derniers noeuds (situés au-dessus de L) définissent les dichotomies successives qui aboutissent aux classes de C et constituent, avec les classes de C, la hiérarchie restreinte retenue. Mais de ce point de vue, rien ne peut aider directement à choisir la ligne L. C'est la partition C elle-même qu'il faut choisir ; plus précisément les classes c de C, d'après l'interprétation que suggère leur contenu.

Mais il faut prendre garde qu'il est rare que toutes les classes c de C soient également satisfaisantes : supposons par exemple que sur l'arbre de la Fig. 1 on désire conserver dans la hiérarchie restreinte les classes 18 et 19. Conserver 18 impose de conserver 9 (classe réduite à un individu) qui est frère de 18 en ce sens que $18 = A[21]$ et $9 = B[21]$. Par le fait, dans la hiérarchie restreinte, on a 21 ; ainsi que son frère 23 et leur père 24 ($A[24] = 21$; $B[24] = 23$) ; ayant 24, on a 20 et 25 ($24 = A[25]$; $20 = B[25]$). De même, avec 19 on a 10 et 22 ; 22 avec 25 (déjà retenu) et 26 ; enfin, avec 26 on doit retenir son frère 14 et 27, père de 26 et de 14. La hiérarchie retenue a donc pour noeuds : 27, 26, 25, 24, 22, 21 ; et pour classes terminales 18, 9, 23, 20, 19, 10, 14. Ainsi pour retenir deux classes 18 et 19, on doit conserver une hiérarchie qui compte 6 noeuds, définissant une partition terminale en 7 classes.

Dans le cas de la CAH sur $I \times T$, il s'impose d'abord de retenir toutes les classes qui se composent des huit profils annuels d'un pays importateur. Ces classes sont au nombre de 6 :

123 : CA3- ω ; 180 : SP3- ω ; 198 : EE3- ω ; 199 : BL3- ω ; 205 : UK3- ω ; 210 : DL3- ω ;

(les années 1973 à 1979 sont désignées par leur dernier chiffre : 3 à 9 ; 1980 par ω : par exemple CA3- ω désigne les 8 années de CA). Retenir ces six classes impose de retenir la hiérarchie restreinte suivante avec 10 noeuds et 11 classes terminales.

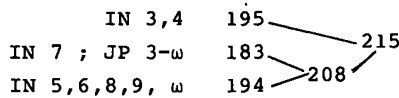


(N.B. : Sur cette esquisse de l'arbre, l'échelle des niveaux n'est pas respectée).

Parmi les classes terminales, six demandées pour elles-mêmes sont marquées d'une étoile ; les cinq autres, introduites comme frères de classes retenues, sont marquées du signe " ? " ; la question étant de décider si on les subdivisera ou non.

■ 171 comprend 6 années du Brésil (3, 5, 7, 8, 9, ω) ; on la gardera telle quelle.

■ 215 comprend les 8 années de l'Inde et les 8 années du Japon ; il n'est pas possible de séparer l'Inde du Japon par une simple dichotomie de la cl. 215 ; mais en coupant 215 en 195 et 208, ses deux successeurs immédiats, puis 208 en 183 et 194, on a une séparation quasi parfaite :

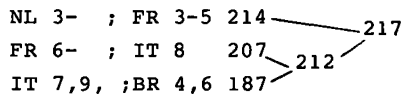


L'importance essentielle de cette subdivision apparaîtra sur le listage VACOR (§ 2.3.3).

■ 213 comprend les 8 années de la Pologne et de la Roumanie. On ne peut séparer Pologne de Roumanie, même par deux ou trois dichotomies successives. Dans un but de simplification on gardera donc 213 comme classe terminale (bien qu'entre 200 = A[213] et 209 = B[213] il y ait une opposition temporelle intéressante : cf. *infra* §§ 3.2 et 3.3.3).

■ 190 comprend 4 années d'Italie : IT3-6.

■ 217 : France 3- ω ; Nederland 3- ω ; Brésil 4, 6 ; Italie 7- ω . Une dichotomie de 217 en 214 et 212 permet de séparer les huit années de Nederland avec toutefois trois années de la France. En étudiant la classe 217 on verra qu'il y a de plus avantage à subdiviser en B(217 = 212 en 207 et 187 (cf. § 3.2).



1.3 Critique de la partition choisie : Il est difficile d'arrêter la liste des dichotomies retenues : même une classe pure correspondant à un seul pays peut mériter d'être scindée. Par exemple :

210 = DL3- ω ; A[210] = 181 = DL3-5 ; B[210] = 192 = DL6- ω .

La dichotomie est intéressante en ce qu'elle atteste une tendance évolutive, puisqu'elle oppose début et fin de période. La partition choisie sera d'autant plus fine qu'on est plus curieux du détail des données ; et même si on ne publie qu'une partition en 11 classes, on peut explorer sur les listages les subdivisions de chacune de ces classes.

Revenons à l'histogramme des niveaux : il se trouve que nous avons retenu les huit noeuds 216 à 223 en accord avec l'histogramme qui suggérerait de couper l'arbre entre les noeuds 215 et 216 (cf. § 1.1) ; mais en plus de ces huit noeuds nous avons également retenu, au-dessous de cette coupure, les noeuds 211, 203, 206, 215, 208 et 212.

D'ailleurs le contenu des classes ne se définit pas seulement en termes de pays ; une classe peut aussi nous intéresser parce qu'elle occupe une partie de la période T, même si elle est hétérogène quant aux pays : des exemples sont publiés dans les C.A.D. (cf. e.g. [OPEP + OCDE] § 4.4 ; CAD Vol VIII n°1) ; ici même, la classe 214 se scinde en :

A[214] = 175 = NL6- ω ; B[214] = 204 = {NL3-5 ; FR 3-5} ; ainsi, la classe 204 apparaît concentrée sur les trois premières années de T (73, 74, 75).

Mais autant et plus que l'inventaire de son contenu plus ou moins structuré (dans le temps, dans l'espace etc.), c'est sa définition par les variables (listage VACOR ; cf. *infra* § 2.2) qui fait l'intérêt d'une classe et engage à retenir celle-ci ou à la subdiviser. C'est pourquoi le choix de la hiérarchie restreinte ne sera arrêté qu'après une lecture descendante de l'arbre faite en consultant les listages FACOR et VACOR.*

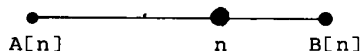
Comme on le verra ci-dessous, ces listages comportent un nombre de lignes proportionnel au nombre des noeuds retenus. Tant que l'étude n'est pas achevée, on ne sait pas exactement à quelle partition on s'arrêtera. Aussi choisit-on d'abord un niveau de coupure assez bas pour pouvoir examiner comment se subdivisent les classes de la partition que l'on pense retenir. Mais ici, afin de montrer au lecteur un listage complet, nous avons demandé un FACOR et un VACOR avec seulement 8 noeuds supérieurs ; ce qui correspond à une coupure entre les noeuds 215 et 216 (cf. § 1.1). Pour l'étude elle-même, nous avons utilisé un FACOR et un VACOR avec 55 noeuds supérieurs : dans l'interprétation, nous aurons parfois recours à ce grand listage non publié ici (cf. § 2.3.1 *in fine* ; § 2.3.3 *in fine*).

2 Définition des classes et des noeuds en termes de facteurs et de variables ; listages FACOR et VACOR

2.0 Représentation de l'arbre dans l'espace ambiant au nuage ; dipô-

les : Comme on l'a expliqué dans [CORR. ECO.], l'ensemble des individus (présentement $I \times T$) est représenté par un nuage de profils sur J. Chaque classe c a un centre de gravité noté également c . Le centre de gravité du nuage des individus qui correspond au sommet de la hiérarchie est noté g ; il se situe à l'origine dans le système des axes factoriels.

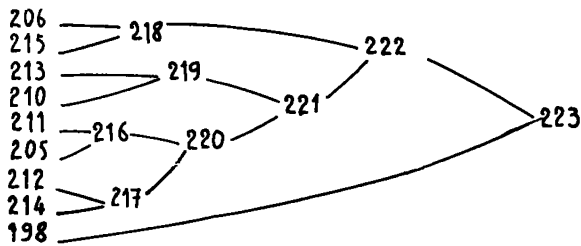
A chaque noeud n est associé un dipôle ($A[n]$, $B[n]$) formé par les centres de ses deux descendants immédiats et (cf. [CORR. ECO.] § 2.2.4) le point n divise le segment ($A[n]$, $B[n]$) dans le rapport inverse des masses des classes $A[n]$ et $B[n]$:



L'ensemble de ces dipôles constitue en quelque sorte une charpente pour le nuage des individus.

L'objet des calculs d'aide à l'interprétation est de préciser la situation dans l'espace d'une part des rayons qui joignent le centre g du nuage aux divers centres de la hiérarchie ; d'autre part des segments joignant $A[n]$, $B[n]$ dans les dipôles. Cette étude sera faite d'abord dans l'espace rapporté au système des axes factoriels (FACOR : § 2.1) ; ensuite dans l'espace rapporté au système des axes définis par les variables elles-mêmes, système dans lequel les coordonnées d'un individu ne sont autres que les composantes de son profil sur J (VACOR § 2.2).

2.1 Le listage FACOR : Le listage restreint (cf. § 1.3 *in fine*) publié ici comprend un tableau relatif aux 8 classes les plus hautes de la hiérarchie : 216 à 223 (les noeuds) ; un second tableau relatif aux classes terminales de cette hiérarchie restreinte : classes 206, 215, 213, 210, 211, 205, 212, 214, 198 ; (classes de la partition) ; et un troisième tableau relatif aux dipôles définis par les noeuds supérieurs n et leurs deux successeurs immédiats $A[n]$ et $B[n]$ (les dipôles des noeuds supérieurs).



§2.1 : Hiérarchie restreinte comprenant 8 nœuds supérieurs.

Les coordonnées qui figurent dans ces tableaux sont calculées dans le système des axes factoriels issus de l'analyse des correspondances du tableau $(I \times T) \times J$.

2.1.1 Les deux tableaux relatifs aux classes, terminales ou non, sont ceux des listages d'analyse des correspondances relatifs aux éléments supplémentaires (cf. [CORR. ECO.] § 7). Une classe c de la hiérarchie restreinte est un ensemble de pays-années ; son centre de gravité n'est autre que le profil de la ligne obtenue en cumulant les lignes du tableau $(I \times T) \times J$ afférentes aux éléments de c [CORR. ECO.] § 2.2) ; cette ligne c est adjointe en ligne supplémentaire au tableau $(I \times T) \times J$. Et le programme calcule pour l'élément supplémentaire c , comme il est expliqué en [CORR. ECO.] § 7, les facteurs F_a , les contributions CTR_a et COR_a (aussi noté $CO2_a$), le poids, l'inertie relative et la qualité de la représentation.

■ Le centre c lui-même est désigné par le n° N du nœud qui lui correspond et auquel on adjoint les n°s de ses successeurs immédiats $A[n]$ et $B[n]$. Ainsi la désignation de c occupe les 3 premières colonnes des deux tableaux.

$QLT(c) = \|0 - pr_1, \dots, 6(c)\|^2 / \|0 - c\|^2$; où 0 désigne l'origine des axes factoriels qui est aussi, rappelons-le, le centre de gravité du nuage, i.e. le centre de la classe la plus haute de la hiérarchie, 223 : $0 = g = 223$; $pr_1, \dots, 6(c)$ désigne la projection de c sur le sous-espace engendré par les six premiers axes factoriels, car sur le listage présenté ici ne figurent que six facteurs. Rappelons que J ayant 8 éléments, l'espace des profils sur J est de dimension 7 (cf. [CORR. ECO.] § 2.1.2). La qualité de représentation par les six premiers facteurs n'est parfaite que pour les centres situés dans le sous-espace engendré par les six premiers axes : 222, 220 (1-er tableau) et 198 (2-ème tableau) pour lesquels $QLT = 1000/1000$. Plus le centre c s'écarte de ce sous-espace et plus QLT descend au-dessous de la valeur 1 : la qualité de représentation est très bonne ($>950/1000$) pour tous les centres sauf 212 (2ème tableau) pour lequel $QLT = 724/1000$.

Le centre 223 (origine des coordonnées factorielles) échappe au calcul de QLT qui se présente pour ce centre sous la forme indéterminée $0/0$: la valeur imprimée $QLT = 0$ est sans signification ; l'origine est au contraire parfaitement représentée sur un sous-espace engendré par un nombre quelconque d'axes factoriels, étant un point commun à tous ces sous-espaces.

REPRÉSENTATION FACTORIELLE POUR LES 8 NOEUDS SUPERIEURS DE L'ARBRE

INCE	A(N)	B(N)	SALT	PDS	INR	F#1	COR.CTR	F#2	COR.CTR	F#3	COR.CTR	F#4	COR.CTR	F#5	COR.CTR	F#6	COR.CTR
'223	'198	'222	'0	1000	0	0	0	0	0	0	0	0	0	0	0	0	0
'222	'221	'218	'1000	922	21	161	897 51	-51	90 7	-13	6 1	5	1 0	0	0	13	6 3
'221	'220	'219	'596	659	60	-25	6 1	-333	969 214	20	7 3	-24	5 3	7	0 0	33	9 13
'220	'216	'217	'1000	444	114	139	59 18	-549	926 392	1	0 0	-50	8 10	31	3 6	-38	4 12
'219	'213	'210	'984	214	32	-365	697 61	113	67 0	84	37 0	31	5 2	-43	9 5	180	169 128
'218	'215	'206	'998	264	175	626	466 220	654	507 330	-115	16 19	76	7 13	-19	0 1	-36	2 6
'217	'212	'214	'986	255	95	224	106 27	-505	537 190	361	275 187	74	12 12	152	49 83	-60	0 17
'216	'211	'235	'997	109	101	24	1 0	-609	546 285	-484	345 240	-216	69 77	-131	25 46	-9	0 0

REPRÉSENTATION FACTORIELLE POUR LES 9 CLASSES DE LA PARTITION

CLAS	A(N)	B(N)	SALT	PDS	INR	F#1	COR.CTR	F#2	COR.CTR	F#3	COR.CTR	F#4	COR.CTR	F#5	COR.CTR	F#6	COR.CTR
'198	'191	'176	'1000	78	250	-1907	897 602	605	90 83	152	6 10	-57	1 2	2	0 0	-158	6 36
'210	'192	'191	'992	90	44	-59	6 1	230	85 14	506	411 129	-137	30 15	-306	150 119	441	311 322
'213	'200	'209	'996	125	43	-506	700 91	29	2 0	-222	112 34	152	52 25	147	49 30	-9	0 0
'206	'171	'123	'999	142	154	750	410 170	84	722 298	53	2 2	-284	59 99	30	1 2	-83	5 18
'215	'195	'200	'997	122	74	483	304 60	427	140 65	-311	126 66	495	320 258	-76	7 10	19	0 1
'214	'175	'204	'962	112	93	184	32 0	-727	496 171	606	349 230	141	19 19	256	62 104	69	4 10
'212	'197	'207	'924	147	37	255	199 20	-734	341 47	170	89 23	22	1 1	71	15 10	-160	70 68
'205	'163	'167	'955	49	49	120	12 2	-734	430 70	-157	20 7	-10	0 0	-686	375 329	-505	118 135
'211	'180	'203	'990	140	86	-9	0 0	-564	406 120	-599	459 282	-209	107 101	65	5 0	124	20 40
'C U M U L				1000	830		953		888		784		519		619		629

Toutes les valeurs imprimées dans ces tableaux sont en millièmes.

■ PDS(c) : la ligne supplémentaire c a pour total la somme des totaux des lignes dont elle est le cumul (cf. *supra*). L'analyse factorielle lui assigne aussi pour poids la somme des poids des éléments (pays-années) qui constituent la classe c. Le centre du nuage 223 a pour poids 1000/1000. On sait en effet que le poids total du nuage est normalisé à 1.

■ INR(c) : inertie de c par rapport au centre du nuage, divisée par l'inertie totale du nuage.

Le centre 223 qui coïncide avec le centre du nuage a une inertie nulle par rapport à ce centre, et INR(223) = 0.

Le centre c, étant un élément supplémentaire, n'est pas compté dans l'inertie totale du nuage ; INR(c) n'est donc pas une part relative à un tout mais seulement une valeur rapportée à une autre (cf. *infra* : ligne CUMUL du 2-ème tableau).

■ F_α(c) est la coordonnée de c sur l'axe α. En valeur absolue, |F_α(c)| mesure la distance qui sépare le centre du nuage et la projection du centre c sur l'axe α. Quel que soit α, F_α(223) = 0, puisque le centre 223 est à l'origine des axes factoriels.

■ COR_α(c) (noté ailleurs CO2_α(c)), cosinus carré de l'angle que fait le rayon (O, c) avec l'axe α :

$$\text{COR}_\alpha(c) = (F_\alpha(c))^2 / \|O - c\|^2$$
 ; c'est encore la part du facteur α dans l'explication de la distance c au centre du nuage, quantité nécessairement inférieure ou égale à 1, qui décroît de 1 à 0 lorsque l'angle de (O, c) avec l'axe α croît de 0 à 90°.

$$\text{On a : } QLT(\underline{c}) = COR_1(\underline{c}) + COR_2(\underline{c}) + \dots + COR_6(\underline{c}) \\ = \sum \{COR_\alpha(\underline{c}) \mid \alpha = 1, 2, \dots, 6\}.$$

Pour le centre 223 = g = 0, quel que soit α , COR_α se présente sous forme indéterminée 0/0, comme QLT (cf. *supra*). La valeur imprimée $COR = 0$ est sans signification.

■ $CTR_\alpha(\underline{c})$, inertie de $\underline{c}$ par rapport à $O = g$ en projection sur l'axe α , divisée par l'inertie totale du nuage projeté sur l'axe α .

$$CTR_\alpha(\underline{c}) = (PDS(\underline{c}) \times (F_\alpha(\underline{c}))^2) / \lambda_\alpha.$$

Ici, comme dans INR (cf. *supra*) ce rapport n'est pas le rapport d'une part à un tout.

Le 2-ème tableau (classes de la partition) comprend une ligne :

■ CUMUL donnant le total des colonnes PDS, INR et CTR. Puisque le tableau considéré est relatif à une *partition*, tout couple pays-année (i,t) appartient à une classe et à une seule du tableau. Donc, la somme des poids des diverses classes (ou, ce qui revient au même, des divers centres) du tableau n'est autre que le poids total du nuage : 1000/1000. En revanche, l'inertie relativement à l'origine O du centre d'une classe $\underline{c}$ n'est pas la somme des inerties (relat. à O) des points (i,t) constituant la classe ; mais elle est égale à cette somme diminuée de l'inertie interne de la classe $\underline{c}$ (ou inertie des points (i,t) qui composent cette classe, relativement au centre $\underline{c}$). C'est pourquoi la colonne INR à un total inférieur à 1 (dans notre exemple, 830/1000). De façon précise, ce total est le quotient par l'inertie totale du nuage, de l'inertie *inter*classe de la partition retenue (ou inertie du nuage des centres des classes de la partition ; cf. [CAH ECO] § 3.2) ; et le complémentaire à 1 de ce total (ici : 170/1000) est le quotient par l'inertie totale du nuage $N(I \times T)$ de l'inertie *intra*classe de la partition (ou somme des inerties internes des classes de la partition). Le total de la colonne CTR_α a une interprétation tout analogue : à ceci près qu'il s'agit d'inertie en *projection* sur l'axe α ; (on sait que l'inertie totale du nuage $N(I \times T)$ en projection sur l'axe α n'est autre que la valeur propre λ_α).

2.1.2 Le tableau des dipôles relatifs aux noeuds supérieurs

REPRESENTATION FACTORIELLE POUR LES 8 DIPOLES DES NOEUDS SUPERIEURS

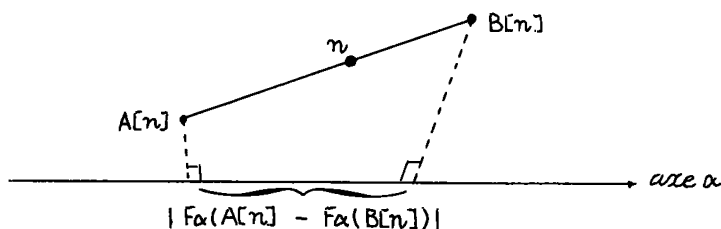
!DIP.!A(N)!B(N)! QALD PDS IND !	! D#1	! COD CTD !	! D#2	! COO CTD !	! D#3	! COD CTD !	! D#4	! COO CTD !	! D#5	! COO CTD !	! D#6	! COD CTD !
!223 !198 !222 !1000 1000 271 !	!-2068 897 653 !	! 654 93 90 !	! 165 6 11 !	! -62 1 2 !	! 2 0 0 !	! -171 6 39 !						
!222 !221 !218 !1997 922 214 !	!-652 295 170 !	! -987 677 538 !	! 143 14 21 !	! -99 7 16 !	! 26 0 2 !	! 68 3 16 !						
!221 !220 !219 !1997 658 87 !	! 504 334 78 !	! -662 576 185 !	! -82 9 6 !	! -81 9 8 !	! 74 7 11 !	! -218 62 126 !						
!220 !216 !217 !1968 444 83 !	! -199 41 9 !	! -104 11 3 !	! -845 742 435 !	! -290 89 79 !	! -223 83 123 !	! 51 3 5 !						
!219 !213 !210 !1999 214 55 !	! -526 207 31 !	! -201 30 6 !	! -720 396 155 !	! 269 62 30 !	! 453 153 151 !	! -452 151 194 !						
!218 !215 !206 !1000 264 52 !	! -267 71 10 !	! -418 173 34 !	! -364 132 49 !	! 779 623 143 !	! -126 11 10 !	! 102 10 13 !						
!217 !212 !214 !1640 255 35 !	! 71 7 1 !	! 380 216 28 !	! -436 272 67 !	! -119 20 0 !	! -195 49 30 !	! -229 75 61 !						
!216 !211 !205 !1968 189 34 !	! -130 -14 1 !	! 170 25 3 !	! 442 165 40 !	! -279 86 25 !	! 751 478 291 !	! 509 220 175 !						
! CUMUL	! 830 !	! 953 !	! 603 !	! 704 !	! 519 !	! 619 !	! 629 !					

■ Les trois premières colonnes : $\underline{n}$, $A[\underline{n}]$, $B[\underline{n}]$ désignent le centre $\underline{n}$ du dipôle et les deux extrémités $A[\underline{n}]$ et $B[\underline{n}]$ qui, dans la hiérarchie, correspondent au noeud $\underline{n}$ et à ses deux successeurs immédiats.

Les colonnes suivantes sont présentées comme dans les deux premiers tableaux. On remarquera que les sigles qui les désignent sont tous (sauf PDS) terminés par la lettre D (Dipôle). Pour un noeud supérieur $\underline{n}$ on trouve dans ce tableau :

■ $D_\alpha(\underline{n}) = F_\alpha(A[\underline{n}]) - F_\alpha(B[\underline{n}])$, différence qui renseigne, non sur la position sur l'axe des projections des centres $A[\underline{n}]$ et $B[\underline{n}]$, mais sur leur position *relative* l'un par rapport à l'autre.

■ $COD_\alpha(\underline{n}) = (F_\alpha(A[\underline{n}]) - F_\alpha(B[\underline{n}]))^2 / \|A[\underline{n}] - B[\underline{n}]\|^2$.



Projection du dipôle $(A[\underline{n}], B[\underline{n}])$ sur l'axe α .

(En général, le dipôle $(A[\underline{n}], B[\underline{n}])$ et l'axe α ne sont pas coplanaires, ce qui gauchit la figure). Par projection sur l'axe α , la longueur du segment $(A[\underline{n}], B[\underline{n}])$ ne peut que diminuer (ou rester constante si le segment est parallèle à l'axe, auquel cas $COD_\alpha(\underline{n}) = 1$). Plus ce segment s'écarte, en direction, de l'axe α et plus $COD_\alpha(\underline{n})$ descend au-dessous de la valeur 1, atteignant la valeur 0 lorsque le segment $(A[\underline{n}], B[\underline{n}])$ est perpendiculaire à l'axe : de façon précise, $COD_\alpha(\underline{n})$ est le cosinus carré de l'angle que font, en direction, l'axe α et le dipôle.

■ $CTD_\alpha(\underline{n})$: par projection sur l'axe α , $A[\underline{n}]$ et $B[\underline{n}]$ conservent leur poids et le centre de gravité du dipôle projeté est la projection de $\underline{n}$. Le coefficient $CTD_\alpha(\underline{n})$ est l'inertie interne du dipôle projeté, divisée par l'inertie totale du nuage projeté sur l'axe α : λ_α . (L'inertie interne d'un dipôle $\underline{n}$ est l'inertie du système de ses deux extrémités $A[\underline{n}]$ et $B[\underline{n}]$ munies de leurs masses, par rapport au centre de gravité $\underline{n}$ du dipôle.

■ $QLD(\underline{n}) = \|pr_{1,\dots,6}(A[\underline{n}]) - pr_{1,\dots,6}(B[\underline{n}])\|^2 / \|A[\underline{n}] - B[\underline{n}]\|^2$;

c'est encore la somme des $COD_\alpha(\underline{n})$: $\sum \{COD_\alpha(\underline{n}) | \alpha = 1, \dots, 6\}$; ou qualité de la représentation du dipôle par sa projection sur le sous-espace engendré par les axes afférents aux facteurs imprimés dans le tableau.

■ $PDS(\underline{n})$: poids de la classe $\underline{n}$, i.e. somme des poids des pays-années entrant dans cette classe. Cette colonne a même signification que dans les deux premiers tableaux.

■ $IND(\underline{n})$: inertie interne du dipôle (cf. *supra* CTD), divisée par l'inertie totale du nuage, i.e., cf. EA n° 2 § 2.2.4 : $D(A[\underline{n}], B[\underline{n}]) / Itot = T(\underline{n})$, cf. EA n° 2 § 4.1.

■ La ligne CUMUL du tableau des dipôles indique le total des colonnes IND et CTD_α . Ici les diverses classes $\underline{n}$ ne constituent pas une partition de I : un même pays-année appartient e.g. à 223, 222, 221, 220,

216 et figure 5 fois dans le cumul de la colonne PDS. C'est pourquoi le total de cette colonne n'aurait pas de sens. Quant aux inerties, rappelons que l'inertie interne du dipôle n n'est autre que $\text{crit}(A[n], B[n])$, ou, dans la construction de la hiérarchie à partir de la base, l'accroissement de l'inertie intraclasse de la partition par agrégation de $A[n]$ et $B[n]$, cf. [CAH ECO] § 3.4. Si l'on fait la somme des inerties internes de tous les dipôles de la hiérarchie, on obtient l'inertie totale du nuage. Donc, si le tableau contenait tous les dipôles de la hiérarchie le total de la colonne IND vaudrait 1. Comme notre tableau ne contient que les dipôles centrés aux 8 noeuds supérieurs, on a un total inférieur à 1 : 830/1000. De façon précise, ce total, déjà trouvé sur la tableau afférent aux classes de la partition retenue, n'est autre que le quotient par l'inertie totale du nuage $N(I \times T)$, de l'inertie interclasse de la partition retenue (qui est l'inertie interne du nuage des centres de la partition). Et de même, le total de la colonne CTD_a du tableau des dipôles est égal à celui de la colonne CTR_a du tableau des classes. (On vérifie, par exemple, que les totaux des colonnes CTD₁ du t. des dipôles et CTR₁ du t. des classes sont égaux à 930/1000).

N. B. L'utilisation du listage FACOR sera expliquée au § 2.3, en même temps que celle du listage VACOR.

2.2 Le listage VACOR : Le but des calculs dont les résultats sont imprimés sur ce listage est simple. A toute classe est associé un système moyen de pourcentages des importations reçues de chaque fournisseur principal (e.g. la classe 205 reçoit des USA 13% de ses importations) : il s'agit de trouver, parmi ces pourcentages, ceux qui de par leur valeur anormalement forte ou faible, caractérisent la classe.

2.2.0 Système d'axes orthogonal défini par les variables : La CAH porte sur un nuage $[I \times T]$ de profils sur J . Dans l'espace P_J des profils sur J , l'analyse des correspondances construit un système de coordonnées orthonormées : l'origine est le centre de gravité du nuage ; les axes sont les axes principaux d'inertie, appelés axes factoriels ; les coordonnées sont les facteurs. Mais avant toute analyse, on a un système de coordonnées pour les profils : c'est le système de leurs composantes ; 8 dans le cas présent, soit : $\mathcal{Z}_{\text{Belgique}}$; $\mathcal{Z}_{\text{USA}}$; $\mathcal{Z}_{\text{Jordanie}}$; $\mathcal{Z}_{\text{Maroc}}$; $\mathcal{Z}_{\text{Sénégal}}$; $\mathcal{Z}_{\text{ToGo}}$; $\mathcal{Z}_{\text{Tunisie}}$; $\mathcal{Z}_{\text{CC(URSS)}}$.

Ces coordonnées n'ont pas pour origine le centre de gravité du nuage, car le 0 est une distribution de masse dont toutes les composantes sont nulles. Mais les axes sont orthogonaux, car même si la formule de la métrique du χ^2 comporte des coefficients, elle n'a pas de termes rectangles. Or, on a :

$$d^2(i, i') = \sum \{ (f_{ij}^i - f_{ij}^{i'})^2 / f_{.j} \mid j \in J \},$$

sans termes dans la somme qui soient des produits $(f_{ij}^i - f_{ij}^{i'}) (f_{ij'}^i - f_{ij'}^{i'})$ avec deux variables j et j' différentes. C'est pourquoi les calculs faits par FACOR relativement aux FACTeurs peuvent être repris dans VACOR relativement aux VARIables (plus précisément aux composantes des profils).

2.2.1 Les centres des classes

■ Les coordonnées : le centre de gravité du nuage n'est pas, nous venons de le rappeler, à l'origine du système d'axes de référence défini par les variables ; ses coordonnées se lisent sur la ligne 223, aux colonnes $\mathcal{Z}_{\text{BL}}$, $\mathcal{Z}_{\text{USA}}$, $\mathcal{Z}_{\text{JR}}$ etc. et ne sont autres que les composantes du profil sur J de la ligne de marge du tableau $(I \times T) \times J$ ou du

VACOR

COR=COSINU AU CARRE DE CETTE CLASSE A CET AXE VARIABLE

CTR=CONTRIBUTION DE CETTE CLASSE A CET AXE VARIABLE

LES CLASSES RETENUS

CLAS	A(N)	B(N)	QALT	INER	POID	EBL	COR	CTR	EUSA	COR	CTR	EJR	COR	CTR	EMR	COR	CTR	ESN	COR	CTR	ETG	COR	CTR
198	191	176	1000	250	78	0	8	26	126	75	98	118	1	2	1184	17	35	10	9	31	10	15	39
205	163	167	1000	49	49	125	1	1	134	110	28	10	21	11	1591	152	61	1201	591	397	18	37	19
206	171	123	1000	154	142	1	22	45	1914	641	515	10	20	32	172	152	190	10	27	56	1	43	70
210	192	181	1000	44	90	151	697	414	1434	28	7	10	44	20	1165	141	51	130	2	1	23	38	17
211	180	203	1000	86	140	14	32	37	153	147	66	15	23	21	1751	648	456	15	34	40	144	6	5
212	187	207	1000	37	143	140	6	3	1312	16	7	113	24	9	1347	1	0	155	27	14	109	120	46
213	200	209	1000	43	125	0	73	42	1169	222	50	162	100	45	1399	26	9	13	69	41	156	1	0
214	175	204	1000	93	112	199	129	161	171	91	44	11	23	23	1329	0	0	192	80	103	1254	588	571
215	195	208	1000	74	122	0	42	42	1580	199	77	1120	400	371	1226	40	29	138	0	2	18	60	46
216	211	205	1000	101	189	110	24	32	140	179	94	14	30	32	1709	605	499	157	16	22	134	17	18
217	212	214	1000	95	255	166	73	94	1250	66	53	18	29	29	1339	0	0	171	69	90	1173	439	438
218	215	206	1000	175	264	1	37	87	1764	550	510	159	43	80	1143	133	189	118	12	28	14	63	115
219	213	210	1000	32	214	163	155	67	1200	83	14	136	14	5	1301	20	5	114	71	31	142	30	10
220	216	217	1000	114	444	142	9	13	1206	191	114	16	51	61	1497	232	216	165	68	106	1114	144	171
221	220	219	1000	60	650	149	73	59	1230	382	119	116	42	26	1433	230	115	140	34	20	190	120	80
222	221	218	1000	21	922	135	0	2	1303	75	0	120	1	0	1350	17	3	140	9	3	166	15	3
223	198	222	0	0	1000	122	0	0	1355	0	0	127	0	0	1337	0	0	137	0	0	161	0	0

CLAS	A(N)	B(N)	QALT	INER	POID	ETH	COR	CTR	ECCP	COR	CTR
198	191	176	1000	250	78	144	0	0	1720	875	705
205	163	167	1000	49	49	141	0	0	18	86	14
206	171	123	1000	154	142	112	16	65	10	79	29
210	192	181	1000	44	90	133	3	4	1164	47	7
211	180	203	1000	86	140	116	20	47	122	87	25
212	187	207	1000	37	143	124	476	475	10	331	37
213	200	209	1000	43	125	147	1	1	1264	507	71
214	175	204	1000	93	112	146	0	1	18	80	26
215	195	208	1000	74	122	111	30	59	10	141	74
216	211	205	1000	101	189	122	13	37	116	116	33
217	212	214	1000	95	255	109	110	203	14	214	66
218	215	206	1000	175	264	112	26	125	10	129	73
219	213	210	1000	32	214	141	0	0	1222	626	65
220	216	217	1000	114	444	161	25	77	19	201	104
221	220	219	1000	60	650	154	30	49	178	72	14
222	221	218	1000	21	922	142	0	0	156	875	60
223	198	222	0	0	1000	142	0	0	1100	0	0

tableau I-J : 32 ; 355 ; 27 ; etc. (cf. [CAH ECO] RAPPEL III ou [CORR. ECO.] § 1.2) en effet cette ligne de marge est le cumul de toutes les lignes du tableau (I×T)×J.

Pour un centre c quelconque, la ligne des coordonnées n'est autre que le profil de la ligne cumulant les lignes des pays-années constituant la classe c , dans le tableau (I×T)×J. Pour voir en quoi c diffère du centre du nuage, on compare ses coordonnées à celles du centre 223, i.e. on considère le signe et l'amplitude des différences:

$\mathcal{E}BL(c) - \mathcal{E}BL(223)$; $\mathcal{E}US(c) - \mathcal{E}US(223)$; etc. . Par exemple, en comparant les coordonnées des centres 218 et 223, on voit que 218 a, par rapport à 223, un déficit en $\mathcal{E}BL$, un excédent très fort en $\mathcal{E}USA$, un excédent en $\mathcal{E}JR$, un fort déficit en $\mathcal{E}MR$ etc. . Mais de toutes ces qualités de la classe 218, certaines seulement sont caractéristiques de la classe : on les reconnaît à la valeur élevée du coefficient COR qui leur correspond (cf. *infra*).

■ COR (ou CO2) cosinus carré de l'angle que fait avec l'axe considéré le rayon joignant le centre $g = 223$ du nuage au centre c de la classe considérée. Par exemple, sur l'axe $\mathcal{Z}_{BL}$,

$$COR_{\mathcal{Z}_{BL}}(c) = \|\text{pr}_{\mathcal{Z}_{BL}}(g, c)\|^2 / \|g - c\|^2.$$

On notera que selon la formule de la distance du χ^2 ,

$$\|\text{pr}_{\mathcal{Z}_{BL}}(g, c)\|^2 \text{ a pour valeur : } (f_{\mathcal{Z}_{BL}}^c - f_{\mathcal{Z}_{BL}})^2 / f_{\mathcal{Z}_{BL}}.$$

Plus le rayon (g, c) est voisin d'un axe en direction, plus grande est la part de cet axe dans l'explication de l'écart du centre de la classe c au centre g . Par exemple, l'écart au centre $g = 223$ du centre 218 est expliqué à 55,8% ($COR = 558$) par son excédent en $\mathcal{Z}_{USA}$; à 13,3% par son déficit en $\mathcal{Z}_{MR}$ et à 12,9% par l'absence totale de $\mathcal{Z}_{CC}$ dans son profil ($\mathcal{Z}_{CC}(218) = 0$).

■ CTR : on projette sur l'axe considéré, e.g. $\mathcal{Z}_{BL}$, le nuage entier et la classe c considérée. Les projections $\text{pr}_{\mathcal{Z}_{BL}}(g)$ et $\text{pr}_{\mathcal{Z}_{BL}}(c)$ des centres g et c sont les centres de gravité des projections du nuage et de la classe c sur l'axe $\mathcal{Z}_{BL}$.

$CTR_{\mathcal{Z}_{BL}}(c)$ est l'inertie du centre $\text{pr}_{\mathcal{Z}_{BL}}(c)$ par rapport à $\text{pr}_{\mathcal{Z}_{BL}}(g)$ divisée par l'inertie totale de la projection $\text{pr}_{\mathcal{Z}_{BL}}(N(I))$ du nuage sur l'axe $\mathcal{Z}_{BL}$.

Par exemple pour la classe 218 et l'axe $\mathcal{Z}_{USA}$ on lit : $CTR_{\mathcal{Z}_{USA}}(218) = 510$; cela signifie qu'en projection sur l'axe $\mathcal{Z}_{USA}$, l'inertie de 218 par rapport au centre du nuage est les 510/1000 de l'inertie totale du nuage. L'inertie du nuage le long de l'axe sert seulement de référence: le centre c ne fait pas partie du nuage.

■ QALT : les 8 coordonnées étant imprimées, la qualité de représentation des centres est parfaite et l'on a $QALT = 1000/1000$ pour tous les centres. Certains programmes comportent une option permettant de ne retenir qu'une partie des variables (e.g. celles qui ont les plus fortes contributions dans l'analyse des correspondances). Dans ce cas la qualité de représentation peut descendre au-dessous de 1 comme dans FACOR lorsqu'on n'imprime pas la totalité des facteurs. Mais dans la pratique on imprime toutes les variables. Lorsque celles-ci sont trop nombreuses on les groupe en classes par CAH comme il est dit plus bas § 3.4.4.

Quoiqu'il en soit, on a toujours :

$$QALT(c) = \|\text{proj}_{[\text{axes}]}(g, c)\|^2 / \|g - c\|^2$$

où $\text{proj}_{[\text{axes}]}$ est mis pour : projection sur le sous-espace engendré par les axes retenus ; QALT est la somme des COR relatifs à ces axes.

■ PDS et INR ont exactement les mêmes définitions dans VACOR et dans FACOR : on peut vérifier que les valeurs imprimées dans ces colonnes coïncident dans les deux listages.

2.2.2 Les dipôles

■ Les coordonnées : comme dans FACOR, pour chaque noeud n et pour chaque axe est donnée la différence sur cet axe des coordonnées (i.e. des composantes du profil) de $A[n]$ et de $B[n]$. Le but est de mettre en évidence ce qui oppose les classes $A[n]$ et $B[n]$ en lesquelles se scinde la classe n . On considérera en priorité les coordonnées qui correspondent à un fort COD (cf. *infra*).

■ COD mesure comme dans FACOR, le cosinus carré de l'angle que fait le segment (A[n],B[n]), avec l'axe considéré ; ou encore, la part de la variable donnant son nom à l'axe dans l'explication de la séparation de A[n] et B[n]:

$$\text{COD}_{\text{BL}}(n) = \|\text{pr}_{\text{BL}}(A[n]) - \text{pr}_{\text{BL}}(B[n])\|^2 / \|A[n] - B[n]\|^2.$$

Par exemple, le partage de la classe 222 en 218 et 221 est expliqué à 55,6% par l'excès en ΣUS de 218 sur 221 :

$$\Sigma\text{US}(218) - \Sigma\text{US}(221) = 533/1000 ;$$

et à 17,3% par le défaut en ΣMR de 218 relativement à 221 :

$$\Sigma\text{MR}(218) - \Sigma\text{MR}(221) = -290/1000.$$

■ CTD donne le rapport de l'inertie du dipôle à l'inertie totale du nuage, en projection sur l'axe considéré.

■ POID, IND, QALD ont la même signification que dans FACOR ; les inerties étant toujours prises par rapport au centre de gravité $\bar{g}$ du nuage, centre qui n'est pas ici à l'origine des axes (à la différence de FACOR).

COD=COSINUS AU CARRÉ ENTRE LE DIPOLE ET AXE
CTD=CONTRIBUTION DU DIPOLE

LES NOEUDS RETENUS

NDEU	A(N)	B(N)	QALT	INER	POID	ESL	COD	CTD	EUSA	COD	CTD	EJR	COD	CTD	EMR	COD	CTD	ESN	COD	CTD	ETG	COD	CTD
223	198	222	1000	271	1000	1-35	8	28	1-357	75	106	1-10	1	2	1-166	17	38	1-40	9	34	1-66	15	42
222	221	218	1000	214	922	1-49	53	143	1-533	556	621	1-43	47	106	1-290	173	301	1-31	18	53	1-86	86	192
221	220	219	1000	87	658	1-21	19	22	1-74	20	9	1-30	43	39	1-196	150	106	1-51	92	110	1-72	112	102
220	216	217	1000	83	444	1-56	102	117	1-102	31	13	1-4	1	1	1-370	421	203	1-15	6	7	1-139	329	284
219	213	210	1000	55	214	1-151	526	389	1-265	140	43	1-62	105	61	1-235	122	55	1-27	14	11	1-33	13	8
218	215	206	1000	52	264	1-1	0	0	1-326	297	81	1-129	52	324	1-154	70	30	1-38	40	29	1-7	1	0
217	212	214	1000	35	255	1-59	151	70	1-142	81	15	1-11	7	2	1-18	1	0	1-38	55	26	1-145	497	179
216	211	205	1000	34	189	1-24	12	5	1-19	1	0	1-5	1	0	1-160	64	18	1-196	889	415	1-36	18	7

NDEU	A(N)	B(N)	QALT	INER	POID	ETN	COD	CTD	ECCP	COD	CTD
223	198	222	1000	271	1000	1-2	0	0	1-672	875	765
222	221	218	1000	214	922	1-43	30	173	1-79	39	27
221	220	219	1000	87	658	1-20	12	29	1-213	552	155
220	216	217	1000	83	444	1-67	109	242	1-12	1	0
219	213	210	1000	55	214	1-14	3	5	1-99	68	12
218	215	206	1000	52	264	1-1	0	0	1-0	0	0
217	212	214	1000	35	255	1-78	206	192	1-9	1	0
216	211	205	1000	34	189	1-24	12	11	1-22	4	0

2.3 Utilisation des listages FACOR et VACOR dans la lecture descendante de l'arbre

2.3.0 Principe de lecture : Le listing FACOR permet de tracer sur les plans d'analyse factorielle la carcasse des dipôles, laquelle, descendue assez bas (e.g. jusqu'à des classes de 20 individus) permet seule de voir le nuage si celui-ci comprend de nombreux individus. Le tableau des dipôles met en évidence les facteurs qui interviennent dans les dichotomies effectuées aux divers noeuds supérieurs de l'arbre. Au noeud

n , la classe n se scinde en les deux classes $A[n]$ et $B[n]$, et pour ce dipôle ($A[n]$, $B[n]$) on repère sur la ligne n , les COD α les plus forts. On voit ainsi sur quels plans se séparent au mieux les classes $A[n]$ et $B[n]$. Tandis que les valeurs élevées de COR α indiquent dans les directions de quels axes une classe s'écarte le plus de l'origine. Mais, quant à l'interprétation, l'explication des dichotomies et la caractérisation des classes par les facteurs suppose qu'on connaisse parfaitement la signification de ceux-ci. En fait on doit pour plus de précision recourir à l'explication directe par les variables d'après le listage VACOR.

Nous ne considérerons pas en détail toutes les dichotomies de l'arbre restreint définissant la partition retenue au § 1.2. Nous nous bornerons à une suite descendante de 3 noeuds à partir du sommet : le noeud 223 (ou sommet de l'arbre), le noeud 222 = $B(223)$ et le noeud 218 = $B[222]$; et avec chaque noeud, on considérera plus particulièrement l'une des classes filles en montrant comment elle se sépare du reste. Au § 3, on résumera l'essentiel de la définition des classes par un étiquetage de l'arbre utilisant le listage INF SUP.

2.3.1 Le noeud 223 et la classe 198 : Au sommet (223) l'ensemble $I \times T$ se scinde en deux classes : $B(223) = 222$ et $A(223) = 198$, très inégales entre elles, car 198 se compose exclusivement des 8 années de JEE, tandis que 222 comprend le reste de $I \times T$. Le point 223, centre de gravité de la classe 223, n'est autre que le centre du nuage $N(I \times T)$ des individus à classer. Le support du dipôle (222, 198) est aussi le support des rayons joignant le centre du nuage aux centres 222 et 198, et les $\cos^2$ des angles formés par cette droite avec les axes des variables (VACOR) ou des facteurs (FACOR) se lisent aussi bien sur les tableaux des centres des classes (lignes 222 ou 198, colonnes COR), que sur les tableaux des dipôles (ligne 223, colonnes COD).

Afin de caractériser complètement la classe 198 (et *ipso facto* sa complémentaire 222), nous lisons la ligne 198 du tableau des centres du listage VACOR. Parmi les COR ressort : celui de $\mathcal{E}CC$ qui vaut 875/1000; ce qui signifie que le rayon (g , 198) qui joint le centre du nuage g au centre 198 fait avec l'axe $\mathcal{E}CC$ un angle très faible, dont le $\cos^2$ est .875. Parmi les parts des diverses origines ($\mathcal{E}BL$, $\mathcal{E}USA$ etc.) des importations de JEE, données par les coordonnées du centre 198 sur les variables $\mathcal{E}BL$, $\mathcal{E}USA$ etc. (coordonnées qui, rappelons-le, sont les composantes du profil du centre 198), $\mathcal{E}CC$ vient bien en tête avec 728 = 72,8%; suivent $\mathcal{E}MR$ (184 = 18,4%) et $\mathcal{E}TN$ (044 = 4,4%).

Quant à la disposition spatiale du dipôle (222, 198) dans le système des axes principaux d'inertie, on lit dans le listage FACOR :

$$COR1(222) = COR1(198) = COD1(223) = 897 ; COD2(223) = 90 ;$$

$$F1(198) = -1,907 ; F2(198) = 0,603 ; CTD1(223) = 653.$$

La séparation du nuage en les deux classes 198 et 222 se voit clairement sur le plan 1×2 ($COD1(223) + COD2(223) = 987$: le dipôle est très proche du plan 1×2) ; de plus, l'inertie totale du nuage sur l'axe 1 est expliquée à 65% par la dichotomie entre ces deux classes ($CTD1(223) = 653$) : en d'autres termes, en projection sur l'axe 1, la partition de $I \times T$ en deux classes : 198 et 222 a une inertie interclasse de : $653 \times \lambda_1$, l'inertie intraclasse (ou somme des inerties de chacune des deux classes par rapport à son centre de gravité) étant donc de : $0,347 \times \lambda_1$ (car en projection sur l'axe 1, on a pour toute partition de $I \times T$: Intra + Inter = λ_1).

En consultant le tableau des données, on voit que le volume des importations de l'Europe de l'Est en provenance de l'URSS varie relativement peu d'année en année ; au contraire, les importations en provenance de l'Algérie peuvent varier du simple au double (e.g. de 73 à 74).

2.3.2 Le noeud 222 et la classe 218 : Le tableau des dipôles du listage VACOR montre que le noeud 222 se scinde dans la direction des axes l'USA et l'Algérie en les classes A[222] = 218 et B[222] = 221.

$$COD_{US}(222) = 556 \quad ; \quad COD_{MR}(222) = 173.$$

La classe 218 est associée à l'USA (suivi de l'Algérie cf. infra) ; la classe 221 est associée à l'Algérie et tous les exportateurs autres que l'USA, l'Algérie et l'URSS, c'est-à-dire tous les exportateurs africains, ont dans la classe 221 un % supérieur à la moyenne générale :

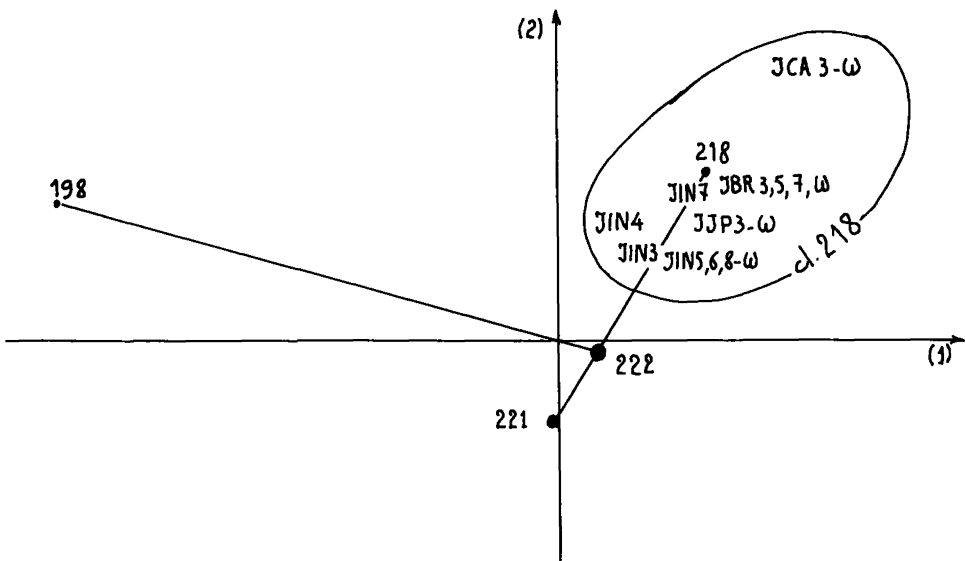
$$COR_{US}(218) = 558^+ ; \%US(218) = 76,4\% ; COR_{MR}(218) = 133^- ; \%MR(218) = 14,3\%$$

$$COR_{US}(221) = 382^- ; \%US(221) = 23\% ; COR_{MR}(221) = 230^+ ; \%MR(221) = 43,3\%$$

$$\%US(223) = 35,5\% \qquad \%MR(223) = 33,7\%$$

Nous avons affecté $COR_{US}(218)$ d'un signe + en indice supérieur parce que la coordonnée correspondante (i.e. le % de US) dépasse la moyenne (i.e. le % de US dans la classe 223) ; pour une raison analogue $COR_{MR}(218)$ est affecté d'un signe - (le % de MR dans la cl. 218 est inférieur à la moyenne).

Quant à la disposition spatiale des classes 218 et 221, le listage FACOR montre que le dipôle 222 est orienté suivant une droite du plan 1x2, (plus proche de l'axe 2 que de l'axe 1) ; les classes 221 et 218 se plaçant respectivement sur la partie négative de l'axe 2 et dans le quadrant ($F_1 > 0, F_2 > 0$) (dans les résultats donnés ci-dessous, le signe - qui affecte $COR_2(221)$ signifie que $F_2(221) < 0$; le signe + de $COR_1(218)$ signifie que $F_1(218) > 0$ etc.) :



$$\text{COD1}(222) = 295 ; \text{COD2}(222) = 677$$

$$\text{COR1}(221) = 006 ; \text{COR2}(221) = 969^-$$

$$\text{COR1}(218) = 466^+ ; \text{COR2}(218) = 507^+.$$

Dans le plan 1x2 la classe 218 apparaît très concentrée ; la classe 221 plus étalée s'étend jusqu'à la masse dense de 218 : nous y reviendrons en énumérant le contenu de la classe 218.

2.3.3 Subdivisions de la classe 218 ; la classe 215 : En moyenne, d'après le profil de son centre de gravité, la classe 218 est caractérisée par une part prépondérante de $\mathcal{E}\text{USA}$. On verra cependant que cette prépondérance ne s'étend pas à toutes les subdivisions, lesquelles s'expliquent principalement par un taux variable de $\mathcal{E}\text{Jordanie}$ (et de $\mathcal{E}\text{MaRoc}$).

Sur la ligne 218 du tableau des dipôles du listage VACOR, on lit :

$$\text{COD}_{\text{US}}(218) = 297 ; \text{COD}_{\text{JR}}(218) = 592 ;$$

et le tableau des centres des classes permet de placer les deux descendants immédiats de 218 sur les axes $\mathcal{E}\text{US}$ et $\mathcal{E}\text{JR}$:

	$\mathcal{E}\text{US}$	COR_{US}	$\mathcal{E}\text{JR}$	COR_{JR}
A[218] = 206	914	641 ⁺	000	020 ⁻
B[218] = 215	588	199 ⁺	128	480 ⁺

On remarque au passage qu'un $\mathcal{E}\text{JR}$ de 12,8 donne à la classe 215 un COR_{JR} de 480 ; tandis qu'un $\mathcal{E}\text{US}$ plus de 4 fois plus élevé que le $\mathcal{E}\text{JR}$, ne donne à cette même classe qu'un COR_{US} inférieur à la moitié de COR_{JR} ; la raison en est que d'une part les COR sont calculés d'après les différences des coordonnées à la moyenne générale de $223 = \text{I} \times \text{T}$; et que d'autre part dans le calcul des COR comme dans la formule de distance du χ^2 , une variable est affectée d'un coefficient d'autant plus important que son poids moyen est plus faible. De ce point de vue il faut considérer les différences :

$$\mathcal{E}\text{US}(215) - \mathcal{E}\text{US}(223) = 588 - 355 = 233 ;$$

$$\mathcal{E}\text{JR}(215) - \mathcal{E}\text{JR}(223) = 120 - 027 = 093 ;$$

les différences élevées au carré étant ensuite affectées des coefficients respectifs $(\mathcal{E}\text{US}(223))^{-1}$ et $(\mathcal{E}\text{JR}(223))^{-1}$; ce deuxième coefficient étant plus de 150 fois plus grand que le premier.

Quant à la disposition spatiale, le partage de la classe 218 se fait suivant l'axe 4 : $\text{COD4}(218) = 603$, axe qui est précisément celui de la Jordanie ($\text{COR4}(\text{JR}) = 785$; $\text{CTR4}(\mathcal{E}\text{JR}) = 817$).

Il vaut la peine de considérer en détail la classe 215 avec ses subdivisions retenues au § 1.2. Nous le ferons en recourant à un listage VACOR plus complet que celui publié ici. Au sein de la cl. 218 dont le centre est si nettement associé à $\mathcal{E}\text{USA}$, la classe 215 = B(218) est caractérisée par la pénétration de la Jordanie ; pénétration qui est maxima dans 195 = A[215], classe pour laquelle la part de $\mathcal{E}\text{US}$ est même inférieure à la moyenne générale, et, fait plus remarquable encore, inférieure à la part de $\mathcal{E}\text{JR}$. C'est encore suivant la pénétration plus ou moins grande de la Jordanie que se scinde la classe 208 = B[215].

De façon précise, on extrait du listage VACOR les valeurs suivantes relatives aux dipôles :

$$\text{COD}_{\text{JR}}(215) = 918 ; \text{COD}_{\text{JR}}(208) = 777 ; \text{COD}_{\text{US}}(208) = 134 ;$$

avec, pour les profils des centres des classes le sous tableau suivant:

	JR	COR _{JR}	US	COR _{US}
A[215] = 195 = {IN3,4}	435	954 ⁺	280	002 ⁻
B[208] = 194 = {IN5,6,7,8,w}	176	719 ⁺	508	059 ⁺
B[215] = 208	110	370 ⁺	606	264 ⁺
A[208] = 183 = {JP3-w;IN7}	082	179 ⁺	648	398 ⁺

En lisant le tableau des données brutes relatives à l'Inde, on voit qu'après les deux années 73 et 74 où JR l'emporte sur US, viennent deux années 75, 76 où le tonnage des importations en phosphates baisse (comme en bien d'autres pays) du fait de la hausse des prix ; puis, à partir de 1977, les importations de l'Inde s'accroissent, surtout celles en provenance des USA. L'ensemble de ces observations justifie pleinement le choix fait plus haut de retenir dans la hiérarchie restreinte les subdivisions de la classe 215.

3 Étiquetage de l'arbre : En analyse des correspondances, la représentation simultanée des deux ensembles I et J fournit à la fois des résultats et leur interprétation. La classification n'a pas cette symétrie : il revient à l'utilisateur de porter sur les classes d'individus, pour expliquer celles-ci, des informations relatives aux variables ; mais cet étiquetage (fondé sur le listage VACOR) suit désormais des règles précises.

3.1 Principes généraux : La représentation arborescente d'une CAH est universellement adoptée. En revanche, l'étiquetage des noeuds et des branches progresse avec les aides à l'interprétation, dont les principales indications sont résumées dans les étiquettes.

Après lecture descendante de l'arbre (cf. § 1.2) on ne retient ici qu'une partition en 15 classes. L'arbre est dessiné en plaçant en marge supérieure le profil du centre de gravité du nuage $N(I \times T)$ sur l'ensemble des 8 exportateurs principaux ; et en marge gauche, le contenu des 15 classes terminales de la hiérarchie restreinte retenue (ou classes de la partition), ainsi que le n° de chacune de ces classes. Ainsi, à quelques variantes près, l'arbre est disposé comme sur le listage issu de l'ordinateur. Mais avant d'achever le tracé et d'entreprendre l'étiquetage, il reste quelques choix à faire. D'une part l'ordre des classes terminales dans la marge gauche n'est pas absolument imposé : il dépend (cf. [CAH ECO] § 4.2) du choix pour chaque noeud de l'ordre dans lequel ses descendants sont appelés *Afiné* et *Benjamin* : même si ce choix ne modifie pas la hiérarchie des classes, il peut aider à disposer des classes dans un ordre plus satisfaisant : par exemple dans la classe 215 (cf. § 2.3.3 *in fine*), il est satisfaisant que les trois subdivisions retenues soient rangées dans l'ordre d'un % JOR-danie croissant : 183 (%JR = 8%) ; 194 (%JR = 18%) ; 195 (%JR = 43,5%) ; etc. . D'autre part, l'échelle de niveau des noeuds doit être assez large pour laisser place à l'étiquetage des branches : dans le cas présent l'échelle linéaire n'est conservée que pour les noeuds de numéro ≤ 221 ; au-delà, des traits coupant les branches rappellent que les dénivellations ont été diminuées ; mais on a noté les τ de quelques noeuds supérieurs afin d'indiquer les proportions exactes des niveaux.

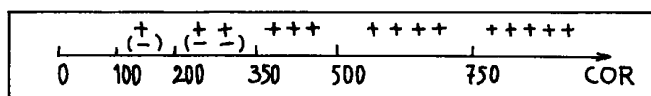
Quant à l'étiquetage de l'arbre, l'usage s'est introduit depuis plusieurs années de marquer d'après le programme VACOR des noms de variables affectés de signes : l'indication JR^{++++} signifiant par exemple dans le cas présent très fortes importations en phosphates de Jordanie ; l'indication CC^{-} , faibles importations reçues d'URSS etc. .

Initialement, l'usage des signes et plus encore leur multiplication a été laissé à l'initiative du dessinateur, sans normes précises. Comment en effet normaliser l'expression simultanée d'informations de deux sortes : des valeurs de variables et des contributions relatives à l'écart du centre de gravité, informations parallèles mais d'échelle différente. Comment avertir le lecteur que recevoir 51% de ses phosphates des USA n'est pas une caractéristique notable de la cl. 194 alors que pour la cl. 187 un pourcentage pourtant inférieur de 48% reçu des USA est une caractéristique notable en ce sens que $COR_{US}(187) = 157$ (alors que $COR_{US}(194) = 059$) ? Cependant que 18% est un niveau très élevé relativement à la Jordanie pour la classe 194 ($COR_{JR}(194) = 719$) et un niveau faible relativement aux USA pour la classe 190 ; etc. .

La solution semble être : noter à la fois les deux sortes d'information. D'une part donner des valeurs précises de % (e.g. $JR = 18\%$ pour la cl. 194 ; ou des inégalités : $TN > 10\%$ pour la cl. 212 dont tous les descendants figurés sur l'arbre reçoivent au moins 10% de leurs phosphates de Tunisie, ce qui est plus de 2 fois supérieur à la moyenne) et d'autre part, ajouter des signes + (e.g. $JR^{++++} = 18\%$; $TN^{++} > 10\%$) pour indiquer dans quelle mesure cette particularité contribue à l'écart entre le centre de la classe considérée et le centre de gravité du nuage.

De façon précise, on suivra les trois principes suivants :

1°) Marquer pour chacune des classes terminales de la hiérarchie restreinte les % afférents à toutes les variables qui apportent à l'écart du centre de la classe considérée au centre du nuage un $COR \geq 100$: l'ordre de grandeur du COR étant indiqué par des signes suivant une échelle d'intensité telle que celle proposée ici (et dont on recommande l'usage général).



Echelle pour l'étiquetage des arbres.

Ainsi, pour la classe 190 on a $MR^{++++} = 70\%$, avec 4 signes parce que $COR_{MR}(190) = 600$ (donc compris entre 500 et 750), et ces signes sont des signes + parce que la part de MR dans le profil de la cl. 190 dépasse le pourcentage moyen (34%) (qui figure dans la partie supérieure du graphique de l'arbre étiqueté).

Pour la classe 205, on a $US^{-} = 13\%$: un seul signe parce que $COR_{US}(205) = 110$; et ce signe est - parce que 13% est inférieur au pourcentage moyen (35,5%) de US.

2°) Transporter dans la mesure du possible vers les noeuds supérieurs de l'arbre les informations valant pour tous leurs descendants retenus. Par exemple, on lit avec le noeud 216 : $MR^{+} > 59\%$ parce que cette mention vaut pour les quatre subdivisions 205, 180, 190, 199 ; mais des % plus forts sont notés sur deux de ces classes : sur 190 , $MR^{++++} = 70\%$ et sur 180, $MR^{++++} = 92\%$.

3°) Outre les % liés à des contributions COR très élevées, on note les % les plus élevés, de telle sorte que pour chaque classe terminale, le total des % notés atteigne 90% ou plus : par exemple, la classe 194 (IN5,6,8,9,w) est d'abord étiquetée JR⁺⁺⁺⁺ = 18% ; on ajoutera US > 51% qui vaut pour la classe 208 et ses descendants ; et MR > 22% qui vaut pour toutes les subdivisions retenues de la classe 215 ; soit au total : 18 + 51 + 22 = 91 > 90. Il est essentiel de noter ainsi les % afférents à de gros fournisseurs (ici, US, MR) qui tout en contribuant généralement peu à la caractérisation des classes ont, en revanche une part prédominante dans la masse totale de celles-ci.

Grâce à ces conventions graphiques fondées sur des calculs statistiques précis, l'arbre étiqueté devrait être une réplique ordonnée, hiérarchisée, directement lisible du tableau de données (cf. § 4). Aussi engageons-nous le praticien à faire cette ultime opération avec la plus grande exactitude. On propose au § 3.2 suivant une lecture de l'arbre étiqueté selon les principes qu'on vient d'énoncer.

3.2 Lecture de l'arbre étiqueté : Deux exportateurs occupent chacun un tiers du marché mondial des phosphates : ce sont les USA et le Maroc. Le reste du marché (Océanie exclue cf. [CORR ECO], § 1.0 Rem 2) est partagé entre cinq producteurs d'importance moyenne (en tête CCCP=URSS, suivi de ToGo, TuNisie, SÉNégal, JoRdanie) et un redistributeur (la BeLgique) ; les exportateurs représentant 2% ou moins du marché (e.g. l'Algérie) n'étant pas considérés ici.

Pour présenter les grandes lignes du commerce mondial des phosphates, nous partons d'une partition en 5 classes chacune caractérisée par un titre.

198 : *prédominance de CC* : la Bulgarie, la Hongrie et la R.D.A. (Allemagne de l'Est) réunis ici sous le sigle EE (Europe de l'Est) rejoignent de l'URSS (CC) les trois quarts de leurs phosphates, le reste venant du Maroc, de Tunisie, mais non des USA (%US = 2,6%).

219 : *présence de CC* : la classe 219 qui comprend tous les profils annuels de trois pays : Deutschland (Allemagne de l'Ouest) RouManie, PoLogne, se distingue par un %CC supérieur à la moyenne mondiale ; avec, selon les cas une contribution majeure des USA (pour DL) ou du Maroc (pour PL et RM). Sur la période étudiée, %US progresse aussi bien pour DL que pour PL et RM ; tandis que %BeLgique diminue pour DL ; %MaRoc et %CC pour RM et PL. Mais le fait le plus remarquable est l'introduction massive des phosphates de JoRdanie en RouManie (dès 1975) puis en PoLogne (dès 1975, mais surtout à partir de 77). Ces faits apparaissent nettement en subdivisant les classes 210 et 213 ou en consultant le listage INF SUP (cf. *infra* § 3.3.3).

218 : *prédominance des USA* (sauf dans 195) : la classe 218, composée de tous les profils annuels des importateurs non européens (à l'exception du BRésil 74 et 76) est caractérisée par le fait que %US dépasse toujours 28%, tandis que %MaRoc est toujours inférieur à 28%. Bien plus, %US est supérieur à 50% si on excepte deux années de l'Inde ({IN3, 4} = 185) pour lesquelles %JoRdanie est exceptionnellement fort, ce qui lie IN3, 4 aux autres années de l'Inde où %JR est toujours relativement élevé.

Pour les profils exceptionnels BR4, 6 (qui manquent à 218), %US dépasse 50%, mais %MR = 36% ; et surtout, (ce qui fait sortir BR4, 6 de la classe 218), %TuNisie ≈ 10% (alors que la moyenne de TN est seulement 4,2%). Il est à peine utile de noter que US est le fournisseur exclusif du Canada en phosphates.

217 : *prédominance de l'Afrique* (sauf dans 187) : la classe 217 comprend tous les profils annuels de France et de Nederland, ainsi que les quatre profils de fin de période d'Italie (IT7- ω) et les deux profils du Brésil déjà cités (BR4, 6). Au sein des subdivisions de la classe 217, la part du Maroc ne s'écarte guère du tiers ($\% \text{Maroc} \approx 33\%$); s'y ajoutent des % relativement importants du ToGo, du Sénégal, et de la Tunisie. La subdivision 187 (IT7,9, ω ; BR4, 6) est particulière en ce que %US y atteint, en moyenne, 48% (avec plus de 50% dans BR4, 6) ; mais 187 s'attache 207 (FR6- ω ; IT8), qui est dans 217, par un fort % Tunisie (10%). A noter dans 214 (NL3- ω ; FR3-5) : ToGo = 25%.

216 : *prédominance du Maroc* : la classe 216 comprend tous les profils annuels de trois pays importateurs (UK, Espagne, Belgique) avec les quatre premières années de l'Italie (IT3-6). La part du Maroc est partout supérieure à 59% ; elle atteint 92% dans la subdivision 180 (Espagne3- ω) : il faut se souvenir que le gisement de Bu-Craa (compris dans un territoire qui sous le nom de Rio de Oro fut par le passé sous la souveraineté de l'Espagne) est aujourd'hui exploité par une société hispano-marocaine. A noter dans 205 (UK-3 ω) : %SN = 20% (alors que la moyenne de % Sénégal est seulement 3,7%).

3.3. *Le listage INF SUP* : L'étiquetage d'un arbre suivant les principes posés au § 3.1 peut se faire d'après le listage VACOR (cf. § 2.2). Toutefois l'étiquetage peut être simplifié et enrichi de nuances en utilisant les listages INF SUP (de Y.L. Cheung) publié et expliqué ci-dessous.

3.3.1 *Spécification du listage demandé* : Le listage INF SUP est relatif à une partition que l'utilisateur choisit à sa guise suivant les principes du § 1.2. Pour spécifier cette partition, il suffit de donner la liste des numéros des classes que l'on désire faire figurer dans la hiérarchie choisie. Ici on a donné la liste {171,194,210,199, 207}. Le programme détermine par lui-même quelle est la hiérarchie la moins fine (i.e. la plus simple : contenant le moins de classes possibles) contenant ces classes ; et les tableaux qu'il construit sont relatifs respectivement l'un aux noeuds, l'autre aux classes terminales de cette hiérarchie retenue.

3.3.2 *Tableau INF SUP relatif aux noeuds* : Ce tableau comprend d'abord un bloc de quatre colonnes initiales donnant respectivement :

- le n° I du noeud ; le n° de A[I] ; le n° de B[I] ; le POIDS du noeud I. Viennent ensuite des blocs de 5 colonnes afférents chacun à l'une des variables. Considérons par exemple le bloc Σ US relatif aux exportations de phosphates des USA.

- Une colonne centrale EUS donne pour chaque noeud la valeur en millièmes de la part des USA dans ses importations de phosphates : on a e.g. 355 = 35,5% pour le noeud 223 (qui correspond au nuage tout entier) ; on a 206 = 20,6% pour le noeud 220 qui ne contient qu'une seule classe (187) où la part des USA est supérieure à la moyenne ; etc..

De part et d'autre de la colonne EUS on a deux colonnes : MIV et MXV, donnant pour chaque noeud N respectivement la valeur MINIMA et la valeur MAXIMA prise par la Variable EUS (plus précisément la composante EUS de profil) sur l'ensemble des descendants du noeud N retenus dans la hiérarchie restreinte. Par exemple, sur la ligne 223 on a MXV (EUS) = 999, valeur de EUS qui est réalisée par la classe 123 (Canada 6- ω) ; et MIV(EUS) = 026 valeur de EUS afférente à la classe 198 (Europe de l'Est6- ω , où prédomine ECC). Sur la ligne 220 on a MXV(EUS) = 484 (réalisé par la classe 187) et MIV(EUS) = 061 (réalisé par la cl. 180 = SP3- ω).

■ Enfin, le bloc EUS est bordé par deux colonnes : à gauche MIC et à droite MXC. Ces colonnes donnent pour le noeud N respectivement le Minimum et le Maximum de la Contribution relative (COR = CO2) de la variable EUS à l'écart au centre de gravité sur l'ensemble des descendants retenus du noeud N. Par exemple, sur la ligne 220 on a MIC(EUS) = 067 (réalisé par la cl. 207) et MXC(EUS) = 161 (réalisé par la cl. 180)

L'intérêt de ce tableau est clair. On y lit en effet sur une seule ligne les indications susceptibles d'être portées sur un noeud sans avoir à s'assurer (en consultant diverses lignes du listage VACOR) que selon le deuxième principe posé au § 3.1 ces indications valent pour tous les descendants conservés de ce noeud.

LES NOEUDS RETENUS

MIV-MINIMUM DE LA VARIABLE POUR LES DESCENDANTS RETENUS DE CE NOEUD

MIV-MAXIMUM DE LA VARIABLE POUR LES DESCENDANTS RETENUS DE CE NOEUD

MIC-MINIMUM DU COS AU CARRE POUR LES DESCENDANTS DE CE NOEUD

MXC-MAXIMUM DU COS AU CARRE POUR LES DESCENDANTS DE CE NOEUD

NOEU(A1):B(1):POID	MIC	MIV	EBL	MIV	MXC	MIC	MIV	EUSA	MIV	MXC	MIC	MIV	EJR	MIV	MXC	MIC	MIV	EMR	MIV	MXC
223 :198 :222 :112 :1	0	32	151	697	12	26	355	999	645	10	0	27	435	954	10	0	337	917	661	1
222 :221 :218 :104 :1	0	35	151	697	12	61	383	999	645	10	0	28	435	954	10	0	358	917	661	1
221 :220 :219 :74 :1	0	49	151	697	12	61	270	484	222	10	0	16	62	180	10	165	433	917	661	1
220 :216 :217 :50 :1	0	42	99	129	167	61	206	484	161	10	0	6	30	65	10	329	497	917	661	1
219 :213 :210 :24 :73	0	63	151	697	12	169	280	434	222	144	0	36	62	180	126	165	381	399	141	1
218 :215 :206 :30 :5	0	1	3	53	12	280	764	999	645	115	0	59	435	954	12	0	143	278	187	1
217 :212 :214 :22 :29	5	66	99	129	167	171	250	484	157	12	1	8	23	25	10	329	339	348	1	1
216 :211 :205 :28 :1	0	18	25	77	110	61	149	225	161	10	0	4	33	65	152	391	789	917	661	1
215 :195 :208 :16 :5	0	0	0	53	12	280	588	648	398	1179	82	128	435	954	12	221	226	278	66	1
212 :187 :207 :11 :29	5	48	54	79	167	246	312	484	157	12	9	13	23	25	11	347	347	348	1	1
211 :180 :203 :20 :15	0	4	15	77	114	61	153	225	161	10	0	5	30	65	1546	615	751	917	661	1
208 :194 :183 :14 :29	0	0	0	57	159	588	686	648	398	1179	92	110	176	719	132	221	223	228	66	1
206 :171 :123 :14 :18	0	1	3	32	1608	786	914	999	645	115	0	0	32	184	0	72	181	187	1	1
203 :190 :199 :12 :15	0	4	15	77	114	191	215	225	121	10	0	0	30	65	1546	615	638	694	688	1

NOEU(A1):B(1):POID	MIC	MIV	ESN	MIV	MXC	MIC	MIV	ET6	MIV	MXC	MIC	MIV	ETN	MIV	MXC	MIC	MIV	ECPP	MIV	MXC
223 :198 :222 :112 :0	0	37	201	591	11	0	61	254	588	10	0	42	132	383	117	0	188	728	875	1
222 :221 :218 :104 :0	0	48	201	591	11	0	66	254	588	10	0	42	132	383	117	0	56	264	587	1
221 :220 :219 :74 :2	0	48	201	591	11	5	98	254	588	10	2	54	132	383	147	0	78	264	587	1
220 :216 :217 :50 :12	0	65	201	591	122	5	114	254	588	10	2	61	132	383	172	0	9	51	365	1
219 :213 :210 :24 :12	3	14	38	69	11	23	42	56	38	11	33	41	47	3	147	164	222	264	587	1
218 :215 :206 :30 :0	0	18	55	43	110	0	4	12	71	12	0	12	33	61	117	0	0	0	178	1
217 :212 :214 :22 :72	0	71	92	88	148	31	173	254	588	10	46	89	132	383	188	0	4	0	365	1
216 :211 :205 :28 :12	0	57	201	591	122	5	34	86	80	10	2	23	45	31	172	0	16	51	172	1
215 :195 :208 :16 :0	7	38	55	0	110	0	8	12	65	12	0	11	33	61	117	0	0	0	178	1
212 :187 :207 :11 :72	0	55	73	75	148	31	189	139	206	1273	181	124	132	383	217	0	0	0	365	1
211 :180 :203 :20 :12	0	5	28	58	124	5	44	86	80	10	2	16	45	31	172	0	22	51	172	1
208 :194 :183 :14 :0	34	48	55	0	154	0	8	12	65	12	3	12	33	61	197	0	0	0	178	1
206 :171 :123 :14 :20	0	0	0	47	132	0	1	1	71	14	0	12	39	23	168	0	0	0	126	1
203 :190 :199 :12 :12	4	9	20	68	126	5	63	86	80	10	19	26	45	31	173	0	36	51	172	1

3.3.3 *Tableau INF SUP relatif aux classes* : Formé comme le précédent de blocs de cinq colonnes afférents chacun à une variable, ce tableau permet de nuancer les informations relatives aux classes terminales

inscrites sur l'arbre. Considérons par exemple la mention $SN^{++++} = 20\%$ afférente à la classe 205(UK3- ω). Dans le bloc de la variable ESN on lit à la ligne 205 que cette valeur moyenne de 201(20%) est encadrée

par deux valeurs extrêmes : d'une part un minimum MIV de 135 (13%) réalisé par le profil de l'individu IMI = UK74 ; d'autre part un maximum MXV de 261 (26%) réalisé par l'individu IMX = UK79. Ici l'année de IMI (74 : la deuxième année de la période étudiée) et celle de IMX (79 : l'avant-dernière année de la période) suggère de plus une certaine tendance à la croissance. Parfois on note des variations de beaucoup plus grande amplitude. Ainsi, au sein de la cl. 213, la part de EUS varie de zéro (pour RM74) à 384 (pour RM79). En descendant, dans la lecture de la CAH, au-dessous du noeud 213, on voit que celle-ci se partage en deux classes 200 et 209 :

$$A[213] = 200 = \{RM3,4; PL3-6\}; B[213] = 209 = \{RM5-w; PL7-w\};$$

lesquelles correspondent respectivement au début et à la fin de la période, la part de US allant croissant.

LES CLASSES RETENUES

CLAS	A(I)	B(I)	POID	IMI	MIV	EBL	MXV	IMX	IMI	MIV	EUSA	MXV	IMX	IMI	MIV	EJR	MXV	IMX	IMI	MIV	ENR	MXV	IMX
123	110	1118	18	CA77	0	0	0	CAB0	CA74	992	999	1000	CA77	CA80	0	0	0	CAB0	CA80	0	0	0	CAB0
171	186	1158	16	BR80	0	3	12	BR73	BR75	724	785	831	BR73	BR80	0	0	0	BR80	BR73	120	181	228	BR75
180	143	1153	18	SP77	1	4	16	SP74	SP78	14	61	124	SP80	SP80	0	0	0	SP80	SP80	819	917	974	SP78
183	168	1164	19	IN77	0	0	0	JP78	JP76	623	648	697	JP75	JP75	41	82	117	IN77	IN77	193	221	252	JP76
187	154	1165	15	IT77	0	5	26	BR74	IT80	418	484	500	BR76	BR76	0	23	51	IT80	IT79	296	348	392	IT77
190	162	1138	14	IT76	4	15	19	IT73	IT76	87	191	281	IT73	IT74	0	30	86	IT76	IT73	621	694	768	IT75
194	157	1149	15	IN80	0	0	0	IN83	IN76	389	508	528	IN80	IN78	148	176	258	IN76	IN75	210	228	261	IN76
195	174	1173	12	IN74	0	0	0	IN74	IN74	228	280	338	IN73	IN73	347	435	513	IN74	IN74	258	278	381	IN73
198	191	1176	18	EE79	0	0	2	EE74	EE73	0	26	51	EE79	EE79	0	18	59	EE76	EE73	119	184	263	EE78
199	179	1140	18	BL80	0	0	0	BL80	BL76	79	225	332	BL78	BL80	0	0	0	BL80	BL79	541	615	746	BL76
285	163	1167	18	UK76	7	25	37	UK75	UK74	46	154	228	UK78	UK80	0	0	3	UK76	UK77	420	591	728	UK74
287	196	1201	16	IT78	7	54	65	FR79	FR76	144	246	314	FR78	FR80	0	9	52	IT78	FR77	263	347	499	IT78
210	192	1181	18	DL80	79	151	252	DL75	DL74	385	434	522	DL80	DL80	0	0	0	DL80	DL75	117	165	249	DL78
213	200	1209	116	RM77	0	0	0	RM80	RM74	0	169	384	RM79	RM74	0	62	185	RM79	RM79	0	399	671	RM74
214	175	1204	111	NL77	24	99	182	FR75	FR74	46	171	377	NL77	NL80	0	1	12	FR75	NL73	234	329	397	NL79

CLAS	A(I)	B(I)	POID	IMI	MIV	ESH	MXV	IMX	IMI	MIV	ETG	MXV	IMX	IMI	MIV	ETH	MXV	IMX	IMI	MIV	ECCP	MXV	IMX
123	110	1118	18	CA80	0	0	0	CAB0	CA80	0	1	0	CA74	CA80	0	0	0	CAB0	CA80	0	0	0	CAB0
171	186	1158	16	BR80	0	0	0	BR80	BR80	0	0	0	BR80	BR79	3	33	72	BR78	BR80	0	0	0	BR80
180	143	1153	18	SP79	0	0	2	SP80	SP78	0	15	53	SP80	SP80	0	2	18	SP78	SP80	0	0	0	SP80
183	168	1164	19	JP76	27	34	47	JP73	IN77	0	12	31	JP73	JP80	0	3	21	IN77	IN77	0	0	1	JP76
187	154	1165	15	IT79	6	8	16	BR76	BR76	0	31	59	IT80	BR76	51	101	118	IT80	BR76	0	0	0	BR76
190	162	1138	14	IT75	3	20	27	IT73	IT75	0	5	25	IT76	IT75	25	45	60	IT76	IT76	0	0	0	IT76
194	157	1149	15	IN80	32	55	181	IN75	IN80	0	0	0	IN80	IN76	0	33	54	IN80	IN80	0	0	0	IN80
195	174	1173	12	IN74	0	7	14	IN73	IN74	0	0	0	IN74	IN74	0	0	0	IN74	IN74	0	0	0	IN74
198	191	1176	18	EE80	0	0	0	EE80	EE80	0	0	0	EE80	EE73	0	44	87	EE74	EE78	631	728	881	EE73
199	179	1140	18	BL80	0	4	29	BL76	BL73	60	86	149	BL74	BL76	9	19	27	BL77	BL78	7	51	139	BL73
285	163	1167	18	UK74	135	281	261	UK79	UK80	0	0	61	UK77	UK75	23	41	57	UK73	UK80	0	0	0	UK80
287	196	1201	16	IT78	7	73	161	FR76	IT78	91	139	228	FR76	FR78	98	132	198	FR77	FR80	0	0	0	IT78
210	192	1181	18	DL76	0	38	65	DL79	DL76	0	23	40	DL80	DL75	11	33	75	DL77	DL78	64	164	246	DL73
213	200	1209	116	RM80	0	3	16	PL75	RM80	0	56	172	PL77	RM76	0	47	83	PL75	RM77	167	264	443	RM75
214	175	1204	111	NL80	0	92	159	FR75	FR75	179	254	313	NL79	NL78	2	46	89	FR73	NL80	0	0	37	NL73

3.4 Obstacles à l'étiquetage : On voudrait tout inscrire sur l'arbre (cf. § 4) : un schéma de structure, l'interprétation de ce schéma, les données elle-mêmes. Il est juste d'appeler l'attention sur certains obstacles, non sans suggérer des moyens de les contourner.

3.4.1 Individus et variables : On a présenté une classification des profils annuels des importateurs de phosphates : il s'impose de faire de même une classification de l'ensemble des exportateurs, i.e., en bref, une CAH sur $J \times T$ considéré comme ensemble des profils sur I . Plus précisément, on restreindra alors I à un sous-ensemble I_l des pays importateurs les plus lourds. Notre but étant ici que d'exposer les principes des aides à l'interprétation, on a considéré une seule CAH : nous renvoyons à [BLE] pour une étude complète du marché d'un produit.

3.4.2 Classes voisines du centre du nuage : Ainsi qu'on l'a dit au § 2.1.1 le listage porte $COR = 0$ dans la ligne afférente à la classe 223 (sommet de l'arbre). Mais pour une classe très proche du centre de gravité du nuage, sans toutefois coïncider avec celui-ci, on aura des valeurs de COR qui peuvent être élevées bien que n'ayant aucune signification réelle. On sera averti de cette anomalie par le fait que pour une telle classe INR est très faible par rapport à PDS (e.g. $INR < PDS/10$) et on se gardera alors de multiplier les signes +, même si $COR = 800$.

3.4.3 Classes et dipôles : Les informations inscrites sur l'arbre sont des valeurs de variables et des COR relatifs aux classes ; les différences et les COD relatifs aux dipôles (cf. §§ 2.1.2 et 2.2.2) n'y figurent pas. Car d'une part il est difficile de tout écrire sans être confus ; d'autre part les informations relatives à un dipôle N peuvent contrairement au 2-ème principe du § 3.1 ne pas valoir pour tous les descendants de $A[N]$ et $B[N]$. Ainsi, pour le noeud 222 on a $COD(US) = 556$; $D(US) = 533$; la classe $A[222] = 218$ s'oppose à la classe $B[222] = 221$ par un excès de US dans 218 relativement à 221 ; il n'en est pas moins vrai qu'au sein de 218 on a la classe 195 où $\%US = 28\%$; et au sein de 221, on a la classe 187 où $\%US = 48\%$! Ces anomalies résultent de ce que l'agrégation se fait en tenant compte de toutes les variables (e.g. c'est le fort taux de JR qui agrège 195 à 208 au sein de 218).

Pratiquement, les informations relatives aux dipôles servent d'abord à saisir rapidement la structure de la CAH ; elles guident ensuite le choix des variables portées sur les classes (les variables à fort COD pour un noeud N ne devant jamais être oubliées dans les descendants de N) ; elles aident enfin à rédiger le commentaire.

3.4.4 Agrégation des variables : Lorsque les variables sont très nombreuses (e.g. plusieurs dizaines) le listage $VACOR$ devient inutile : les $\%$ sont très faibles ; les COR et les COD sont généralement émiettés en fractions inférieures à 100... On peut alors se contenter d'une explication synthétique par le seul listage $FACOR$ ou employer les procédés suivants (ici, on note $I \times J$ le tableau de correspondance de départ) :

1) Agréger les variables par CAH en un nombre de classes assez petit pour rendre $VACOR$ utilisable. D'où un nouveau tableau de correspondance croisant I et la partition sur J que nous noterons CJ : ce tableau $I \times CJ$ s'obtient en cumulant les colonnes j du tableau $I \times J$ rentrant dans une même classe c de la partition CJ retenue sur J . On fait alors d'après ce tableau un $VACOR$ ayant CJ pour ensemble de variables. En termes concrets, au lieu de considérer les $\%$ des importations reçues d'un pays j , on considérera les $\%$ des importations reçues d'une classe c de pays qui occupent sur le marché des positions voisines.

2) A partir du tableau $I \times J$ et de CAH sur I et sur J , constituer un nouveau tableau de correspondance croisant les partitions CI et CJ retenues respectivement sur I et sur J . Ce tableau s'obtient en cumulant les lignes du t . $I \times J$ afférentes à une même classe de CI et en cumulant les colonnes du tableau $CI \times J$ ainsi obtenu afférentes à une même classe de CJ . L'analyse factorielle de ce tableau $CI \times CJ$ donne en général une explication satisfaisante des classes de CI par les classes de CJ . Comme dans toute analyse de correspondance, on peut être amené à placer en supplémentaires des classes qui monopolisent un axe (cf. Amalberti et coll. in CAD Vol IX n° 4 § 4.3 p. 393 et Ait Hamlat in CAD Vol VIII n° 2 § 4 p. 202 sqq).

4 Conclusion : nécessité des aides à l'interprétation

Vers 1970, nous regardions avec intérêt des arbres non étiquetés que nous considérons presque aujourd'hui comme des schémas inintelligibles. Mais déjà alors, en dessinant les contours des classes sur des plans issus de l'a. des correspondances, nous avions des aides à l'interprétation.

A partir de 1974, (cf. M. Sadaka, thèse ; et M. Jambu) les calculs de contributions ont permis de chiffrer cette interprétation ; et avec les listages FACOR et VACOR (cf. J.P. Benzécri et M.O. Lebeaux in CAD Vol V ; 1980) on a pu reporter commodément les contributions sur l'arbre.

Puis les thèses de A. et L. Alawieh (cf. [ACIER], [HOUILLE], [BLE] in CAD Vol IX et X, 1984-85) il est apparu possible (malgré certains obstacles : cf. § 3.4) d'utiliser dans l'étiquetage de l'arbre les données mêmes du tableau soumis à l'analyse des correspondances et à la CAH.

Aujourd'hui un arbre étiqueté suivant les principes posés au § 3.1 (principe dont le listage INF SUP facilite la mise en oeuvre) peut rendre intégralement disponible l'information très riche mais à peine lisible que recèlent les tableaux ; des détails ponctuels (tel pays a en telle année fait tel achat) échappant seuls à l'étiquetage ; avec à l'horizon le projet d'une lecture hiérarchique parfaite des banques de données qui se multiplient de nos jours mais ne sont que ponctuellement consultées et non saisies dans leur ensemble.

5 Appendice : représentation du contenu des classes en CAH : Dans l'exemple traité, la CAH a abouti à construire de nombreuses classes pures contenant la suite des profils annuels d'un importateur. On peut avoir à présenter un système de classes dont la structure n'est pas aussi simple. Nous nous bornerons ici à faire quelques suggestions en renvoyant à des exemples publiés, ou en cours d'élaboration.

5.1 Histogrammes à base temporelle et histogramme à base spatiale : Considérons d'abord le cas où l'ensemble des individus à classer est comme dans le premier article, un produit $[I \times T]$, où I est un ensemble de pays, de régions, de produits etc. ; et T une suite de quelques années. Il se peut qu'une classe sans être de la forme $I \times \{t\}$, (i.e. sans coïncider avec l'ensemble des profils individuels afférents à une année t) , se compose par exemple quasi exclusivement de couples (i, t) avec pour t l'une des trois dernières années de la période considérée, cette particularité apparaîtra si on représente la classe q par



un histogramme de base T au-dessus de chaque année t, un créneau dont la hauteur est égale au nombre de paires (i,t) afférentes à cette année et contenues dans q. On pourra même inscrire dans les cellules de ce créneau les sigles des individus i : e.g. sur la figure, la classe q apparaît composée de (a,71) ; (a,74) ; (c,74) ; (d,74) ; (a,75) ; (c,75) ; (a,76) ; (d,76) ; (e,76). On peut aussi faire des histogrammes dont la base est l'ensemble I, à condition d'ordonner celui-ci en respectant les similitudes entre les i, par exemple suivant l'ordre d'un 1-er facteur. Pour des exemples cf. : G. Mourad [OPEP - OCDE], § 4.4 in CAD, Vol VIII p. 59 (1983) ; et S. MAIZA et H. Maklouf [MARIA-GES] § 3 in CAD Vol IX p. 107 (1984).

5.2 Classifications sur Mois x Ans : Ici Mois est l'ensemble des 12 mois m ; Ans une suite de quelques années a. La classification sur Mois x Ans est faite, e.g. (El Moussaoui ; thèse à paraître) en associant à chaque couple (m, a) la consommation de l'essence en France dans chaque département au cours du mois m de l'année a. L'ensemble Mois x Ans a une structure naturelle de produit : e.g., on considère 8 années consécutives ; c'est un rectangle 12 x 8. (le seul arbitraire qui reste dans la représentation de ce rectangle est le choix du mois auquel on fait débiter l'année). On peut reporter sur ce rectangle les classes de la partition retenue, en conservant les cellules (m, a) de chaque classe par des hachures de type et d'orientations différents ; comme on le fait sur une carte.

5.3 Cartographie : Si l'ensemble I des individus à classer est un ensemble d'unités territoriales, il n'est que naturel, de reporter les classes q sur la carte : e.g. si I est l'ensemble des dép. français, et que la classe q comprend 7 départ. on pourra choisir de couvrir ceux-ci de hachures verticales ; tandis que pour une autre classe q' on aura des hachures horizontales ; ou un semis de triangles etc. (cf. S. Tibeiro [SRU.] § 2.2.1 in CAD Vol IX p. 410 (1984)). La règle d'or du choix des trames ou des couleurs d'une carte étant "de choisir une représentation des classes par des couleurs de telle sorte qu'à des classes voisines correspondent des teintes voisines ; et qu'on ne suggère pas d'annulation erronée en attribuant des teintes vives à des classes très distinctes". ce qui est dit teintes valant aussi bien pour les hachures et autres trames en noir et blanc ; cf. M.-M. Thomassin [ANA. CART.] § 3.1 in CAD Vol X pp 45 sqq (1985).

5.4. Histogrammes des profils des classes : Enfin aux classes correspondent des ensembles d'individus décrits par des variables telles que âge, dépense, durée de maladie... niveau d'instruction ; on peut à chaque classe associer un ou plusieurs histogrammes en décrivant le contenu quant à ces variables ; avec comme au § 5.1, la possibilité de porter des informations sur les cellules dont se compose chaque créneau de l'histogramme : cf. I. & R.-M. Cohen Hallaleh [IRAN] CAD Vol III 58,59 (1978) M. Sleiman [ENQ. MAROC] CAD Vol VIII pp 415, 417 (1983) J.J. Denimal [PRESC. BIOL.] in CAD Vol VIII (1983) et [PATHOLOGIE PECHE] in CAD Vol IX (1984).