LES CAHIERS DE L'ANALYSE DES DONNÉES

F. BENZÉCRI

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

Les cahiers de l'analyse des données, tome 10, n° 3 (1985), p. 279-302

<http://www.numdam.org/item?id=CAD_1985__10_3_279_0>

Les Cahiers de l'Analyse des Données
Vol. X - 1985 - n° 3 - pp. 279-302

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

[C.A.H. ECO.]

par F. Benzécri *

Le présent exposé fait suite à [*CORR. ECO.*] d'une part en ce qu'il suppose que le lecteur est familier avec les notions géométriques de l'analyse des correspondances, d'autre part en ce qu'il utilise pour les calculs numériques le même tableau de correspondance I × J, (14 × 8), relatif au commerce mondial des phosphates. L'essentiel de ces présupposés est rappelé dans 3 cadres inclus dans le texte.

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 pour l'exemple la présentation d'une CAH sur les listages issus de l'ordinateur. Un exposé ultérieur ([*AIDE CAH ECO*]) est consacré à l'interprétation d'une CAH d'après des calculs statistiques auxiliaires.

1 *Notions 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 × 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 masses dans l'espace euclidien des profils sur J où la distance est celle du chi2 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

(*) *Docteur-ès-sciences.*

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, il peut être amené à adjoindre à I des éléments 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 classificiation niérarchique, 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) :

{ JBL, JIT, JSP, JUK, JDL, JPL, JRM, JFR, JNL}; { JCA, JJP, JBR, JIN} ;{ JEE}.

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 ×J (cf. *infra* le dessin de l'arbre).

L'arbre comprend à sa base les 14 individus à classer et, aboutissant à ces individus, des branches se raccordant entre elles par des *noeuds* ; à ces noeuds aboutissent de nouvelles branches qui se raccordent entre elles par de nouveaux noeuds ; ainsi de suite jusqu'au *sommet*. Les individus portent le numéro qu'ils ont dans le tableau I×J (de 1 à 14) ; les noeuds 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 les deux branches allant l'une vers le noeud 26, l'autre vers JEE = 14 ; et de même, le noeud 26 se subdivise par ramifications successives juqu'aux individus encore appelés *terminaux*.

Un tel arbre définit un système emboîté de classes : plus exactement un système *dichotomique* parce que de chaque noeud 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 JBL à JIN, à la base du graphe) ; l'autre branche B(27) est réduite à un seul individu terminal JEE. 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

RAPPEL I : Le tableau de correspondances I × J croise les ensembles :

I : 14 pays importateurs de phosphates :

Belgique, Canada, France, Deutschland, Italie, Japon,
JBL JCA JFR JDL JIT JJP

Nederlend, Espagne, United Kingdom, Indes, Brésil
JNL JSP JUK JIN JBR

Pologne, Roumanie, autres pays d'Europe de l'Est
JPL JRM JEE = {RAD, Bulgarie, Hongrie}

la 1-ère lettre de ces 14 sigles rappelle qu'il s'agit de pays importateurs ;

J : 8 pays exportateurs (ƶ):

Belgique, USA, Jordanie, Maroc, Sénégal, Togo, Tunisie,
ƶBL ƶUS ƶJR ƶMR ƶSN ƶTG ƶTN

URSS (CCCP).
ƶCC

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.

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

noeuds 21 et 23 définissant les deux classes des individus qui leur sont respectivement rattachés :

$$21 = \{ \text{JBL} \ ; \ \text{JIT} \ ; \ \text{JSP} \ ; \ \text{JUK}\} \ ; \ 23 = \{ \text{JDL} \ ; \ \text{JPL} \ ; \ \text{JRM}\},$$

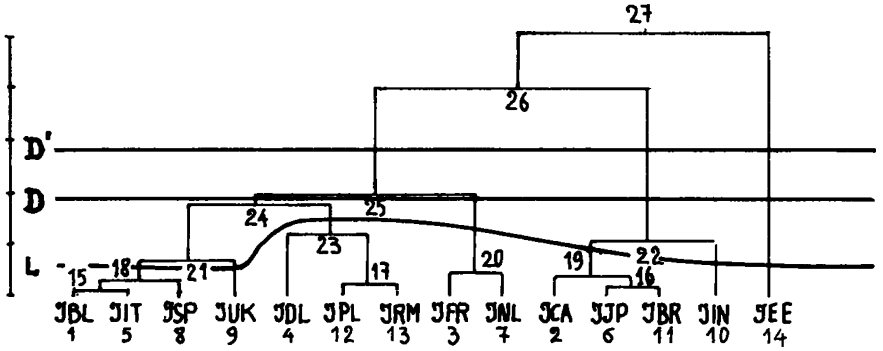
et la classe 24 étant la réunion de celles-ci, ce qu'on écrira :

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

A gauche de l'arbre, une échelle graduée à partir de 0 situé à la base, indique le *niveau* de chaque noeud. On peut dire, en bref, que le niveau d'un noeud représente par un nombre le degré de généralité de la classe définie par ce noeud. 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 noeud 24 est à un niveau supérieur à celui du noeud 22. En particulier, comme le montre clairement le dessin, les noeuds A(n) et B(n) descendants immédiats (aîné et benjamin) du noeud n, sont à des niveaux inférieurs à celui de n. On vérifiera que le numérotage des noeuds est fait dans l'ordre des niveaux croissants . Suivre dans l'arbre un *chemin descendant* c'est considérer une suite de noeuds ou individus : n1, n2, n3... telle que n(p+1) soit descendant immédiat de np : par exemple, 25, 24, 23, 17 { JPL} = 12, est un chemin descendant ; le chemin inverse : { JPL}, 17, 23, 24, 25, un *chemin ascendant*. Quand on parcourt un chemin descendant, les n°s des noeuds 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 noeud 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 noeud 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 noeu 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 noeuds 24 et 25, on a une partition S en 4 classes :

$$I = 27 = 24 \cup 20 \cup 22 \cup \{ \text{JEE} \} \; ; \; S = \{24 \; ; \; 20 \; ; \; 22 \; ; \{ \text{JEE} \}\} \; ;$$

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

$$I = 27 = 25 \cup 22 \cup \{ \text{JEE} \} \quad ; \quad S' = \{25 \; ; \; 22 \; ; \{ \text{JEE} \}\} .$$

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 :

$$I = 27 = 18 \cup \{ \text{JUK} \} \cup 23 \cup 20 \cup 19 \cup \{ \text{JIN} \} \cup \{ \text{JEE} \} \; ;$$

$$C = \{18 \; ; \{ \text{JUK} \} \; ; \; 23 \; ; \; 20 \; ; \; 19 \; ; \{ \text{JIN} \} ; \{ \text{JEE} \}\} .$$

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 (cf. Fig. 1). 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 (Fig. 2).

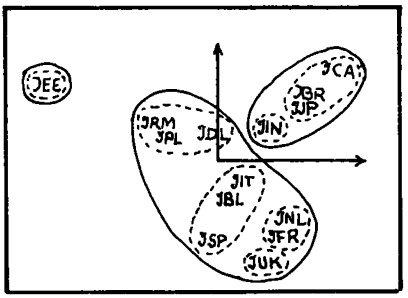


Fig. 1

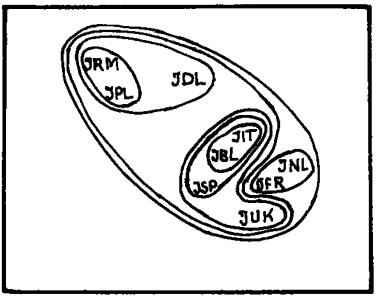


Fig.2

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 noeud 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' = 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 noeud I = 27 se scinde en ses deux descendants immédiats A(27) = 26 et B(27) = 14 = {JEE} ; le noeud 26 se scinde en A(26) = 25 et B(26) = 22 ; etc. . La *lecture ascendante* part de la base : les individus 1 et 5 (JBL et JIT) s'*agrègent* pour former la classe 15 dont ils sont les deux descendants immédits : 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 classes 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 constituée ; puis des classes entre elles, créant ainsi des noeuds 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 XVII^{e} 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'oeuvre d'Adanson et de Linné au milieu du XVIII^{e} siècle... Voila 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.

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 des calculs complémentaires (cf. [*AIDES CAH*]) 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 struture 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.

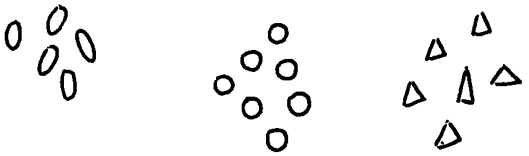
2.1 Qualités d'une classification : Une classification hiérarchique s'interprète en terme de partition (§ 1.2) et se construit par un algorithme ascendant comme une suite de partitions de moins en moins fines (§ 2.3) : il nous suffira donc ici de considérer le cas d'une partition.

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

---

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

(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 multidimensiinnel (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.

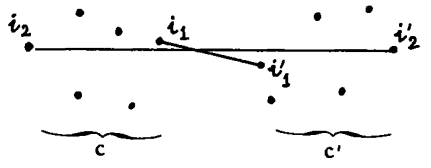
[C.A.H. ECO.] *§2.1. On a suggéré par des cadres deux façons de grouper les classes A, B, C, D, E, F, G, H.*

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 (§ 3) assure à la classification les qualités requises.

2.2.1 *Critère du saut minimum* : $D_{saut}(c,c')$ 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 et l'autre à c' et le plus proches possible).



[CAH ECO] §2.2
$D_{saut}(c,c') = D(i_1,i'_1)$
$D_{diam}(c,c') = D(i_2,i'_2)$

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* : $D_{diam}(c,c')$ est la distance maxima entre un point de c et un point de c'. En agrégeant en priorité les paires de classes 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 :

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

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

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

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.

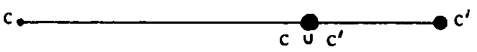
2.2.4 *Critère de l'inertie* : Pour calculer ce critère, on doit supposer que I est un ensemble de points munis de masse d'un espace euclidien (c'est le cas dans l'exemple du commerce des phosphates qui sert de base au présent exposé). Comme au § 2.2.3, on note respectivement f_c, $f_{c'}$, les masses 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 :

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

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 g(c,c') de ce nuage à 2 points.

Du fait de l'association de la construction du centre de gravité, g(c,c') est aussi le centre de gravité de la réunion $c \cup c'$ des classes c et c' : on peut donc le noter $c \cup c'$. Et on a la figure suivante qui dans la suite sera appelée *dipôle*.



$c \cup c'$ est le point du segment (c,c') qui divise ce segment dans le rapport inverse des masses de c et c' :

$\|(c \cup c')-c\|/\|(c \cup c')-c'\|$ = poids de c'/poids de c.

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 satisfaisant 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_{saut} ni par D_{diam}. En effet, soit c et c' deux parties distinctes non vides ayant un point commun. On aura $D_{saut}(c,c') = 0$ bien que $c \neq c'$; et on aura $D_{diam}(c,c) \neq 0$ si c comprend au moins deux points car $D_{diam}(c,c)$ 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 :

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

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.

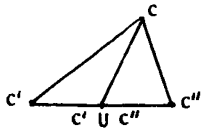
2.3.2 Axiomes de la médiane : De même qu'une distance, tout critère D(c,c') devra être une fonction positive (ou nulle) ne dépendant pas de l'ordre dans lequel on prend les deux classes c et c' :

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

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' ∪ 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 :

$$\forall\ c, c' c" \subset I : D(c',c") \leq \inf(D(c,c'), D(c,c"))$$

$$\Rightarrow \inf(D(c,c'),D(c,c")) \leq D(c, c' \cup c")$$



Ainsi, si dans la suite, on doit agréger c à n'=(c'∪ c") pour créer une classe n = c ∪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 noeud n' ; ce qui correspond sur le dessin de l'arbre au fait que tout noeud 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'∪ c" (assimilé à un centre de gravité de classe) sera placé au milieu du côté (c',c"). Dès lors (c, c' ∪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 *Algorithme de base et algorithmes accélérés* : L'algorithme de CAH, sous la 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.

ETAPE 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 noeud 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 noeud 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.

ETAPE 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 noeud 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 2e étape (Card I - 2) sommets. Ceux-ci constituent une nouvelle partition de I.

ETAPE 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)ème noeud qui n'est autre que I tout entier et reçoit le n° (2Card I - 1). Cet ultime noeud 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 $(\text{Card I})^3$. Des algorithmes accélérés édifient la même CAH en un temps de l'ordre de $(\text{Card I})^2$: en bref, ces algorithmes procèdent en agrégeant plusieurs paires de sommets à la fois, ou en découvrant des paires de sommets à agréger sans

RAPPEL II : I : ensemble de 14 pays importateurs ;

J : ensemble de 8 pays exportateurs .

Pour chaque couple (i,j) d'un pays i de I et d'un pays j de J, le tableau I × J donne le nombre k(i,j) de milliers de tonnes de P_2O_5 exportés 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 × J, tout pays importateur est décrit par

la ligne i : $\{k(i,j) \mid j \in J\}$;

tout pays exportateur est décrit par

la colonne j : $\{k(i,j) \mid i \in I\}$.

La *ligne de marge* du tableau I × 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 × J, noté k.

j

i \ j	Belgique	USA	Jordanie	Maroc	Sénégal	Togo	Tunisie	URSS	marge
Belgique		1305		3573	25	500	110	293	5806
Canada	2	8335				8			8345
France	1311	2691	70	4891	1484	2526	1697	4	14674
Deutschl.	1322	3808		1445	261	200	288	1442	8767
Italie	42	1883	194	2881	67	195	493	ε	5755
Japon		4426	522	1540	239	93		ε	6819
Nederl.	299	1483		1853	249	1584	59	84	5611 ← k(i)
Espagne	20	339		5073	2	85	11		5531
Un. King.	122	645	2	2852	971	36	197		4825
Indes		2559	996	1149	218		134		5057
Brésil	29	4918		1398	15		241		6602
Pologne	ε	1284	271	3311	33	540	548	1996	7984
Rouman.	ε	768	483	1541	6	139	22	1206	4165
Eur. Est	2	201	136	1398			333	5533	7603
marge	3151	34646	2673	32905	3571	5906	4134	10559	97545 ← k

k(i,j) ↑ k(j) ↑

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

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

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

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

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

j

I \ J	ℰBL	ℰUS	ℰJR	ℰMR	ℰSN	ℰTG	ℰTN	ℰCC	POIDS
JBL		225		615	4	86	19	51	60
JCA		999				1			86
JFR	89	183	5	333	101	172	116		150
JDL	151	434		165	30	23	33	164	90
JIT	7	327	34	501	12	34	86		59
JJP		649	77	226	35	14			70
JNL	53	264		330	44	282	11	15	58
JSP	4	61		917		15	2		57
JUK	25	134		591	201	8	41		49 ← $\frac{k(i)}{k} = f_i$
JIN		506	197	227	43		26		52
JBR	4	745		212	2	68	36		68
JPL		161	34	415	4	33	69	250	82
JRM		184	116	370	1		5	290	43
JEE		26	18	184			44	728	78
marge	32	355	27	337	37	61	42	108	1000 ← k/k

f^i_J ; $f^i_j = \frac{k(i,j)}{k(i)}$

$k(j)/k = f_j$ (Les nbs. de ce tabl. sont des millièmes)

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

Conformément au principe d'équivalence distributionnelle, la distance entre profils est définie par la distance euclidienne du χ^2 :

$$\|f^i_J - f^{i'}_J\|^2 = \Sigma\{(1/f_j)(f^i_j - f^{i'}_j)^2 \mid j \in J\}$$

Dans le système de coordonnées fourni par l'analyse factorielle la distance prend la forme classique :

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

éventuellement, on utilisera une formule approchée telle que :

$$\|f^i_J - f^{i'}_J\|^2 = \Sigma\{(F_\alpha(i) - F_\alpha(i'))^2 \mid \alpha = 1,\ldots, 4\}.$$

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 partiton 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) effectuées 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 en 3 classes S' = (25, 22, 14 = { ƱEE}) 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, {ƱEE}) la somme des inerties internes des classes 25, 22 et {ƱEE} 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 { ƱEE}. On a :

inertie interne de la cl. 25 = .398 ;
inertie interne de la cl. 22 =.093 ;
inertie interne de la cl. {ƱEE}= 0 ;

d'où l'inertie intraclasse de la partition S' = (25, 22, {ƱEE}) :

Intra(S') = .398 + .093 + 0 = .491.

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

$$F_\alpha(g(22)) = (.086\ F_\alpha(\text{ƱCA}) + .070\ F_\alpha(\text{ƱJP}) + .068\ F_\alpha(\text{ƱBR}) + .052\ F_\alpha(\text{ƱIN}))/.276$$

où .086, .070 etc. sont des poids de ƱCA, ƱJP etc. (cf. tableau des profils RAPPEL III) ; .276 est la somme des poids des 4 pays ƱCA, ƱJP, ƱBR et ƱIN ; $F_\alpha(\text{ƱCA})$ etc. les $\alpha^{èmes}$ facteurs des 4 pays,

facteurs que l'on trouve au tableau des facteurs sur I issu de l'a. des c. du tableau I × J ; voici les valeurs de $F_\alpha(g(22))$ ainsi calculées :

	F1	F2	F3	F4	F5	F6	F7
g(22)	.591	.657	-.118	.067	-.009	-.055	-.002

Puis on calcule les 4 distances au carré :

$\|g(22) - \mathfrak{J}CA\|^2$, $\|g(22) - \mathfrak{J}JP\|^2$, $\|g(22) - \mathfrak{J}BR\|^2$, $\|g(22) - \mathfrak{J}IN\|^2$:

$\|g(22) - \mathfrak{J}CA\|^2 = \Sigma\{(F_\alpha(g(22)) - F_\alpha(\mathfrak{J}CA))^2 \mid \alpha = 1,2,\ldots,7\}$;

etc. ;

Voici les valeurs des 4 distances au carré :

i	$\mathfrak{J}CA$	$\mathfrak{J}JP$	$\mathfrak{J}BR$	$\mathfrak{J}IN$
$\|g(22) - i\|^2$	.366	.076	.145	.930

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

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

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 les listages 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, {$\mathfrak{J}EE$}) a, nous l'avons vu plus haut, une inertie intraclasse de .491 ; celle en 4 classes (24, 20, 22, {$\mathfrak{J}EE$}) 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, { JEE}). Les 3 centres de gravité de ces classes : g(25), g(22) et g(JEE) munis chacun du poids de sa classe, constituent un nuage dans l'espace où est défini N(I) : c'est le *nuage des centres des 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, { JEE}) l'inertie du nuage des centres {g(25), g(22), g({ JEE})} 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, { JEE}) =

((poids de la cl. 25) × $\|O-g(25)\|^2$) + ((poids de la cl. 22) × $\|O-g(22)\|^2$)

+ ((poids de JEE) × $\|O-g\{JEE\}\|^2$) = .621.

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

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

Intra (C) + Inter (C) = Itot.

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

partition C	Intra (C)	+	Inter (C)	= total
14 classes à 1 élément		+	1.112	= 1.112
S = (24,20,22, {JEE})	.375	+	.737	= 1.112
S' = (25, 22, {JEE})	.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 Huygens qui s'énonce ainsi :

Théorème de Huygens : 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 :

$$\Sigma\{f_i \|i-h\|^2 | i\epsilon c\} = \Sigma\{f_i \|i-g(c)\|^2 | i\epsilon c\} + f_c \|h-g(c)\|^2 ;$$

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 l'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 appliquée 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 $D_{inert}(s,s')$, ou encore crit(s,s'). On a donc :

$$Intra(S') = Intra(S) + crit(s, s') ;$$

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 :

$$S = (24, 20, 22, \{ \mathfrak{JEE}\}) ;$$

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

$$S'=(25 = s \cup s', 22, \{\mathfrak{JEE}\})$$ (cf.§1.2 : arbre)

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

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

Calculons maintenant crit(24,20) :

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

24 = {ℑBL,ℑIT, ℑSP, ℑUK, ℑDL, ℑPL,ℑRM}

20 = {ℑFR,ℑNL}

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

$$f_{24} = .060 + .059 + .057 + .049 + .09 + .082 + .043 = .44 ;$$
$$f_{20} = .150 + .058 = .208;$$

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 (cf. [*AIDES CAH*] § 2) qui donne les différences $D_\alpha = F_\alpha(24) - F_\alpha(20)$ pour $\alpha = 1,\ldots, 7$:

NOEUD	AINE	BJMN	D1	D2	D3	D4	D5	D6	D7
25	24	20	-.341	.401	-.661	-.193	.228	.140	-.013

on a : $\|24-20\|^2 = D^2_1 + D^2_2 + D^2_3 + D^2_4 + D^2_5 + D^2_6 + D^2_7$

$$= .120 + .161 + .437 + .037 + .052 + .020 + .000$$
$$= .827$$

D'où pour crit(24,20) :

$$crit(24,20) = .14 \times .827 = .116,$$

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 des niveaux des noeuds (§ 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 noeuds) : 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.

ETAPE 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.

ETAPE 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 noeud 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] = crit(A[15], B[15])= crit(1,5) = .008.

Ainsi qu'on l'a dit au § 1.3, ii 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 noeud 15) et l'on a pour cette partition en 13 sommets :

Intra = 0 + D[15] = .008 ; Inter = Itot - 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 noeud 15 sont créées à l'étape 1

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

ETAPE 2 : Création du noeud 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 noeud 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 noeuds déjà créés. On a pour la partition en 12 noeuds :

Intra = 0 + D[15] + D[16] = .018 ; Inter = Itot-Intra = 1.094;

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

. . .

ETAPE 12 : Création du noeud 26. Tous les individus, à l'exception de 14(JEE) 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 créées (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 existe 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] = 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 :

Intra = 0 +D[15] +D[16]+...+D[26] = .770.

Inter = Itot - Intra = .342 ;

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

ETAPE 13 : Création du noeud 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]= crit(26,14) = .342 ;
Intra =0 +D[15] +... +D[27]= 1.112 =Itot ; 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 noeuds : 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'inerties correspondants $\tau 1 = \lambda 1/Itot$; $\tau 2 = \lambda 2/Itot$ 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 noeuds (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 noeud 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 noeuds suivant l'algorithme de base on a (en vertu de l'axiome de la médiane cf. § 2.3) :

$$\nu 15 \leq \nu 16 \leq \ldots \leq \nu 27 \ ;$$

Les noeuds 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 noeud : colonne N ;

le niveau du noeud : colonne D[N];

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

le taux d'inertie τ afférent au noeud : τ[N] = D[N]/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 .11125E + 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 noeud N/inertie totale du nuage) en millièmes

n°s des classes qui s'agrègent au noeud N

niveau du noeud N en centièmes

n° du noeud N

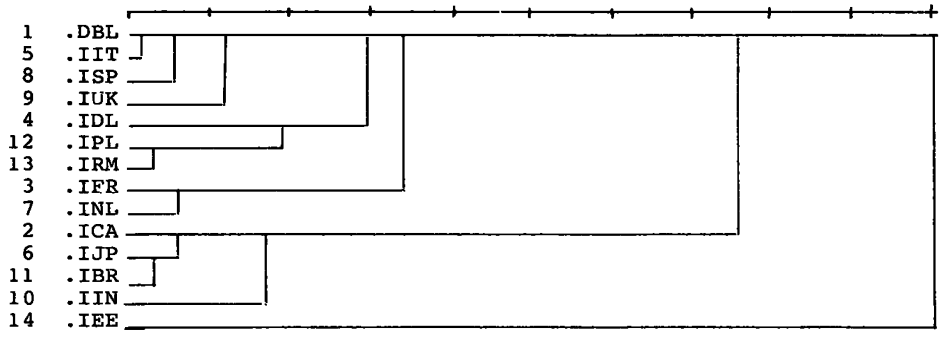
Remarque : La CAH définit sur chacune des classes n de la hiérarchie une structure hiérarchique, représentée graphiquement par la branche suspendue au noeud n et dont les noeuds sont, outre n, les noeuds de la CAH qui sont les descendants immédiats ou non de n. Par exemple, la hiérarchie définie sur la classe 22 compte 3 noeuds : 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 :

Inertie interne de 22 = D[22] + D[19] + D[16] = .093

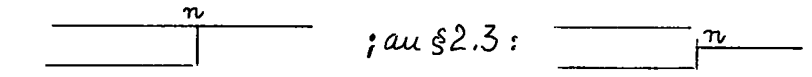
C'est à cette formule qu'on fait allusion au § 3.

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 noeuds sont tracés de façon dissymétrique ;



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°s des noeuds 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 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 JCA, correspond une seule ligne, la 10^e ; etc. . En général, si on note DEB(c) le rang de la 1-ère 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 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 et DEB[B(27)]= 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	2	9	10	11	12	13	14

On achève alors aisément le tracé. D'une part le sigle du i-ème 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 = JUK. D'autre part, pour chaque noeud n on doit tracer deux traits horizontaux occupant respectivement les lignes de rang DEB[n] et DEB[n] +P[n] = DEB[B(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 tracés plusieurs fois).

4.3 *Le tableau du contenu des classes* : Dans ce tableau divisé en 6 colonnes successives, chaque noeud occupe une ligne ou plusieurs (selon l'effectif de la classe).

Dans la colonne de gauche on lit les n°s N des noeuds : de 15 à 27 dans notre cas ; les 3 colonnes suivantes donnent 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 :

* JBL * JIT * JSP * JUK * JDL * JPL * JRM

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.) :

(* JBL * JIT * JSP(* JUK)) (* JDL * JPL * JRM)

n° du noeud N
niveau du noeud N en millièmes
n°s des classes qui s'agrègent au noeud 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 HIERARCHIE
15	8	1	5	2	*IBL*IIT
16	10	6	11	2	*ISP*IBR
17	10	12	13	2	*IPL*IRM
18	23	15	8	3	*IBL*IIT*ISP
19	24	2	16	3	*ICA*IJP*IBR
20	25	3	7	2	*IFR*INL
21	43	18	9	4	*IBL*IIT*ISP*IUK
22	59	19	10	4	*ICA*IJP*IBR*IIN
23	69	4	17	3	*IDL*IPL*IRM
24	104	21	23	7	*IBL*IIT*ISP*IUK*IDL*IPL*IRM
25	116	24	20	9	*IBL*IIT*ISP*IUK*IDL*IPL*IRM*IFR*INL
26	271	25	22	13	*IBL*IIT*ISP*IUK*IDL*IPL*IRM*IFR*INL *ICA*IJP*IBR*IIN
27	342	26	14	14	*IBL*IIT*ISP*IUK*IDL*IPL*IRM*IFR*INL *ICA*IJP*IBR*IIN*IEE

§ 4.3 : le tableau du contenu des classes.