JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

J.-P. BENZECRI

F. BENZECRI

Introduction à la classification ascendante hiérarchique d'après un exemple de données économiques

Journal de la société statistique de Paris, tome 126, nº 1 (1985), p. 14-34

<http://www.numdam.org/item?id=JSFS_1985__126_1_14_0>

INTRODUCTION A LA CLASSIFICATION ASCENDANTE HIÉRARCHIQUE D'APRÈS UN EXEMPLE DE DONNÉES ÉCONOMIQUES (*)

J.-P. BENZECRI et F. BENZECRI
Laboratoire de statistique, Université Pierre-et-Marie- Curie (**)

Partant d'un ensemble d'individus, la classification ascendante hiérarchique construit un système de classes emboîtées dont l'hétérogénéité s'accroît avec la taille. Cette construction dépend de la formule mathématique choisie pour calculer la quantité critère, mesure de cette hétérogénéité. Le présent exposé considère plusieurs critères, en montrant les avantages du critère de l'inertie, lequel est choisi pour traiter complètement un petit exemple.

Starting from a set of individuals, a nested system of classes is built, according to a criterion chosen as the measure of the diversity of a class. Several criteria are introduced among which the criterion of inertia is selected to complete the classification for a small scale example.

Par le format des données et l'interprétation géométrique qui en est faite, la classification ascendante hiérarchique (CAH) ne se sépare pas de l'analyse des correspondances. Elle en utilise même le plus souvent les résultats. Comme nous l'avons fait ailleurs (1) pour l'analyse des correspondances, nous exposerons la CAH sur un exemple simple. Celui dont nous nous servirons ici est extrait d'une étude sur le commerce mondial des phosphates (2).

La donnée de base est un tableau de correspondance $I \times J$ croisant l'ensemble I des 14 principaux pays importateurs de phosphates avec l'ensemble J des 8 principaux pays exportateurs. On se propose d'effectuer une CAH sur l'ensemble I.

L'exposé débute par un § 0 consacré à la présentation des données et au rappel des principes géométriques de l'a. des c. Après avoir expliqué au § 1 ce qu'on entend par Classification Ascendante Hiérarchique, CAH, on considère au § 2 les critères d'après lesquels est construite une hiérarchie de classes; et plus particulièrement, au § 3, le critère de l'inertie utilisé ici. Le § 4 montre sur l'exemple la présentation usuelle d'une CAH sur les listages issus de l'ordinateur.

0. LES DONNÉES ET LE CADRE GÉOMÉTRIQUE

0.1. *Le tableau de correspondances $I \times J$* recense les flux de P_2O_5 entre les principaux importateurs et exportateurs, à l'exclusion du marché quasi fermé de l'Océanie (Australie, Nouvelle-Zélande et producteurs insulaires de l'Océan Indien et Pacifique : Christmas, Banaba et Nauru) :

I : ensemble de 14 pays importateurs :

Belgique,	Canada,	France,	Deutschland,	Italie,	Japon,	Nederland,	Espagne	United Kingdom,
*I*BL	*I*CA	*I*FR	*I*DL	*I*IT	*I*JP	*I*NL	*I*SP	*I*UK

Inde,	Brésil,	Pologne,	Roumanie,	autres pays d'Europe de l'Est;
*I*IN	*I*BR	*I*PL	*I*RM	*I*EE = RDA, Bulgarie, Hongrie

la 1re lettre I des sigles de ces 14 pays rappelle que ce sont des pays importateurs.

J : ensemble des 8 pays exportateurs (E) :

Belgique,	U.S.A.,	Jordanie,	Maroc,	Sénégal,	Togo,	Tunisie,	U.R.S.S. (C.C.C.P.).
*E*BL	*E*US	*E*JR	*E*MR	*E*SN	*E*TG	*E*TN	*E*CC

Pour chaque couple (i, j) d'un pays i de I et d'un pays j de J, le tableau donne le nombre $k(i,j)$ de milliers de tonnes de P_2O_5 exportées par le pays j vers le pays i durant la période des 8 années 1973-1980.

$$I \times J = \{\ k(i,j) \mid i \in I, j \in J\}.$$

Dans ce tableau $I \times J$,
tout pays importateur i est décrit par la *ligne* i : $\{\ k(i,j) \mid j \in J\}$;
tout pays exportateur j est décrit par la *colonne* j : $\{\ k(i,j) \mid i \in I\}$

La *ligne de marge* du tableau $I \times J$ contient les 8 totaux des 8 colonnes du tableau; la *colonne de marge* contient les 14 totaux des 14 lignes du tableau :
ligne de marge $= \{\ k(j) \mid j \in J\}$ *avec* $k(j) = \Sigma\ \{\ k(i,j) \mid i \in I\}$
colonne de marge $= \{\ k(i) \mid i \in I\}$ *avec* $k(i) = \Sigma\ \{\ k(i,j) \mid j \in J\}$.

A la croisée de la ligne et de la colonne de marge figure le total général du tableau $I \times J$, total noté k.

0.2. *Le tableau des profils des lignes du t·I × J*

Afin de comparer entre eux les divers pays de I, on rapporte tous les nombres d'une même ligne i au total de cette ligne : $k(i)$; on obtient ainsi le *tableau des profils des lignes* du t·$I \times J$.

$$\text{profil de la ligne } i = f_J^i = \{\ k(i,j)/k(i) \mid j \in J\};$$

le profil de la ligne de marge figure également; c'est la ligne :

$$f_J = \{\ k(j)/k \mid j \in J\}.$$

A ce tableau à 8 colonnes (une col. par pays exportateur j), on adjoint une colonne POIDS donnant pour chaque profil de ligne i, le poids attaché à ce profil : $k(i)/k = k = f_i$.

colonne j ↓

	EBL	EUS	EJR	EMR	ESN	ETG	ETN	ECC	POIDS	
IBL		225		615	004	086	019	051	060	
ICA	ε	999				001			086	
IFR	089	183	005	333	101	172	116		150	
IDL	151	434		165	030	023	033	164	090	
IIT	007	327	034	501	012	034	086		059	
IJP		649	077	226	035	014			070	
INL	053	264		330	044	282	011	015	058	
ISP	004	061		917		015	002		057	
ligne i→ IUK	025	134	ε	591	201	008	041		049	←$k(i)/k$
IIN		506	197	227	043		026		052	
$f_J^i = \frac{k(i,j)}{k(i)}$ IBR	004	745		212	002		036		068	
IPL	ε	161	034	415	004	068	069	250	082	
IRM	ε	184	116	370	001	033	005	290	043	
IEE	ε	026	018	184			044	728	078	
marge	032	355	027	337	037	061	042	108	1 000	←k/k

↑ $k(j)/k = f_j$

(Tous les nombres de ce tableau sont des millièmes.)

0.3. *La métrique du chi2*

On considère dans l'espace à 8 dimensions rapporté aux 8 axes : EBL, EUS, ..., ECC chaque profil comme un point ayant pour coordonnées les 8 rapports $\{\ k(i,j)/k(j) \mid j \in J\}$. Par exemple, les coordonnées de IBL sont : 0; .225; 0; .615; .004; .086; .019; .051.

Dans cet espace à 8 dimensions on choisit une *formule de distance* qui rapproche au mieux les profils qui se ressemblent : *la formule du* χ^2 présente le double avantage d'être *euclidienne* et de respecter le principe d'équivalence distributionnelle :

$$\| f_J^i - f_J^{i'} \|^2 = \Sigma\ \{\ (1/f_j)\ (f_j^i - f_j^{i'})^2 \mid j \in J\}.$$

0.4. *Le système othonormé des axes factoriels*

L'ensemble des 14 points pesants $\{ (f_J^i, f_i) \mid i \in I \}$ dans l'espace euclidien à 8 dimensions muni de la formule de distance du χ^2 constitue le *nuage N(I)*.

On peut, soit garder les axes primitifs : *E*BL, *E*US, etc., soit utiliser les axes factoriels issus de l'analyse des correspondances, auquel cas un profil f_J^i est défini par ses 7 facteurs (rappelons que l'espace de départ ayant 8 dimensions, les profils, astreints à avoir la somme de leurs coordonnées égale à 1 sont confinés dans un sous-espace à 7 dimensions) :

$$f_J^i = \{ F_\alpha (i) \mid \alpha = 1, \ldots, 7 \} ;$$

et la formule du χ^2, par changement de coordonnée devient :

$$\| f_J^i - f_J^{i'} \|^2 = \Sigma \{ (F_\alpha (i) - F_\alpha (i'))^2 \mid \alpha = 1, \ldots, 7 \}$$

On peut même, lorsqu'on a reconnu que les derniers facteurs ne correspondent à aucune réalité mais rendent compte de fluctuations aléatoires (ce qui n'est pas le cas de notre exemple) ne garder que la projection du nuage sur le sous-espace des axes significatifs. Par exemple, si on ne garde que 4 facteurs, on aura :

$$f_J^i = \{ F_\alpha (i) \mid \alpha = 1, 2, 3, 4 \}$$

$$\| f_J^i - f_J^{i'} \|^2 = \Sigma \{ (F_\alpha(i) - F_\alpha (i'))^2 \mid \alpha = 1, 2, 3, 4 \} .$$

En ce qui concerne les 14 pays importateurs de phosphates, nous avons gardé la description complète des profils par les 7 facteurs .

1. Notion de Classification Ascendante Hiérarchique (C.A.H.)

Nous expliquerons successivement en quel sens nous entendons les trois termes de classification (§ 1.1), hiérarchique (§ 1.2), ascendante (§ 1.3).

1.1. *Classification et classement*

Dans l'exemple sur lequel nous avons choisi de fonder le présent exposé, on part d'un ensemble I de 14 individus i, les pays grands importateurs de phosphates; ces pays sont décrits (en tant qu'importateurs) par le tableau $I \times J$. Plus précisément, selon les représentations géométriques propres à l'analyse des correspondances, l'ensemble I est identifié à un nuage $N(I)$, ou ensemble de points munis de masse dans l'espace enclidien des profils sur J où la distance est celle du χ^2 de centre f_J. Faire une classification sur I, ce sera édifier un système de classes ou parties de I, d'après cette représentation géométrique. Le terme de classification est appliqué à la fois à un procès : l'édification des classes, et à un état : le résultat de ce procès, le système des classes.

Il importe de comparer la notion mathématique de classification automatique avec l'emploi commun du mot de classification. En un sens, tout langage implique une classification, dans la mesure où certains mots délimitent plus ou moins clairement une classe d'objets ou de situations. Plus précisément, dans les sciences de la nature, les êtres vivants sont répartis suivant une classification ou *taxinomie*. Mais il y a plusieurs grandes différences avec la classification mathématique. D'une part la description des individus n'est pas donnée au naturaliste sous un format fixe : les lignes d'un tableau; il doit découvrir avant de classer, les traits d'une description adéquate; d'autre part le naturaliste considère non un ensemble fini I, mais un ensemble potentiellement infini : tous les vivants, ou même seulement tous les mammifères, ensemble dont les représentants (les individus vivant aujourd'hui des espèces actuelles...) se renouvellent sans cesse. C'est pourquoi, à supposer que soit édifiée une classification satisfaisante, le problème se pose toujours devant un individu d'en faire la détermination ou *classement*, i.e. de décider de la classe à laquelle il appartient.

Ensemble potentiel et classement ne sont cependant pas étrangers au travail du mathématicien : en travaillant sur I fini, il envisage souvent un champ plus vaste indéfini; et comme en analyse factorielle, peut être amené à adjoindre à I des élements supplémentaires.

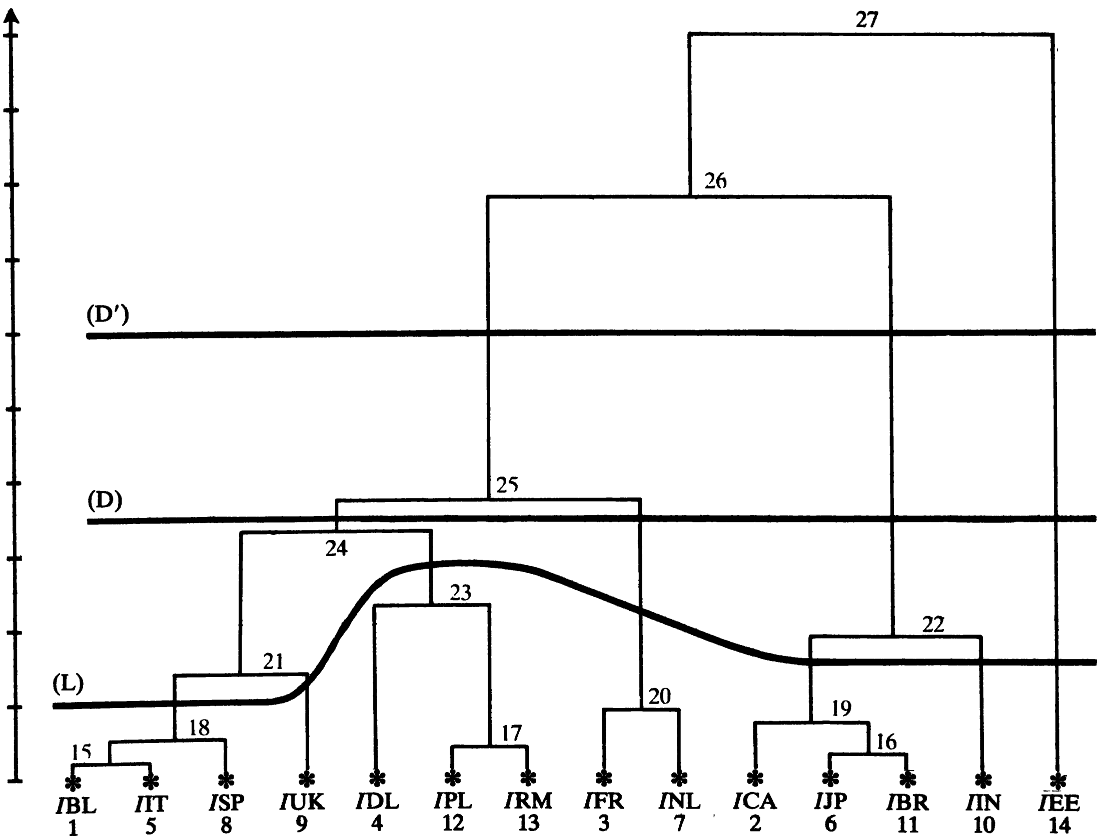
1.2. *Hiérarchie et partition*

La forme la plus simple de classification est la *partition* : on partage I en un système de classes non vides, de telle sorte que tout individu i appartienne à une classe et une seule. Mais le terme de classification sert aussi à désigner un système emboîté ou *hiérarchie* de classes, comme on en voit en Sciences Naturelles : les êtres vivants sont partagés en deux grands règnes, animal et végétal; et chacun de ces règnes est lui-même divisé en embranchements : ainsi, les animaux sont partagés en vertébrés, arthropodes, mollusques, ...; les vertébrés sont à leur tour subdivisés en classes : mammifères, oiseaux, reptiles, batraciens et poissons; etc. On parle alors de classification hiérarchique, ou hiérarchie de classes.

Il est facile de représenter graphiquement, ou de décrire formellement une partition. Voici par exemple une partition de I en trois classes; (Europe Centrale et Occidentale; Europe de l'Est; reste du monde) :

IBL, IIT, ISP, IUK, IDL, IPL, IRM, IFR, INL ; ICA, IJP, IBR, IIN ; IEE

La structure d'une hiérarchie est à la fois plus riche et plus complexe. Considérons d'emblée, présenté comme un *arbre*, le résultat de la CAH effectuée sur l'ensemble I d'après le tableau de correspondance $I \times J$.



L'arbre comprend à sa base les 14 individus à classer et, aboutissant à ces individus, des branches se raccordant entre elles par des *nœuds*; à ces nœuds aboutissent de nouvelles branches qui se raccordent entre elles par de nouveaux nœuds; ainsi de suite jusqu'au *sommet*. Les individus portent le numéro qu'ils ont dans le tableau $I \times J$ (de 1 à 14); les nœuds sont numérotés de 15 à 27 (le dernier numéro : 27 est celui du sommet). Si on veut rattacher cette terminologie à une image familière, il faut retourner le dessin en sorte que le sommet 27 se place à la base d'où partent deux branches allant l'une vers le nœud 26, l'autre vers IEE = 14; et de même, le nœud 26 se subdivise par ramifications successives jusqu'aux individus encore appelés *terminaux*.

Un tel arbre définit un système emboîté de classes : plus exactement un système *dichotomique* parce que de chaque nœud partent deux branches. Ainsi, du sommet 27 partent deux branches : à l'une, 26, disposée à gauche sur le dessin et encore notée $A(27)$ (A est l'initiale d'*aîné*, terme emprunté à la généalogie, non à la botanique!) se rattachent 13 individus (de IBL à IIN, à la base du graphe); l'autre branche $B(27)$ est réduite à un seul individu terminal IEE. La lettre B attribuée à la branche de droite est l'initiale de *benjamin,* terme qui répond à A, aîné, et on dit que $A(27)$ et $B(27)$ sont les deux *descendants immédiats* de 27. De même, on a $A(24) = 21$, $B(24) = 23$; les nœuds 21 et 23 définissant les deux classes des individus qui leur sont respectivement rattachés :

$$21 = \{ I\mathrm{BL}; I\mathrm{IT}; I\mathrm{SP}; I\mathrm{UK} \} ; 23 = \{ I\mathrm{DL}; I\mathrm{PL}; I\mathrm{RM} \},$$

et la classe 24 étant la réunion de celles-ci , ce qu'on écrira :

$$24 = A(24) \cup B(24) = 21 \cup 23.$$

A gauche de l'arbre, une échelle graduée à partir de 0 situé à la base, indique le *niveau* de chaque nœud. On peut dire, en bref, que le niveau d'un nœud représente par un nombre le degré de généralité de la classe définie par ce nœud. De même qu'en Sciences Naturelles un *ordre,* e.g. les carnivores, est plus général qu'une *espèce,* e.g. le chat, nous dirons que le nœud 24 est à un niveau supérieur à celui du nœud 22. En particulier, comme le montre clairement le dessin, les nœuds $A(n)$ et $B(n)$ descendants immédiats (aîné et benjamin) du nœud n, sont à des niveaux inférieurs à celui de n. On vérifiera que le numérotage des nœuds est fait dans l'ordre des niveaux croissants. Suivre dans l'arbre un *chemin descendant* c'est considérer une suite de nœuds ou individus : $n1$, $n2$, $n3$ ... telle que $n(p+1)$ soit descendant immédiat de np : par exemple, 25, 24, 23, 17, { IPL } = 12, est un chemin descendant; le chemin inverse : { IPL } , 17, 23, 24, 25, un *chemin ascendant*. Quand on parcourt un chemin descendant, les n^{os} des nœuds vont en décroissant ($25 > 24 > 23 > 17 > 12$); c'est le contraire dans un chemin ascendant ($12 < 17 < 23 < 24 < 25$).

Les individus, ou classes réduites à un seul élément, sont au niveau 0. Le niveau d'un nœud n est généralement noté $\nu(n)$, ou encore $D(n)$, la lettre D étant l'initiale de diamètre et aussi de distance : en effet, plus le nœud n est élevé, plus la classe est grande (ce qui explique diamètre), et plus aussi les deux classes $A(n)$ et $B(n)$ dont il est composé ont de chance d'être éloignées l'une de l'autre (ce qui explique distance). En général, le *niveau* $D[n]$ du nœud n sera calculé comme l'écart entre ses deux descendants immédiats $A(n)$ et $B(n)$. Mais pour donner un sens précis à tout ce que nous évoquons ici, il faut attendre les §§ 2 et 3 : la définition du critère d'agrégation.

En coupant l'arbre à un niveau donné par une droite horizontale D on a au-dessous de celle-ci plusieurs branches séparées, définissant une partition de I. Par exemple si D passe entre les nœuds 24 et 25, on a une partition S en 4 classes :

$$I = 27 = 24 \cup 20 \cup 22 \cup \{ I\mathrm{EE} \} ; S = \quad 24; 20; 22; \{ I\mathrm{EE} \} \ ;$$

avec une droite D′ passant entre les nœuds 25 et 26, on a la partition en 3 classes proposée au début du § 1.2 et notée ici S' :

$$I = 27 = 25 \cup 22 \cup \{ I\mathrm{EE} \} ; S' = \{ 25; 22; \{ I\mathrm{EE} \} .\}.$$

Plus généralement, une ligne continue, qui peut être sinueuse, astreinte à couper une fois et une seule tout chemin descendant partant du sommet de l'arbre et aboutissant à un pays, définit une partition. Ainsi, sur la figure, la ligne (L) définit une partition en 7 classes :

$$I = 27 = 18 \cup \{ /UK \} \cup 23 \cup 20 \cup 19 \cup \{ /IN \} \cup \{ /EE \} ;$$

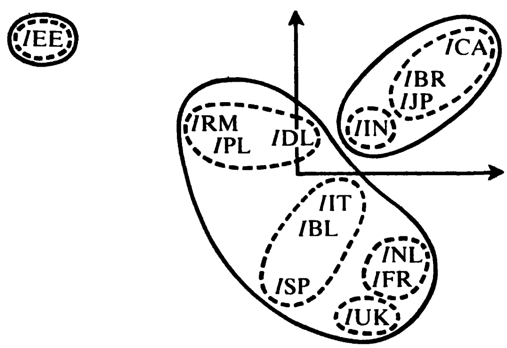
$$C = \quad 18; \{ /UK \}; 23; 20; 19; \{ /IN \}; \{ /EE \} .$$

Dans cette partition trois classes sont réduites à un seul élément (qu'on a ici comme plus haut, selon l'usage mathématique, placé entre accolades parce qu'il est considéré comme une partie).

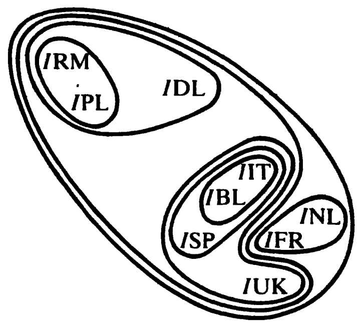
Ainsi, à partir d'une classification hiérarchique dichotomique, on dispose d'un grand nombre de partitions : en ce sens, le résultat de la construction mathématique, l'arbre laisse au spécialiste des données traitées la liberté de choisir : ce dialogue entre calcul et réflexion est l'essence même de l'analyse des données dont la philosophie est de s'opposer à la traduction irréversible des choses en nombres...

Sur le plan (1 × 2) issu de l'a. des c. on a délimité par des contours, pleins ou tiretés, les classes des deux partitions S' et C.

Plan 1, 2 d'a. des c.
— part en 3 cl. S'
... part en 7 cl. C



Ce dessin suggérera au lecteur de critiquer la représentation arborescente que nous lui proposons. Ne pourrait-on pas montrer plus clairement les classes emboîtées en traçant leurs contours? En effet, voici, en se bornant à la classe 25 et à ses subdivisions, le système des dichotomies emboîtées.



L'inconvénient de ce schéma est qu'il ne se prête pas à une formalisation complète, nécessaire au mathématicien, et plus encore à l'ordinateur. Au contraire, l'arbre peut facilement (moyennant des modifications minimes *cf.* infra § 4.2) être tracé par une imprimante; et sa description complète ne requiert qu'un tableau de nombres à trois lignes A, B, D.

	15	17	18	20	21	23	24	25
A	1	12	15	3	18	4	21	24
B	5	13	8	7	9	17	23	20
D	.008	010	.023	.025	.043	.069	104	.116

Par exemple on lit dans la colonne 24 que le nœud 24, situé au niveau $D(24) = .104$ a pour aîné (branche de gauche) $A(24) = 21$ et pour benjamin (branche de droite) $B(24) = 23$.

Nous terminons ce § sur une dernière critique : pourquoi avoir placé 21 à gauche et 23 à droite? Il n'y a aucune raison à cela : le choix $A'(24) = 23$; $B'(24) = 21$ conviendrait tout aussi bien. Mais il est indispensable de faire un choix, aussi bien pour la commodité de la représentation numérique (au sein de l'ordinateur) que pour l'impression du graphique (destiné à l'utilisateur).

1.3. *Classification ascendante et classification descendante*

Il y a deux façons de lire l'arbre hiérarchique présenté au § 1-2 : l'une descendante, l'autre ascendante.

La *lecture descendante* part du sommet : le nœud $I = 27$ se scinde en ses deux descendants immédiats $A(27) = 26$ et $B(27) = 14 = \{ I\mathrm{EE} \}$; le nœud 26 se scinde en $A(26) = 25$ et $B(26) = 22$; etc. La *lecture ascendante* part de la base : les individus 1 et 5 (IBL et IIT) *s'agrègent* pour former la classe 15 dont ils sont les deux descendants immédiats : $A(15) = 1$, $B(15) = 5$; les individus 6 et 11 s'agrègent de même en 16 ($A(16) = 6$; $B(16) = 11$); 12 et 13 s'agrègent en 17; la classe 15 s'agrège à l'individu 8 pour donner la classe 18 ($A(18) = 15$; $B(18) = 8$); etc. Finalement, 26 s'agrège à 14 pour donner 27 ($A(27) = 26$; $B(27) = 14$).

De même, on peut concevoir deux types d'algorithme de classification, c'est-à-dire deux types de méthodes pour édifier progressivement une hiérarchie de classe emboîtées. Un *algorithme descendant* part du tout qu'il scinde en deux classes; à nouveau, il scinde chacune de ces deux classes en deux et ainsi de suite jusqu'à isoler les individus. Un *algorithme ascendant,* tel que celui de la CAH, part des individus et d'un *critère de ressemblance* des individus qui s'étende aux classes, agrège en priorité les individus qui se ressemblent le plus; puis il agrège soit deux autres individus soit un individu et une classe déjà constituée; puis des classes entre elles, créant ainsi des nœuds n dont le niveau $D[n]$ (*cf.* supra 1.2) se calcule comme l'écart entre $A(n)$ et $B(n)$; et ainsi de suite jusqu'au sommet qui est I tout entier.

Procéder par voie descendante suppose que l'on soit assuré d'avoir reconnu les variables ou les caractères auxquels il faut recourir pour définir les divisions supérieures de la hiérarchie, i.e. que l'on ait une vue juste de ce que, depuis Jussieu, les taxinomistes appellent *hiérarchie des caractères.* Or l'histoire de la botanique ou de la zoologie montre que cette hiérarchie n'est connue qu'au terme d'un long progrès. Les espèces végétales, par exemple, sont connues dès le début du XVIIe siècle. La gloire de Tournefort est d'avoir, à la fin de ce siècle, groupé des centaines d'espèces en des genres dont la plupart ont été admis par la suite (*). L'agrégation des genres en familles fut l'œuvre d'Adanson et de Linné au milieu du XVIIIe siècle... Voilà pourquoi en classification automatique nous préférons les algorithmes *ascendants* : dans la mesure où le calculateur, procédant sans information *a priori*, se trouve dans la position du botaniste au temps où cette science était dans l'enfance.

(*) « Tournefort († 1708) a été pour la nomenclature des genres ce que G. Bauhin († 1624) fut pour les espèces » *in* R. Dughi : Tournefort, Museum d'Histoire Naturelle, Paris 1957; p. 175.

L'arbre une fois constitué (par voie ascendante), on aura recours, pour l'interprétation, aux deux procédés de lecture : descendant et ascendant, en s'aidant de calculs complémentaires afin de dégager les caractères propres aux principales classes et choisir en définitive une partition (voire deux, concurremment) d'après laquelle on rendra compte de la structure de I.

Cependant, restent dans le vague la structure de l'algorithme et plus encore la notion de *ressemblance* ou *critère d'agrégation* entre classes sur laquelle repose la CAH. D'où le titre du § 2.

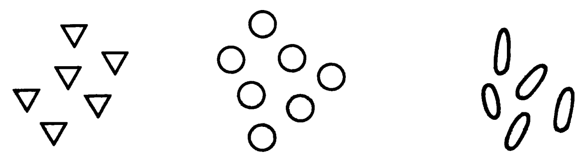
2. Critères d'agrégation et algorithmes de CAH

Les qualités qu'on exige d'une classification (§ 2.1) suggèrent plusieurs critères simples (§ 2.2) dont la définition part d'une distance entre points et qui sont analogues à une distance entre parties. Cependant, du point de vue axiomatique, un critère ne se définit pas comme une distance (§ 2.3), le critère étant seulement conçu pour permettre le déroulement d'un algorithme ascendant.

2.1. *Qualités d'une classification*

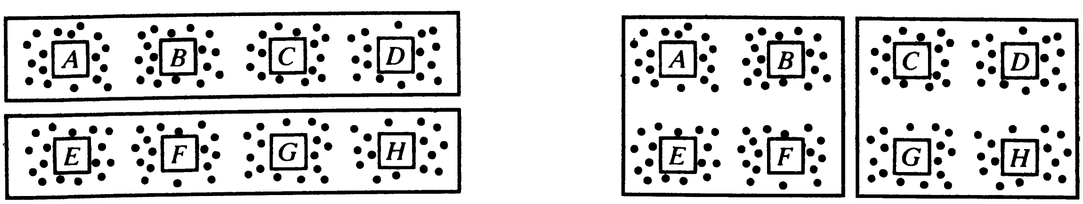
Une classification hiérarchique s'interprète en terme de partition (§ 1.2) et se construit par un algorithme ascendant comme une suite de partitions de moins en moins fines (§ 2.3) : il nous suffira donc ici de considérer le cas d'une partition.

Une partition n'est intéressante que dans la mesure où les classes sont nettement individualisées; c'est-à-dire d'une part, forment chacune un tout cohérent bien caractérisé (nous parlerons de *compacité* des classes); et d'autre part sont distinctes les unes des autres *(séparabilité)*. Par exemple, si l'ensemble I des individus à classer est un ensemble d'objets plats de formes diverses, on pourra en constituer trois tas selon que la forme est oblongue, ronde ou triangulaire.



Mais si l'ensemble I comprend des objets de forme intermédiaire, il se prête moins bien à une classification.

Puisque l'analyse des données permet de traduire la description des individus par un point placé dans un espace multidimentionnel (le nombre des dimensions étant celui des facteurs retenus selon l'interprétation de l'analyse), on peut encore proposer le schéma d'une situation où l'on ait le choix entre plusieurs façons de grouper les individus.



On a suggéré par des cadres deux façons de grouper les classes A, B, C, D, E, F, G, H.

Une fois reconnues les 8 classes A B C D E F G H, assemblera-t-on (A, B, C, D) en raison du peu d'espace qui sépare A de B, B de C et C de D; et de même pour (E, F, G, H), ce qui produirait finalement deux grandes classes allongées? Une telle partition en deux classes est assez satisfaisante du point de vue de la séparabilité, car une bande vide assez large passe entre $(A, B, C, D,)$ et (E, F, G, H); mais ces deux classes qui s'étirent ne sont guère compactes. Au contraire, si l'on groupe d'une part (A, B, E, F) et d'autre part (C, D, G, H), les deux classes obtenues sont bien ramassées, compactes; mais elles sont mal séparées, car B et C se touchent presque; et de même F et G. Les deux exigences de compacité et de séparabilité apparaissent ici contradictoires.

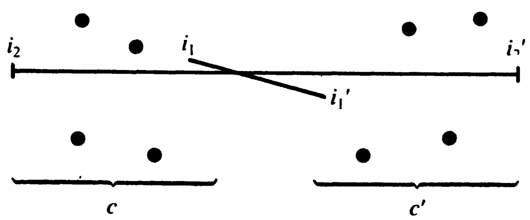
Certes un schéma plan (à 2 dimensions) n'est aucunement réaliste. Il faut insister sur le fait que seule l'épreuve des données réelles multidimensionnelles permet d'apprécier les mérites d'une méthode de CAH. De plus, le résultat définitif dépend grandement de la traduction géométrique préalable, du codage spatial auquel on a soumis les données brutes : de ce point de vue, la CAH est, selon nous, inséparable de l'analyse des correspondances. Mais le schéma proposé suffit à nous rappeler que la classification automatique ne peut faire mieux que de découvrir les séparations qui existent réellement dans les données. Et c'est l'un des rôles majeurs de l'interprétation que de préciser parmi toutes les dichotomies d'une CAH, celles qui correspondent à des divisions géométriquement bien tranchées et conceptuellement interprétables (ainsi, dans le schéma ci-dessus, les classes A, B, C, D, E, F, G, H s'imposent; leurs subdivisions sont irrelevantes; les agréger entre elles est embarrassant).

2.2. *critères usuels*

Soit c et c' deux parties finies quelconques de l'ensemble I des éléments à classer : le calcul de la valeur $D\,(c, c')$ du critère D pour le couple (c, c') se fonde presque toujours sur une distance usuelle $D\,(i, i')$ entre éléments de I; mais utilise aussi dans les cas qui nous intéressent, un système de masses positives f_i attribuées aux éléments de I. Nous présentons ici quatre critères classiques; notre but étant de montrer dans quelle mesure leur utilisation dans un algorithme de CAH (§ 2.3) assure à la classification les qualités requises.

2.2.1. *Critère du saut minimum*

$D_{\text{saut}}\,(c, c')$ est la distance minima entre un point de c et un point de c' (i.e. la distance entre deux points i et i' appartenant l'un à c l'autre à c' et le plus proches possible).



$$D_{\text{saut}}\,(c, c') = D(i_1, i_{1'})$$
$$D_{\text{diam}}\,(c, c') = D(i_2, i_2')$$

En agrégeant en priorité les paires de classes entre lesquelles l'écart D_{saut} est le plus faible, on crée des classes bien séparées entre elles : car si elles ne l'étaient pas, on aurait décidé de les agréger (plus précisément d'agréger celles de leurs parties au niveau desquelles se réalise le saut minimum). Mais on peut ainsi construire des classes allongées, voire filiformes : on dit alors qu'il y a effet de chaînage; ce qui correspond au premier choix proposé sur la figure 2 du § 2.1 : constituer les classes (A, B, C, D) et (E, F, G, H).

2.2.2. *Critère du diamètre*

$D_{\text{diam}}\,(c, c')$ est la distance maxima entre un point de c et un point de c'. En agrégeant en priorité les paires de classes entre lesquelles l'écart D_{diam} est le plus faible, on crée des classes ramassées, compactes, dont le diamètre (défini comme la plus grande distance entre deux points de la classe) est minimum. Mais on peut ainsi s'interdire de réunir des classes qui ne sont pas séparées mais sont presque en contact l'une avec l'autre : ce qui correspond au deuxième choix proposé sur la fig. 2 du § 2.1.

2.2.3. *Critère de la distance moyenne*

D_{moy} (c, c') est la moyenne des distances séparant un point i de c et un point i' de c', chaque segment $D(i, i')$ ayant pour poids le produit $f_i f_{i'}$ des masses de ses extrémités. De façon précise, on a :

$$D_{moy}(c, c') = (1/(f_c f_{c'}))\ \Sigma\ \{ f_i f_{i'}\ D(i, i') \mid i \in c, i' \in c' \}\ ;$$

on a noté f_c et $f_{c'}$ les masses totales respectives des classes c et c' :

$$f_c = \Sigma\ \{ f_i \mid i \in c \}\ ;\ f_{c'} = \Sigma\ \{ f_{i'} \mid i' \in c' \}\ .$$

Cette formule apparaît comme un compromis entre D_{saut} et D_{diam}, en ce qu'elle tient compte à la fois de tous les segments $D(i, i')$, les plus petits comme les plus grands.

2.2.4. *Critère de l'inertie*

Pour calculer ce critère, on doit supposer que I est un ensemble de points munis de masse d'un espace euclidien (c'est le cas dans l'exemple du commerce des phosphates qui sert de base au présent exposé). Comme au § 2.2.3, on note respectivement f_c, $f_{c'}$, les masses totales des classes c et c' de plus, on note ici simplement c' le centre de gravité de la classe c; et de même pour c'. On pose :

$$D_{inert}(c, c') = (f_c f_{c'} \div (f_c + f_{c'}))\ \|c-c'\|^2,$$

où $\|c-c'\|^2$ désigne le carré de la distance euclidienne entre les centres des classes c et c'.

Cette valeur D_{inert} n'est autre que l'inertie d'un nuage très simple réduit aux deux points c, c' munis des masses $f_{c'}$ et f_c (inertie prise par rapport au centre de gravité de ce nuage à 2 points). Dans le cas particulier où les classes c et c' sont chacune réduites à un élément i et i' le critère D_{inert} (i, i') diffère du carré de la distance euclidienne $\|i-i'\|^2$ par un coefficient de masse. Le critère D_{inert} étant seul utilisé dans la suite, on le notera encore crit (c, c') : son intérêt apparaîtra au § 3. On remarquera dès maintenant que ce critère s'accorde avec les constructions géométriques de l'a. des corr.. Quant à la hiérarchie des classes construites, D_{inert} et D_{moy} fournissent dans la pratique des résultats généralement bons, à la différence de D_{saut} et D_{diam} dont on a montré ci-dessus les inconvénients.

2.3. *Propriétés axiomatiques des critères*

2.3.1. *Critère et distance*

Pour soutenir l'intuition, nous avons dit que $D(c, c')$ était comme une distance (ou, pour employer un terme moins précis, un écart), entre les deux parties c et c' de I. Il importe de noter d'abord que des 4 critères usuels cités au § 2.2, seul D_{moy} est une véritable distance; puis on verra au § 2.3.2 que la propriété axiomatique qu'il faut exiger d'un critère de CAH *n'est pas* d'être une distance.

En termes mathématiques, on dit qu'un ensemble I (fini ou infini) est un espace métrique si est défini pour tout couple (i, i') de points de I un nombre réel positif ou nul $D(i, i')$ appelé distance entre i et i' et satisafaisant aux axiomes suivants :

a) symétrie : $D(i, i') = D(i', i)$.

b) positivité stricte : $D(i, i')$, est strictement positif si $i \neq i'$; nul si et seulement si $i = i'$.

c) inégalité du triangle : quels que soient les trois points i, i', i''.

$$D(i, i'') \leq D(i, i') + D(i, i'');$$

ce qu'on peut paraphraser : le chemin direct (i, i'') est inférieur ou égal à la somme des deux segments (i, i') et (i', i'') du chemin passant par i'.

La question se pose de savoir si, sur l'ensemble des parties non vides de I, un écart $D(c, c')$ définit ou non une structure d'espace métrique.

Les critères satisfont tous à la condition de symétrie. En revanche, la condition de positivité n'est satisfaite ni par D_{sant} ni par D_{diam}. En effet, soit c et c' deux parties distinctes non vides ayant un point commun. On aura $D_{\text{saut}}(c, c') = 0$ bien que $c \neq c'$; et on aura $D_{\text{diam}}(c, c) \neq 0$ si c comprend au moins deux points; car $D_{\text{diam}}(c, c')$ n'est autre que le maximum d'un point de c à un point de c. Toutefois, dans le déroulement de l'algorithme, ces particularités n'apparaissent pas, car les calculs d'écart se font exclusivement entre parties disjointes.

Le critère D_{inert} ne satisfait pas à l'inégalité du triangle; posons par exemple :

$$f_c = f_{c''} = 10^{-1};\, f_{c'} = 10^{-6};\, \| c - c' \|^2 = \| c - c'' \|^2 = \| c' - c'' \|^2 = 1;$$

en appliquant la formule du § 2.2.4, on aura, en contradiction avec l'inégalité du triangle :

$$D_{\text{inert}}(c, c'') = 0{,}5\ 10^{-1};\ D_{\text{inert}}(c, c') = D_{\text{inert}}(c', c'') \simeq 10^{-6}.$$

Ce qu'on peut résumer en disant que la classe c' très légère est à la fois très proche des deux classes lourdes c et c'', elles-mêmes séparées par une distance notable.

2.3.2. *Axiome de la médiane*

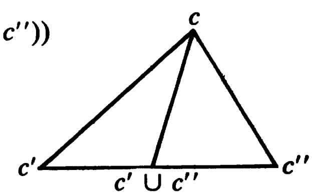
De même qu'une distance, tout critère $D(c, c')$ devra être une fonction positive (ou nulle) ne dépendant pas de l'ordre dans lequel on prend les deux classes c et c' :

$$D(c, c') = D(c, c') \geqslant 0.$$

Mais l'inégalité du triangle est indifférente au déroulement de l'algorithme de CAH. En revanche, comme les classes sont constituées par étapes successives, en agrégeant d'abord les paires de classes qui rendent D minimum, on impose à D une condition assurant que l'agrégation de deux classes ne remette pas en cause les agrégations précédentes. Pour tout triplet de classes c, c', c'' tel que c' et c'' soient à agréger avant c et c' ou c et c'', on demande que la classe $c' \cup c''$ (créée par agrégation de c' avec c'') ne soit pas moins écartée de c que ne l'était la plus proche des deux classes préexistantes c' et c''.

Ce qu'on écrira en formules :

$$\forall\, c, c', c'' \subset \mathrm{I} : D(c', c'') \leqslant \inf (D(c, c'), D(c, c''))$$
$$\Rightarrow \inf (D(c, c'), D(c, c'')) \leqslant D(c, c' \cup c'')$$



Ainsi, si dans la suite on doit agréger c à $n' = (c' \cup c'')$ pour créer une classe $n = c \cup n'$, cette agrégation se fera à un niveau qu'on calcule comme il est expliqué au § 3 : $D[n] = D(A(n), B(n)) = D(c, n')$, supérieur au niveau $D[n'] = D[c', c'']$ auquel est créé le nœud n'; ce qui correspond sur le dessin de l'arbre au fait que tout nœud est à un niveau supérieur à celui de ses descendants.

Symboliquement, on peut se représenter c, c', c'' par un triangle dont (c', c'') est le plus petit côté; $c' \cup c''$ (assimilé à un centre de gravité de classe) sera placé au milieu du côté (c', c''). Dès lors $(c, c' \cup c'')$ est représenté par une médiane. C'est pourquoi la formule ci-dessus est appelée « axiome de la médiane ». Cet axiome est satisfait par les 4 critères du § 2.2.

2.3.3. *Algorithmes de base et algorithmes accélérés*

L'algorithme de CAH, sous sa forme la plus simple (qui n'est pas la plus rapide), crée successivement toutes les partitions qui peuvent être obtenues en coupant l'arbre par une droite horizontale.

Étape 1 : On part de la partition la plus fine de I, dont chaque classe est constituée par un individu unique i; ces classes sont encore appelées *sommets* parce que, présentement, chacune n'est comprise dans aucune classe plus grande qu'elle. On calcule le tableau des valeurs du critère choisi $D(i, i')$ pour tout couple de classes (sommets) $\{ i \}, \{ i' \}$. On agrège alors la paire i, i' réalisant le minimum de D

pour créer le nœud $n = \{i, i'\}$ avec $A(n) = i$, $B(n) = i'$, qui est numéroté après les individus de I et reçoit le n° Card $I + 1$. Il y a maintenant Card $I - 1$ sommets (ou classes maximales) constituant une nouvelle partition de I : d'une part $n = \{i, i'\}$; d'autre part les (Card $I - 2$) classes réduites à un élément i'' de I autre que i et i'.

On prendra garde qu'au cours du déroulement de l'algorithme, toute classe créée par agrégation de deux autres joue le rôle de sommet jusqu'à ce qu'elle soit elle-même agrégée; à la différence du terme de nœud qui exprime une qualité permanente (le fait qu'une classe a été créée par agrégation de deux autres) le terme de sommet n'a de sens que relativement au déroulement de l'algorithme.

Étape 2 : On calcule les écarts $D(n, i'')$ du nouveau sommet n aux (Card $I - 2$) sommets préexistants (entre lesquels les écarts D sont déjà connus), et on agrège la paire de sommets réalisant le minimum de D : ce qui entraîne qu'un nouveau nœud est créé; il reçoit le n° (Card $I + 2$) et prend le rôle de sommet, cependant que deux classes cessent d'être sommets. Il y a donc au terme de cette 2^e étape (Card $I - 2$) sommets. Ceux-ci constituent une nouvelle partition de I.

Étape 3 : Comme précédemment, on calcule les écarts D entre le dernier sommet créé et les sommets préexistants, etc.

Et ainsi de suite jusqu'à ce que par réunion des deux derniers sommets subsistants soit créé le (Card $I - 1$)^{e} nœud qui n'est autre que I tout entier et reçoit le n° (2 Card $I - 1$). Cet ultime nœud garde définitivement le nom de *sommet*.

L'inconvénient de cet algorithme dit *alg. de base* est qu'il requiert qu'à chaque étape on passe en revue l'ensemble du tableau des écarts D entre paires de sommets pour découvrir la valeur la plus faible et agréger les sommets correspondants : ainsi, le temps requis pour édifier une CAH sur I est de l'ordre de (Card $I)^3$. Des algorithmes accélérés édifient la même CAH en un temps de l'ordre de (Card $I)^2$: en bref, ces algorithmes procèdent en agrégeant plusieurs paires de sommets à la fois, ou en découvrant des paires de sommets à agréger sans revoir à chaque étape l'ensemble du tableau des écarts entre sommets; la seule contrainte étant de n'agréger en une étape que des paires de sommets qui sont plus proches voisins réciproques (i.e. dont chacun réalise le minimum de l'écart à l'autre). Nous nous bornons à dire ici que les algorithmes accélérés ne donnent le même résultat que l'algorithme de base que si est vérifié l'axiome de la médiane; de plus, du point de vue de l'encombrement de la mémoire et de la complexité des calculs d'écart, le critère de l'inertie l'emporte nettement sur celui de la distance moyenne. Comme D_{saut} et D_{diam} ne donnent des résultats satisfaisants que dans les cas simples, il reste, comme nous l'avons annoncé, le critère D_{inert} = crit, dont l'explication détaillée fait l'objet du § 3.

3. Le critère de l'inertie.

Au § 3.1 on définit pour toute partition C de l'ensemble I une quantité Intra (C) appelée inertie intraclasse de la partition C : Intra (C) est d'autant plus faible que les classes de la partition C sont plus compactes. Au § 3.2 on définit l'inertie interclasse de la partition C : Inter (C) qui est, en un certain sens, d'autant plus élevée que les classes de C sont mieux séparées. Ces deux manières globales d'évaluer les qualités d'une partition sont strictement complémentaires (§ 3.3). Elles conduisent à adopter le critère de l'inertie pour décider des agrégations entre classes maximales (ou sommets) effectués successivement par l'algorithme de CAH (§ 3.4). Ainsi, points de vue local et global se rejoignent (§ 3.5).

3.1 *Inertie intraclasse d'une partition et compacité des classes*

Prenons l'exemple de la partition en 3 classes $S' = (25, 22, 14 = \{I\,\text{EE}\})$ déjà considérée au § 1.2. Chaque classe c de la partition constitue un sous-nuage de $N(I)$ et ce sous-nuage a une inertie totale par rapport à son centre de gravité propre, inertie que l'on peut appeler *inertie interne* de la classe c.

Définition : On appelle inertie intraclasse de la partition $S' = (25, 22, \{I\ \text{EE}\})$ la somme des inerties internes des classes 25, 22 et $\{I\ \text{EE}\}$ constituant la partition. Cette somme est notée Intra (S').

Comme nous le verrons plus loin (§ 4.1 Remarque), on peut trouver d'après le listage de CAH les inerties internes des classes 25, 22 et $\{I\ \text{EE}\}$. On a :
inertie interne de la cl. 25 = .398;
inertie interne de la cl. 22 = .093;
inertie interne de la cl. $\{I\ \text{EE}\} = 0$;
d'où l'inertie intraclasse de la partition $S' = (25, 22, \{I\ \text{EE}\})$:

$$\text{Intra}(S') = .398 + .093 + 0 = .491.$$

Mais on peut calculer directement les inerties internes des classes 25, 22 et $\{I\ \text{EE}\}$. Il est d'abord évident que pour la cl. $\{I\ \text{EE}\}$ constituée d'un seul point qui se confond avec le centre de gravité g ($\{I\text{EE}\}$)de cette classe, l'inertie interne est nulle car la distance $||g(\{I\text{EE}\}) - I\ \text{EE}||^2$ est nulle. Prenons maintenant la classe 22. On déterminera les facteurs du centre de gravité g (22) de cette classe par la formule :

$$F_\alpha(\text{g}(22)) = (.086\ F_\alpha(I\ \text{CA}) + .070\ F_\alpha(I\ \text{JP}) + .068\ F_\alpha(I\ \text{BR}) + .052\ F_\alpha(I\ \text{IN}) / .276$$

où .086, .070. etc. sont les poids de I CA, I JP etc. (*cf.* tableau des profils § 0.2); .276 est la somme des poids des 4 pays I CA, I JP, I BR et I IN; $F_\alpha(I\ \text{CA})$ etc. les α^{es} facteurs des 4 pays, facteurs que l'on trouve au tableau des facteurs sur I issu de l'a. des *c.* du tableau $I \times J$; voici les valeurs des $F_\alpha(\text{g}(22))$ ainsi calculées (en millièmes) :

	F_1	F_2	F_3	F_4	F_5	F_6	F_7
g (22)	591	657	−118	067	−009	−055	−002

Puis on calcule les 4 distances au carré :

$$||g(22) - I\ \text{CA}||^2, \quad ||g(22) - I\ \text{JP}||^2, \quad ||g(22) - I\ \text{BR}||^2, \quad ||g(22) - I\ \text{IN}||^2 :$$

$$||g(22) - J\ \text{CA}||^2 = \Sigma\{(F_\alpha(\text{g}(22)) - F_\alpha(I\ \text{CA}))^2\ |\alpha = 1, 2, \ldots, 7\}; \text{etc...}$$

Voici les valeurs des 4 distances au carré :

i	ICA	IJP	IBR	IIN
$d^2(\text{g}(22), i)$	.366	.076	.145	.930

Enfin on calcule l'inertie interne de la classe 22 comme somme des inerties des 4 points I CA, I JP, I BR, I IN munis de leurs poids, par rapport au centre g (22) :

$$\begin{aligned}\text{inertie int. de } 22 &= (.086 \times .366) + (.070 \times .076) + (.068 \times .145) \\ &\qquad + (.052 \times .93) \\ &= .095\,.\end{aligned}$$

Facteurs et poids utilisés dans nos calculs sont des valeurs approchées imprimées sur les listages, et nos résultats se trouvent entachés d'erreur. La valeur de cette même inertie interne de la classe 22 d'après le listage de CAH est, *cf.* supra, .093.

On peut calculer de la même façon l'inertie interne de la cl. 25; nous ne le ferons pas ici.

Plus les classes c d'une partition C sont compactes, i.e. moins elles sont dispersées autour de leurs centres respectifs, plus l'inertie interne de chacune d'elles est faible et plus l'inertie intraclasse de la partition C est faible. A la limite, la partition en 14 classes à 1 élément (confondu avec le centre de gravité de la classe) a une inertie intraclasse nulle. La partition la moins fine qui soit : celle qui n'a qu'une seule classe identique au nuage $N(I)$, a pour inertie intraclasse l'inertie du nuage $N(I)$: 1.112 (valeur qui s'obtient en faisant la somme des valeurs propres issues de l'analyse des correspondances, et qui d'autre part, nous le verrons, figure sur le listage de CAH). La partition en 3 classes (25, 22, {*I* EE}) a, nous l'avons vu plus haut, une inertie intraclasse de .491; celle en 4 classes (24, 20, 22, {*I* EE}) a pour inertie intraclasse .375, comme nous le verrons plus bas à l'aide des listages de CAH. Nous reviendrons sur l'inertie intraclasse au § 3.4 après avoir défini l'inertie interclasse. Remarquons seulement ici que l'on a

$$0 < .375 < .491 < 1.112;$$

de la partition la plus fine en 14 classes, à la partition la moins fine en une seule classe, l'inertie intraclasse varie en croissant, au fur et à mesure que certaines classes sont agrégées entre elles.

3.2. *Inertie interclasse d'une partition et séparation des classes :*

Considérons à nouveau la partition en 3 classes (25, 22, {*I* EE}). Les 3 centres de gravité de ces classes : g (25), g (22) et g ({*I* EE}) munis chacun du poids de sa classe, constituent un nuage dans l'espace où est défini $N(I)$: c'est le *nuage des centres de classes de la partition.* Ce nuage des centres a même centre de gravité que le nuage $N(I)$, en vertu de l'associativité de l'opération qui consiste à prendre le centre de gravité de plusieurs points (on peut remplacer une partie de ces points par leur centre de gravité muni de la somme de leurs poids).

Définition : On appelle *inertie interclasse* de la partition C = (25, 22, {*I* EE}) l'inertie du nuage des centres $\{$g (25), g (22), g ({*I* EE})$\}$ par rapport au centre de gravité de ce nuage (qui est aussi le centre de $N(I)$, origine des axes factoriels, que nous noterons O), cette inertie est notée Inter (C). On a : Inter (25, 22, {*I* EE}) = (poids de la cl. 25) × $\|O-\text{g}(25)\|^2$ + (poids de la cl. 22) × $\|O-\text{g}(22)\|^2$ + (poids de {*I* EE}) × $\|O-\text{g}\{I\ \text{EE}\}\|^2$ = .621.

(g {*I*EE}) n'est autre que *I* EE, cette classe étant réduite à un point).

On calcule de même l'inertie interclasse de la partition en 4 classes (24, 20, 22,{*I* EE}) : .737.

L'inertie interclasse mesure la séparation des classes de la partition en ce sens que plus le nuage des centres se disperse autour de O (centre de $N(I)$) et plus l'inertie interclasse de la partition est grande. A la limite, pour la partition en 14 classes à un seul élément, l'inertie interclasse coïncide avec l'inertie du nuage $N(I)$: 1.112. Au contraire, pour la partition en une seule classe 27 = $N(I)$, le nuage des centres se réduit au point O et l'inertie interclasse est nulle. Il faut toutefois souligner que la séparation des centres des classes ne suffit pas à assurer la séparation des classes elles-mêmes, car celle-ci requiert de plus qu'il y ait entre les classes un espace vide aussi large que possible.

3.3. *Complémentarité de l'inertie intraclasse et de l'inertie interclasse d'une même partition*

On a vu que de la partition de I la plus fine à la partition la moins fine l'inertie intercl. et l'inertie intracl. varient en sens opposé. De façon précise, on a la proposition suivante :

Pour toute partition C de I, la somme des inerties intraclasse et interclasse de cette partition est égale à l'inertie totale du nuage $N(I)$:

$$\text{Intra}(C) + \text{Inter}(C) = I_{\text{tot}}.$$

On vérifie cette proposition sur les quelques résultats numériques donnés plus haut :

Partition C	Intra (C) +	Inter (C) =	Total
14 classes à 1 élément	0 +	1.112 =	1.112
S = (24, 20, 22, {*I* EE})	.375 +	.737 =	1.112
S' = (25, 22, {*I* EE})	491 +	.621 =	1.112
Une seule classe	1.112 +	0 =	1.112

Quant à la démonstration, nous nous bornerons à dire qu'elle résulte immédiatement du théorème de Huyghens qui s'énonce ainsi :

Théorème de Huyghens : soit c un ensemble de points munis de masses dans un espace euclidien (i.e. c est un nuage); $g(c)$ le centre de gravité de c; f_c, la masse totale du nuage c; h un point quelconque de l'espace. Alors, on a :

$$\Sigma\, f_i \, \|i-h\|^2 \mid i \in c \;=\; \Sigma\, f_i \, \|i-g(c)\|^2 \mid i \in c \;+ f_c\, \|h-g(c)\|^2;$$

autrement dit : l'inertie du nuage c par rapport au point h est égale à la somme de l'inertie de c relativement à son centre de gravité $g(c)$ et du produit par la masse f_c du carré de la distance entre h et $g(c)$.

De la relation de complémentarité, il résulte qu'il suffit de calculer l'une des deux inerties Intra ou Inter pour connaître l'autre : dans la suite, nos calculs porteront principalement sur l'inertie Intra.

3.4. *Variation de l'inertie intraclasse par agrégation de deux classes*

Ainsi qu'on l'a expliqué au § 2.3, la construction ascendante d'une hiérarchie de classes peut se décomposer en une suite d'étapes élémentaires dont chacune consiste à agréger deux classes s et s' de la partition de I constituée par l'ensemble S des classes maximales (ou sommets) du système déjà construit. Ainsi, de la partition S on passe à la partition S' :

$$S' = (S - \{s, s'\}) \cup \{s \cup s'\};$$

S' diffère de S par la suppression de s et s' (en tant que sommets) et la création de $s \cup s'$: S' compte donc au total une classe de moins que S. Afin de comparer les partitions S et S', nous comparerons Intra (S) et Intra (S').

Au § 3.1, on a défini l'inertie intraclasse (Intra) d'une partition comme la somme des inerties internes des classes de cette partition; or, S et S' comportent les mêmes classes à 3 exceptions près qui sont s, s' et $s \cup s'$. On a donc :

$$\text{Intra}(S') = \text{Intra}(S) + I \text{ interne de } (s \cup s') - I \text{ interne de } s. \\ - I \text{ interne de } s'.$$

La différence : Intra (S') – Intra (S) qui vaut :

$$I(s \cup s') - I(s) - I(s'),$$

se calcule immédiatement par la formule de complémentarité du § 3.3 appliqué au nuage $s \cup s'$. En effet, (s, s') constitue une partition en deux classes de $s \cup s'$. L'inertie intraclasse de cette partition est :

$$\text{Intra}(s, s') = I(s) + I(s');$$

la différence qui nous intéresse s'écrit donc encore :

$$I(s \cup s') - \text{Intra}(s, s');$$

or, la formule de complémentarité affirme :

$$I(s \cup s') = \text{Intra}(s, s') + \text{Inter}(s, s');$$

d'où il résulte que la différence Intra (S') – Intra (S) n'est autre que Inter (s, s'), c'est-à-dire par définition l'inertie (relativement à son centre de gravité) du système des deux points $g(s)$ et $g(s')$ (centres de gravité des cl. s et s') munis des masses f_s et $f_{s'}$ (poids des classes s et s'). C'est précisément ce qu'au § 2.2.4 on a noté $D_{\text{inert}}(s, s')$ ou encore crit (s, s'). On a donc :

$$\text{Intra}(S') = \text{Intra}(S) + \text{crit}(s, s');$$

en d'autres termes : en agrégeant deux classes s et s' de la partition S, on obtient une partition S' dont l'inertie intraclasse est supérieure à celle de S d'une quantité qui ne dépend que des deux classes agrégées (et non du reste de la partition) : crit (s, s').

Par exemple, considérons à nouveau la partition en 4 classes :

$$S = (24, 20, 22, \{I\,\text{EE}\});$$

en agrégeant les classes $s = 24$ et $s' = 20$, on obtient la partition en 3 classes :

$$S' = (25 = s \cup s', 22, \{I\,\mathrm{EE}\})\ (cf.\ \S\ 1.2 : \text{arbre})$$

Intra (S) = .375; Intra (S') = .491 (*cf.* § 3.1)

Intra (S') – Intra (S) = .491 – .375 = .116.

Calculons maintenant crit (24, 20) :

$$\text{crit}\,(24, 20) = (f_{24}\,f_{20} \,/\, (f_{24} + f_{20}))\ \|\,24\text{-}20\,\|^2$$

$$24 = \{I\ \mathrm{BL},\ I\ \mathrm{IT},\ I\ \mathrm{SP},\ I\ \mathrm{UK},\ I\ \mathrm{DL},\ I\ \mathrm{PL},\ I\ \mathrm{RM}\}$$

$$20 = \{I\ \mathrm{FR},\ I\ \mathrm{NL}\}$$

f_{24} (resp. f_{20}) est la somme des poids des pays constituant la cl. 24 (resp. 20).

$$f_{24} = .060 + .059 + .057 + .049 + .09 + .082 + .043 = .44;$$

$$f_{20} = .150 + .058 = .208$$

d'où : $f_{24}\,f_{20} \,/\, (f_{24} + f_{20}) = .14$.

Pour calculer $\|24 - 20\|^2$, on peut se dispenser d'effectuer les soustractions en utilisant le listage FACOR qui donne les différences $D_\alpha = F_\alpha\,(24) - F_\alpha\,(20)$ pour $\alpha = 1, \ldots, 7$:

Nœud	Aine	BJMN	*D1*	*D2*	*D3*	*D4*	*D5*	*D6*	*D7*
25	24	20	–.341	401	–.661	–.193	.228	.140	–.013

on a :

$$\|24\text{–}20\|^2 = D_1^2 + D_2^2 + D_3^2 + D_4^2 + D_5^2 + D_6^2 + D_7^2$$

$$= .120 + .161 + .437 + .037 + .052 + .020 + .000$$

$$= .827$$

D'où pour crit (24, 20) :

$$\text{Crit}\,(24, 20) = .14 \times .827 = .116,$$

ce qui est bien la valeur de Intra (S') – Intra (S).

3.5 *Points de vue local et global en classification ascendante hiérarchique.*

De la complémentarité des inerties, il résulte que pour autant qu'on les évalue d'après Intra (C) et Inter (C), les deux qualités de compacité et de séparabilité des classes d'une partition C sont non seulement compatibles, mais équivalentes : en effet, plus Intra (C) est faible, plus les classes sont compactes; et simultanément plus Inter (C) = (Itot ($N(I)$) – Intra (C)) est grand et, donc, meilleure est la séparation entre les classes (ou tout au moins entre leurs centres).

Donc, la partition idéale serait celle de I en classes réduites à un seul élément (14 dans l'exemple), pour laquelle Intra = 0. Mais une telle partition n'offre aucun intérêt puisqu'elle s'identifie à I lui-même. Pour être utile, une partition C doit constituer une schématisation des données. L'intérêt d'un schéma est d'être simple et fidèle. Il est d'autant plus simple que le nombre des classes est plus petit; il est d'autant plus fidèle que chaque classe peut être assimilée à un point; autrement dit que Intra (C) est plus faible.

Ceci suggère de regarder d'un point de vue nouveau l'algorithme de CAH. Agréger deux sommets s et s' c'est substituer à la partition S une partition S' qui est *plus simple* en ce qu'elle compte une classe de moins que S, mais *moins fidèle* en ce que Intra (S') est supérieur à Intra (S). La simplification se paye d'un prix qui est la différence Intra (S') – Intra (S) = crit (s, s'). Pour que la fidélité *globale* du schéma soit aussi peu altérée que possible, il faut choisir d'agréger les deux classes s et s' entre lesquelles *localement* se réalise le minimum du critère crit (s, s').

Ainsi, le critère de l'inertie, introduit d'un point de vue local comme une mesure de l'écart entre deux classes, se justifie globalement en ce qu'il conduit par agrégation binaire à des partitions successives... S, S'..., certes de moins en moins fines, mais dont la fidélité aux données diminue le moins possible à chaque étape.

4. *Résultats de la CAH sur les listages issus de l'ordinateur*

Nous considérerons successivement l'histogramme de niveaux des nœuds (§ 4.1), l'arbre de la CAH (§ 4.2) et le tableau du contenu des classes (§ 4.3) en expliquant sommairement comment ces sorties graphiques sont préparées par l'algorithme (§ 4.0 et § 4.2 *in fine*).

4.0. *Déroulement de l'algorithme*

Nous suivrons le déroulement de l'algorithme ascendant sur 4 tableaux (d'une ligne) dont les cases sont numérotées de 15 à 27 (comme les nœuds) : d'une part, les tableaux A, B, D (Aîné, Benjamin, niveau) déjà introduits au § 1.2, d'autre part un tableau P (cardinal) dont l'utilité apparaîtra au § 4.2. Ces tableaux donnent les résultats de la CAH; d'autres tableaux sont créés au sein de l'ordinateur pour déterminer à chaque pas quels sont les sommets à agréger : nous n'en dirons rien ici.

ÉTAPE 0. L'algorithme part de la partition la plus fine qui soit : chacun des 14 pays importateurs (numérotés de 1 à 14) constitue une classe maximale ou sommet. L'inertie intraclasse est nulle; l'inertie interclasse est égale à l'inertie totale de $N(I)$: Itot = 1.112.

ÉTAPE 1. Parmi les 14 sommets, la paire qui réalise le minimum du critère est (1, 5), avec crit (1, 5) = .008. On doit donc agréger 1 et 5, pour créer un premier nœud qui, étant numéroté à la suite des 14 individus, reçoit le n° 15; on écrit :

$$A[15] = 1;\ B[15] = 5;\ D[15] = .008;\ P[15] = 2;$$

l'agrégation se faisant à un niveau $D[15]$ qu'on calcule comme au § 1.3 :

$$D[15] = \text{crit}\,(A[15], B[15]) = \text{crit}\,(1,5) = .008.$$

Ainsi qu'on l'a dit au § 1.3, il importe que l'un ou l'autre des individus agrégés reçoive le titre d'Aîné ou de Benjamin (on aurait pu poser $A[15] = 5$, $B[15] = 1$). $P[15] = 2$ parce que la classe 15 compte 2 individus. Il y a présentement 13 sommets (les individus sauf 1 et 5 et le nœud 15) et l'on a pour cette partition en 13 sommets :

$$\text{Intra} = 0 + D[15] = .008;\ \text{Inter} = \text{Itot} - \text{Intra} = 1.104$$

	15	16	17	18	19	20	21	22	23	24	25	26	27
A	1	6	12	15	2	3	18	19	4	21	24	25	26
B	5	11	13	8	16	7	9	10	17	23	20	22	14
D	.008	.010	.010	.023	.024	.025	.043	.059	.069	.104	.116	.271	.342
P	2	2	2	3	3	2	4	4	3	7	9	13	14

les informations relatives au nœud 15 sont créées à l'étape 1

les informations relatives au nœud 16 sont créées à l'étape 2

ÉTAPE 2 : création du nœud 16. Parmi les 13 sommets, la paire qui réalise le minimum du critère est constituée des deux individus 6 et 11; on a : crit (6, 11) = .010. On écrit donc pour le nœud 16 :

$$A[16] = 6;\ B[16] = 11;\ D[16] = .010;\ P[16] = 2$$

Il y a présentement 12 sommets (ou classes maximales) : les 10 individus qui n'ont pas encore été agrégés; et les deux nœuds déjà créés. On a pour la partition en 12 nœuds :

$$\text{Intra} = 0 + D\,[15] + D\,[16] = .018;\ \text{Inter} = \text{Itot} - \text{Intra} = 1.094;$$

en effet, en agrégeant 6 et 11 pour créer le nœud 16, on a augmenté l'inertie intraclasse de crit (6, 11) = [16] (*cf.* § 3.4).

etc.

ÉTAPE 12 : création du nœud 26. Tous les individus, à l'exception de 14 (*I* EE) ont été successivement agrégés soit par paires (cas de (1, 5), (6, 11), (12, 13), (3, 7), soit à des classes déjà créées (8 à 15; 2 à 16; 9 à 18; 10 à 19; 4 à 17). Quant aux 11 classes successivement créées par l'algorithme (classes de 15 à 25) deux seulement (25 et 22) n'ont pas été agrégées entre elles ou à des individus pour créer des classes plus grandes. Il subsiste donc 3 sommets : 14, 22 et 25. La paire qui réalise le minimum du critère est (22, 25) avec : crit (22, 25) = .271. On écrit donc :

$$A\,[26] = 25;\ B\,[26] = 22;\ D\,[26] = \text{crit}\,(25, 22) = .\,271;\ P\,[26] = 13;$$

le nombre P [26] des éléments de la classe 26 est égal à la somme de P [25] et P [22]. Il ne subsiste que deux sommets 26 et 14; on a pour la partition en deux classes :

$$\text{Intra} = 0 + D\,[15] + D\,[16] + \ldots + D\,[26] = .770;$$
$$\text{Inter} = \text{Itot} - \text{Intra} = .342;$$

le calcul de Intra se faisant à chaque étape en ajoutant l'écart des deux classes agrégées (§ 3.4).

ÉTAPE 13 : création du nœud 27. Puisqu'il ne reste plus que deux sommets, l'agrégation à effectuer s'impose; on a :

$$A\,[27] = 26;\ B\,[27] - 14;\ D\,[26] = \text{crit}\,(26, 14) = .342;$$
$$\text{Intra} = 0 + D\,[15] + \ldots + D\,[27] = 1.112 = \text{Itot};\ \text{Inter} = 0.$$

La CAH proprement dite est achevée; il reste des calculs complémentaires à effectuer pour le dessin de l'arbre (§ 4.2).

4.1 *L'histogramme des niveaux des nœuds*

De même que le listage d'analyse des correspondances, le listage de CAH commence par un histogramme; sur celui-ci sont portés les tableaux *A, B, D*.

Dans le listage d'a. des corr. l'histogramme des valeurs propres est présenté en un tableau à autant de lignes qu'il y a de v.p., celles-ci étant rangées de haut en bas par valeurs décroissantes ; $\lambda 1 > \lambda 2 > \ldots$; avec les pourcentages d'inertie correspondants $\tau 1 = \lambda / \text{Itot}$; $\tau 2 = \lambda 2 / \text{Itot}$ etc.; la somme des $\tau\alpha$ étant égale à 1 parce que la somme des $\lambda\alpha$ n'est autre que l'inertie totale du nuage.

En CAH (avec le critère de l'inertie) l'inertie totale du nuage est égale à la somme des niveaux des nœuds (*cf.* § 4.0 dernière étape); et c'est pourquoi, par analogie avec la lettre λ affectée aux valeurs propres en a. des corr., le niveau d'un nœud est souvent désigné par la lettre ν (au lieu de D). Mais avec le numérotage adopté qui est celui de l'ordre de la création des nœuds suivant l'algorithme de base, on a (en vertu de l'axiome de la médiane, *cf.* 2.3) :

$$\nu\,15 \leqslant \nu\,16 \leqslant \ldots \leqslant \nu\,27;$$

les nœuds sont donc rangés de haut en bas selon l'ordre décroissant de leurs numéros : 27, 26... 15 qui est aussi l'ordre décroissant des valeurs ν.

Chaque ligne donne successivement :

- le numéro du nœud : colonne N;
- le niveau du nœud : colonne D [N];

– les numéros de l'Aîné et du Benjamin : col. *A* [N] et *B* [N];

– le taux d'inertie τ afférent au nœud : $\tau\ (N) = D\ [\mathrm{N}]\ /\ \mathrm{Itot}$;

– un segment de longueur proportionnelle à *D* [N], l'échelle étant choisie de telle sorte que le segment le plus long (première ligne) prenne toute la largeur disponible.

le 1[er] chiffre est celui des entiers

INERTIE TOTALE DU NUAGE 11125 E + 01

N	*D* [*N*]	*A* [*N*]	*B* [*N*]	*T* (*N*)
27	34	26	14	307
26	27	25	22	244
25	11	24	20	104
24	10	21	23	93
23	6	4	17	62
22	5	19	10	53
21	4	18	9	39
20	2	3	7	22
19	2	2	16	22
18	2	15	8	21
17	1	12	13	9
16	1	6	11	9
15	0	1	5	7

(niveau du nœud N / inertie tot. du nuage) en millièmes

n[os] des classes qui s'agrègent au nœud *N*

niveau du nœud *N* en centièmes

n° du nœud *N*

Remarque : la CAH définit sur chacune des classes *n* de la hiérarchie une structure hiérarchique, représentée graphiquement par la branche suspendue au nœud *n* et dont les nœuds sont, outre *n*, les nœuds de la CAH qui sont les descendants, immédiats ou non de *n*. Par exemple, la hiérarchie définie sur la cl. 22 compte 3 nœuds : 22, *A* [22] = 19, *B* [19] = 16 et 4 individus terminaux : *B*[22] = 10, *A* [19] = 2, *A* [16] = 6, *B* [16] = 11. A cette hiérarchie s'applique la formule du calcul de l'inertie intraclasse de *n* comme somme des niveaux; on a ici :

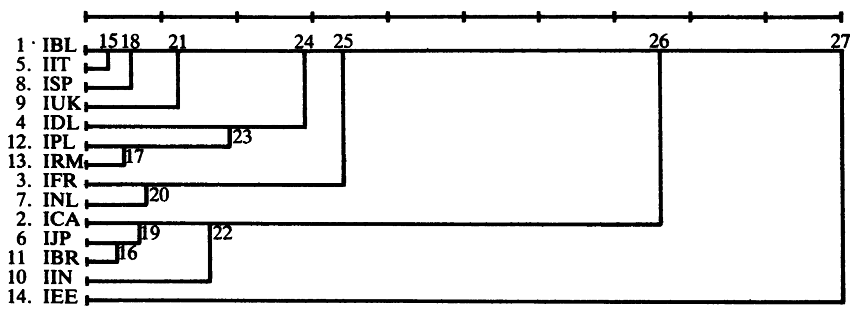
$$\text{Inertie interne de } 22 = D\ [22] + D\ [19] + D\ [16] = .093$$

C'est à cette formule qu'on fait allusion au § 3.1.

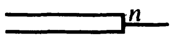
4.2 *L'arbre de la CAH*

Le graphique imprimé par l'ordinateur correspond à celui présenté au § 1.2, mais avec quelques différences que nous justifierons avant d'expliquer la construction de l'arbre.

Première différence : le graphique est tourné d'un angle droit; les sigles des individus (qui étaient à la base au § 1.2) sont ici à la marge de gauche. La raison : certaines classifications portent sur des centaines d'individus et l'on ne pourrait écrire tous les sigles sur la largeur d'une page; en revanche le déroulement du listage permet d'imprimer une colonne aussi longue qu'il le faut.



Deuxième différence : les nœuds sont tracés de façon dissymétrique.

ici :  ; au § 2.3 : 

la raison : la précision des graphiques imprimés étant limitée par l'espacement des caractères, on doit simplifier le tracé au maximum pour éviter que les points et les traits ne se superposent.

Troisième différence : pour la même raison, la plupart des programmes de tracé n'inscrivent pas les n[os] des nœuds qui doivent être reportés manuellement; et comme, malgré ces simplifications, la partie inférieure de l'arbre est souvent peu lisible, on doit compléter le graphique par un tableau donnant explicitement le contenu des classes (*cf.* § 4.3).

Quant au tracé automatique de l'arbre, on remarquera d'abord que du fait que le graphique est composé par une imprimante qui est une sorte de machine à écrire, les lettres et traits élémentaires dont est formé le graphique se rangent nécessairement sur une suite de lignes régulièrement espacées. A toute classe c, il correspond un ensemble d'individus dont les sigles sont inscrits à gauche sur un bloc de lignes consécutives dont le nombre est égal à P (c), cardinal de la classe; mettre en place ces blocs est la tâche essentielle de l'algorithme de tracé.

Par exemple, à la classe 21 correspond le bloc des 4 premières lignes; à la classe 23 le bloc des lignes suivantes (5[e], 6[e], 7[e]); à la classe 22 le bloc des lignes (10[e], 11[e], 12[e], 13[e]); à la classe 2 composée du seul élément I CA, correspond une seule ligne, la 10[e], etc. En général, si on note $DEB\ (c)$ le rang de la 1[re] ligne du bloc afférent à la classe c, le rang de la dernière ligne du bloc sera $DEB\ [c] + P[c] - 1$.

Pour emplir le tableau des valeurs de DEB, on remarque d'abord que $DEB\ [27] = 1$, puisque la classe 27 qui est I tout entier occupe les 14 lignes. D'autre part, avec $DEB\ (c)$ on connaît la place des deux descendants immédiats de c qui occupent des blocs consécutifs.

$$DEB\ [A\ (c)] = DEB\ [c];\ DEB\ [B\ (c)] = DEB\ [c] + P\ [c].$$

D'où le tableau de la fonction DEB : après $DEB\ [27] = 1$, on écrit :

$$DEB\ [A\ (27)] = DEB\ [26] = 1 \text{ et } DEB\ [B\ (c)] = DEB\ [14] = 1 + P\ [26] = 14.$$

etc.

	15	16	17	18	19	20	21	22	23	24	25	26	27
DEB	1	11	6	1	10	8	1	10	5	1	1	1	1
1	10	8	5	2	11	9	3	4	13	12	6	7	14
1	2	3	4	5	6	7	8	9	10	11	12	13	14

On achève alors aisément le tracé. D'une part, le sigle du i[me] individu doit être inscrit sur la marge gauche de la ligne de rang $DEB\ [i]$; par exemple $DEB\ [9] = 4$: on écrit à la 4[e] ligne le sigle du pays importateur 9 = IUK. D'autre part, pour chaque nœud n on doit tracer deux traits horizontaux occupant respectivement les lignes de rang $DEB\ [n]$ et $DEB\ [n] + P\ [n] = DEB\ [B(\mathrm{n})]$; ces traits débutent à gauche immédiatement après la colonne des sigles et ont une longueur proportionnelle au niveau $D\ [n]$; enfin leurs extrémités droites sont réunies par un trait vertical. (Nous n'insistons pas sur le fait que selon nos instructions simplifiées certaines parties de traits horizontaux sont tracées plusieurs fois).

4.3 *Le tableau du contenu des classes*

Dans ce tableau divisé en 6 colonnes successives, chaque nœud occupe une ligne ou plusieurs (selon l'effectif de la classe).

Dans la colonne de droite on lit les n[os] des nœuds : de 15 à 27 dans notre cas; les 3 colonnes suivantes dans les nombres D [N], A [N], B [N] déjà rencontrés sur l'histogramme § 4.1. La 5[e] colonne donne l'effectif P [N] de la colonne N (e.g. P [27] = 14 : la cl. 27 est I tout entier). Enfin, sous le titre : « Description des classes de la hiérarchie », on trouve la liste des individus de chaque classe, rangés dans l'ordre où ils sont imprimés en marge de l'arbre. Par exemple, pour la cl. 24, on a :

✻*I* BL ✻ *I* IT ✻ *I* SP ✻ *I* SP ✻ *I* UK ✻ *I* DL ✻ *I* PL ✻ *I* RM

Viennent d'abord les 4 éléments de la classe A [24] = 21, puis les 3 éléments de B [24] = 23; ce qu'on peut préciser à la main par des parenthèses; éventuellement, on peut noter aussi des subdivisions par des parenthèses emboîtées (e.g. de la cl. 21 en la cl. 18 et l'individu 9, etc.).

(✻ *I* BL ✻ *I* IT ✻ *I* SP (✻ *I* UK)) (✻ *I* DL ✻ *I* PL ✻ *I* RM)

n° du nœud N

niveau du nœud en millièmes

n[os] des classes qui s'agrègent au nœud N

nombre de pays constituant la cl. N

contenu de la cl. N.

N	D [N]	A [N]	B [N]	P [N]		Description des classes de la hiérarchie
15	8	1	5	2		✻*I*BL✻*I*IT
16	10	6	11	2		✻*I*SP✻*I*BR
17	10	12	13	2		✻*I*PL✻*I*RM
18	23	15	8	3		✻*I*BL✻*I*IT✻*I*SP
19	24	2	16	3		✻*I*CA✻*I*JP✻*I*BR
20	25	3	7	2		✻*I*FR✻*I*NL
21	43	18	9	4		✻*I*BL✻*I*IT✻*I*SP✻*I*UK
22	59	19	10	4		✻*I*CA✻*I*JP✻*I*BR✻*I*IN
23	69	4	17	3		✻*I*DL✻*I*PL✻*I*RM
24	104	21	23	7		✻*I*BL✻*I*IT✻*I*SP✻*I*UK✻*I*DL✻*I*PL✻*I*RM
25	116	24	20	9		✻*I*BL✻*I*IT✻*I*SP✻*I*UK✻*I*DL✻*I*PL✻*I*RM✻*I*FR✻*I*NL
26	271	25	22	13		✻*I*BL✻*I*IT✻*I*SP✻*I*UK✻*I*DL✻*I*PL✻*I*RM✻*I*FR✻*I*NL ✻*I*CA✻*I*JP✻*I*BR✻*I*IN
27	342	26	14	14		✻*I*BL✻*I*IT✻*I*SP✻*I*UK✻*I*DL✻*I*PL✻*I*RM✻*I*NL ✻*I*CA✻*I*JP✻*I*BR✻*I*IN✻*I*EE

§ 4.3 : *le tableau du contenu des classes*